Implementierung der FAIR-Prinzipien im Forschungsdaten-management: eine Terminologiebasierte Strategie für die inhaltliche Beschreibung numerischer Faktendatensätze

Giacomo Lanza, Joachim Meier, Thomas Wiedenhöfer[1]
Ulrich Schwardmann[2]

[1]Physikalisch-Technische Bundesanstalt, Referat Q.11 „Wissenschaftliche Bibliotheken“, Braunschweig und Berlin

[2]Gesellschaft für Wissenschaftliche Datenverarbeitung Göttingen

Zusammenfassung

In der Open Science-Ökonomie stellen numerische Faktendaten eine große Herausforderung für die praktische Umsetzung der vier FAIR-Prinzipien dar. Dies resultiert einerseits aus einer unüberschaubar großen Anzahl von Datensätzen und andererseits aus einer großen Vielfalt an in unterschiedlichen Disziplinen verwendeten Datenstrukturen. Diese Heterogenität erschwert den Vergleich von Forschungsdatenstrukturen unterschiedlichen Ursprungs, sowie die Festlegung eines einheitlichen Standards zu ihrer Beschreibung mittels Metadaten. Beispielsweise sieht das gebräuchliche DataCite-Metadatenschema keine Felder für eine detaillierte Beschreibung zusätzlich zur Angabe frei zu vergebender – und damit unkontrollierter – Schlagworte vor. Vor diesem Hintergrund ist bereits die erste Stufe der FAIR-Prinzipien, die Auffindbarkeit (*Findability*), nur unzureichend zu realisieren. Zielgerichtetes, feingranulares Suchen und präzises Finden auf Datenrepositorien-übergreifender Ebene ist aktuell nicht möglich.

Zur Lösung des beschriebenen Problems ist es sinnvoll, bei typischen Eigenschaften numerischer Faktendaten anzusetzen. Diese Faktendaten sind gekennzeichnet durch Größen, Maßeinheiten, numerische Wertebereiche, Rollen (z.B. Messgröße, Messvariable, Messparameter) und, bei quantitativ bewerteter Zuverlässigkeit der Faktendaten, zusätzlich die Messunsicherheitsangabe und sofern bekannt das zugrundeliegende Messunsicherheitsmodell. Als Module eines aufzubauenden „Metrology Terminology Directory“ (MTD) werden in abgegrenzten Namensräumen kontrollierte Vokabulare für Messgrößen, Maßeinheiten, Messverfahren und verschiedene Charakteristiken von Messobjekten mehrsprachig entwickelt und jeweils über spezifische „Persistent Identifiers“ in einer sogenannten Data

Type Registry sprachübergreifend adressierbar gemacht. In neu entwickelten Faktendaten-spezifisch strukturierten Metadatenmodulen dienen diese Vokabulare zur Beschreibung der Faktendatensätze. Auf diese Weise werden die wesentlichen Eigenschaften numerischer Faktendatensätze für komplexes Suchen und Finden mittels geeigneter Retrieval-Werkzeuge zugänglich gemacht.

Durch Implementierung von Metadatenschema-Modulen seitens der Hersteller von digitalen Messgeräten, digitalen Sensoren oder Messverarbeitungs-Software, könnte zukünftig erreicht werden, dass eine vereinheitlichte Metadatenbeschreibung schon bei der erstmaligen analog/digital-Wandlung von Messdaten begonnen und über die weiteren Schritte der Messdatenverarbeitung „halbautomatisch“ sukzessive angereichert werden kann. Hierdurch würde nicht nur die Dokumentationsarbeit erleichtert, sondern es bestünde auch die Möglichkeit, die FAIR-Prinzipien vollständig umzusetzen.

Eine Pilotrealisierung der MTD und ausgewählter Metadaten-Module werden im Vortrag vorgestellt.

Abstract

Within the Open Science economy, a major challenge for the practical application of the FAIR principles is represented by numerical factual data. This is a result of the overwhelming big quantity of data sets, as well as of the big variety of used data structures currently used in the different disciplines. This heterogeneity makes it difficult to compare data structures of different origins, as well as the establishment of unique standards for their description through metadata: for instance, the common metadata schema DataCite doesn't include any fields for the detailed description of the study object, other than the inclusion of free-text – thus non controlled – keywords. Under these conditions there are limited possibilities to apply even the first of the FAIR principles, Findability: a directed advanced search of data over a plurality of repositories is currently not feasible.

We are seeking a solution to this problem in relation to the typical characteristics of numerical factual data, which are commonly described by reporting quantity names, units, numerical ranges, roles (measured quantity,

variable or parameter) and some estimate of the data reliability, such as the value uncertainty and the uncertainty model. Within a newly defined "Metrology Terminology Directory" (MTD), we are developing a collection of multilingual controlled vocabularies for physical quantities, measuring units, experimental techniques and selected information about research objects; the correct designation of each item takes place in a language-independent manner via a persistent identifier. These vocabularies are then applied within an *ad hoc* defined metadata module for numerical data for a thorough description of a data set. That way, the relevant features of numeric factual data are made accessible for a complex search and finding with suitable retrieval tools.

