


Towards a Systematic Data Harmonization to Enable AI Application in the Process Industry

Michael Wiedau¹, Gregor Tolksdorf¹, Jonas Oeing^{2,*}, and Norbert Kockmann²

DOI: 10.1002/cite.202100203

 This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Current methods of artificial intelligence may often prove ineffective in the process industry, usually because of insufficient data availability. In this contribution, we investigate how data standards can contribute to fulfill the data availability requirements of machine learning methods. We give an overview of AI use cases relevant in the process industry, name related requirements and discuss known standards in the context of implicit vs. explicit data. We conclude with a roadmap sketching how to bring the results of this contribution into practical application.

Keywords: Artificial intelligence, Data integration, KEEN project, Ontology, Process industry

Received: May 27, 2021; *revised:* October 18, 2021; *accepted:* October 20, 2021

1 Introduction

In the process industry, a huge variety of data is produced and has to be managed in several different software tools, databases and documents. Looking at the currently existing landscape of data, tools and services, several issues appear. The often discussed problem of data silos [1] is still ongoing. This comes with the lack of harmonized data structures. Each software system defines its own data structure and, furthermore, the existing software tools are highly configurable, which leads to a high degree of diversity in configurations. Since all of these systems have grown over time and have developed their own configurations, it is necessary to take action to harmonize these systems. Why is that necessary? The answer is that the current software landscape evolves from standalone desktop software towards data handling platforms that can be accessed, e.g., over the web and are mainly database driven, including the concept of small but powerful apps, which are focusing on specific tasks (such as calculation of overall equipment effectiveness (OEE) [2]).

This contribution focuses on the problem of a diverse software landscape that holds different forms of data in different data silos with differing configuration, hindering the chemical process industry to efficiently apply artificial intelligence (AI) driven technologies for their assets in large scale. Here the focus is to improve the data integration across the whole asset life cycle from process design over the functional design and asset specification up to the operation of the actual assets, building a basis for data availability for machine learning (ML) methods.

While former publications [3] have described how to combine the research on data integration in the project “Energieeffizienz in der Prozessindustrie (ENPRO)” with the definition of open data standards by the Data Exchange

in the Process Industry (DEXPI) Initiative, this paper will demonstrate a more application-oriented way of good practices together with the outlook of using methods of AI in the process industry.

2 Areas of AI/ML Applications

Before describing the different existing standards in the industry and how they can fulfil the requirements to improve the handling of industrial data over the asset life cycle, possible applications for data management in AI have to be defined.

To illustrate the application of AI methods we start with sketching a scenario bringing together everything that will be discussed in this contribution. Afterwards we will have a closer look into the areas of application of AI methods.

2.1 Ideal-World Scenario for the Application of AI Methods

The scenario describes the vision of having full semantic knowledge graph principles and applied standards with flexible extensions everywhere. In this scenario for each subdomain of the process industry across the asset life cycle

¹Michael Wiedau, Dr.-Ing. Gregor Tolksdorf
Evonik Operations GmbH, Paul-Baumann-Straße 1, 45128 Marl, Germany.

²Jonas Oeing, Prof. Dr.-Ing. Norbert Kockmann
jonas.oeing@tu-dortmund.de
TU Dortmund University, Department of Bio- and Chemical Engineering, Emil-Figge-Straße 68, 44227 Dortmund, Germany.

basic semantic data models exist that have been standardized using a common reference data library (RDL) (this is the 'everywhere' aspect of above vision). These standards describe data in a semantic way, making it not just machine-readable (document vs. data) but also machine-understandable, by using, e.g., semantic web technologies (this is the 'semantic knowledge graph' aspect).

All the data and metadata respectively context data is explicitly available and not hidden in some human internal memory, very long attribute names, or flat, unconnected data sheets. The different aspects in focus of the different standards are connectable in an overarching way via a suitable top-level ontology *semantic data model* for the chemical process industry, covering buildings as well as process plants. Computer programs can access and interpret semantically annotated data sets and directly match the source data to an integration model and go on from the integration model into the target structure (this is the 'full semantic knowledge graph' aspect). A separate mapping is not necessary anymore, as data points are automatically mapped by having the same semantic annotation. As standards so far mostly cover what is already known and harmonized, we must ensure to have a mechanism that allows for a fast integration of new specialized concepts (this is the 'flexible extension' aspect of the vision).

Within the sketched scenario the semantic data model allows for an easy extension of the RDL on company level, always using the general principles already defined in the basic top-level ontology. Data handover is supported independent of facing a one-time-handover, using a request-based server client architecture / data platform, or a direct event-based machine-to-machine communication (see asset administration shell description for more information). In this ideal scenario we can leverage the application of AI methods in the process industry by providing high quality context and meta data, making the available information actually meaningful to the machine learning algorithms.

2.2 Application Areas in the Process Industry

The process industry consists of many data domains that are all playing a connected role. In this contribution we will focus on three main aspects of the asset life cycle of a chemical production plant which are *chemical process optimization*, *chemical engineering* and *plant maintenance*. There are other areas or domains that are not covered in this paper, such as *supply chain optimization*, *site management*, *R&D and product development*, *human resource management*, *marketing processes* and many more.

2.2.1 Application in Chemical Process Optimization

By *chemical process optimization* we here understand methods and algorithms that optimize a chemical production process, either when a real-world plant already exists, or a

new process plant is in the very early conceptional phase. Optimization methods based on mathematical models, but also ML methods can be used in this phase to find optimal production points and a lot of research exists in this area. A good overview on basic and advanced optimization algorithms has been given by Biegler and Grossmann [4] some years ago while a newer update can be found in the work of Grossmann [5].

