NIST Advanced Manufacturing Series
NIST AMS 100-65

# Metadata Modeling for Manufacturing Enterprise Integration

Yan Lu
Boonserm Kulvatunyou
Dimitrije Milenković
James Wilson
Michael Figura
David Noller
Josh Ki

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AMS.100-65

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

**NIST Advanced Manufacturing Series**
**NIST AMS 100-65**

# Metadata Modeling for Manufacturing Enterprise Integration

Yan Lu
*System Integration Division*
*Engineering Laboratory*
*National Institute of Standards and Technology*

Boonserm Kulvatunyou
*System Integration Division*
*Engineering Laboratory*
*National Institute of Standards and Technology*

James Wilson
*Open Application Group, Inc*

Dimitrije Milenković*
*System Integration Division*
*Engineering Laboratory*
*National Institute of Standards and Technology*

Michael Figura
*Open Application Group, Inc*

Josh Ki
*Lockheed Martin*

David Noller
*Virginia Tech. University*

**Former NIST Guest Researcher; all work for this publication was done while at NIST.*

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AMS.100-65

January 2025


U.S. Department of Commerce
*Gina M. Raimondo, Secretary*

National Institute of Standards and Technology
*Charles H. Romine, performing the non-exclusive functions and duties of the Under Secretary of Commerce for Standards and Technology and Director, National Institute of Standards and Technology*

NIST AMS 100-65
January 2025





**Publication History**
Approved by the NIST Editorial Review Board on 2024-12-16

**How to Cite this NIST Technical Series Publication**
Lu Y, Kulvatunyou B, Milenković D, Wilson J, Figura M, Noller D, Ki J (2024) Metadata Modeling for Manufacturing Enterprise Integration, (National Institute of Standards and Technology, Gaithersburg, MD) Advanced Manufacturing Series (AMS) NIST AMS 100-65. https://doi.org/10.6028/NIST.AMS.100-65

**Author ORCID iDs**
Yan Lu: 0000-0003-2927-5860
James Wilson: 0009-0001-9809-7212
Boonserm Kulvatunyou: 0000-0000-0000-0000

**Contact Information**
yan.lu@nist.gov

**Abstract**

The continuous digitalization in manufacturing directly leads to a significant increase in "data volume", while digital transformation is expanding the exchange scenarios of such data. Managing metadata, the description of data, is a foundational aspect of successful enterprise digital transformation. By providing clarity and structure of data, metadata models and standards ensure that data can be discovered, accessed, shared and used effectively across the organization, driving better decision-making, improving operational efficiency, and supporting compliance. This report examines common manufacturing enterprise metadata exchange scenarios, analyzes and compares existing metadata standards using the Findable, Accessible, Interoperable and Reusable (FAIR) principles. Based on the metadata standards reviewed and a FAIR data metadata stack model, we propose a uniform, general, and extendable metadata model which not only captures essential metadata elements for Findability, Accessibility, and Interoperability but also uses an extendable metadata model for (Re)usability. An extension of the proposed Uniform Metadata Model for Manufacturing Enterprise Integration is demonstrated for industrial measurement dataset metadata representation. Additional general metadata elements are identified for measured, sampled data, and derived stream descriptions. The report concludes with a direction for manufacturing metadata standards development.

**Keywords**

Manufacturing enterprise, metadata, metadata modeling, FAIR data, data exchange, digital transformation

## Table of Contents

## 1. Introduction

Metadata is data that defines and describes "*data*" based on ISO/IEC 19763[1], while data is defined as a formalized representation of information that can be interpreted, communicated, or processed. The scope of the "*data*" to be addressed in this paper includes persistent digital objects such as design models, images, measurement datasets, machine learning models, and manufacturing and supply chain documents, which are common in manufacturing enterprises.

Metadata management is a key component of digital transformation [1] in that it provides information about what a digital object is, where it can be found, who its creators are, how it was created, how it was updated, what parties are authorized to access it, etc. Business users who work closely with data (e.g., analysts, data scientists, and IT teams) rely on metadata to give them crucial context for deriving insights about various data objects. However, many organizations do not recognize the importance of metadata management for effective strategic and tactical decision-making. Metadata-related capabilities and practices that are often lacking in the enterprise include:

1) understanding of how data can be repurposed and reused,
2) workflows that collect metadata, log metadata collection, and support the execution of business processes,
3) digital rights management, including authorization delegation,
4) trust in the quality of data and metadata from business parties,
5) interoperable metadata definition that supports its sharing , and
6) technical mechanism and tools for metadata capture.

Standard metadata models and modeling approaches are key to improving metadata management, data object usage, and addressing the deficiencies listed above. Well-developed metadata models specify rich metadata elements for various data object types. Industries face challenges in metadata modeling due to the vast variety of the data object types within and across enterprises, and due to the number of metadata elements needed to satisfy every data use and exchange scenario. Therefore, an effective metadata modeling architecture is required.

Several organizations, including International Organization for Standardization (ISO)[2], Object Management Group[3], Research Data Alliance (RDA)[4], and International Electronics Commission (IEC)[5] etc., are working on digital objects, research and scientific data, and enterprise metadata definitions. More specifically, ISO/IEC JTC1/SC 32/WG 2 was established in 1998 to develop and maintain domain-agnostic standards that facilitate specification and management of metadata. Their scope includes a framework for specifying and managing metadata, and detailed specifications on how to describe data elements, value domains, and other reusable semantic and representational information objects about the meaning and technical details of a data item.

[1] https://www.iso.org/standard/84749.html
[2] https://www.iso.org/
[3] https://www.omg.org
[4] https://www.rd-alliance.org/
[5] https://www.iec.ch

GO FAIR is a bottom-up, stakeholder-driven and self-governed initiative started by the Open Science community that aims to implement the FAIR data principles, making data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable[1]. To implement FAIR, the Research Metadata Schemas Working Group under RDA[6] are developing guidelines for the data-sharing communities whose needs are not addressed by existing metadata schema such as schema.org, and provides guidelines on proposing extensions. In addition, there are multiple interoperability-based standards development organizations working on promoting business process interoperability for both inter-enterprise and intra-enterprise business processes based on open standards and tools[17][18][19]. For example, connectSpec (formerly, OAGIS, [2]) from OAGi, a non-profit industry consortium, defines a common content model and common messages for operational or engineering information exchange. These standards and guidelines greatly improve enterprise business-process integration.

Meanwhile, enterprise metadata management software options are increasing in number and capabilities. They provide the technology necessary to ensure that the metadata across the enterprise adds value to that enterprise's data. The market for metadata management solutions comprises vendors that include one or many metadata management capabilities such as metadata repositories, business glossary, data lineage, semantic modeling, and metadata ingestion and translation. Metadata management is also typically a built-in function in data catalogs, data lake products, and data warehouse products. However, the lack of standards leads to the difficulty in integrating metadata-management software tools.

Continuous digitalization and digital transformation are expanding enterprise information exchange scenarios that span product data, process data sets, and supply chain documents. This technical report aims to provide a foundation for enterprise metadata standards development. We first introduce a few enterprise-data object exchange use cases and their needs of metadata data (Section 2). Section 3 reviews existing standards for metadata modeling and state-of-art metadata technology. In Section 4 we present a Uniform Metadata Modeling approach for Manufacturing Enterprise Integration, named UMM4Mei. It is a novel, extensible, hierarchical metadata modeling approach which can be used for CAD model, process data set and supply chain document metadata modeling. Section 5 concludes the paper with standards development plan.

[6] https://www.rd-alliance.org/groups/research-metadata-schemas-wg/members/all-members/

## 2. Metadata Use Cases and Metadata Types

There are various use cases of metadata by manufacturing enterprises, for product development, production system control, operation and maintenance, and supply chain management. In this section, some representative digital object types, their metadata and associated use cases are described.

### 2.1. CAD model metadata

Three-dimensional (3D) CAD models are the data set that contains 3D geometric elements representing the object/part to be manufactured. *CAD Models* can be generated by commercial CAD solid modeling software such as SolidWorks, NX, Creo, and AutoCAD. 3D Design is typically an iterative process that involves many iterations of refinements in the shape and dimension of the initial model. It outputs an optimized model according to given objectives and boundary conditions to meet design requirements. Designers frequently use simulations during this process to validate design improvements.

CAD models can be exported from the CAD software in many formats. Some of the file formats are native to the specific program but include more information and features. Some file formats are neutral, allowing interoperability among different CAD software. Using the CAD software, users can add additional metadata such as a Title, Subject, Author, Keywords and comments as file properties. They can also add custom properties such as the drawing number and revision. For example, users can add metadata to an AUTOCAD DWG file with the '**DWGPROPS**' function [4]. Design intents, design rules, and various versions of the CAD models and the pedigree information can be collected and managed as well. Additional metadata for CAD models make the model more useful and can be reused by various stakeholders during a product life cycle.

### 2.2. Agriculture Metadata use cases

Data generated by global agriculture systems are transforming food production towards a safe, sustainable, resilient, and high-quality future [3]. Industry benefits from awareness of and linkage between agricultural data sets at the producer/farmer, retailer, processor and up- and downstream channel partner levels. Supply chain transparency, sound farm business management, and consumer confidence through improved traceability all rely on the ability to clearly and consistently link data. A specific use case is to provide the structure to link summary metadata about Critical Tracking Events (CTE), such that raw data can be found for analyses[4]. Metadata will describe the raw data specific to its scope, provide a URI link to the raw data, and potentially known related predecessor metadata that may exist. Agriculture data set metadata:

- Allows awareness and linkage of related agricultural data,
- Addresses reality that individuals working to solve a specific of the agricultural value chain often work with more than one data repository,
- Assists in product recalls, traceability, and consumer confidence, and
- Enables farmers to share their data with a best-of-breed tool for gaining insights.

### 2.3. Process Data Metadata Exchange

Industrial systems generate large amounts of process data, which usually are in time series, collected, stored and can be retrieved from the software known as historian. Process data usually describe measurements over time, which capture the quantification of attributes of a manufacturing object, such as equipment, part or process, or an event on that object.

Process data are the most valuable assets for manufacturing and other enterprise performance improvements. They offer insight into how a system has performed and varied over time and enable operators to make data-driven decisions. The data are usually archived in a long-term historian that helps track, over time, manufacturing system problems and product defects caused by the problems [5]. To identify causes of product quality problems or issue a product recall, the time series data in historians should be packaged and shared with stakeholders in the supply chain [6].

Metadata offers interpretability for insights from process data into resolving manufacturing operation problems or aiding enterprise decision making. However, metadata for process data is a challenge in manufacturing context, since data are not standardized across due to a wide range of products, machines and production lines used. A more precise, standardized description of data is essential in order to extract meaningful metadata from the data.

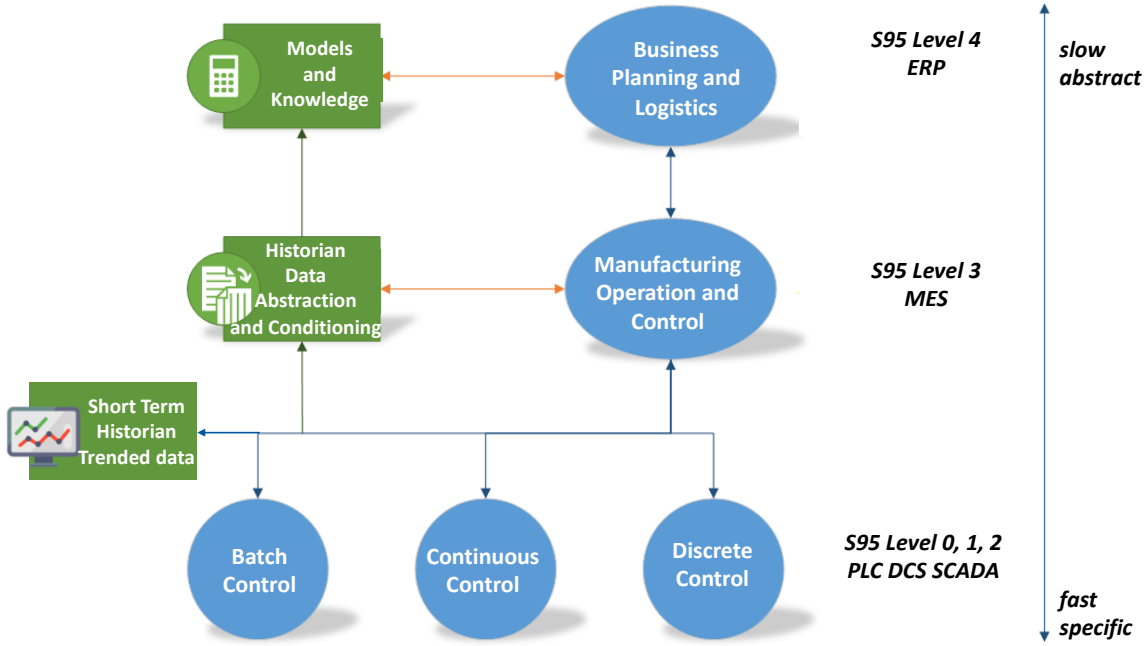


Figure 1. Process data and ISA 95 Hierarchy (Adapted from [9])

The ISA 95 describes the flow of manufacturing process data as illustrated in the Figure 1. Moving from low levels to higher levels, data are compressed, aggregated and abstracted for decision making. The context of data generation, as well as the pedigree information about the data transformation, must be captured for the operational or business intelligence at Level 3 and 4, which are specified by the ISA 95 activity model and object model. ISA 88 captures some

metadata for process data including their formats, associated product data, material data, equipment data, personnel, other process data and business data. The most important metadata are time and location, which place process measurements in context of time and space. The combination of the process model, the equipment model and time can put time series data in context of other events in the enterprise. For example, video data can be combined with time series data by representing the video camera in the equipment model and combining the timestamps from the video stream and the time series data.

Traditionally, ISA 95 equipment model and ISA 88 batch data model, as well as PACKML, OPC UA are used for transient data modeling and enable the data exchange between machines, SCADA and MES or ERP systems. In smart manufacturing, machine learning is more frequently applied to persistent process data stored in clouds. These machine learning models are used for product/process development or value chain optimization, beyond typical operational use in MES and ERP. There is a need to re-examine the method to model the metadata for process time series data.

## 2.4. Supply Chain Metadata

As goods move through the Supply Chain, a plethora of information about the goods, custody of the goods and movement/handling (including division into boxes, bottles or packaging into cartons or pallets) of the goods is collected along the way. Until recently, this information was collected in "siloed" systems owned by various participants in the supply chain (e.g. the manufacturer, the shipping entity, or the receiving entity). The trend today is towards "chain of custody" systems that can gather up much of this information and store it in one universally available system so as to provide a "big picture" view of the information, named metadata.

The data (and metadata) requirements for "Provenance", or "Chain of Custody" for goods moving through the Supply Chain are clearly different from an "open supply chain" (e.g. to support Cloud Manufacturing).

Consider the example of moving food from farm to table, as described by [9].


Figure 2: Tracking Food from Farm to Table [9]

The figure shows that there are typically multiple partners involved in such a Supply Chain, including, Farm, Packer, Transporter, Customs, Processor, Distribution Centers, Retailer and Customer.
Potentially, all these partners may have siloed systems collecting data on their portion/involvement in the Supply Chain processes. Information collected during the process includes, for example, Packing Lists, Bill of Lading, Transportation Documents, Export license, Customs documents, Invoice, Freight Bill, Waybills, Inspection documents. These documents in the supply chain are important for customers and can affect things like price (Organics command a higher price today, sometimes much higher) and recalls (it is important to know the source of food and where it went for situations like recalls in the case of contamination).
Likewise, contract Service Level Agreements (SLAs) may require specified times for delivery (to ensure freshness) and temperature control as well, so IoT data and business events are used to collect data that can be used to help ensure SLA compliance.
In order to understand metadata requirements, it is first necessary to understand the data collection requirements for the use case. In summary, the following categories of data are collected:

1. Business Events – includes events such as packing, unpacking, transportation, location, change of custody. These are typically captured via digital business events based on standards such as EDI, OAGIS or GS1/EPCIS which are self-describing events (no metadata really required outside of the events).
2. Documents – many documents are generated and exchanged during the life of the food in the supply chain, some of which are described in the Use Case Description section. These documents can be paper or digital (e.g. PDF, JPG) but must be stored in a secure way to ensure provenance/chain of custody for the goods. Metadata required here would include, for example, Creation Timestamp, Format, Document ID, Document Type, Location, Creator, Device, Edits and URI.
3. IoT Data (e.g. Temperature, Humidity, Velocity, Shock, Location). Required metadata includes Device, Timestamp, Data Type, Quantity kind and unit, and the aboutness/context of the data, e.g., container number or location from which the data were collected.

### 2.5. Additive Manufacturing Big Data Set Metadata

Additive manufacturing (AM) is a smart manufacturing technique that fabricates components directly from 3-D models by selectively point-to-point or layer-upon-layer joining materials. Different from the traditional subtractive manufacturing, a vast variety of big data sets are generated through AM development lifecycles. The amount (~TB), type (images), and speed (~GB/sec) of the collected data are unprecedented. The data sets are created and collected to capture and characterize various factors that could affect AM part quality, including those from material design, machine and process development, and process monitoring and control, as well as part inspection and testing. In the details of these activities, a large range of overlapping data needs exist that may benefit from sharing and reusing data.