2.2.2 Application in Chemical Engineering

The field of chemical engineering offers a wide range of applications for AI methods. In general, basic engineering and detail engineering should be considered separately from each other since the data differ and thus also the required AI algorithms and models. For example, basic engineering uses material and process data, while detail engineering preferably uses planning and topology data. Already in the year 2008 some researchers have worked on semantic data integration of process engineering design data in order to link the most different planning data [1].

In the following, the applications of AI in engineering will be discussed in more detail. In basic engineering, process data from simulation results can be used to predict suitable separation units for separating the simulated process streams with the help of ML algorithms. First research results show that this is possible for binary as well as multi-substance systems and can provide the basis for future automated process development [6].

In detail engineering, there are first approaches and results in which machine-readable plant topologies are processed with the help of artificial neuronal networks (ANNs). The resulting models enable automated consistency checks as well as prediction of subsequent equipment to support and accelerate the drawing of piping and instrumentation diagrams (P&IDs) [7].

2.2.3 Application in the Maintenance Process

While process engineering is focusing on planning and building a chemical plant, the phase of the plant where it is running and producing products is far longer on the overall time scale of the asset life cycle. During this phase, maintenance plays an important role as every minute of the phase where the plant is not producing can cost thousands of euros. In this context, we understand maintenance as the process of operating the plant, maintaining its components (like pumps, sensors, pipelines, etc.) and by managing the different work-processes (like roundtrips and work-order management). One example for an AI method in this area is the *predictive maintenance* where historical data is used to train a model that can predict when an apparatus or machine (like a pump) will fail and when it has to be maintained so that plant's production does not have to stop.

2.3 Ontology Mapping

In the former sections we have seen areas where AI methods can be used in process systems engineering. Connecting this to our described optimal-world example, we see that different software systems are configured differently. This configuration is defined in data models or other structures. Many different of these structures exist, either as national or international norms, as company-wide standards or, and that is in fact the most frequent case in the authors experience, without any standards just by definitions inside the appropriate tool, system or platform. For the standardization and harmonization of data, it is necessary to combine these structures with each other and define mappings between these structures so that data can be transferred from one structure to the other.

For an alignment of these structures, computer science does use the term “ontology alignment” or “ontology matching”. Euzenat et al. [8] stated that “ontology matching aims at finding correspondences between semantically related entities of different ontologies”. A further literature review has been given by Otero-Cerdeira, Rodríguez-Martínez, Gómez-Rodríguez [9] and Ochieng, Kyanda [10]. A more industrial application view on ontology matching has been given by Kharlamov et al. [11].

Using methods of AI can help to improve this process of ontology mapping and could lead to faster results in integrating different data silos with different data structures [12].

Within the described KEEN project, an idea is to investigate the current state of the art in AI-based ontology alignment. Furthermore, it is planned to make use of this technology to harmonize structures, e.g., from DEXPI, Capital Facilities Information Handover Specification (CFIHOS) and ISO (ISO 15926-2).

2.4 From AI Applications to Requirements

ML methods can be applied in a wide variety across the asset life cycle in the process industry. In the next section we will move the focus from the goals of these applications to the potentially different requirements on data and its availability to apply these methods effectively.

3 Data Requirements for AI/ML Method Application

In this section we will name the key requirements we are looking for to be confident that certain data models or data structures are beneficial to make the application of AI methods a success. We will not look at requirements regarding hardware or software, but on appropriate data structures as one aspect of data quality. So, after giving the first general requirement by introducing the concept of explicit

data as a contrast to implicit knowledge, we will go through the areas of AI applications in the process industry given in the previous section and name relevant requirements in the context of data.

3.1 Implicit Knowledge versus Explicit Data

When looking at classes and attributes of available data models and standards we experience a first step of making implicit knowledge available. Implicit knowledge (or tacit knowledge) comprises, e.g., personal experience and wisdom. By writing it down and assigning properties (attributes) to concepts (classes) we make some aspects of this personal wisdom understandable for other humans. So when we write something like “a ‘centrifugal pump’ has the property ‘material number of the casing’” a human reader can conclude that there exists some kind of pump having a casing of a certain material that can be identified via a number. Why is this only the first step in making implicit knowledge available?

Because it is still not explicit enough for a computer. The fact that a pump has a casing and that this casing has the property of being made of a certain material that can be identified by a number is still implicit and not accessible for a computer.

In order to be really explicit, all the components of classes representing the apparatuses and machines we use in the chemical process industry had to be explicitly modeled. Naming the components of a whole are only one aspect. Another aspect is the fact that no apparatus or machine has a minimum or maximum temperature when we are talking about specification of assets we need in our process plants. So, when we have a data sheet of a centrifugal pump mentioning a ‘minimum allowable operating temperature of chamber two’ we are most likely talking about a property (i.e., temperature) of the fluid stream flowing through a sub-component of the pump (i.e., chamber 2) that corresponds to the condition that its value must not be lower in order to guarantee the safety and quality of the process when continuously operating the asset (i.e., centrifugal pump). In the extreme explicit case, all components of an asset and all limits and all operation conditions are made explicit and set into context by using associations (e.g., a centrifugal pump consists of chambers, a stream has a temperature, a chamber is connected to a stream, an asset is operated under certain conditions, etc.). The more explicitly and more detailed classes and relations/associations are modeled, the smaller the number of attributes per class will be.

Short attribute names are a good indicator for explicit knowledge, or, the other way around, you should be suspicious when you see property names composed of several different elements and concepts (e.g., ‘minimum allowable operating temperature of chamber two’). Our rule of thumb: The longer the name, the more information is

encoded in the name. This is not how it should be when trying to set data into explicit context and making it machine understandable. To give an example referring to the next section where we will investigate some relevant industry standards: As ISO 15926 is based on semantic technologies (e.g., resource description framework (RDF)) it is much more explicit than the database table definitions of CFIHOS; integrating information of the whole asset life cycle needs more advanced concepts than the single hand-over of equipment data sheets.