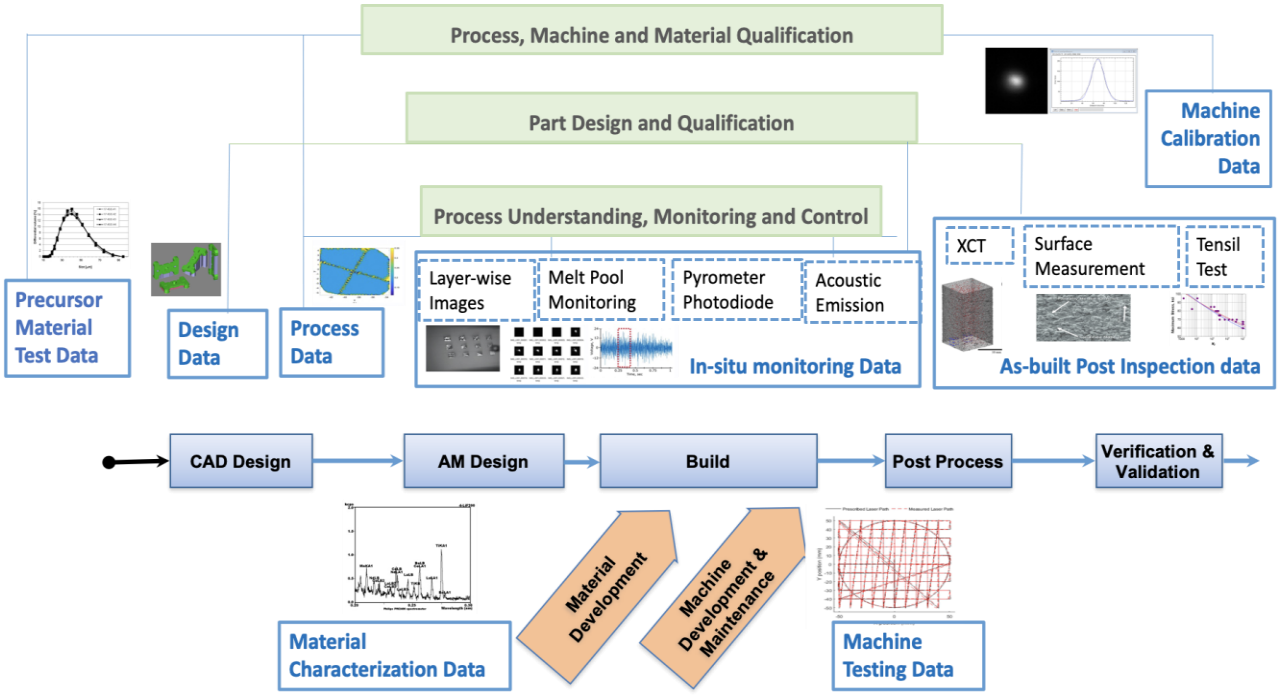


Figure 3: AM data set and use scenarios.

As shown in Figure 3, design data, machine data, in-process monitoring data, as well as measurements from material characterization, inspection and test are common across many AM activities. Three categories of data-driven AM engineering decisions exist: AM process monitoring and control, AM part qualification and development, and AM process, machine and material qualification and development. The latter decisions can rely on a broad spectrum of data beyond individual stakeholder's data generation capability. For example, in-process monitoring data generated for part quality control can be used for material and machine property understanding and reduce the need for experiments by the material and machine vendors for material and machine development.

To enhance the usability and reusability of the data sets, metadata, "description of data", must be collected and accessible by AM ecosystem users. The AM metadata should be organized to support on-demand data discovery and retrieval, as well as data exchange. For big data, especially those from in-process monitoring or structure inspection, metadata is required to support partial data retrieval, e.g., locating data based on material, process, part criticality and (in-process/NDT) data type, retrieve the in-process monitoring data for the first 5 layers and the last five layers for a specific part. Five types of metadata are identified for AM big data objects, including general description, activity related information, instrument related information and measurement related information.

## 2.6. BioPharma

Pharmaceutical product lifecycle includes drug development, clinical trial, regulatory filing, process development (scale up and drug product design), production, and disposal. All these

lifecycle stages generate a lot of data. Supply chain, logistics, and traceability are also important activities and can also generate more data.

Taking the production lifecycle stage as an example, the whole pharma industry is moving from batch toward continuous manufacturing because of potential benefits such as smaller capital investment, smaller footprint, easier to scale up, lower risk of waste, more suitable for distributed manufacturing. However, the process control in continuous manufacturing is more complex because it needs a holistic control model instead of individual control for each unit operation. That means the number of process parameters is at least a multiplication of all unit operations combined and using a feedback control or obtaining a mechanistic control model is rather difficult. Scientists have to rely on a data-driven or hybrid control model. While data-driven control can be used to improve performance in batch manufacturing, scientists and process engineers have to stitch relevant data sets together to develop a control model for continuous manufacturing.

Metadata are needed to help scientists retrieve data sets about the same or similar drug substance and drug products and their related batch runs to create a data-driven control model. Metadata is also required for scientists to perform process scale up effectively as well as to figure out how data from smaller scales can be utilized to improve (train/pre-train) the control model. Finally, metadata plays a pivotal role for regulatory filing as it helps ensure traceability and integrity of the data sets. Example metadata are batch identifier, drug substance, drug product, device id, time of last calibration, reference material used and batch control recipe. While the batch control recipe is the key to other data elements relevant to analytical modeling, selected data elements may be helpful for scientists to find and reuse data better. Examples of these data elements may be the critical quality attributes, critical process control parameters and key performance indicators.

The exemplary metadata elements given above are essential for process development and manufacturing control. When considering other lifecycle activities, additional metadata elements may be identified.

### 2.7. General data-driven manufacturing enterprise applications

To summarize, metadata and data objects play an important role in industrial decision making. While the use cases described above reflect various levels of use of digital object metadata for manufacturing enterprise applications, the figure below, the CRISP framework, summarizes the role of metadata in data-driven industrial decision-making processes for smart manufacturing.

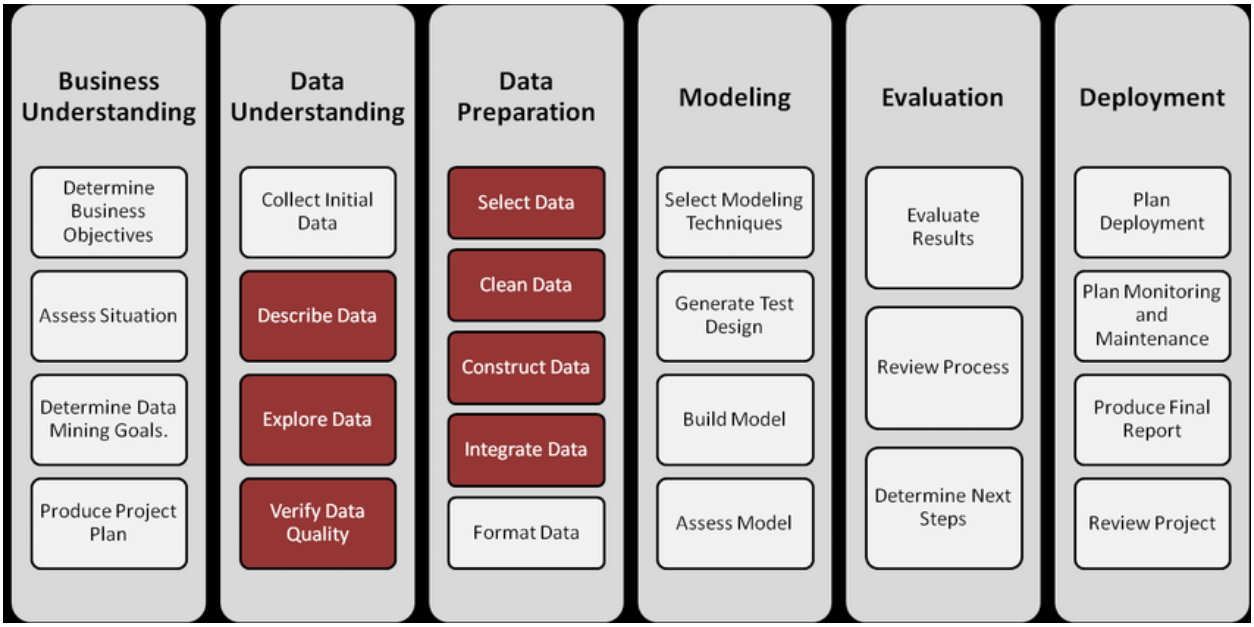


Figure 4: CRISP framework for data-driven business applications [7]

In Figure 4, among the six phases tackling data-driven problems in various applications, business understanding, and data understanding identify the business problems and the available data sets to solve the problems, which are critical in fusing data and formulating a data driven problem for decision making. Hence the information that captures the business context and the provenance constitutes the most important part for metadata. Data understanding also includes data quality evaluation, data lineage tracking, and the reference uses of the data objects, which are also essential in data fitness analysis for problem solving.

In smart manufacturing, additional interoperability is demanded to share machine learning models, during the later phases outlined in the CRISP framework. Existing standards such as Predictive Modeling Markup Language (PMML) by Data Mining Group provides syntax to represent various types of machine learning models, which also capture some metadata for the model. NIST and the Industry Ontology Foundry (IoF) has an ongoing effort to develop ontology to represent machine learning models, the contextual information of the training data and the pedigree of the model training processes [8]. Systematic identification of metadata requirements for machine learning models are necessary for the use and reuse of various ML/AI models, including large language models.

# 3. Existing Metadata Standards, Technology, and Activities

Standards in various domains help capture and structure relevant metadata for interoperability. In this Section, we review existing metadata standards, their development activities, as well as the technology for metadata management. The scope of metadata standards spans from those generally applicable to all kinds of digital objects to those specific to the research/scientific and manufacturing domains.

## 3.1. Metadata Standards for digital Archive and Library Collection

The concept of metadata, or 'data about data', can be traced back to early library cataloging systems. The development of metadata standards also started from the effort to for expressing bibliographic data. Library metadata standards have evolved significantly over the years, reflecting the changing landscape of information organization and retrieval. The development of standards such as MARC (Machine-Readable Cataloging), introduced in the 1960s, marked a significant advancement in the way libraries cataloged materials, enabling efficient data sharing and interoperability among systems. Over time, other standards emerged, including Dublin Core, which focuses on simplicity and ease of use for digital resources, and more complex frameworks like RDA (Resource Description and Access), which aims to improve cataloging practices by emphasizing user needs and the relationships between resources.

Adoption of these standards varies widely across institutions and regions. While MARC remains a staple in many traditional library systems, there is a growing shift towards more flexible and interoperable standards, particularly as libraries increasingly engage with digital collections and linked data. The adoption of standards like BIBFRAME, which seeks to transition libraries into a linked data environment, reflects ongoing efforts to modernize metadata practices. Overall, the landscape of library metadata standards is marked by both continuity and change, as libraries strive to balance established practices with emerging technologies and user expectations.

## 3.2. IEC/ISO 11179– Metadata Registry Standard

The ISO/IEC 11179 includes a set of metadata registry (MDR) standards developed by ISO JTC1 SC32 WG2, for registration of the descriptions of digital objects, including data specification, concept system, data set and (information) models. It is through these descriptions that an accurate understanding of the semantics and a useful depiction of the digital objects are found. There are six general parts in ISO/IEC 11179: Part 1 on Framework, Part 2 on Classification, Part 3 on Metamodel for common registry facilities, Part 4 on Formulation of data definitions, Part 5 on Naming principles, Part 6 on Registration. In addition, there are five newest parts created for basic attribute definitions of a metadata item, metamodel for data specification registration, metamodel for data set registration, metamodel for computable data, metamodel for concept system registration and metamodel for model registration.

Part 33 of the standards series was published in 2023, which defines the model for the registration of metadata about data sets. The figure below shows the model defined to represent data sets.

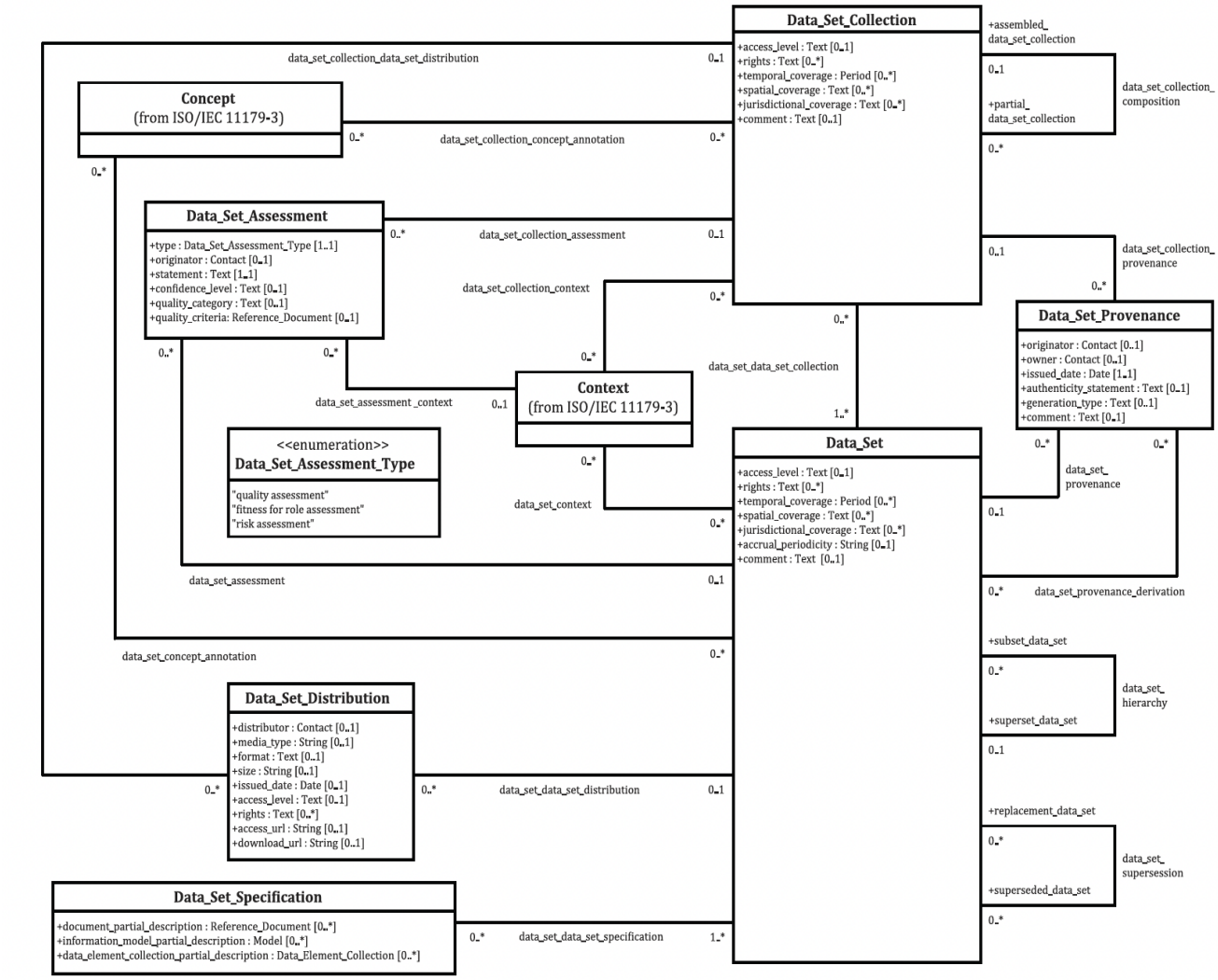


Figure 5: ISO 11179 Data set Metadata Model [7]

**Data Set has the following general attributes, including** access level, rights of type, temporal coverage start date, temporal coverage end date, spatial coverage, accrual periodicity and comments.

**Data Set has the following associations:**

— data set concept annotation uses "Concept" to annotate a data set or a data set collection so as to describe the theme or category of the data set or data set collection.

[7] https://www.iso.org/standard/78915.html

— data_set_context defines a particular program, project or business area that uses a data set, or a project or business area that is used as the context for a quality assessment of a data set. “data_set_usage” which references a Context.
— data_set_data_set_distribution including attributes such as *distributor*, *media_type*(IANA), *format, size, issued_date*, *access_level*, *rights*, access_url, download_url
— data_set_data_set_collection: the collection this data set belongs to
— data_set_specification: with an attribute of document_partial_descriptions, and associations of Data_Element_Collection that describes each data_Elements such as *data_element_precision,* data_element_meaning, data_element_domain, exemplification, Derivation_input,and derivation_output and an association of Information Model,
— data_set_hierarchy to indicate the subset_data_set or superset_data_set of this data set
— data_set_provenance includes the information of originator, issued_date, ownership_statement and generation_type
— data_set_assessment include originator the quality assessment, quality statement, category, confidence_level , quality_criteria
— data_set_supersession describes the version control including replacement_data_set Data_Set and/or superseded_data_set.

The adoption of ISO/IEC 11179 in both the commercial sector and the public sector is successful[8]. The United Nations and the US Government refer to and use the 11179 standards. 11179 is strongly recommended on the U.S. government's XML website and is promoted by the Open Group as a foundation of the Universal Data Element Framework. The Open Group is a vendor-neutral and technology-neutral consortium working to enable access to integrated information within and between enterprises based on open standards and global interoperability.

### 3.3. Geographic Information metadata

ISO 19115 is a key standard for geographic information metadata that provides a framework for documenting the quality, lineage, and other important characteristics of geospatial datasets. Part 1 of the standard outlines the metadata elements needed for geographic information, focusing on the content and structure necessary to adequately describe spatial datasets. This includes elements like identification, quality, spatial representation, and distribution, which help users understand the data's purpose, accuracy, and accessibility.

Part 2, known as ISO 19115-2, extends the original framework by providing additional metadata elements including data acquisition and processing information, specifically for imagery and gridded data. It addresses the unique aspects of remote sensing and raster datasets, ensuring that users can effectively evaluate and utilize these types of geographic information. Part 3, ISO 19115-3, further refines the standard by incorporating metadata for the data lifecycle and data processing, focusing on the management and preservation of geospatial data over time.

[8] https://en.wikipedia.org/wiki/ISO/IEC_11179

Together, these parts create a comprehensive system for metadata that enhances data discovery, sharing, and interoperability in the field of geographic information.

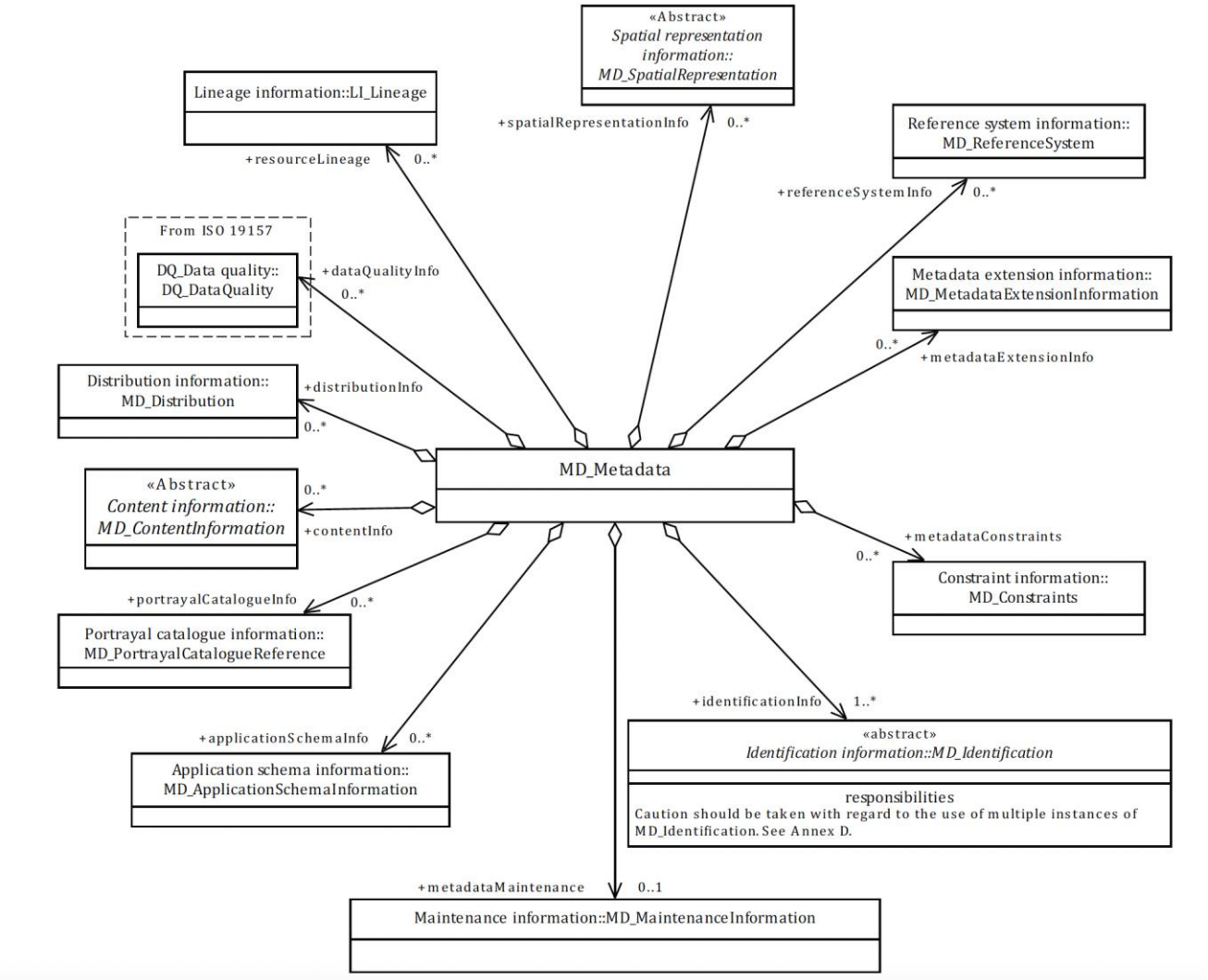


Figure 6: Geographic Information Metadata Model

### 3.4. Dublin Core Metadata Element Set

"The Dublin Core", also known as the Dublin Core Metadata Element Set, is a set of fifteen "core" elements (properties) that are broad and generic, usable for describing a wide range of resources. The fifteen core element is part of a larger set of metadata vocabularies and technical specifications maintained by the Dublin Core™ Metadata Initiative (DCMI). The full set of vocabularies, DCMI Metadata Terms, also includes sets of resource classes, vocabulary encoding schemes, and syntax encoding schemes, in Figure 7. The terms in DCMI vocabularies are intended to be used in combination with terms from other, compatible vocabularies in the context of

application profiles and on the basis of the DCMI Abstract Model. Table 1 shows the definitions of the Dublin Core metadata elements.

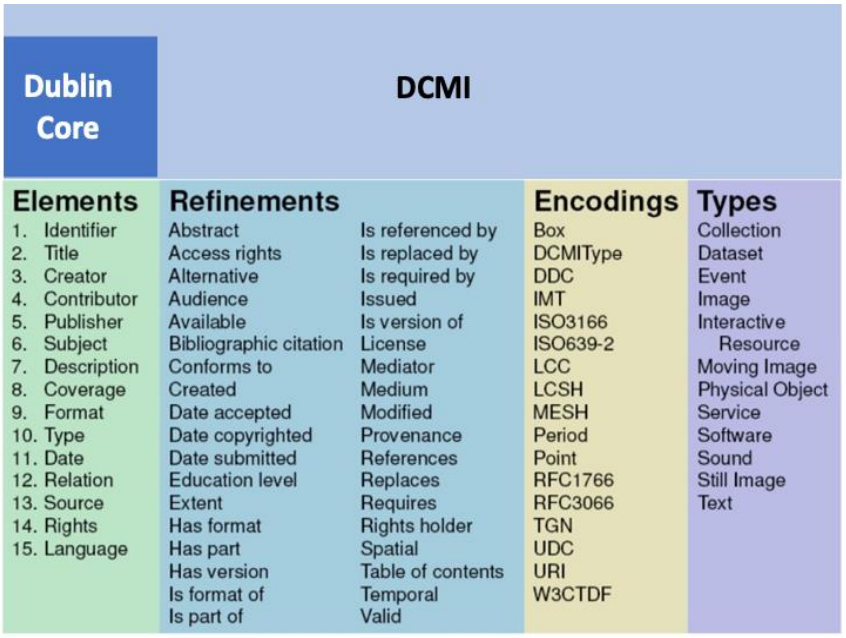


Figure 7: DCMI Metadata Terms [[10]]

Table 1: Dublin Core Data Elements and Definitions

| Name | Definition |
|---|---|
| contributor | An entity responsible for making contributions to the resource. |
| coverage | The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. |
| creator | An entity primarily responsible for making the resource. |
| date | A point or period of time associated with an event in the lifecycle of the resource. |
| description | An account of the resource. |
| format | The file format, physical medium, or dimensions of the resource. |
| identifier | An unambiguous reference to the resource within a given context. |
| language | A language of the resource. |
| publisher | An entity responsible for making the resource available. |
| relation | A related resource. |
| rights | Information about rights held in and over the resource. |
| source | A related resource from which the described resource is derived. |
| subject | The topic of the resource. |

| | |
|---|---|
| title | A name given to the resource. |
| type | The nature or genre of the resource. |

### 3.5. Data Catalog Vocabulary (DCAT)

Data Catalog Vocabulary (DCAT) is W3C recommended standard for data catalogue representation, including models to describe data sets and data services [[11]]. Leveraging on the terms defined by Dublin Core and other metadata standards, DCAT is designed using RDF triples to facilitate interoperability between data catalogs published on the Web and facilitate the consumption and aggregation of metadata from multiple catalogs. The current version of DCAT, DCAT 3, has a structure shown in the figure below.

DCAT has been profiled by various countries or regions to express constraints or usages of DCAT classes and properties to meet to their local requirements. In the United States, DCAT-US is the schema used by some of the data catalogs reported to the https://data.gov/ website[[12]]. DCAT-US is based on version 1.1 of the DCAT, and DCAT-US submissions are in JSON while DCAT is RDF. However, it is possible to interconvert between these various data representation syntaxes. The DCAT-US 3.0 Profile (DCAT-US 3.0) is under development now [[13]]. Other DCAT profiles include DCAT-AP, DCAT-AP-DK, DCAT-Australia etc.

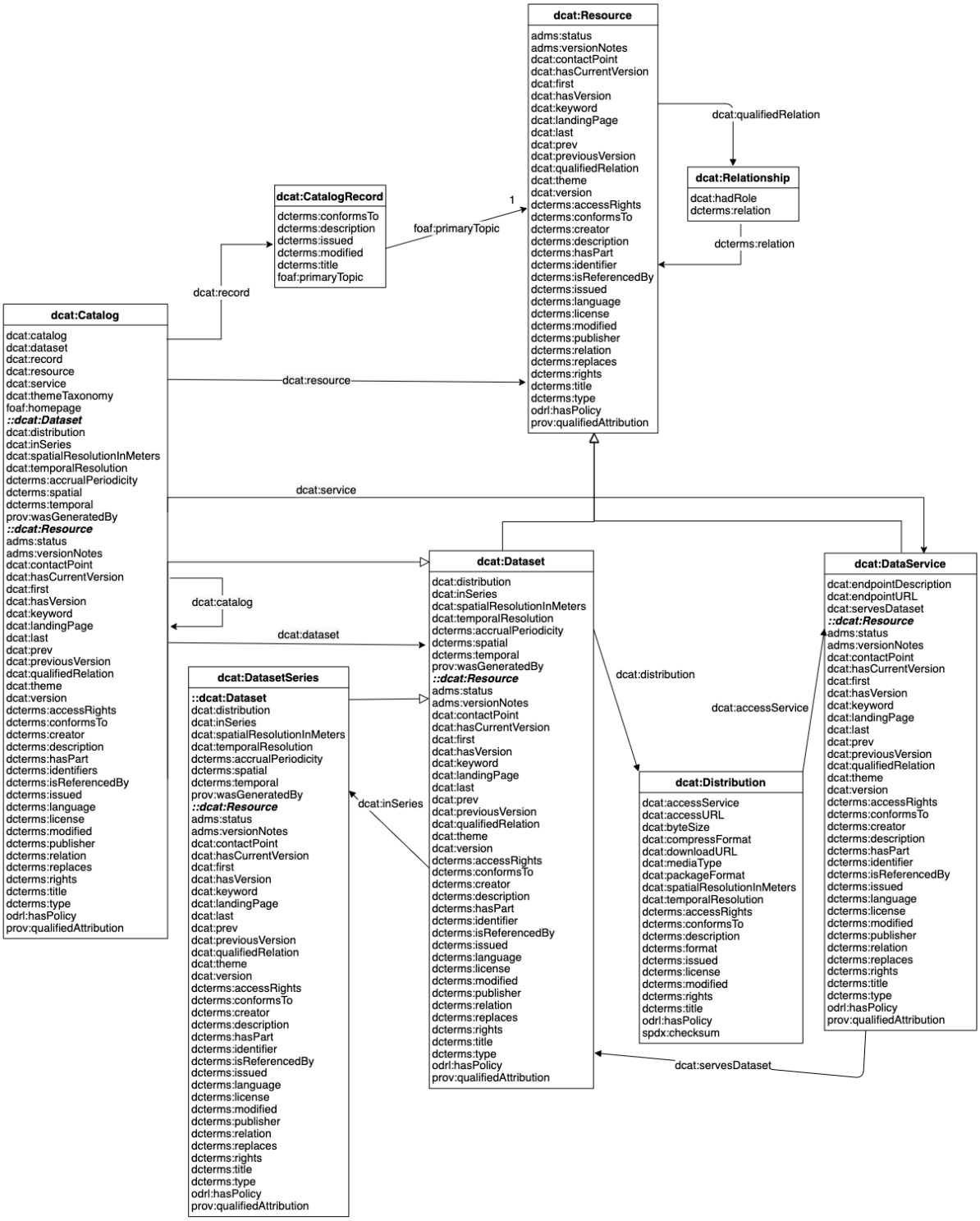


Figure 8: DCAT Version 3 Data Model [9]

[9] https://www.w3.org/TR/vocab-dcat-3/

### 3.6. Content Management from OASIS

**Content Management Interoperability Services** (**CMIS**) from OASIS specifies a core data model that defines the persistent information entities that are managed by a content management system for typed files/documents and folders, and policy for management with generic properties that can be set or read. The data model can be adopted to define structural metadata for complex digital objects. The figure below summarizes the CMIS data model [14].

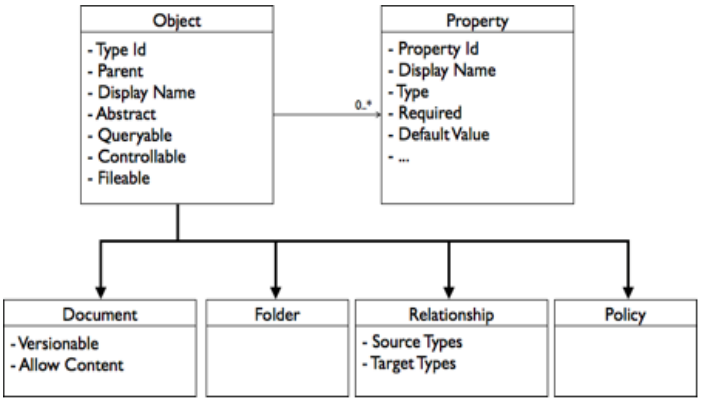


Figure 9: CMIS Data Model

CMIS has the potential to do for Enterprise Content Management what SQL did for relational databases [[15]].

### 3.7. Common Warehouse Metamodel from OMG

The **common warehouse metamodel** (**CWM**) from OMG defines a specification for modeling metadata for relational, non-relational, multi-dimensional, and most other objects found in a data warehousing environment [[16]]. The purpose of CWM is to provide a means by which tools may exchange commonly understood descriptions of relational schemas, record structure, generic representation of a multidimensional database and XML metamodel as well as metadata model for business/process/operation context, transformations between data sources, and applications for data mining/analytical modeling. CWM also allows for the possible inclusion of tool-specific extensions.

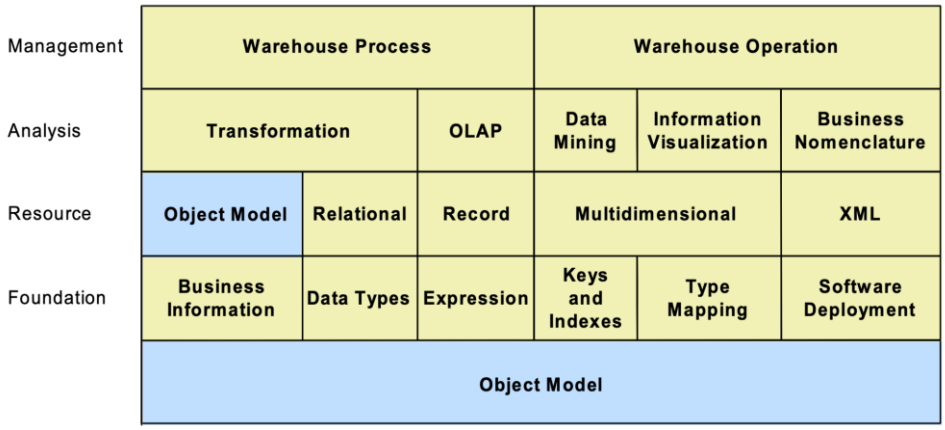


Figure 10: CWM Metamodel

CWM Relational metamodel is based on the SQL-99 standard and is compatible with Java Database Connectivity. The model is centered on the notion of column set and the subsequent derivation of table, view, and query from the column set. However, the concepts of keys (unique keys, foreign keys) and indexes are defined in the CWM as CWM Foundation meta classes, so they have general applicability to other data models within the CWM, not just the CWM Relational.

### 3.8. IEC/ISO Common Data Dictionary

The IEC Common Data Dictionary (IEC CDD) is a metadata registry providing product classification and formalized product descriptions that can be used in various industrial/technical domains based on the methodology and the information model of IEC 61360 series. The information model captures unambiguously identifiable classes and properties, and their relations, using commonly accepted terminology and definitions based on accepted sources such as IEC International Standards, other International Standards, industry standards, or public authorities. The hierarchies of concepts/properties enable users to appropriately characterize their products and services, relevant conditions and constraints, if necessary, on possible values of characteristics, technical representation of concepts including units and data types, and their identification.

The IEC CDD hosts different product classifications (based on international standards) for

- process automation equipment (based on IEC 61987),
- low voltage switchgear and control gear (based on IEC 62683),
- electro-electronic components (provided by IEC TC47),
- optics (based on ISO 23584),
- measuring instruments (based on IEC ISO 13584-501), and
- environmental declaration (based on IEC 60721).

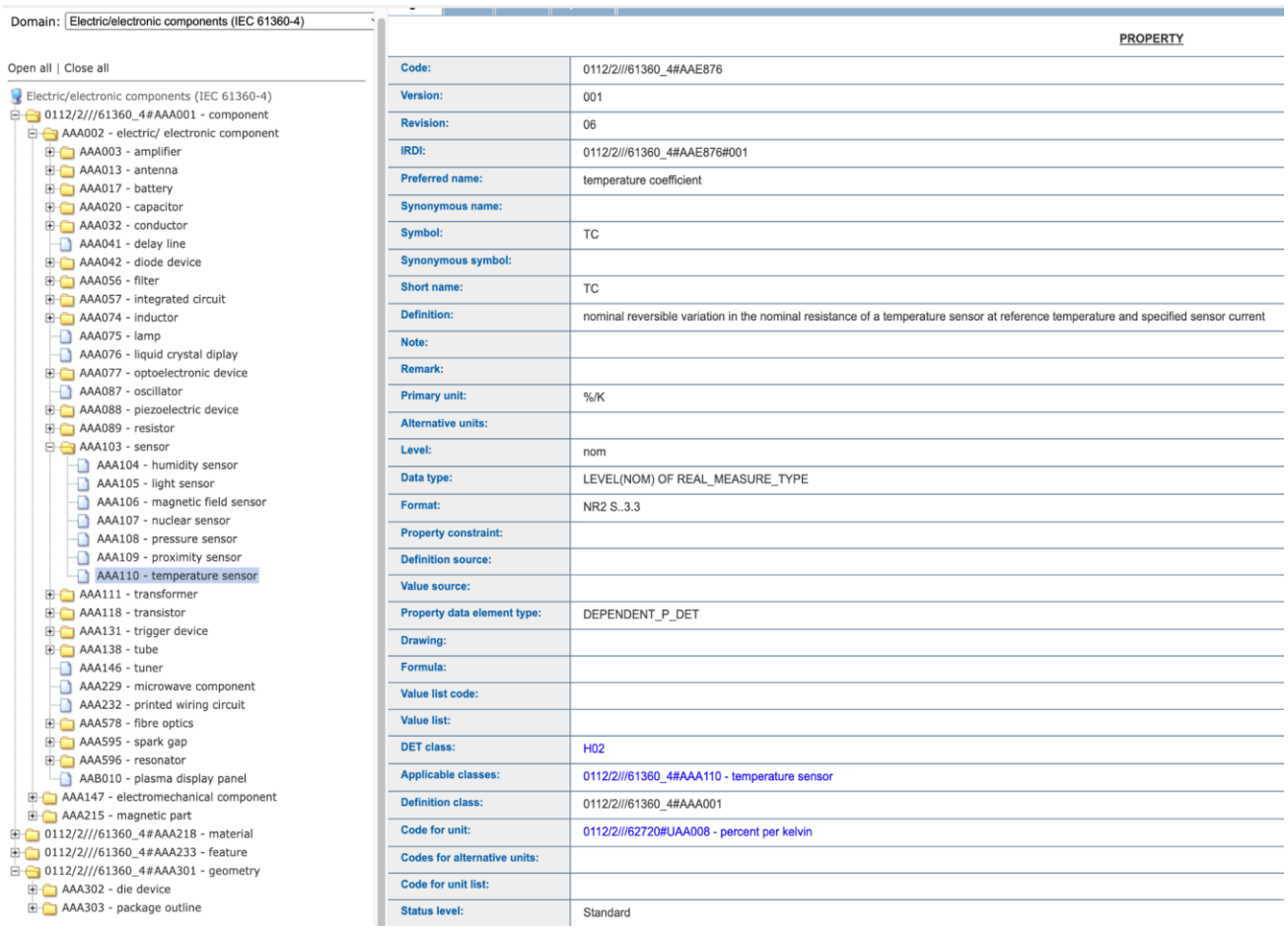


Figure 11: A snapshot of IEC CDD

## 3.9. MTConnect

MTConnect defines a data model which supports manufacturing data integration. It separates Streams Information Model from Device information model. The metadata for MTConnect is defined by the Data Entity type which has an XML schema **MTConnectDevices XML** to describe data that can be reported by a piece of equipment and that are associated with Device and Component Structural Elements. In Figure 12, the Data Item type describes the data that can be reported by a piece of equipment in the MTConnectDevices document, the actual data values are provided in the Streams Information Model.