We argue that most of the meaning and business-value for the application of data lies in its explicit context (associations), not in its name, not in its single measured or defined value. Without proper data models and structure there will be no high-quality context available for machine learning in the chemical process industry.

This is reflected in a recently published German blog entry provided by Verein Deutscher Ingenieure (VDI), giving an opinion from industry [14].

3.2 Requirements for the Process Industry in General

When applying AI in the process industry, we require an understanding of the domains the consumed data and applied or trained models are valid for. So, we need to consider the domain-scope of different standards and data models. Another general requirement is the ability to identify and classify our objects or systems of objects that are the entities we mostly associate the data with. The identification requirement includes the identification of the structural context, e.g., the technical object X of type “Pump” belongs to plant section A, serves function B, and is located in area C. In addition, as many data points are created by measurements, we need the information of the measured value’s physical type and unit of measurement. The last general requirement we want to explicitly name here is having access to relevant properties connected to the assets in an operated plant. The degree of relevance itself is depending on the actual use case, but nevertheless having access to properties like the material of construction, the date of the last maintenance, or the model of the machine can lead to additional insights backed up by data.

3.3 Requirements for Chemical Process Optimization

As mentioned above, the optimization of chemical processes represents a broad field of application for AI methods. To enable application of these methods, requirements for the data need to be fulfilled regarding to optimization steps. For optimization substance data, measurement data and laboratory data will be considered in more detail. The requirement for these data and the availability regarding

their application in the field of AI will be discussed in more detail in the following.

Substance data

Substance data is available in many different, accessible sources. For example, the webbook of the National Institute of Standards and Technology [15] or the GESTIS substance database [16]. With the help of python packages such as *Request*, this data can be accessed and extracted for AI applications. It becomes more complex when substance data will be recorded and measured by third parties, as these are not available in databases. If these are to be made available, additional metadata about the quality of the data must be stored. It is of particular interest who produced the data and under which boundary conditions it was generated.

Measurement data

While a large number of process and OT-related communication protocols have already been developed for the exchange of information (e.g. Profibus, Modbus, OPC, etc.) [17], there are currently no standards for the uniform description of contained measurement data. For the use of measurement data in AI applications, it is essential that it can be interpreted. Measurement data without its meaning and the context of its generation cannot be used by third parties. In order for publicly available measurement data to be reliably harmonized for, further research needs to be done to satisfy the desire for a metadata standard.

Laboratory data

At the current state we do not know anything about the use of laboratory journals in AI applications. Well-logged lab journals have a high information content. Thus, preserving this knowledge in a central exchange platform has a high potential. However, it should always be considered how the data can be made available for AI. The question must be clarified whether current, well-documented protocols are sufficient or whether further standards and regulations are required in which the structuring of laboratory journals is more standardized.

3.4 Requirements for Plant Engineering

In plant engineering, a large amount of different information merges. Substance data and operating conditions have to be determined and simulation results used for the design. At the same time, data is generated to describe the plant topology, such as P&IDs and equipment lists. Although these data are mutually dependent, they are structured in a wide variety of formats and compiled manually, since it is often not possible to exchange the information across application boundaries. There is also a lack of universally valid, machine-readable standards, which further complicates the application of AI. The main requirement in this context is the creation of interfaces that enable the exchange of infor-

mation across the boundaries of different planning phases and thus lay the foundation for AI applications in process engineering.

3.5 Requirements for the Maintenance Process

Maintenance in process industry places high demands on safety and the constant availability of safety-relevant data. Due to their information content, safety data is becoming increasingly attractive as input for AI methods. However, current document formats (e.g., .doc, .pdf etc.) do not have sufficient machine readability and require very complex and costly preprocessing to extract the information they contain. More harmonization is needed in this area, as well as uniform metadata describing the context of the data collection, which allows verification of use by third parties.

4 Investigation of Relevant Industry Standards

We will in the following have an exploratory look at already known data models in the chemical process industry (and closely related fields) that could provide formats potentially suitable for machine learning, giving a first estimation to which degree they can already fulfill aspects related to the requirements given in the previous section. We will conclude this section with a short summarizing overall picture of the data models.

4.1 Data Models in the Process Industry

In this section we take a closer look on data models, exchange formats, and structures in the area of chemical process plants and information modeling. Despite the fact that none of the data models described in the following was developed to foster AI or ML applications, describing these data models will help identifying common ideas on how to structure and annotate data directly when it is created.

The following aspects are considered when looking at the different data models as a basis for giving structure for ML applications to meet the general requirements given in the previous section are:

- intended domain (e.g., instrumentation, apparatus/machine),
- identification of objects in the context or as part of a process plant (e.g., hierarchy elements for the plant structure),
- classification of objects related to the measurement of process and plant conditions (e.g., temperature measurement vs. pressure measurement),
- description of relevant properties for the assets in an operated process plant (e.g., date of last maintenance, material of construction).

4.1.1 Asset Administration Shell

The Asset Administration Shell (AAS) is meant to be the container for all data related to the digital representation of an asset, being a basis for its digital twin. For a current definition of the digital twin for process industries, please refer to [18]. The “Details of the asset administration shell” [19] give an excellent overview on the ideas and implementation behind the concept of the AAS. The reference architecture model is defined the international standard IEC PAS 63088 [20], and the standards proposal for the administration shell was accepted by IEC/TC 65 [21]. As a container format, the AAS can serve as a standardized entry point to access asset data and context data, including data structured according to more domain specific standards.

4.1.2 AutomationML

AutomationML (automation markup language (AML)) is standardized in the IEC 62714 [22] series. It is an engineering data exchange format for use in industrial automation systems engineering. It is designed for the automation domain, more specifically for the automation planning. AML is a data format defined by an XML schema integrating the data format computer aided engineering exchange (CAEX) as its core, which is itself standardized in IEC 62424 [23].

4.1.3 CFIHOS

The CFIHOS is driven by the International Association of Oil and Gas Producers (IOGP). In the first place, it is meant to be an industry standard for the handover of data and documents from an engineering, procurement & construction company (EPC) contractor to an owner operator. The information is specified by a relational database model and a RDL linked to ISO 15926-4 [24]. Currently a focus on oil and gas content can be observed; classification and properties for breakdown structures, apparatuses, machines, and instrumentation equipment are provided, distinguishing between planning/specification (tag concept) and actual/physical implementation (equipment concept), see Fig. 1.

4.1.4 DEXPI

The DEXPI initiative is composed of owner operators, EPCs, software vendors, and research institutions dealing with the development and dissemination of an industry standard for the exchange of information contained in P&IDs. Its current data model and implementation specification version 1.3 was released in 2021. The DEXPI specification brings together different international standards (e.g., ISO 15926 [13], ISO 10628 [26], IEC 62424 [23], ISO 10209 [27], IEC 62264) that are relevant to describe the engineering content of P&IDs (i.e., plant breakdown structure, properties of apparatuses and machines, instrumentation requests, piping topology). The conceptual information model is specified using the widely used Unified Modeling

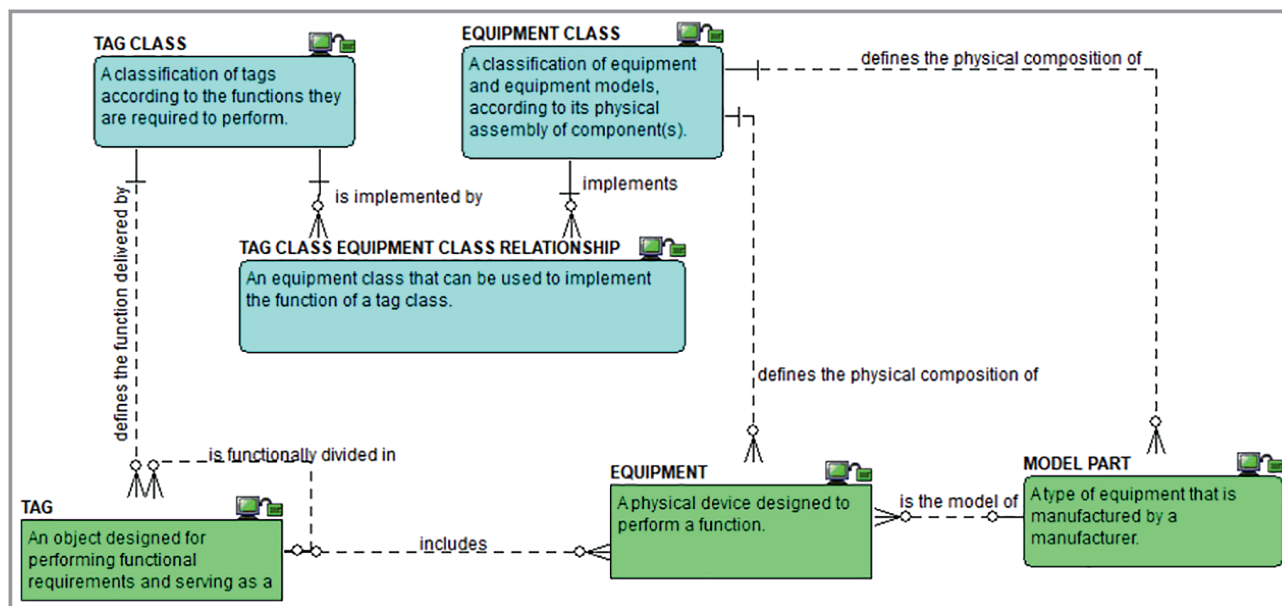


Figure 1. CFIHOS data model excerpt for tag and equipment classes, taken from [25].

Language (UML) class diagrams. Information according to the DEXPI information model can be exchanged using the Proteus XML schema [28]. The plant breakdown structure, the classification of objects, and the mechanism to share the topology make DEXPI the structural basis for information integration. It is the plant structure submodel of the AAS promoted by Platform Industrie 4.0. More information on DEXPI and the related, structured way of working with the complete life cycle (Fig. 2) is given by [3].

4.1.5 IEC/ISO 81346 Designation System Model

The IEC/ISO 81346 [29] is a standard giving guidance on how to systematically designate the elements of your indus-

trial systems. It defines the three main inherent decomposition aspects of a system answering three different questions:

- function – what? – pumping
- location – where? – building 1024
- product – how? – centrifugal pump

The concept of explicitly distinguishing between these three aspects proved to help a lot in discussions about the aspect-dependent best way to decompose a system. The aspect of identification over time (when? – from 2019 to 2021) is not as explicitly covered as with ISO 15926, so on a conceptual level the combination of IEC/ISO 81346 [29] and ISO 15926 [13] should be able to sufficiently cover the most relevant data modeling aspects (identification, classification, operating context incl. time) we need.