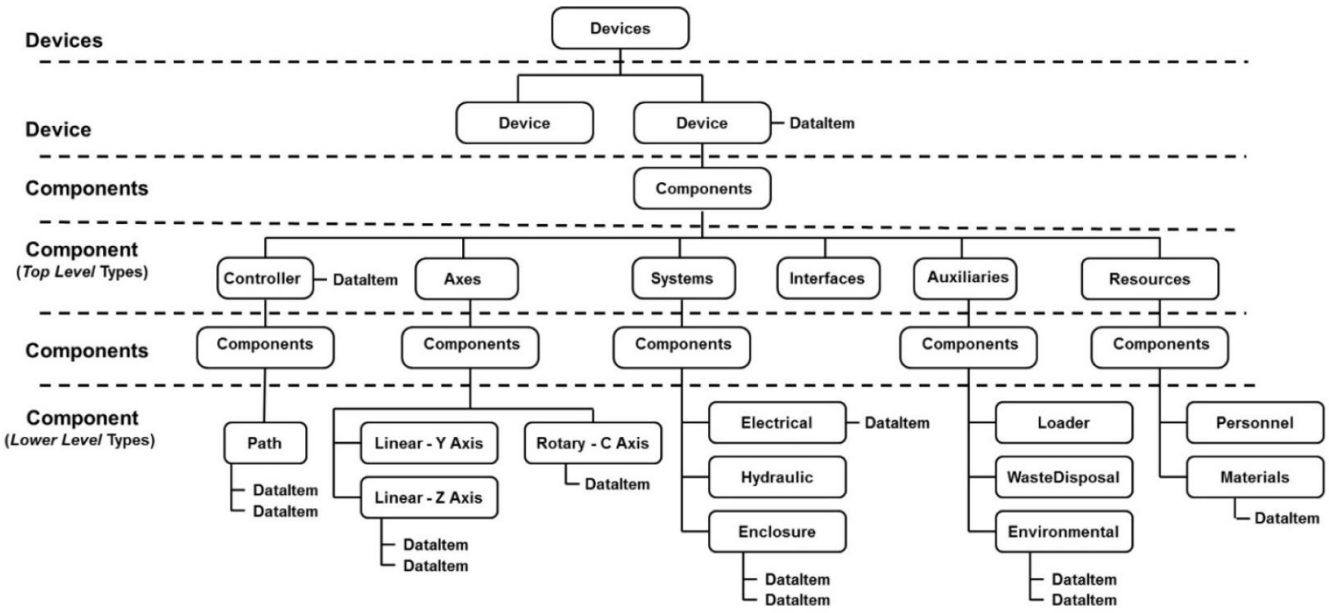


Figure 12: MTConnect DataItem defined in the device model from MTConnect Version 1.6

The table below captures all the metadata elements defined in MTConnect Version 1.6 for DataItem.

Table 2: MTConnect data Item metadata elements

| Name | Definition |
|---|---|
| **name** | The name of the data item, as an additional human readable identifier for this data item in addition to the id |
| **id** | The unique identifier for this element. id is a required attribute. |
| **type** | The type of data being measured. type is a required attribute, e.g., POSITION, VELOCITY, ANGLE, BLOCK, and ROTARY_VELOCITY. |
| **subtype** | A sub-categorization of the data item type, eg, ACTUAL or COMMANDED for POSITION. Not all type attributes have a subType. |
| **statistics** | Describes the type of statistical calculation performed on a series of data samples to provide the reported data value. statistic is an optional attribute. Examples of statistic are AVERAGE, MINIMUM, MAXIMUM, ROOT_MEAN_SQUARE, RANGE, MEDIAN, MODE, and STANDARD_DEVIATION. |
| **units** | The unit of measurement for the reported value of the data item. |
| **nativeUnits** | The native units of measurement for the reported value of the data item. |
| **nativeScale** | nativeScale MAY be used to convert the reported value to represent the original measured value. |

| | |
|---|---|
| **category** | Specifies the kind of information provided by a data item, with options of {Sample, Event, Condition}. |
| **coordinateSystem** | The available values for coordinateSystem are WORK and MACHINE. |
| **compositionID** | The identifier attribute of the Composition element that the reported data is most closely associated. |
| **sampleRate** | The rate at which successive samples of a data item are recorded by a piece of equipment. |
| **representation** | Representation defines the unique format for each set of data. |
| **significantDigit** | The number of significant digits in the reported value. |
| **discrete** | An indication signifying whether each value reported for the Data Entity is significant and whether duplicate values are to be suppressed. |
| **Source** | Source is an optional XML element that identifies the Component, DataItem, or Composition representing the area of the piece of equipment from which a measured value originates. |
| **Constraints** | Constraints is an optional container that provides a set of expected values that can be<br>reported for this DataItem. |
| **Filters** | An optional container for the Filter elements associated with this DataItem element. |
| **InitialValue** | InitialValue defines the starting value for a data item as well as the value to be set for the data item after a reset event. |

Data Representation Type can be *VALUE* which captures the measured value of the sample data, *DATA_SET* which structures the data as a set of key-value pairs; *TIME_SERIES* which defines a series of sampled data with fixed period.

## 3.10. connectSpec Message Model and Metamodel

### 3.10.1. connectSpec Message Model

connectSpec data models, previously named OAGIS, are enterprise interoperability standards that specifies a protocol for passing data in and out of business applications. Under the OAGIS standard, related information is bundled into a structure called a business object document (BOD). The high-level architecture of BODs is illustrated graphically in Figure 13. A BOD contains two areas: one devoted to application and the other to business and engineering data. The Application Area contains information needed by the communication infrastructure to deliver and track the message. It also contains the business process context information that the receiver application may need to process the data correctly. Examples include the engineering or business process it is a part of, and whether it is a production or a test message. The Data Area contains the message content, which comprises Verbs and Nouns. The Verb indicates the action to be

performed on the Nouns; the Noun is an object conveying business or engineering data or instruction to be acted upon or performed by the receiver application.

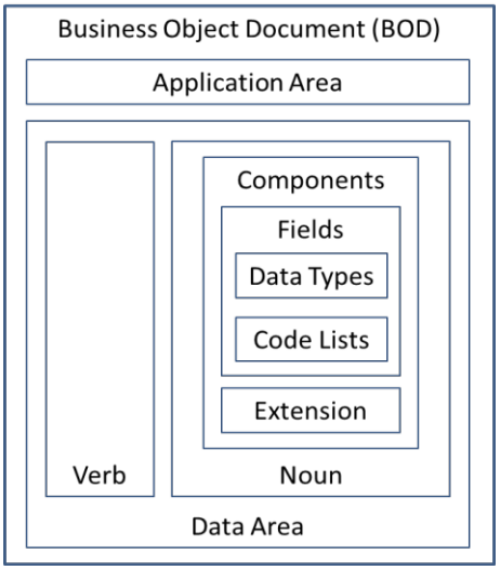


Figure 13. OAGIS message model/architecture.

Nouns are made up of reusable elements – Components. connectSpec is currently defined using the Core Component Specification (CCS) meta model, which is an ISO 15000-5 standard [[17]]. The standard is currently managed using the NIST/OAGi developed software tool called Score, which also known as connectCenter. In a nutshell, there are 7 types of Core Components for using Score. They are as follows: Aggregate Core Component (ACC), Association Core Component Property (ASCCP), Association Core Component (ASCC), Basic Core Component Property (BCCP), Basic Core Component (BCC), Business Data Type (BDT), Code List (CL), and Agency Identifier List.

Although BODs begin with the word 'business', they also support transactions in various engineering functional areas throughout a single enterprise or across multiple enterprises. Example transactions include design, manufacturing, supply chain, finance, sales, and accounting. However, today's data exchange using connectSpec is not limited by BODs. With the Score tool, messages can be constructed from any Nouns or Components.

connectSpec has adopted a model-driven approach (MDA), which separates the models from language-specific implementations. connectCenter is a browser-based tool for developing and managing data standards, creating standard profiles, and expressing the profiles in ways that make developers most productive. And instead of working with full schemas of those messages and objects that can include hundreds of thousands of data element the tool generates only the necessary data elements (and documentation) in a syntax of user choice.

### 3.10.2. ConnectSpec Data set Metadata model

Metadata component models were created for OAGIS 10.7. Figure 14 shows an inheritance view of connectSpec. The properties defined for Metadata Base component type leverage Dublin core. Additional properties were identified including the associated process category, quantity kind, device source, collection time information and geospatial information for the data collection etc. Further specialization can be identified for Additive manufacturing data set metadata modeling.

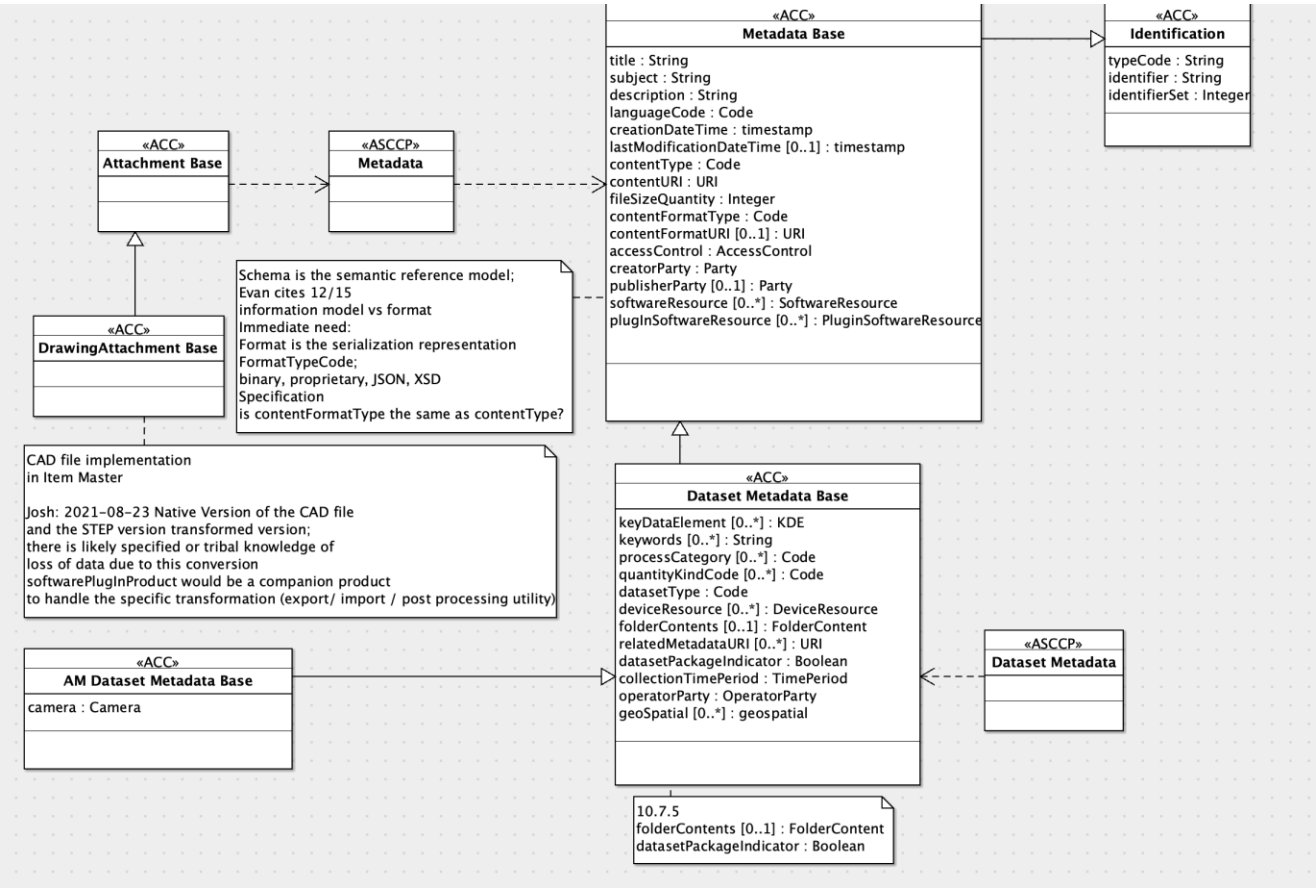


Figure 14: connecSpec Metadata Model

### 3.11. QIF

QIF (Quality Information Framework) [[18]] is a quality framework used on shop-floor manufacturing devices, especially computer-aided processes and engineering applications, shown in Figure 15. It's an ISO standard and uses a CAD neutral file format represented in XML form. QIF offers a feature-based, characteristic-centric ontology of manufacturing quality information. Through the use of both QIF Plans and QIF Results, as shown in the figure below under 6 and 2 respectively, quality information can be applied before and after a product is manufactured. QIF covers the full spectrum of features and characteristics of PMI (Product

Manufacturing Information), including all of the measurements related to each individual characteristic and the measurements within allowed tolerances to ensure proper quality result information is produced.


Figure 15: Quality Information Framework [19])

Additionally, QIF covers the resources involved in manufacturing (3 above), as well as the rules applied during manufacturing. It's also important to note that QIF has its own custom model-based definition representation, similar to that found in STEP AP 242. This is shown above in section 1. This allows for interoperability with AP 242, as well as various CAD file formats. Finally, QIF provides support for a statistical framework for better process model control, as shown in section 7. QIF is part of the big data movement, enabling companies all over the world to bring the power of information together for improved efficiency and quality, reduced manufacturing costs, and an overall increase in safety and sustainability.

QIF digital object is comprehensive, as a container, which holds a group of data together, with all the relevant data along a digital thread present in one location, and the characteristics can be properly identified and communicated between people and machines. The QIF digital objects, including design, test results, statistical analysis report, can be easily traced back to the resources and processes.

### 3.12. MIMOSA CCOM Model

MIMOSA CCOM (Common Conceptual Object Model) serves as an information model for the exchange of asset lifecycle information required for operation and maintenance of plants and complex facilities. The information model covers engineering, asset, configuration, lab testing, and operation and condition information, among others.

CCOM follows prior MIMOSA standards of the layered architecture specified by ISO 13374-2, which separates implementation models, such as XML Schema and JSON schema, from the logical model (defined in UML).

Of particular interest in the area of metadata modeling in MIMOSA CCOM is the “Measurement” class, including various types of measurement data such as single measurements, time series, images, etc, as shown in Figure 16. Metadata are defined to associate each measurement to the operating context in which that measurement was taken, including the virtual or physical measurement location, the system or ‘measurement source that collected and aggregated the measurement; the transducer that made the measurement, the asset associated with the measurement at the time the measurement was taken, and/or the segment associated with the measurement at the time it was taken. The term segment is a generic term representing some logical element of a model such as a functional location in a plant design, an element of a business process model or its instance, or a logistics/supply-chain segment, etc. The exact segment type is defined through an appropriate industry specific reference data set. Furthermore, the measurement location, asset, and segment are connected to a much broader context, for example, the entire process plant and its design considering a temperature transmitter in a plant. This association to the context in which a measurement was taken supports the traceability of measurement data throughout an interconnected set of systems and supports advanced analytics, such as those required by CBM and CBP.