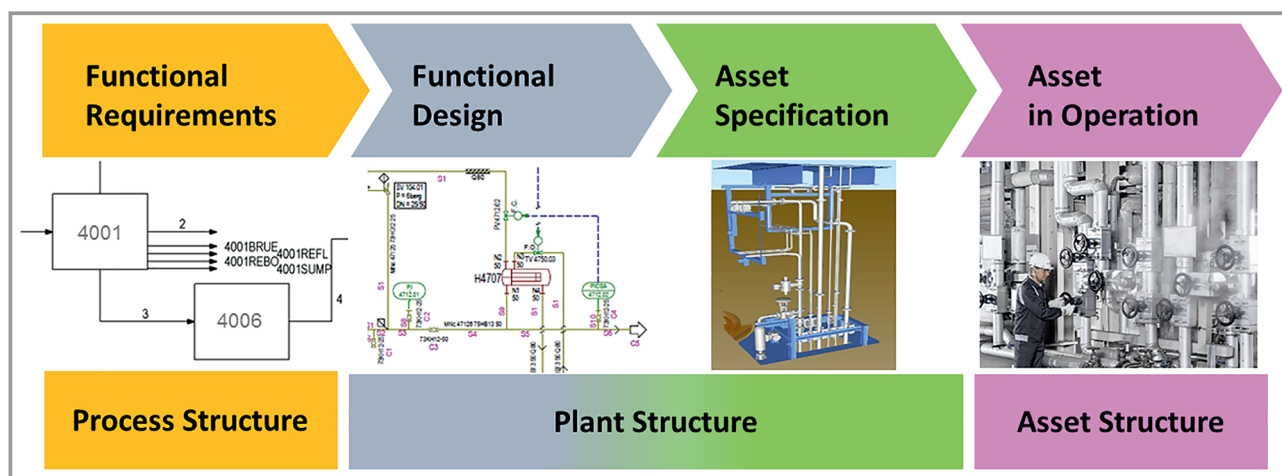


Figure 2. DEXPI life cycle phases according to ENPRO, defined by [3].

4.1.6 Industry Foundation Classes

The concept of the Industry Foundation Classes (IFC) is standardized in ISO 16739 [30]. It was developed for data sharing in the construction and facility management industries. It is the data exchange backbone of Building Information Models (BIM). As BIM is a hot topic in construction industries, several application norms exist guiding the implementation (e.g., VDI 2552 series [31]). It is currently only slowly being applied in non-construction and non-facility management areas, so there is not much experience in applying it to operating chemical process plants.

4.1.7 ISO 15926 Semantic Life Cycle Data Integration Model

The ISO 15926 series defines a framework for the integration of life cycle data for process plants. Its first parts were released in the early 2000s, but it has a history dating back at least ten more years. This standard comprises of a data model (referring to ISO 10303), a reference data library (RDL), and parts describing the implementation using templates and semantic web technologies. The ISO 15926 part 4 RDL [24] is referred to by DEXPI and CFIHOS as the common industry RDL for terms relevant in the asset life cycle (from the first process descriptions via design, construction, and operations, to the final dismantling of a plant). In order to be an integration model, the ISO 15926 [13] is a high level, abstract model. On the other hand, due to the semantic technologies, it is the most explicit model in this non-exhaustive list of data models when it comes to implementations. Especially the aspect of "identification over time" is a concept that is tackled by ISO 15926 [13] much better than in other models (for a technical introduction see e.g. [32]).

4.2 Overall Picture of Data Models

The data models and standards in the previous subsections already show that the different domains and overarching initiatives are already aware of the need of harmonized descriptions of data and data structures. Nevertheless, the application of AI methods cannot be expected to have been the reason for why these models were developed in the first place. To conclude this section, we will shortly comment on our view of the suitability of the mentioned data models, respectively standards, for meeting our requirements.

– The ISO 15926 series ("integration of life cycle data for process plants") with its semantic technologies is at the core of what we think should be used when bringing data of different source data models together in order to give the data more meaning. The main features we value most are the available top-level ontology and the RDL. Its main drawback from our point of view is the currently experienced lack of published or advertised implementations in commercial software used for data creation

or management at the different stages of the asset life cycle.

- The classes of the DEXPI data model are included in the ISO 15926 related RDL. For working with data contained in intelligent P&IDs in a tool independent way, there is no known alternative to DEXPI. In order to make a full-scale use of the topological plant information that is described on P&IDs, the topological part of the DEXPI model still must be integrated into the semantic description of ISO 15926 as well.
- The data model behind CFIHOS is for describing a snapshot of data at the point of an information handover between a principal and a contractor. An advantage of CFIHOS is its available integration into ISO 15926 technologies, building the bridge to the semantic technologies. A drawback of CFIHOS may currently be its focus on oil-and-gas industry and therefore its relative incompleteness regarding the non-oil-and-gas industry.
- We see IEC/ISO 81346 not as a data model itself, but as a technique (rooted in systems engineering) helping to model our systems. We neither use the symbols of the designation system nor the proposed classification system, that is not as elaborated for the process industry as the ISO 15926 RDL. It is the theory and philosophy behind this standard that makes it valuable to understanding our technical systems.
- We recognize IFC to be of importance for anything related to civil engineering, mainly due to regulatory requirements, but we cannot judge yet how likely it is that this standard can be connected to the standards in use for process plants, especially ISO 15926. Here we may have great potential regarding data model integration.
- The AAS itself is just a container and defining the container for things does not standardize the contained things itself that you will receive when you unpack the shell. So AAS shifts the problem of integrating information and specific data models by one level without solving the root cause – unaligned models covering the same domains. The AAS can nevertheless serve as a guidance for data models that want to be a submodel for the AAS, bringing them to a minimum level of quality and compatibility.
- AutomationML is following an XML approach that nowadays can be considered old-fashioned. Despite of its potential usefulness for the automation planning domain, we cannot judge yet to which degree it can be successfully integrated into the more modern, more semantic data models.

In the end, the expectation of not being able to find a single standard that can fulfill all the requirements was met. It is ongoing work to investigate the applicability of these standards in practice and potential implementation guidelines when using them as data backbone for AI.

5 Roadmap

To maintain an economic and innovative lead, it is essential to increase flexibility and agility throughout the whole life cycle of process plants using new, digital technologies such as artificial intelligence [33]. For this purpose, the KEEN project [34] is investigating the use of AI methods in the process industry. In this context, it is important to create approaches in which the quality of accessibility is improved for all three areas of application (chemical process optimization, chemical engineering, maintenance process).

The low and diverging availability of data as well as the high demand for explainable AI solutions in the process industry indicates that off-the-shelf deployments of AI are unlikely to play a role in future development. For this reason, it is important that data is reliably described and harmonized in order to gain the necessary confidence of security for possible AI algorithms.

When AI and ML shall systematically enter the process industry, the following questions must be openly discussed and addressed, because their results form the basis for ensuring that data structures offer the best fit for their specific field of application.

- Which data formats do machine learning models need?
- How do we get from the existing formats to the formats we need?
- Which of the existing formats are already suitable for machine learning?
- What are feasible criteria for assessing whether a format is suitable for machine learning?

The last question has to be further differentiated. In particular, a distinction must be made between data and meta-data [35], and the semantic context of an entry must be defined. For example, it is possible that a data table has a "value" column, but information about units of measurement (UOMs) (kg, m, Pa, ...) or *QuantityTypes* (mass, length, pressure, ...) may be missing. The *QuantityTypes* allow to compare different quantities and they are defined in the definition 3.2 of the ISO 80000-1 standard [36]. Furthermore, well-structured data lays the ground for better results in optimization of data, which can be also shown in the recent work of Schweidtmann et al. [37].

5.1 Standards-Development Roadmap

In recent decades the development of standards in the process industry often followed its own timeline, content and specific focus area. Nowadays, digital workflows and production raise the demand for the integration of different systems. This leads to a rethinking of the way of how standards are developed.

A practical example is the former mentioned ISO 10628 [26]. This norm defines the symbols on P&IDs and is a norm that has its origins in a completely analogue world. Keeping that in mind, the evolution of P&IDs software tools

is subject to a digital transformation (from drawing *lines* on a sheet to a P&ID-object handling software platform), which also leads to a transformation from ISO 10628 as an analogue standard to the digital foundation of an engineering objects classification hierarchy. This example demonstrates that former norms and standards can, and must, be adopted to the digital world as digital workflows are always based on existing workflows, configuration and a huge history of existing standards.

5.2 Implementation Roadmap

Furthermore, in order to achieve the goals, it is extremely helpful if a central repository is established to store the data. Within the KEEN project, a dedicated work package aims at implementing a data exchange platform to make data available for AI applications along the entire life cycle of a process plant, covering engineering and operations. The goal is to create a framework that specifies the structural organization of data so that it can be processed "cleanly" from its source. There are three main aspects such a framework must consist of:

- 1) The systems & platforms view that describes how software systems, tools and platforms are interconnected, which interfaces to other platforms they have and how different layers are defined.
- 2) The data view that defines how data is structured, which kinds of data exist and how different data types are related to each other.
- 3) The process view, where a definition of structured work process inside the company is performed and the actors in the process know when to fulfill which process step.

Even if the process view is the most crucial view in the whole framework, this topic is not covered here, as the definition of working processes is always specific for the use case and the company it is designed for. For a general view, we recommend having a look at the work of Theißen, Hai, Marquardt [38]. In the following, we will have a deeper look into the *systems & platforms view* as well as the *data view*.

5.2.1 The Systems & Platforms View

The topic of how to structure the architecture of systems for data integration is a wide field where people from science, consulting companies, software firms and industry discuss and suggest a lot of different approaches [39–41]. In summary, all parties define an architecture where different *areas* (aka *layers*) are interconnected to each other. Fig. 3 shows an abstraction of such a framework where the source systems deliver their data to the ingest layer. Here, the data is transformed into a common target structure and loaded into the storage area. The serve area can then provide the data to other applications by defining an abstraction layer so that consuming applications (like visualizations, data

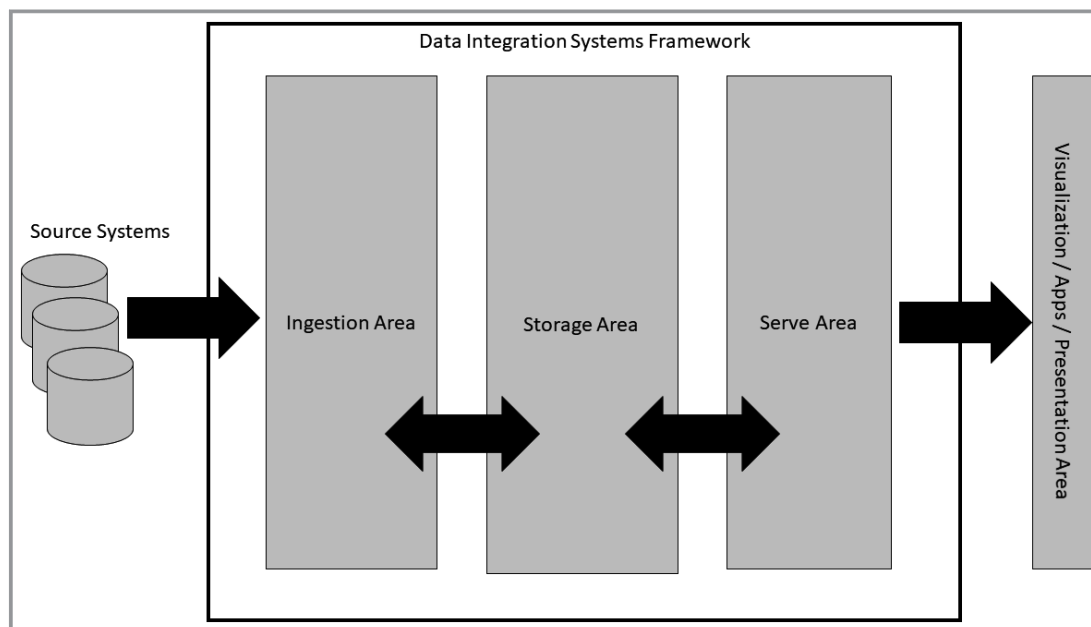


Figure 3. Interaction of computer systems areas for data integration derived from Gill, Vaidyanatha, Belur [39–41] and others.

analytics frameworks and other platforms) can access the integrated data in the same way.