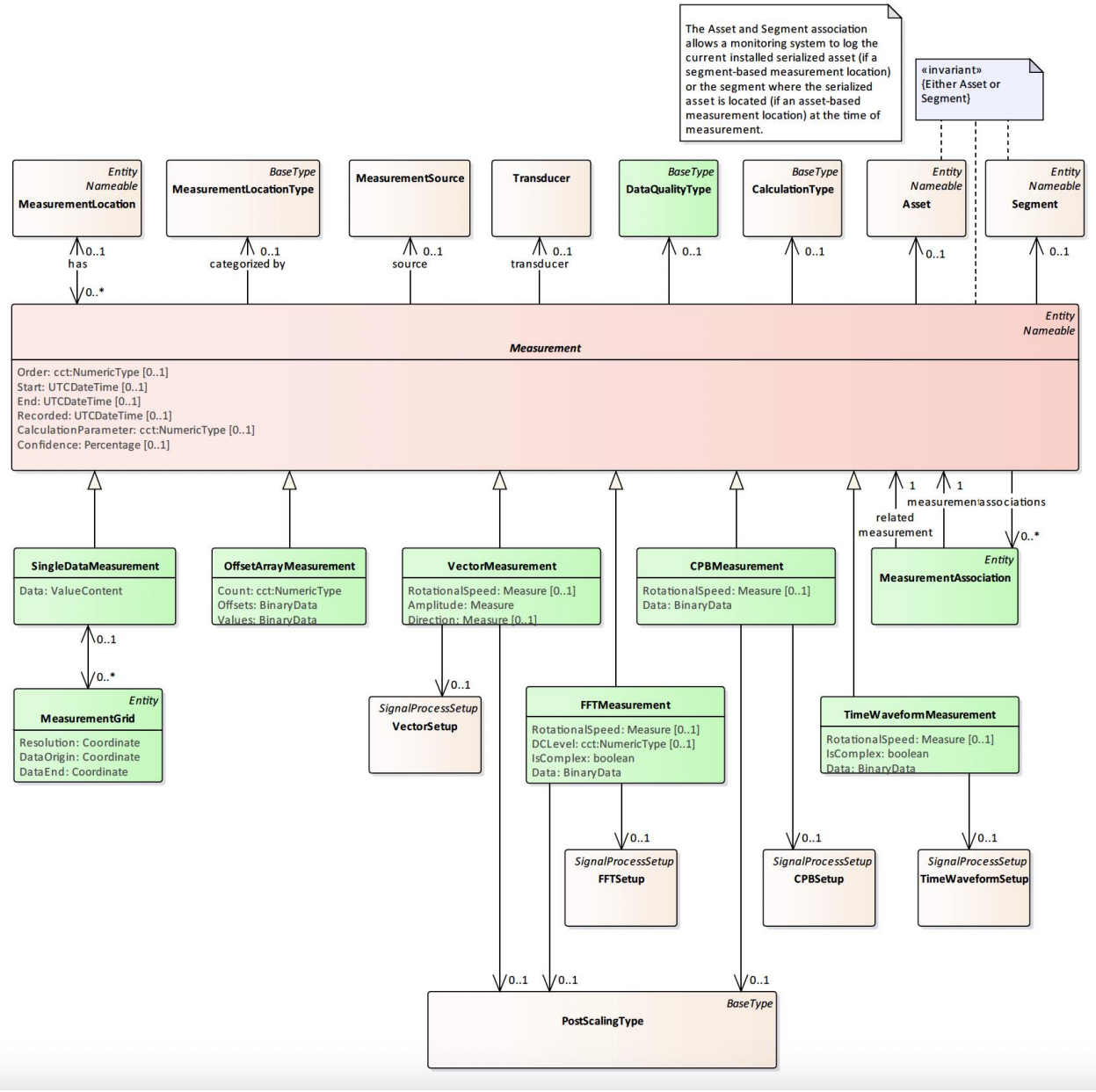


Figure 16: MIMOSA CCOM Measurement Data Model[10]

## 3.13. Ontology for Technical Metadata

Sensor, Observation, Sample, and Actuator ontology (SOSA), developed by W3C and OGC, provides a flexible yet coherent framework for metadata modeling for measurement data sets, covering the entities, relations, and activities involved in sensing, sampling, and actuation [20].

[10] https://www.mimosa.org/mimosa-ccom/ccom-and-other-xml-schemas/

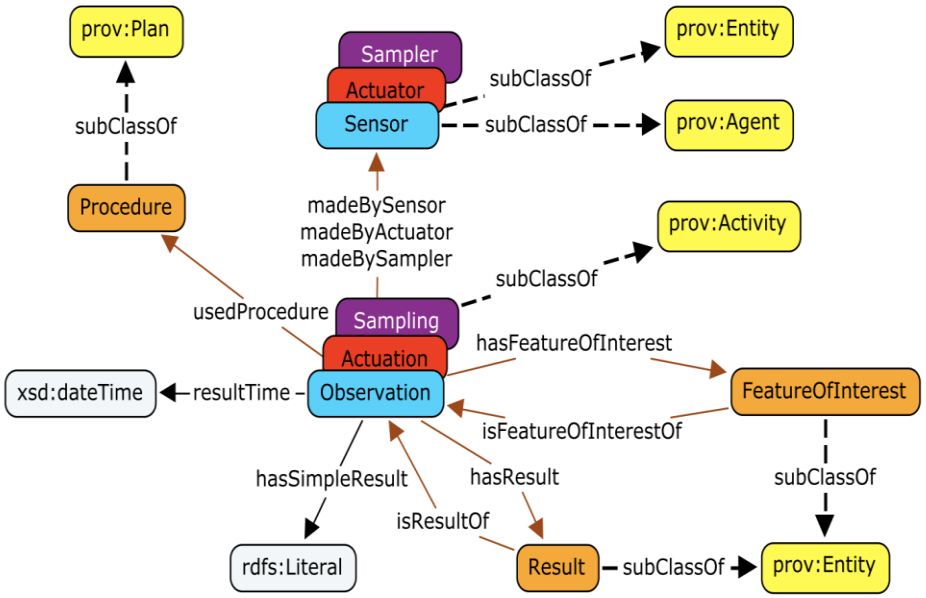


Figure 17: The core structure of SOSA [20]

Instead of focusing on data set – "*Result* "in Figure 17, SOSA takes an event-centric perspective and revolves around observations, sampling, actuations, and procedures. The model can be applied to multiple domains to describe data sets generated to capture "FeaturesOfInterest" of materials or processes.

In addition to SOSA, the QUDT (Quality, Unit, Dimension and Type) ontology was originally developed for the NASA Exploration Initiatives Ontology Models (NExIOM) project by defining a set of vocabularies representing the various quantity and unit standards (qudt.org). Figure 18 below shows the concept model of QUDT.

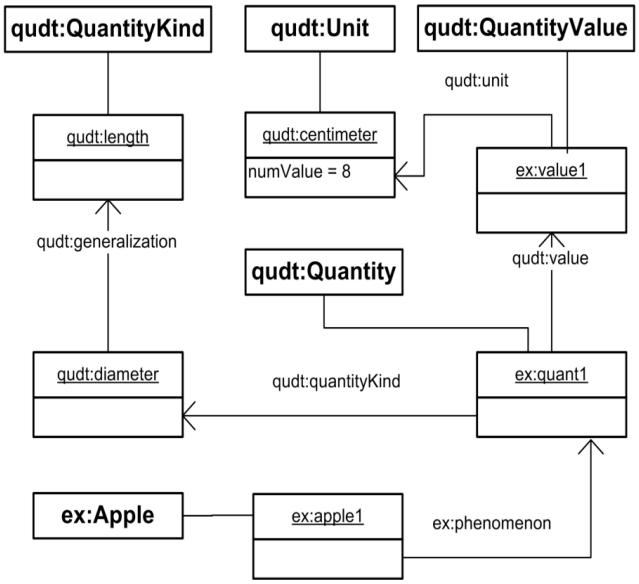


Figure 18: The core structure of QUDT [[19]]

### 3.14. Industrial Ontologies Foundation

The Industrial Ontologies Foundry (IOF) is a unit of the Open Applications Group, Inc. Its aim is to develop a suite of reference ontologies that span the entire digital manufacturing domain. The current release of IOF ontology suite currently includes Core ontology, Supply Chain Ontology, and Maintenance Ontology, but IOF has other working groups that plan to publish material, product service system, production planning and control, system engineering, and MTConnect ontologies. IOF release also includes a metadata model module called Annotation Vocabulary. The Annotation Vocabulary provides metadata model for the ontology, and its classes, and properties. The Annotation Vocabulary reuses elements from Dublin Core and W3C SKOS. It is very comprehensive and reused by other ontology development bodies such OMG and the Industrial Data Ontology group[ref]. Unique metadata element in the Annotation Vocabulary includes Natural Language Definition, First Order Logic Definition, Semi-formal Natural Language Definition, First Order Logic Axiom, Semi-formal Natural Language Axiom, Is Primitive, and Primitive Rationale, Abbreviation, Adapted From, Example, Explanatory Note, and more. Finally, classes and properties in IOF ontologies may be used to create semantic network of metadata for digital objects of interest described in prior sections. The resulting semantic network may allow for reasoning to discover or understand relationships between data sets described by the metadata.

### 3.15. FAIR and Research Data Metadata template

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in *Scientific Data* [[21]]. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets for the research societies. Since the conception, the principle has been adopted by multiple disciplinary areas in the community. Details of these principles may be found at [1] and is summarized in the table as follows [[22]].

Table 3: The FAIR Principles [23]

| **Findable** | **Accessible** |
|---|---|
| Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of data sets.<br>F1. (Meta)data are assigned a globally unique and persistent identifier<br>F2. Data are described with rich metadata (defined by R1 below)<br>F3. Metadata clearly and explicitly include the identifier of the data they describe<br>F4. (Meta)data are registered or indexed in a searchable resource | After the user finds the required data, she/he needs to know how it can be accessed & authenticated.<br>A1. (Meta)data are retrievable by their identifier using a standardized communications protocol<br>A1.1. The protocol is open, free, and universally implementable<br>A1.2. The protocol allows for an authentication and authorization procedure, where necessary<br>A2. Metadata are accessible, even when the data are no longer available |
| **Interoperable** | **Reusable** |
| The data must be integrated with other data. The data needs to interoperate with applications, etc.<br>I1. (Meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation<br>I2. (Meta)data use vocabularies that follow FAIR principles<br>I3. (Meta)data include qualified references to other (meta)data | In order to optimize the reuse of data, metadata and data must be well-described so that it can be replicated and/or combined in different settings.<br>R1. (Meta)data are richly described with a plurality of accurate and relevant attributes<br>R1.1. (Meta)data are released with a clear and accessible data usage license<br>R1.2. (Meta)data are associated with detailed provenance<br>R1.3. (Meta)data meet domain-relevant community standards |

The Research Data Alliance offers a Directory of Metadata Standards that are developed based on the FAIR principles [[23]]. Some discipline-specific metadata standards are described here.

**Wind energy** researchers added 5 additional metadata entries on top of the Dublin Core to describe data set generated for wind study, including “External Conditions”, “Activity”, “Instrument”, “Model”, “Material” to make the taxonomy more useful within their field, as shown in Figure 19.

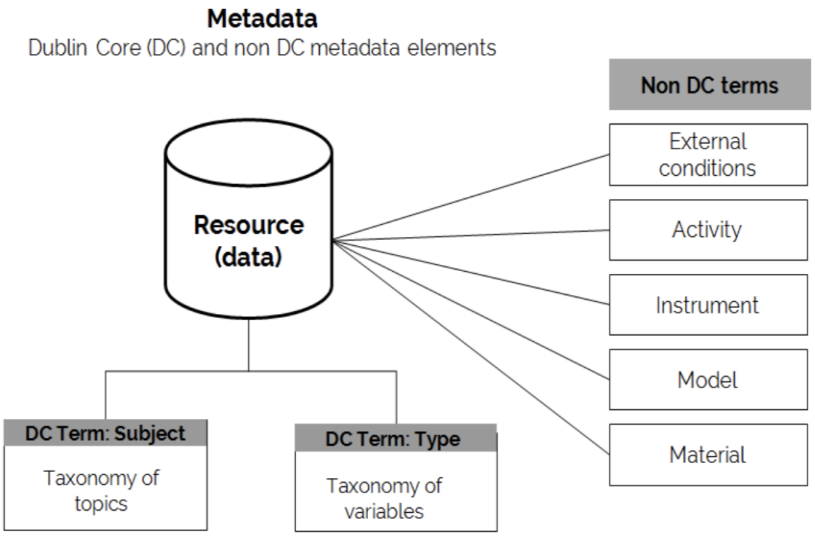


Figure 19: Wind Energy FAIR metadata elements [25]

**NeXus** reis another effort by an international group of scientists motivated to define a common data exchange format for neutron, X-ray, and muon experiments [[24]]. NeXus is built on top of the scientific data format HDF5 and adds domain-specific rules for organizing data within HDF5 files in addition to a dictionary of well-defined domain-specific field names. The NeXus data format has three purposes:

1. *raw data*: NeXus defines a format that can serve as a container for all relevant data associated with a scientific instrument or beamline. This is a very important use case. This includes the case of streaming data acquisition, where time stamped data are logged.
2. *processed data*: NeXus also defines standards for processed data. This is data which has undergone some form of data reduction or data analysis. NeXus allows storing the results of such processing together with **documentation** about how the processed data was generated.
3. *standards*: NeXus defines standards in the form of *application definitions* for the exchange of data between applications. NeXus provides standards for both raw and processed data.

NeXus data files contain four types of entity: groups, fields, attributes, and links.

- Groups - Groups are like folders that can contain a number of fields and/or other groups.
- Fields - Fields can be scalar values or multidimensional arrays of a variety of sizes (1-byte, 2-byte, 4-byte, 8-byte) and types (characters, integers, floats). Fields are represented as HDF5 *data sets*.

- Attributes - Extra information required to describe a particular group or field, such as the data units, can be stored as a data attribute. Attributes can also be given at the file level of an HDF5 file.
- Links - Links are used to reference the plottable data when the data is provided in other groups.

**Frictionless Data** defines a data package for describing a coherent collection of data in a single 'package' [[26]]. It provides the basis for convenient delivery, installation and management of data sets, including metadata that describes:

a. the structure and contents of the package
b. Resources such as data files that form the contents of the package

**DIF** is a de-facto standard used to create directory entries which describe a group of scientific data developed by NASA [[27]]. A DIF consists of a collection of fields which detail specific information about the data. Five fields are required in the DIF, including Directory Entry Identifier, Directory Entry Title. The others expand upon and clarify the information that is necessary for users to decide whether a particular data set would be useful for their needs. Other fields include: Data Set Citation, Personnel, Discipline, Keyword, Sensor Name, Source Name, Temporal Coverage, Paleo Temporal Coverage, Data Set Progress, Spatial Coverage, Location, Data Resolution, Project, Quality, Access Constraints, Use Constraints, Data Set Language, Originating Center, Distribution, Multimedia Sample, Reference, Related URL, Parent DIF, IDN Node, DIF Creation Date, Last DIF Revision Date, DIF Revision History and Future DIF Review Date.

In addition, there are **Data Documentation Initiative (DDI)** for survey and observational data description (Social Sciences), and the Text Encoding Initiative (TEI) for textual data description (Digital Humanities).

### 3.16. Commercial metadata models

#### 3.16.1. Google Storage Metadata Model

Metadata in Google Storage identifies properties of the object, as well as specifies how the object should be handled when it's accessed [[28]]. Metadata exists as *key:value pairs*.

Fixed metadata include Access control, Cache-Control ( e.g.,public, private, no-cache, no-store, max-age, no-transform), Content-Disposition (attachment, inline), Content-Encoding (gzip), Content-Language (defined using ISO 639-1 language codes), Content-Type (specifies that Media Types listed by the Internet Assigned Numbers Authority), Custom-Time (user-specified date and time), Object holds (to place object hold), Generation, Metageneration and Updated.
Users can define their own metadata. To create custom metadata, you specify both a key and a value.

### 3.16.2. Amazon AWS metadata definitions

Amazon S3 defines a set of system metadata for each object stored in a bucket. Amazon S3 processes the following system metadata: Cache-Control, Content-Disposition, Content-Length, Content-Type, Last-Modified, ETag, server-side-encryption, checksum, version-id, delete-marker, storage-class, website-redirect-location, server-side-encryption-aws-kms-key-id, server-side-encryption-customer-algorithm, and tagging. For more information, see Amazon S3 user guide [[29]].

Similar to Google, uploading an object to AWS, users can also assign metadata to the object. You provide this optional information as a name-value (key-value) pair. But allowing for an infinite extension of additional attributes makes interoperation difficult.

## 3.17. Metadata Management

According to Data Management Body of Knowledge (DMBOK2), Metadata Management is the planning, implementation, and control activity that enables access to high-quality, integrated metadata [[30]]. Presently, manufacturing enterprises use data and metadata management systems implemented either as a single enterprise software package (e.g., data catalog, data intelligence platform) or as a combination of several complementary open-source solutions. The final goal is to achieve better data management, anchored on good metadata management.

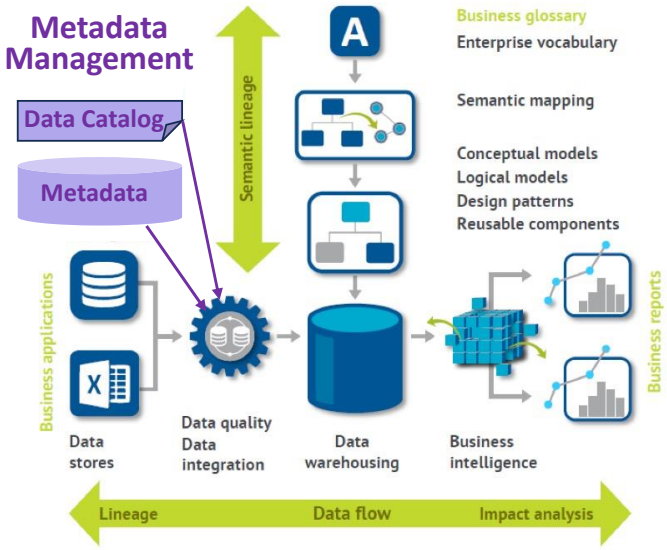


Figure 20: Metadata management for data value generation (Adapted from [32])

In the context of data inventory management, managing the metadata refers to the process of keeping up-to-date versions of the definition of all data sets that exist in the organization and providing users with an easy way to use it. The following software functions of metadata management are usually found.

**Data Discovery** is the core function of every metadata management system (e.g., data catalog software) that works with big data systems. From a software architecture perspective, this refers to the capabilities to create connectors to a wide range of data storage systems (RDBS, NoSQL Storage, Cloud, Sheets, etc.), automatically index all data resources, load the metadata about them, and schedule synchronization with various data storages.
**Data set Searching** is the function that relies upon the Data Discovery to help the data users find the right data set more efficiently. It takes into consideration metadata and crowd-sourced knowledge gained by the data users. It can enhance results by taking into account, for example, the frequency of usage.
**Data/Metadata Curation** refers to creating, organizing and managing a collection of data sets and their metadata to meet the application specific needs and interests of a specific group of people. For example, organizing and managing metadata about all data assets to support data warehousing for business intelligence. Organizing metadata for data sets easy to find, understand, and access is one of the purposes of data curation.