5.2.2 The Data View

The top priority in this context is the FAIR principle (findable, accessible, interoperable, and reusable) to ensure loss-free processing of the data by all users [42]. During development, it is essential to agree on uniform structures depending on the use case. At the same time, comprehensive metadata must be stored for the description to ensure that the stored data sets can be identified and verified for the appropriate AI area of application.

5.2.3 Implementation Outlook

As described in this section, systems and platforms together with data play an important role in the consolidation of information to enable methods of AI. Combining the systems, the data and the standards is a key component in the current research of the publicly funded KEEN project.

6 Summary

The current situation in the industry shows that many workflows are still using paper to handle and transport information from one person to the other. The usage of methods of AI to improve such processes is hindered by the inaccessibility of such data.

While the global pandemic in 2020 and 2021 has improved the digitization of processes, there is still a huge

requirement for integrated data driven processes. By these means, the exchange of data using file formats like portable document Format (PDF), extensible markup language (XML) or comma-separated value file format (CSV) is a good step in the right direction, but the state of the art and future technology must and will be the exchange of data between software platforms by direct data exchange to improve the optimization of processes by methods of AI.

In this contribution we gave an overview of the current initiatives dealing with data standardization and have shown that there is no single standard fulfilling all current requirements to effectively apply methods of AI in process industry in the areas of process optimization, process engineering and plant maintenance by fostering data availability and data integration. We highlighted that the KEEN project aims at investigating the next steps in making data available by defining standards for data exchange platforms considering the platform and data view including data standards.

The BMWi is acknowledged for funding this KEEN project initiative. Open access funding enabled and organized by Projekt DEAL.

Abbreviations

AAS	Asset administration shell
AI	Artificial intelligence
AML	Automation markup language
ANN	Artificial neuronal network
BIM	Building information models

CAEX	Computer aided engineering exchange
CFIHOS	Capital facilities information handover specification
CSV	Comma-separated value file format
DEXPI	Data exchange in the process industry
ENPRO	Energieeffizienz in der Prozessindustrie
EPC	Engineering, procurement & construction company
IFC	Industry foundation classes
IOGP	International Association of Oil and Gas Producers
ML	Machine learning
OEE	Overall equipment effectiveness
OT	Operational technology
P&ID	Piping and instrumentation diagram
PDF	Portable document format
RDF	Resource description framework
RDL	Reference data library
UML	Unified modeling language
UOM	Unit of measurement
VDI	Verein Deutscher Ingenieure
XML	Extensible markup language

References

- A. Wiesner et al., Wissensbasierte Integration und Konsolidierung von heterogenen Anlagenplanungsdaten, *atp edition – Automatisierungstechnische Praxis* **2010**, 52 (4), 48–58.
- S. Nakajima, *Introduction to TPM: total productive maintenance* (Translation), Productivity Press, New York **1988**, 129.
- M. Wiedau et al., Enpro data integration: extending DEXPI towards the asset lifecycle, *Chem. Ing. Tech.* **2019**, 91 (3), 240–255.
- L. T. Biegler, I. E. Grossmann, Retrospective on optimization, *Comput. Chem. Eng.* **2004**, 28 (8), 1169–1192. DOI: <https://doi.org/10.1016/j.compchemeng.2003.11.003>
- I. E. Grossmann, *Advanced Optimization for Process Systems Engineering*, Cambridge Series in Chemical Engineering, Cambridge university Press **2021**.
- J. Oeing, F. Henke, N. Kockmann, Machine Learning Based Suggestions of Separation Units for Process Synthesis in Process Simulation, *Chem. Ing. Tech.* **2021**, 93, (12), in press. DOI: <https://doi.org/10.1002/cite.202100082>
- J. Oeing et al., Uniform data bases as a driver for future process development (data, repositories and application examples), in *ECCE – European Congress of Chemical Engineering*, Berlin, September **2021**.
- J. Euzenat, P. Shvaiko, *Ontology matching*, Springer, Berlin **2007**.
- L. Otero-Cerdeira, F. J. Rodríguez-Martínez, A. Gómez-Rodríguez, Ontology matching: A literature review, *Expert Syst. Appl.* **2015**, 42 (2), 949–971. DOI: <https://doi.org/10.1016/j.eswa.2014.08.032>
- P. Ochieng, S. Kyanda, Large-scale ontology matching: state-of-the-art analysis, *ACM Comput. Surv. (CSUR)* **2018**, 51 (4), 1–35.
- E. Kharlamov et al., Capturing Industrial Information Models with Ontologies and Constraints, in *The Semantic Web – ISWC 2016* (Eds: P. Groth et al.), Springer, Cham **2016**, 325–343.
- M. A. Khoudja, M. Fareh, H. Bouarfa, Ontology matching using neural networks: survey and analysis, in *Proc. of the 2018 International Conference on Applied Smart Systems (ICASS)*, IEEE, Piscataway, NJ **2018**, 1–6.
- ISO 15926-2, *Industrial automation systems and integration – Integration of life-cycle data for process plants including oil and gas production facilities – Part 2: Data model*, Beuth Verlag, Berlin **2013**.
- W. Otten, N. Kiupel, Welchen Wert haben Daten?, VDI-Blog, Düsseldorf **2021**. <https://blog.vdi.de/2021/04/welchen-wert-haben-daten/> (accessed on May 17, 2021)
- NIST Chemistry WebBook, National Institute of Standards and Technology, **2021**. <https://webbook.nist.gov/chemistry/> (accessed on May 12, 2021)
- GESTIS-Stoffdatenbank, Deutsche Gesetzliche Unfallversicherung e.V., Berlin **2021**. <https://gestis.dguv.de/> (accessed on May 12, 2021)
- Bussysteme in der Automatisierungs- und Prozesstechnik: Grundlagen, Systeme und Anwendungen der industriellen Kommunikation* (Eds: G. Schnell, B. Wiedemann), Springer Fachmedien Wiesbaden **2019**. DOI: <https://doi.org/10.1007/978-3-658-23688-5>
- A. Bamberg et al., The Digital Twin – Your Ingenious Companion for Process Engineering and Smart Production, *Chem. Eng. Technol.* **2021**, 44 (6), 954–961. DOI: <https://doi.org/10.1002/ceat.202000562>
- Details of the Asset Administration Shell – Part 1*, Plattform Industrie 4.0, Federal Ministry for Economic Affairs and Energy, Berlin **2020**. www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Details_of_the_Asset_Administration_Shell_Part1_V3.html (accessed on May 17, 2021)
- IEC PAS 63088, *Smart manufacturing – Reference architecture model industry 4.0 (RAMI4.0)*, VDE Verlag, Berlin **2017**.
- German Standardization Roadmap Industrie 4.0*, DIN e.V., Berlin **2020**. www.din.de/resource/blob/65354/1bed7e8d800cd4712d7d1786584a7a3a/roadmap-i4-0e-data.pdf (accessed on May 17, 2021)
- IEC 62714-1, *Engineering data exchange format for use in industrial automation systems engineering – Automation Markup Language – Part 1: Architecture and general requirements*, VDE Verlag, Berlin **2018**.
- IEC 62424, *Representation of process control engineering – Requests in P&I diagrams and data exchange between P&ID tools and PCE-CAE tools*, International Electrotechnical Commission, Geneva **2016**.
- ISO 15926-4, *Industrial automation systems and integration – Integration of life-cycle data for process plants including oil and gas production facilities – Part 4: Initial reference data*, International Organization for Standardization, Geneva **2013**.
- CFIHOS Data Model, Version 1.4.1, **2020**. www.jip36-cfihos.org/wp-content/uploads/2021/01/C-DM-001-CFIHOS-Data-Model-version-1.4.1.pdf (accessed on May 17, 2021)
- ISO 10628-2, *Diagrams for the chemical and petrochemical industry – Part 2: Graphical symbols*, International Organization for Standardization, Geneva **2012**.
- ISO 10209, *Technical product documentation – Vocabulary – Terms relating to technical drawings, product definition and related documentation*, International Organization for Standardization, Geneva **2012**.
- Proteus Schema for P&ID Exchange*, ProteusXML, **2017**. <https://github.com/ProteusXML/proteusxml> (accessed on May 17, 2021)
- ISO/IEC 81346-1, *Industrial systems, installations and equipment and industrial products – Structuring principles and reference designations – Part 1: Basic rules*, International Organization for Standardization, Geneva **2009**
- ISO 16739-1, *Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries – Part 1: Data schema*, International Organization for Standardization, Geneva **2018**.

- [31] VDI 2552, *Building information modeling – Fundamentals*, VDI Verlag, Düsseldorf **2020**.
- [32] *Temporal Parts*, ISO 15926 Portal, **2021**. <https://15926.org/topics/temporal-parts/index.htm> (accessed on May 25, 2021)
- [33] P. Buxmann, H. Schmidt, Wettbewerbsvorteile durch Künstliche Intelligenz, in *Künstliche Intelligenz: Mit Algorithmen zum wirtschaftlichen Erfolg* (Eds: P. Buxmann, H. Schmidt), Springer, Berlin **2019**, 197–201. DOI: https://doi.org/10.1007/978-3-662-57568-0_13
- [34] www.keen-plattform.de
- [35] S. Dustdar et al., Quality-Aware Service-Oriented Data Integration: Requirements, State of the Art and Open Challenges, *SIGMOD Rec.* **2012**, *41* (1), 11–19. DOI: <https://doi.org/10.1145/2206869.2206873>
- [36] ISO/TC 12 ISO 80000-1, *Quantities and units – General*, International Organization for Standardization, Geneva **2009**.
- [37] A. M. Schweidtmann et al., Obey validity limits of data-driven models through topological data analysis and one-class classification, *Optim. Eng.* **2021**. DOI: <https://doi.org/10.1007/s11081-021-09608-0>
- [38] M. Theißen, R. Hai, W. Marquardt, A framework for work process modeling in the chemical industries, *Comput. Chem. Eng.* **2011**, *35* (4), 679–691.
- [39] N. S. Gill, *Data Ingestion: Pipeline, Architecture, Tools, Challenges*, Xenonstack, Punjab **2021**. www.xenonstack.com/blog/big-data-ingestion/ (accessed on May 14, 2021)
- [40] G. Vaidyanatha, *Architecting Serverless Data Integration Hubs on AWS for Enterprise Data Delivery*, Towards Data Science, Canada **2020**. <https://towardsdatascience.com/architecting-serverless-dataintegration-hubs-on-aws-for-enterprise-data-delivery-2020-a27b42569518> (accessed on May 14, 2021)
- [41] V. Belur, *Kappa Architecture – Easy Adoption with Informatica End-to-End Streaming Data Management Solution*, Informatica, Redwood City, CA **2020**. www.informatica.com/blogs/adopt-a-kappa-architecture-for-streaming-and-ingesting-data.html (accessed on May 14, 2021)
- [42] M. D. Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data* **2016**, *3* (1), 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>