To support integrated data analytics, today's metadata management software usually offers an extended set of functions, including:

**Cross-dataset Querying** is recognized as the feature of data management systems that help analysts conduct ad hoc analysis. Sethi et al. have provided an open source distributed query engine that supports much of the SQL analytics on top of the various data sources, i.e. Hadoop data warehouses, RDBMSs, NoSQL systems, and stream processing systems [[32]]. Such a system can be used both for ad hoc analytics and ETL jobs that prepare the data for further usage. Some of the enterprise data management solutions offer the function of querying on top of the raw data. Even though this function makes the system more effective, it can cause serious issues in the speed and responsiveness of the system if it's not set up properly, i.e., if the database systems that serve operational needs can be directly queried for the analytical purposes.
**Data Access Interface** allows users, instead of handling many different DB technologies in which raw data is stored, access and authorization procedures, to use a single system to retrieve any data set and long-term synchronization with the raw sources. Previous functions deal mainly with providing easy access to data, while this function contributes to both this and the interoperability of the system, which is recognized as a complex challenge that influences an organization's performance and the development of data products. [[33]]
**Data Lineage** or provenance refers to the job of describing where data came from, how it was derived, and how it was updated over time. Lineage is exploited in tasks ranging from data verification in curated database [[34]] to confidence computation in probabilistic databases [[35]]. For instance, for data warehouses, it is important to be able to find a lineage of anomalous view data to identify incorrect database data [[36]], and confidence computation is important for AI products developed on top of the probabilistic database. Ikeda and Widom have formalized and categorized lineage implementations in their Data Lineage survey [[37]]. Enterprise data management systems usually offer the part of this function, and implementers can use the authors' survey to understand which level of lineage is part of the interests for the company.

### 3.18. OPEN-SOURCE metadata/data catalog tools

Under the Linux Foundation AI and Data, Amundsen is established as a data discovery and metadata engine for improving the productivity of data analysts, data scientists, and engineers when interacting with data, marking itself as “Google search for data”, as shown in Figure 21.

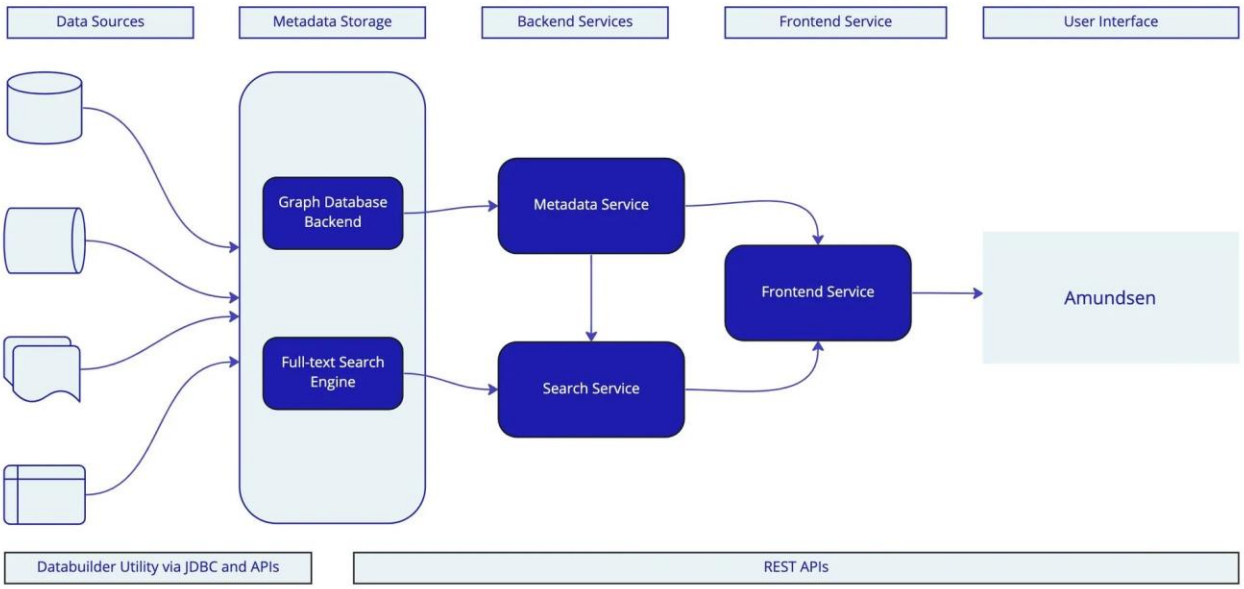


Figure 21 Schematic architecture of Amundsen

CKAN is another significant data catalog project that powers the US, Swiss, and Canadian open data platforms, but since it does not offer an automatic data discovery function, it differs from the enterprise data catalog solutions discussed in the other works mentioned. Magda is a federated, open-source data catalog with many diverse tool packages that can be used by other organizations.

With the further development of enterprise data catalog software, the tendency for data catalogs to converge towards a universal analytical tool is rising. Such a tool would, in addition to describing the data, also offer analytics in the tool itself, creating reports, exporting data, as well as writing queries over data sets coming from different sources. For example, Alation offers direct querying using standard SQL. Companies can increase productivity using the standard SQL, since the data users don’t need to worry about varying query languages used in various storage technologies.
On the other hand, introducing analytical features to data catalog comes with a significant price of complexity, with a process making simpler models more complex, slowing up the system or losing some information.

In the domain of Big Data query, significant progresses are made in the last years. Presto, an open-source distributed SQL query engine runs interactive analytic queries against data sources of all sizes ranging from gigabytes to petabytes, which is being developed under the Linux

Foundation. *Apache Druid* is open-source project which enables data-driven teams to develop analytics application for real-time analytics on top of the different data sources, both streaming as Apache Kafka or Kinesis, and data lakes such as HDFS and Amazon S3.
Apache Superset offers easy data exploration and visualization on top of different data sources and even more some data catalog uses Superset underneath to meet Data Overview functionality. And at the end, there is open-source BI and data lake software, which enable all-in-one analytics with developed dashboards and reporting systems, but on a higher level of abstraction, for business analyst, managers and C-level employees to derive business intelligence.

## 4. Metadata Model Mapping and Analysis

While Dublin Core or ISO 11179 are generic metadata standards that aim to describe a digital object with a minimum set of data elements, ISA 88, MTConnect, and QIF are standards specific for manufacturing data modeling. The question is how to model metadata that is sufficient to find and access datasets easily on the one side and provides 'specific' information describing context and content for the data set to be interpretable and reusable. A few basic metamodeling methods are developed to address the question.

Defined by the Research Data Management Service Group at Cornell [[38]], metadata essentially describes the WHO, WHAT, WHEN, WHERE, WHY and HOW of your data in the context of your domain and should provide enough information so that users know what can and cannot be done with your data. The National Information Standards Organization (NISO) classifies metadata into 4 categories, Descriptive, Administrative, Structural and Markup languages [[39]]. Based on NISO, Descriptive metadata provides information about the content of a resource that aids in finding or high-level understanding. Administrative metadata is an umbrella term referring to the information needed to manage a resource or that relates to its creation. Within the administrative metadata sphere is technical metadata, information about digital files necessary to decode and render them, such as **file type**; preservation metadata supporting the long-term management and future migration or emulation of digital files, for example, **a checksum or hash**; and rights metadata, such as a Creative Commons **license**, which details the intellectual property rights attached to the content. Descriptive and administrative metadata are considered distinct from structural metadata, which describes the relationships of parts of resources to one another; examples include pages in a sequence, **a table of contents** with pointers to the beginnings of milestone sections and connecting different resolutions or bit depth representations of identical content. Markup languages mix metadata and content together, with flags inserted in the content to denote notable features, for example, marking as paragraphs or flagging words with semantic information—that the word is a place name or a certain part of speech, for example, or providing formatting information, such as italics.

Table 4 NISO metadata types and definitions

| Metadata Type | Definition |
|---|---|
| Descriptive metadata | For finding or understanding a resource |
| Administrative Metadata<br>Technical metadata<br>Preservation metadata<br>Rights metadata | <br>⇒ For decoding and rendering files<br>⇒ Long-term management of files<br>⇒ Intellectual property rights attached to content |
| Structural metadata | Relationships of parts of resources to one another |
| Markup language | Integrates metadata and flags for other structural or semantic features within content |

ISO 11179's metadata structure provides another way to categorize metadata, shown in the table below:

Table 5: ISO 11179 data set metadata categories and definitions

| Metadata Type | Definition |
|---|---|
| Basic metadata | access_level /rights of type; temporal_coverage; Spatial_coverage and accrual_periodicity defines the frequency of revision |
| Concept metadata | to annotate a data set so as to describe the theme or category |
| Context metadata | To define a particular program, project or business area |
| Distribution metadata | Defines: *distributor, media-type, formats, size, issue date,* access-level/right, access/download url, |
| Provenance metadata | Defines the revision/processing history of the data |
| Specification metadata | Defines either a human-readable document, or data element descriptions or information model that describe the data |
| Assessment metadata | Defines the quality of the data |

The FAIR principles metadata structure, however, group all the metadata elements into 4 categories based on how they support levels of data use, defined below:

Table 6: FAIR data metadata types and Definitions

| Metadata Type | Definition |
|---|---|
| Findability metadata | Description of a data object for finding or identify it |
| Accessibility metadata | Description of a data object on how to access it<br>⇒ link to the data source<br>⇒ authentication and authorization |
| Interoperability metadata | Description of a data object that makes the resource interoperable, for example, for importing and exporting. |
| Usability/Reusability metadata | Description of a data object for using and reusing it |

The FAIR principles capture all the metadata types defined by NISO and ISO 11179, and at the same time it provides a multi-level metadata stack based on how data will be consumed. To support the C-level decision making about the existence of a certain type of data objects, the Findability metadata should be sufficient. To move data objects around, the Accessibility metadata will be additionally needed. To operate on or manipulate the data objects, the metadata for data object interoperability are required. For data analytics to support engineering activities, the metadata should cover the full stack to include those elements for findability, accessibility, interoperability, and usability. Figure 22 shows the FAIR metadata stack we would like to adopt for enterprise metadata modeling.

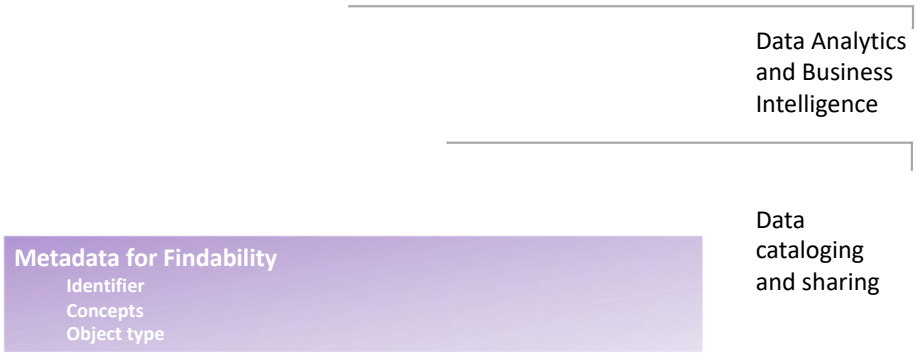


Figure 22: FAIR data Metadata stack model

### 4.1. Metadata Standards Mapping

The table below maps the capability of the surveyed metadata models into the FAIR metadata stack they serve. Detailed exemplar mappings are conducted for a selected set of those models shown in Appendix B.

Table 7: FAIR data metadata types and Definitions

| Metadata Model | Domain | Owner | FAIR Metadata Stack | | | |
|---|---|---|---|---|---|---|
| | | | Findability | Accessibility | Interoperability | (Re)usabi |
| ISO 11179-7 | General | ISO/IEC JTC1 SC32 | x | x | x | x |
| Dublin Core | Digital resources (video, images, web pages, etc.) as well as physical resources such as books or works of art. | DCMI | x | x | x | |
| DC Metadata Initiative | Extended from DC Core elements, to describe software | DCMI | x | x | x | x |
| CMIS | Digital Content | OASIS | x | x | x | |
| CWM | Databases | OMG | | | | |
| IEC CDD | Industrial/Technical | | x | | x | |
| MTConnect | Manufacturing/Machinery | AMT | x | | x | x |
| QIF | Manufacturing/Dimensional | DMSC | x | | x | x |
| CBM | Process industry | MIMOSA | x | | x | x |
| SOSA | Sensor | W3C and OGC | x | | | x |
| QUDT | Quantity | QUDT.org | | | | x |

| | | | | | | |
|---|---|---|---|---|---|---|
| FAIR | Research | GoFAIR.org | x | x | x | x |
| NeXus | Research | nexusformat.org | x | | x | x |
| DIF | Space | NASA | x | x | | x |
| Frictionless Data | Research | frictionlessdata.io | x | x | x | |
| Google Cloud | General Digital Object | Google | x | x | x | |

The mapping analysis shows that the existing metadata standards for manufacturing enterprise are great for data set sharing and use by covering one or more FAIR metadata stacks. Some of those standards are very general and usually too abstract for scientific or engineering use. Other standards are very domain-specific and detailed, which makes them not easily interoperate with each other. At the same time, the need for rich metadata is becoming generally accepted. As mentioned by authors from the EOSC Pilot project [[40]] "Minimum and common metadata is useful for data discovery and data access. Rich metadata formats can be complex to adopt but have the advantage of making data more "usable" by both humans and machines." The highlights from the standards mapping exercise are summarized below.

1) Most metadata models are built on top of Dublin core and DCMI data elements to provide findability, accessibility, and interoperability of digital objects.
2) Domain specific models, such as MTConnect, and QIF, are developed for data exchanges within specific domains focusing on the interoperability and usability of the data, but lack of metadata elements for findability and accessibility beyond the domain. The data representation can be as simple as a single value sample or as complex as highly structured XML.
3) A lot of manufacturing data models, for example, MTConnect, support transactional data exchange (sample, events, and conditions, etc.), instead of persistent data management, which limits their use within their subdomains.
4) Abstract metadata models such as ISO 11179-7, and SOSA well defines metadata element groups and complementarily cover the usability/reusability type of metadata elements. For example, 11179-33 defines the usability metadata elements related to provenance, quality, and context. And SOSA defines the physical process associated with measurement data sets, the features of interest, the resources used for the measurements, as well as data acquisition information.
5) ConnectSpec [42] *Data set Metadata* takes an alternative to only explicitly define the key usability metadata elements while captures other related metadata elements for usability in a reference link.

### 4.2. A Uniform Metadata Model

The better the metadata model, the more effective the use of data sets. Cherradi M. et al. have developed the DLDS approach (Data Lake Description Service), which is based on XML and allows data engineers to represent the metadata of each data set in the data lake, which can be used further for better browsing, understanding, and validating the data lake's documents [[41]]. Bellomarini L. et al. (2019) have discussed knowledge graphs as the data model for the goal of data cataloging that will enable enterprise AI technology [[42]]. University of Washington developed a metadata repository based on a graph database named Knowledge Navigator to keep users informed and engaged throughout the enterprise system migration by providing self-service access to conceptual and technical descriptions, definitions, lineage, interactive relationship maps, and impact analysis information [[43]]. Subramaniam P. et al. proposed a model to organize metadata to achieve better scalability, which is important in big data systems [[44]]. There are several approaches to organize data for efficient usage such as XML schemas, knowledge graphs, ontologies, and respective standards. Some of the enterprise data catalogs uses multiple ones to support multiple types of data sets.
While such plural systems provide better supports to use the data for the original purpose of data generation and domain specific tasks, there is a need for a common metadata model that supports enterprise-wide integration and data driven-business intelligence, which usually spans a broad scope of domains, organizational units and data management technologies.

Based on the metadata models reviewed and the FAIR data metadata stack model, we propose a uniform, general, and extendable metadata model which not only captures essential metadata elements for Findability, Accessibility, and Interoperability but also uses an extendable metadata model for (Re)usability.
Figure 23 shows the proposed uniform metadata model for manufacturing enterprise integration (UMM4Mei), a hierarchical method which is expandable. The top level of the UMM4MEi model prescribes metadata elements in the FAIR categories that generally describes all types of digital objects in manufacturing enterprises. For metadata usability and reusability which support data analytics and business intelligence using various domains of data sources, a modular and hierarchical structure offers a solution allowing for the addition of domain-specific or application-specific extensions. It refers to the specific level of the well-recognized *generic - specific - occurrence* modeling paradigm. For any digital object, **type object** usability metadata elements define the specific meta information about a type of digital objects. For hierarchical modeling, subtypes are defined to include additional data elements for the domain.
An alternative can be reference **element set definitions** to define shareable and extensible property sets attachable to occurrences of objects. The property set is regarded as a partial type of information as it establishes a subset of common shared property information among all digital objects.

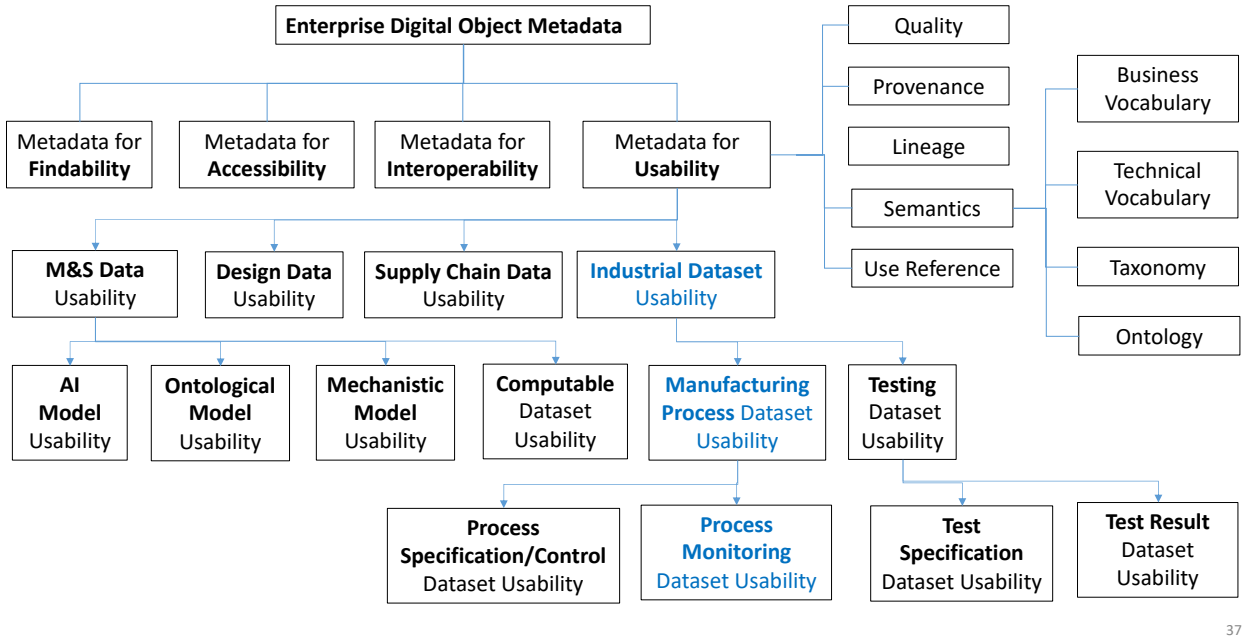


Figure 23: A Uniform Metadata Model for Manufacturing Enterprise Integration (UMM4Mei)

### 4.2.1. General UMM4Mei metadata elements

The top level of the UMM4Mei model prescribes general metadata elements for manufacturing enterprise digital objects covering all the FAIR data metadata stacks. The metadata at this level are built from the DCMI elements and incorporate additional data elements that enrich metadata to support general findability, accessibility, interoperability, and usability/reusability. Table 8 lists a set of essential metadata elements for findability and accessibility:

Table 8: UMM4Mei: essential metadata elements for findable and accessible digital objects

| **Findability** | | **Accessibility** | |
|---|---|---|---|
| **Metadata Element** | **Definitions** | **Metadata Element** | **Definitions** |
| identifier | An unambiguous reference to the resource within a given context. (DCMI) | access rights | Information about who access the resource or an indication of its security status.(DCMI) |
| title/designation | A name given to the resource. (DCMI) | license requirements | “License” is defined as a legal document giving official permission to do something with the resource. (DCMI) {Open, |

| | | | |
|---|---|---|---|
| | | | Free License, License with A Fee} |
| creator | An entity primarily responsible for making the resource. (DCMI) | license issuer | The contact to obtain a license |
| contributor | An entity responsible for making contributions to the resource. (DCMI) | (date) available | Date that the resource became or will become available. (DCMI) |
| publisher | An entity responsible for making the resource available. (DCMI) | rights holder | A person or organization owning or managing rights over the resource.(DCMI) |
| subject/keyword/ concepts | The topic of the resource. Note: Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary. (DCMI) | download url | url(s) to the data source(s) |
| description/definition | An account of the resource. (DCMI) | metadata url | url(s) of the metadata |
| Spatial coverage | The spatial applicability of the resource, or jurisdiction under which the resource is relevant. Note: Spatial topic and spatial applicability may be a named place, or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named | | |

| | | | |
|---|---|---|---|
| | administrative entity or a geographic place to which the resource applies. Recommended practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods may be used in preference to numeric identifiers such as sets of coordinates or date ranges. | | |
| temporal coverage | The temporal topic of the resource | | |
| type | The nature or genre of (the resource)The digital object (DCMI) Comment: Recommended practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMI-TYPE]. To describe the file format, physical medium, or dimensions of the resource, use the property Format. | | |
| source | The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification | | |

| | | | |
|---|---|---|---|
| | system. | | |
| language | A language of the resource. | | |
| version | A related resource that is a version, edition, or adaptation of the described resource | | |
| audience | A class of agents for whom the resource is intended or useful | | |
| designation | The sign which denotes a source if it is a designatable item | | |
| registrationID | Registration Identifier in a registry if the source is registered | | |
| adminstrationID | Adminstration Identifier of a registered source if it is administrated | | |

For interoperability, the essential metadata elements are listed in Table 9. The top eight elements define the format of the digital objects which allows the users to open and manipulate the data with right software. However, if a data object is complex in structure and customizable freely by using compressing tools, for example, zip, there are additional metadata elements needed to describe that structure. Here are some general metadata elements that describe data set structure:

Table 9: UMM4Mei: essential metadata elements for interoperable digital objects

| **Interoperability** | |
|---|---|
| **Metadata Elements (exchange: read/write)** | **Definitions** |

| | |
|---|---|
| format | The file format, physical medium, or dimensions of the resource.(DCMI)<br>Comment: Recommended practice is to use a controlled vocabulary where available. For example, for file formats one could use the list of Internet Media Types [MIME]. Examples of dimensions include size and duration.<br>Note: for example, zip, CSV, JSON, XML, or HDF5 etc |
| data schema | Data Schema could be a formal metadata element to describe of the data set structure, which can be expressed using a schema language such as JSON Schema, XML Schema, or SQL DDL. (OAGI) |
| data Schema URI | For those data schemas not well known, a URI of the data schema can be specified. The metadata for interoperability can also provide a list of software or plugins to open a data set. (OAGI) |
| conformance | An established standard to which the described resource conforms. (in DCMI ConformsTo)<br>Comment: specifications or abstract models |
| encodings | Comment: we would like to keep this as an element, but need definitions |
| software source | Deployed software resource that was used to create, collect and exchange the data set or can be used to read/write the data object (OAGi) |
| software plugin | Deployed plugin for the software resource that was used to create, collect and exchange the data set or can be used to read/write the data object (OAGi) |
| metadata version | The version number or date of the metadata schema, in case updates are made to the metadata structure. |
| data reduction method | Information about any data compression techniques or algorithms used to reduce the data set's storage size.<br>Note: specific algorithms to compress the data such as swing door method, or big data reduction methods. |
| data partitioning | A description of any data partitioning or sharding strategies used to distribute the data set across multiple storage devices or nodes. |

| | |
|---|---|
| Indexing | Information about any indexes or indexing structures used to optimize data access, such as primary keys, foreign keys, or spatial indexes. |

Additional metadata are needed for data usability and re-usability cases, where data can be interpreted, linked and fused for inferences and decision makings. The table below lists some of the representative metadata elements to interpret the data.

Table 10: UMM4Mei: essential metadata elements for usable and reusable digital objects

| **Essential Metadata Elements for Usability** | |
|---|---|
| **Provenance** | A statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation. (DCMI) |
| originator | The individual or organization responsible for the origination of the digital object which is the subject of this provenance. (ISO 11179-3) |
| owner | The individual or organization that claims ownership of the the digital object which is the subject of this provenance. |
| Is required by | A related resource that requires the described resource to support its function, delivery, or coherence. Note: This property is intended to be used with non-literal values. For example: FDA, FAA, NRC etc |
| Is useful for | A class of agents for whom the resource is useful, (besides the intended audience.) (DCMI) Note: Recommended practice is to use this property with non-literal values from a vocabulary of audience types; |
| **Lineage** | |
| Is part of | A related resource in which the described resource is physically or logically included. (DCMI) Note:This property is intended to be used with non-literal values. For example, NIST AMBench 2022 has many data sets, each of which is independently available. |
| references | A related resource that is referenced, cited, or otherwise pointed to by the described resource.(DCMI). For example, a data set refers to an earlier calibration data set |
| requires | A related resource that is required by the described resource to support its function, delivery, or coherence. (DCMI) |
| **Reference Use** | |
| Is referenced by | A related resource that references, cites, or otherwise points to the described resource.(DCMI) (an active metadata element) |

| | |
|---|---|
| Is (re)used by | A related people/group that downloads and uses the described resource. (an active metadata element) |
| **Data Quality** | |
| quality originator | A specification of the individual or organization responsible for the assessment of the quality of the data set or data set collection. (ISO 11179-Part 7 (2019)) |
| statement | a statement of the assessment of the quality of the data within the data set assessed using the quality dimension of this assessment. (ISO 11179-Part 7 (2019)) |
| category | a statement of the quality dimension against which the assessment of the quality of the data within the data set is assessed. (ISO 11179-Part 7 (2019)) |
| confidence_level | a statement of the confidence level, if any, of this assessment of the quality of the data within the data set assessed using the quality dimension of this assessment. (ISO 11179-Part 7 (2019)) |
| quality_criteria | a document that specifies the criteria against which this assessment of the quality of the data within the data set has been assessed. (ISO 11179-Part 7 (2019)) |
| **Semantic** | |
| business vocabulary | The vocabulary that can be used for business understanding of the data and the business context. Could be a URL link. (IRI) |
| technical vocabulary | The vocabulary that can be used for technical understanding of the data and the business context. Could be a URL link. (IRI) |
| domain ontology | The ontology that can be used to describe the concepts and their relationship that support the understanding of the data and the use of the data. Could be a URL link. (IRI) |

### 4.3. Metadata Model for Industrial Data set

Industrial data set refers to is a logical grouping of one or more data streams generated or used by industrial equipment. The primary sources of the massive complex data for data-driven-enterprises are measured, sampled data, and derived streams, industrial data set metadata modeling is hence very important. Here are some general metadata elements identified for Industrial Data set Metamodel in Table 11.

Table 11: Metadata elements for industrial data sets

| Metadata Element Name | Definitions |
|---|---|
| Data set Type | Raw vs Derived vs Synthetic (AI generated) |
| Quantity Kind | Definition of the kind of data element that can be measured using defined and unrestricted units of measurement, for example, length, weight, and velocity etc. |
| Data Element Name | The name of the data element, unit of **data** that is considered in context to be indivisible |
| Data Type | The type of the data element, for example, integer, float, string, or image. |
| Permissible Values | The allowable value of the data element, for example, positive, negative or enumerations. |
| Unit of Measure | The unit of a quantity kind, may be measured by meter, kilometer, or foot units. |
| Process Temporal | Temporal characteristics of the data element |
| Process Spatial | The spatial characteristic of the data element |

Moving to the next level of specificity, there are some general requirements to model the metadata for measurement data sets from manufacturing processes. The analysis of the measurement data sets using Statistical Process Control (SPC) and Root Cause Analysis (RCA) offers opportunities to improve the manufacturing processes continuously. A manufacturing data set can be very complicated. It usually is composed of one or more data streams from measurements. Each measurement kind is associated with activity, sensing and acquisition-specific metadata, which describes the context in which the measurement was taken, including information about the process, device, measurement method, and uncertainty. Each measurement type has specific parameters that define the measurement. Measurement results can be organized by data format, which specifies the structure and organization of the measurement data set.
Hereafter, a set of rules are summarized to further describe of quality industrial data set metadata model:

1. Define extensibility mechanisms: Specify mechanisms for extending the metadata model with new elements, attributes, or relationships as needed. These mechanisms can

include subclassing, sub-properties, or composition, which allow users to build upon the base metadata schema and create custom extensions that suit their specific needs.

2. Provide clear guidelines and best practices: Develop clear guidelines and best practices for using and extending the metadata model, including naming conventions, documentation, and examples. This guidance will help users create consistent and high-quality metadata that effectively supports their specific domains and applications.
3. Ensure interoperability and compatibility: Make sure the metadata model is compatible with widely used data formats and serialization formats, such as JSON, XML, or RDF. This compatibility will facilitate data exchange and integration across different systems and platforms.
4. Test and validate the metadata model: Test the metadata model against a variety of use cases and requirements, ensuring it meets the needs of different domains and applications. Perform validation checks to ensure that the metadata is consistent, accurate, and complete.

By following these steps, one can define a general metadata model that is flexible, extensible, and adaptable to various domains and applications. This approach will enable users to create rich and meaningful metadata that enhances the usability, discoverability, and interoperability of their data resources.

Based on the principles of, in addition to those defined for describing general industrial data set, a metadata model for manufacturing process measurement data sets should include the following groups of metadata elements:

- Instrument: The specific instrument, device, or equipment used to perform the measurement, including make, model, and calibration information.
- Measurement Technique: The methodology or technique used to perform the measurement, such as optical microscopy, mass spectrometry, or tensile testing.
- Uncertainty: Information about the measurement uncertainty, including error estimates, confidence intervals, or any applicable statistical information.
- Resolution: The smallest detectable change or increment in the measured quantity.
- Range: The range of values that the instrument or measurement technique can accurately measure.
- Sampling Method: A description of the sampling method or procedure used to obtain representative samples from the population or material under investigation.
- Sample Size: The number of samples or data points used in the measurement.
- Sample Description: Detailed information about the sample, such as material composition, physical properties, or preparation methods.
- Environmental Conditions: Information about the environmental conditions during the measurement, such as temperature, humidity, or pressure.

- Data Processing: A description of any data processing, analysis, or transformation methods applied to the raw measurement data.
- Validation: Information about the validation of the measurement data or method, including comparisons to reference materials, inter-laboratory studies, or theoretical predictions.
- Metadata Version: The version number or date of the metadata schema, in case updates are made to the metadata structure.
- Data Citation: Recommended citation format for the measurement data, including authors, title, publication date, and any applicable identifiers (e.g., DOI).

The table below summarize the metadata for measurement data metadata based on the Uniform Metadata model:

Table 12: Metadata elements for process monitoring dataset

| **Reference Extension for Process Monitoring Data set** | |
|---|---|
| **Measurand** | The name of the measurement data element |
| **Measurement** | |
| Measurement Type | The type of the measurement, for example, tensile, SEM, EBSD, XCT |
| Device Identity | The identifier of the measurement device. |
| Operator | The person/organization who conducted the measurement |
| Device Type | The type of the measurement device. |
| Device factory settings | Data that captures device factory setting/data; for example, camera intrinsic parameters, focus x, y, center x, y, detector resolution etc |
| Device installation configuration | Data that captures how a device/sensor is installed that can be used to describe the relative position of the sensor to the object. For example, camera extrinsic parameters |
| Device calibration data | Data which captures the calibration result of the measurement device |
| Data acquisition method | Automated or Manual |
| Data integration | Common protocols include camera link, CoaXPress, USB, pr |

| **Reference Extension for Process Monitoring Data set** | |
|---|---|
| **Measurand** | The name of the measurement data element |
| **Measurement** | |
| Measurement Type | The type of the measurement, for example, tensile, SEM, EBSD, XCT |
| Device Identity | The identifier of the measurement device. |
| Operator | The person/organization who conducted the measurement |
| Device Type | The type of the measurement device. |
| protocol | customized protocols |
| Data acquisition rate | The frequency of the data acquisition |
| **Specimen** | |
| Parent material | The portion of material where the specimen is sampled/is extracted from |
| Sampling method | The method to collect the specimen. |
| Quantity | The quantity of the specimen, for example, by weight, volume or mass. |
| Treatment method | The treatment method for the specimen to be measured. |
| **Procedure** | |
| Standards | The standards this measurement is based on, or compliant to, for example, ASTM E8 for tensile strength measurement. |
| description | The description of the operation procedure. |

## 5. Summary and Conclusions

Digitalization enables manufacturers to increase efficiency and productivity. However, managing vast amounts of data from various sources such as IoT devices, production lines and supply chains is a tremendous challenge. Metadata management is foundational for manufacturing enterprises to navigate digital transformation. Digital transformation helps manufacturing enterprises to better manage their digital objects, enabling better utilization, integration and governance, which is essential for value creation. A unified metadata model ensures that metadata is consistent and standardized across manufacturing value chains which reduces ambiguity in data use and reuse and therefore increases data sharing for accelerated product development, optimal production performance and streamlined value chain decision making. This report captures typical uses of metadata for manufacturing operations and enterprise decision making and provides a survey of existing metadata standards for findable, accessible, interoperable and reusable (FAIR) manufacturing data. At the end, we propose, a Unified Metadata Model for Manufacturing Enterprise Integration (UMM4Mei). It applies FAIR and provides an extensible metadata architecture for within and across enterprises, providing a common basis for the coordination of standards development for the purpose of metadata modeling and management. The extensibility mechanism allows for existing and emerging standards to be placed into perspective of (re)usability of the data. For future standard development, the Industrial Ontology Foundry (IOF) plans to ensure that semantics of metadata elements described in the UMM4Mei are covered and axiomatically defined as much as possible to maximize precision and commonality of their semantics.

## 6. Reference

[1] Keith Foote, "Nine Key Benefits of Metadata Management", https://www.dataversity.net/nine-key-benefits-of-metadata-management/, Accessed on Feb. 25, 2024
[2] FAIR Principles, https://www.go-fair.org/fair-principles/, Accessed on Feb. 25, 2024
[3] connectSpec, https://oagi.org/pages/connectspec, Accessed on Feb. 25, 2024
[4] Jan Top, Sander Janssen, Hendrik Boogaard, Rob Knapen, Görkem Şimşek-Şenel, "Cultivating FAIR principles for agri-food data, Computers and Electronics in Agriculture", Volume 196, 2022, https://doi.org/10.1016/j.compag.2022.106909.
[5] Farhad Ameri, Evan K. Wallace, Reid Yoder and Frank Riddick, "Enabling Traceability In Agri-Food Supply Chains Using An Ontological Approach", ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, August, 2020.
[6] Wayne Matthews, "How Process Historians Turn Data Into Actionable Information", https://blog.isa.org/process-historians-turn-industrial-data-actionable-information#:~:text=Process%20historians%20are%20being%20used,)%2C%20and%20asset%20management%20systems.
[7] Reuben Schuitemaker, Xun Xu, "Product traceability in manufacturing: A technical review", Procedia CIRP, Volume 93, 2020, https://doi.org/10.1016/j.procir.2020.04.078.
[8] Corrales, David Camilo et al. "A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal." *J. Comput.* 10 (2015): 396-405.
[9] Drobnjakovice, M., Chareonwut, P., Nikolov, A., Oh, H., and Kulvatunyou, B (2024). An Introduction to Machine Learning Lifecycle Ontology and its Applications, International Conference on Advances in Production Management Systems, Chemnitz Germany.
[10] Carlo Gutierrez, "Blockchain at Walmart: Tracking Food from Farm to Fork", https://www.altoros.com/blog/blockchain-at-walmart-tracking-food-from-farm-to-fork/
[11] https://glennas.wordpress.com/2010/01/31/dublin-core-metadata-initiative-dcmi-learning-resources/, Accessed on Oct. 28, 2024.
[12] Data Catalog (DCAT) Version 3.0, https://www.w3.org/TR/vocab-dcat-3/#fig-dcat-all-attributes
[13] DCAT-US Schema v1.1, https://resources.data.gov/standards/catalog/dcat-us/
[14] Data Catalog Application Profile for the United States of America, Candidate Recommendation Snapshot https://doi-do.github.io/dcat-us/
[15] Alfresco Content Services 7.0, https://docs.alfresco.com/content-services/7.0/develop/repo-ext-points/content-model/
[16] "Introduction to CMIS", http://www.oldschooltechie.com/blog/2009/11/23/introduction-cmis#:~:text=CMIS%20has%20the%20potential%20to,Programming%20language%20neutral
[17] Common Warehouse Metamodel (CWM) Specification, https://www.omg.org/spec/CWM/1.1/PDF/
[18] ISO 15000-5 Core Component Specification, also known as UN/CEFACT Core Component Technical Specification, https://unece.org/trade/uncefact/ccts

[19] QIF (Quality Information Framework) 2024 Definitive Guide, https://www.capvidia.com/blog/qif-quality-information-framework-definitive-guide
[20] Krzysztof Janowicz, Armin Haller, Simon J.D. Cox, Danh Le Phuoc, Maxime Lefrançois, "SOSA: A lightweight ontology for sensors, observations, samples, and actuators," Journal of Web Semantics, Volume 56, 2019, https://doi.org/10.1016/j.websem.2018.06.003.
[21] Rijgersberg, Hajo, van Assem, Mark, and Top, Jan. 'Ontology of Units of Measure and Related Concepts'. 1 Jan. 2013 : 3 – 13.
[22] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[23] Frazier W E, Lu Y, Witherell P, Kitt A (2023) FAIR Additive Manufacturing Data Management Principles. Additive Manufacturing Design and Applications, eds Seifi M, Bourell DL, Frazier W, Kuhn H (ASM International), pp 171–176. https://doi.org/10.31399/asm.hb.v24A.a0006979
[24] RDA Metadata Standards Directory, http://rd-alliance.github.io/metadata-directory/.
[25] Sempreviva Anna Maria, Vesth Allan, Bak Christian, Verelst David Robert, Giebel Gregor, Danielsen Hilmar Kjartansson, Mikkelsen Lars Pilgaard, Andersson Mattias, Vasiljevic Nikola, Barth Stephan, Sanz Rodrigo Javier, Gancarski Pawel, Reigstad Tor Inge, Bolstad Hans Christian, Wagenaar Jan Willem, & Hermans Koen W. (2017). Taxonomy and metadata for wind energy Research &Development. Zenodo. https://doi.org/10.5281/zenodo.1199489
[26] Nexus, https://rdamsc.bath.ac.uk/msc/m25, Accessed on Oct. 31, 2024
[27] Frictionless Project, https://frictionlessdata.io/standards/#standards-toolkit
[28] Direct Interexchange Format, Writer's Guide, Version 8, https://web.archive.org/web/20011116121741/
[29] Google: Object metadata, https://cloud.google.com/storage/docs/metadata
[30] Amazon: Working with object metadata, https://docs.aws.amazon.com/AmazonS3/latest/userguide/UsingMetadata.html
[31] Data Management Body of Knowledge (DMBOK2), DAMA International, published by Technics Publications, July 2017; refer to: https://dama.org/content/body-knowledge
[32] "Metadata Management 101: What, Why and How", https://medium.com/@Talend/metadata-management-101-what-why-and-how-70edbc56b5cc
[33] Sethi, R., Traverso, M., Sundstrom, D., Phillips, D., Xie, W., Sun, Y., ... & Berner, C. (2019, April). Presto: SQL on everything. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (pp. 1802-1813).
[34] A. Kadadi, R. Agrawal, C. Nyamful and R. Atiq, "Challenges of data integration and interoperability in big data," *2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 2014, pp. 38-40, doi: 10.1109/BigData.2014.7004486.
[35] Peter Buneman, Adriane Chapman, and James Cheney. 2006. Provenance management in curated databases. In Proceedings of the 2006 ACM SIGMOD international conference on Management of data (SIGMOD '06). Association for Computing Machinery, New York, NY, USA, 539–550. https://doi.org/10.1145/1142473.1142534
[36] Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Shubha Nabar, Tomoe Sugihara, and Jennifer Widom. 2006. Trio: a system for data, uncertainty, and lineage. In

Proceedings of the 32nd international conference on Very large data bases (VLDB '06). VLDB Endowment, 1151–1154.

[37] Yingwei Cui, Jennifer Widom, and Janet L. Wiener. 2000. Tracing the lineage of view data in a warehousing environment. ACM Trans. Database Syst. 25, 2 (June 2000), 179–227. https://doi.org/10.1145/357775.357777

[38] Ikeda, Robert and Jennifer Widom. "Data Lineage: A Survey." (2009), http://ilpubs.stanford.edu:8090/918/1/lin_final.pdf

[39] Metadata and Describing Data, https://data.research.cornell.edu/content/writing-metadata

[40] Understand Metadata, What is Metadata and What is it for, https://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf

[41] Quimbert, E., Jeffery, K., Martens, C., Martin, P., Zhao, Z. (2020). Data Cataloguing. In: Zhao, Z., Hellström, M. (eds) Towards Interoperable Research Infrastructures for Environmental and Earth Sciences. Lecture Notes in Computer Science(), vol 12003. Springer, Cham. https://doi.org/10.1007/978-3-030-52829-4_8

[42] Cherradi M., EL Haddadi A., Routaib H. "Data Lake Management Based on DLDS Approach", Networking, Intelligent Systems and Security: Proceedings of NISS 2021

[43] L. Bellomarini, D. Fakhoury, G. Gottlob and E. Sallinger, "Knowledge Graphs and Enterprise AI: The Promise of an Enabling Technology," *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, Macao, China, 2019, pp. 26-37, doi: 10.1109/ICDE.2019.00011.

[44] Pieter Visser, "Advantages of a Graph-Based Metadata Repository", https://neo4j.com/blog/advantage-graph-based-metadata-repository/

[45] Subramaniam, P., Ma, Y., Li, C., Mohanty, I., & Fernandez, R.C. (2021). Comprehensive and Comprehensible Data Catalogs: The What, Who, Where, When, Why, and How of Metadata Management. *ArXiv, abs/2103.07532*.

# 7. Appendix

## 7.1. A: Terms and definitions

**Data -** Reinterpretable representation of information in a formalized manner suitable for communication, interpretation or processing, [SOURCE:ISO 2382-1:2015, 2121272 — Notes to entry modified]
Note 1 to entry: Data can be processed by humans or by automatic means.

Note 2 to entry: Data may also be described using the terminological notions defined in ISO 1087-1:2000 and the computational notions defined in ISO/IEC 11404:2007. A datum is a designation of a concept with a notion of equality defined for that concept.

**digital object**: object that has been digitally created or digitized by the **museum** (2.1.5) or has been acquired in digital form, [SOURCE: ISO 18461:2016(en)]

Note 1 to entry: Documents are included.

**Interoperability:** ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality *Understanding Metadata*, 2004, NISO Press

**persistent data:** data which is retained in the information system for more than one data management session

**transient data:** data which is either flowing into and out of an information system, or, in the case of a distributed system, between two computer systems

**data management:** the activities of defining, creating, storing, maintaining and providing access to data and associated processes in one or more information systems, ISO 10032

**data management system:** a system which is concerned with the organization and control of data

**common data management services:** common data management services such as those required to define, store, retrieve, update, maintain, backup, restore, and communicate applications and dictionary data. ISO 10032

**metadata**

1. **data** (3.2.6) that defines and describes other data, ISO 11179-1
2. structured information which describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource, US National Information Standards Organization (NISO)

**data model:** graphical and/or lexical representation of **data** (3.2.6), specifying their properties, structure, and inter-relationships, ISO 11179-1

**metamodel:** data model that specifies one or more other models, such as data models, process models, ontologies, etc, ISO 11179-1

**metadata object:** object type defined by a **metamodel** (3.2.20), ISO 11179-1

**metadata item:** instance of a **metadata object** (3.2.18), ISO 11179-1

**metadata registry:** As such, metadata can be stored and managed in a database, often called a registry or repository. (https://www.bls.gov/osmr/research-papers/2000/pdf/st000010.pdf)

**metadata item** recorded in a **metadata registry** (3.2.19)

**data element:** unit of **data** (3.2.6) that is considered in **context** (3.3.7) to be indivisible

[SOURCE:ISO/IEC 15944-1:2011, 3.16]

Note 1 to entry: The definition states that a data element is “indivisible” in some context. This means it is possible that a data element considered indivisible in one context (e.g., telephone number) may be divisible in another context (e.g., country code, area code, local number).

**data element concept:** concept that is an association of a property with an object class  Note 1 to entry: A data element concept is implicitly associated with both the property and the object class whose combination it expresses.

Note 2 to entry: A data element concept may also be associated with zero or more conceptual domains each of which expresses its value meanings.

Note 3 to entry: A data element concept may also be associated with zero or more data elements each of which provides representation for the data element concept via its associated value domain.

**metadata element:** resource property name that can be used in metadata and that can be given a value (ISO 24622)

## 7.2. Appendix B: Exemplar metadata mapping


Figure 24: ISO 11179 metadata element mapping to the FAIR stack

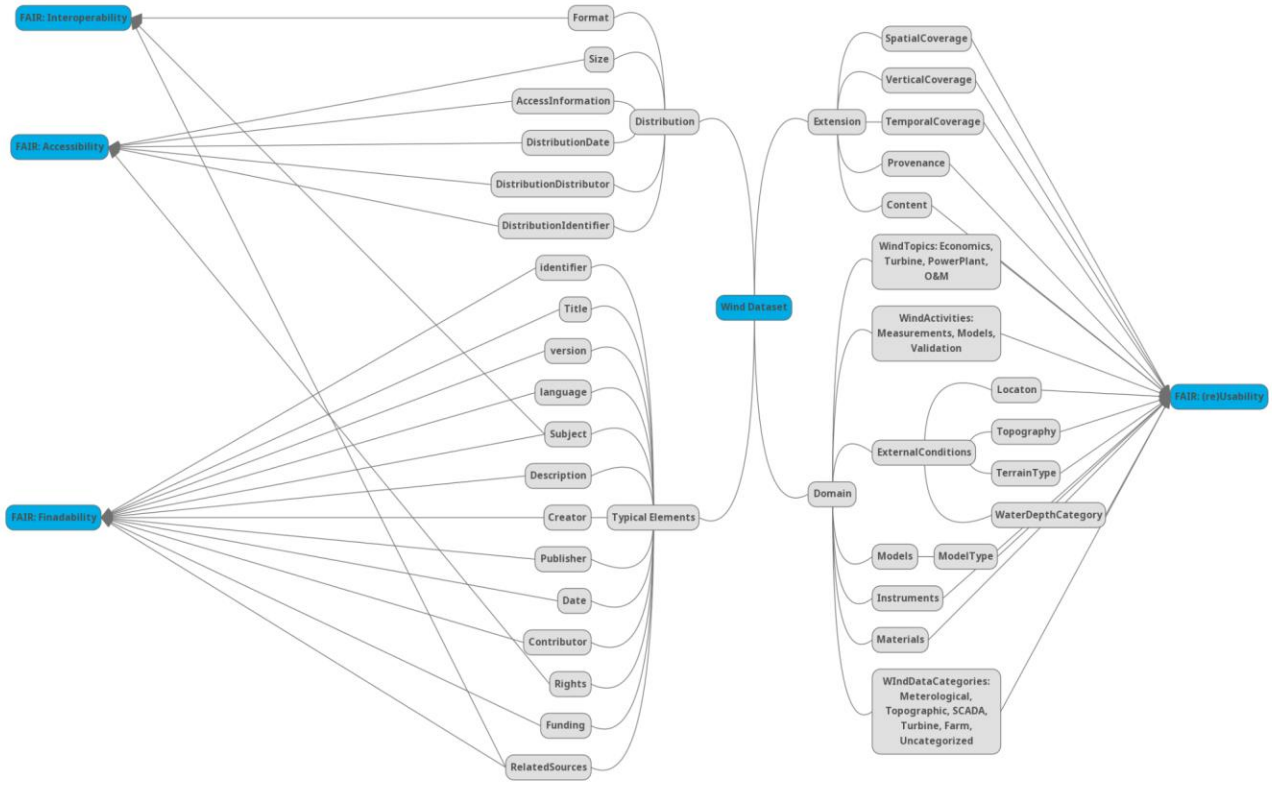


Figure 25: Wind Energy metadata element mapping to the FAIR stack

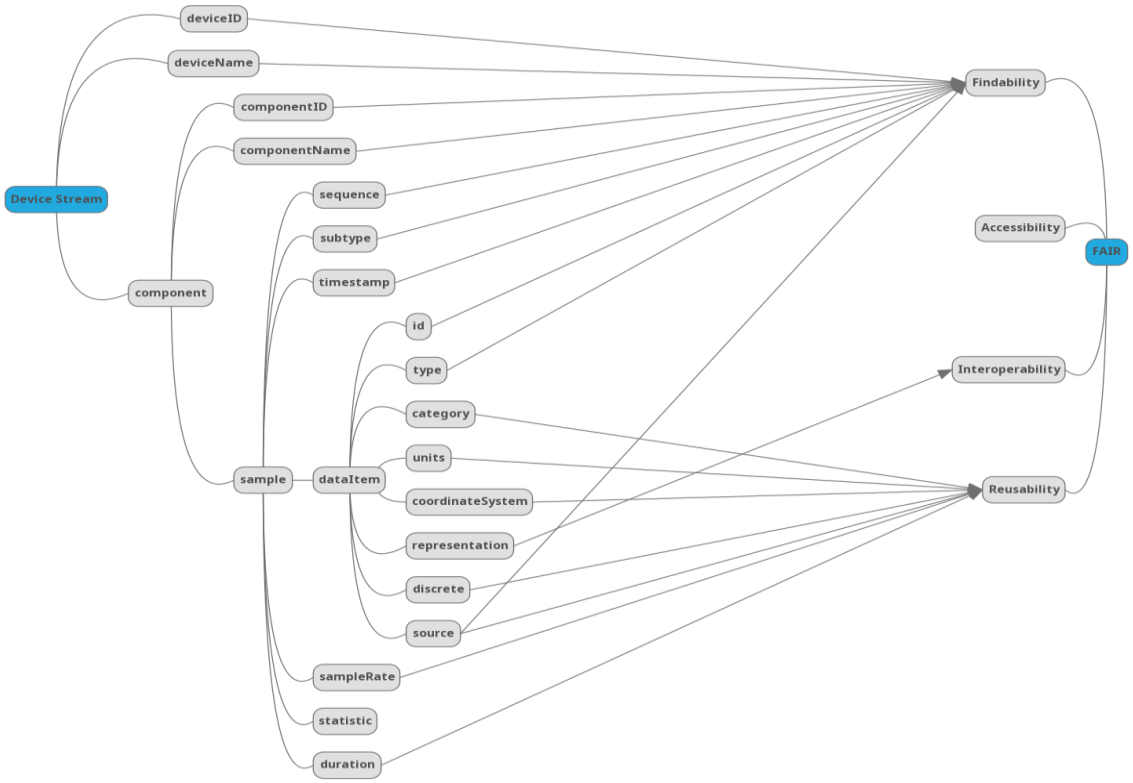


Figure 26: MTConnect data model mapping to the FAIR stack

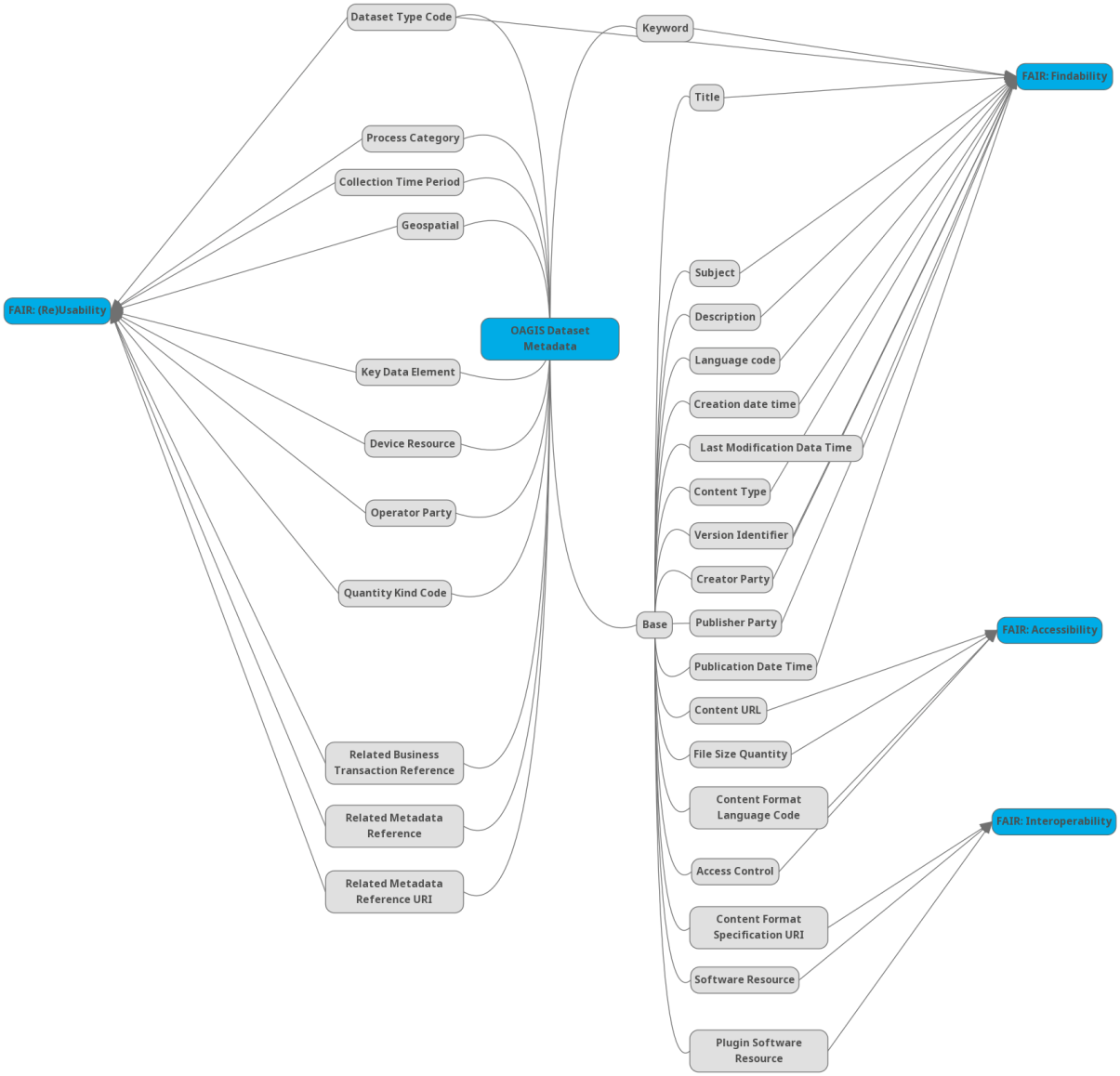


Figure 27: ConnectSpec metadata model mapping to the FAIR stack