If the proposed metadata module is adopted and implemented by the producers of digital measuring instruments, digital sensors and software for data analysis, it will be possible in the future to implement a standardised metadata description already at the point of the first analog-to-digital conversion, and then propagate and enrich it somehow automatically during all data transformation steps. This would make documentation work tremendously easier, as well as contribute to the complete realisation of the FAIR principles.

Within the talk a pilot realisation of the MTD will be presented, along with selected metadata modules.

Hintergrund und Problemstellung

Der freie Austausch von Informationen und Wissen ist eine der tragenden Säulen der Wissenschaft. Mit zunehmender Digitalisierung in der Wissenschaft und mit der Verbreitung der Open-Science-Philosophie werden in wachsendem Ausmaß auch Forschungsdaten der Öffentlichkeit zur verfügbar gemacht. Diese Daten fallen in einer Vielzahl offener und proprietärer Formate an. Für deren Beschreibung werden Metadaten verwendet, deren Umfang i. d. R. von den Möglichkeiten der Datenportale oder von Fachgemeinschaften bestimmt wird. Die Qualität der Metadatenerfassung ist dabei abhängig von der fachlichen Expertise und dem Vollständigkeitsanspruch der einzelnen Kuratoren.

Eine Nachnutzung dieser Daten setzt die Fähigkeit voraus, sie mittels geeigneter Suchmaschinen selektiv recherchieren und filtern zu können. Eine feingliedrige, „erweiterte“ Suche bedarf einer gemeinsamen „Sprache“ zwischen den Suchmaschinen und den Datenrepositorien. Das wurde zum Teil erreicht mit der Festlegung standardisierter Metadatenschemata (DataCite) und Protokolle für das *Metadata Harvesting* (OAI-PMH); darüber hinaus wenden einige Fachdisziplinen zusätzliche Metadatenmodule an, die eine feingranulare Beschreibung ermöglichen. In gewissen Fällen wird auch eine Suche im Datenbestand selbst angeboten.

Angesichts der wachsenden Datenmengen und der anfallenden multidisziplinären Fragestellungen wird der Bedarf immer offensichtlicher an einer interoperablen Struktur für die dokumentarische Beschreibung der Daten, die das zuverlässige Finden, Filtern und Vergleichen von Forschungsdaten unterschiedlichen Ursprungs ermöglicht. Die weitgehend KI-unterstützten Methoden der *Big Data*-Analyseund des *Machine Learning* erfordern, dass die Daten nicht nur für Menschen zugänglich, sondern auch maschinenlesbar und -verständlich sind. Dies erfordert eine Auswahl standardisierter, langlebig lesbarer Datenformate, sowie eine eindeutige Kodierung der Metadatenbeschreibungen mit Standardisierung sowohl der Attributfelder als auch deren zulässiger Inhalte (Attributwerte). Die für Forschungsdatenmanagement eingeführten vier FAIR-Prinzipien, Die Daten sollen auffindbar (*Findable*), zugänglich (*Accessible*), interoperabel (*Interoperable*) und wiederverwendbar (*Reusable*) sein, geben die Ziele vor, machen aber keine Angaben für Lösungen dieser nicht trivialen Aufgabenstellung.

Sehr stark automatisierungs- und messtechnischgeprägte Industrieunternehmen sind sich dieser Herausforderungen seit längerem bewusst. Firmeneigene Schemata und Protokolle für die digitale Übertragung von Informationen sind bereits entwickelt und z. B. im Bereich künstlicher Intelligenz im Einsatz.

Lösungsansatz

Forschungsdaten ohne eine standardisierte, feingranulare, den

Suchmaschinen zugängliche Metadaten-Beschreibung sind schwierig zu finden bzw. mit zunehmender Komplexität schwer oder nicht immer eindeutig interpretierbar.

Um die Interoperabilität von Forschungsdaten zu gewährleisten, ist die Einführung einer gemeinsamen Grundlage für die Metadaten-Beschreibung unentbehrlich. Diese entsteht aus (1) einer definierten Palette von Attributen (Metadatenfelder) sowie (2) je einer kontrollierten Liste zulässiger Werte, die eine eindeutige, reproduzierbare und zweifelsfreie Erfassung ermöglicht. Das Ganze sollte international und womöglich fachübergreifend abgestimmt werden. Ein solcher Standard besteht aus folgenden Bausteinen:

- Ein Metadatenschema, also eine hierarchische Anordnung kontrollierter Metadatenfelder mit festgelegten Benennungen und vorgegebenen Regeln (Variabeltyp, Freitext / kontrollierter Text). Erstrebenswert ist hier ein modularer Aufbau, bei dem neben einem gemeinsamen Kernschema (z.B. eine Erweiterung des aktuell verwendeten DataCite-MDS) unterschiedliche fachspezifische oder methodenspezifische Metadaten-Module optional verwendet werden können.
- Mehrere kontrollierte Vokabulare, die für ausgewählte Metadatenfelder eine Liste zulässiger Werte (Terme) zur Verfügung stellen. Je nach Komplexität können die Vokabulare als Thesauri oder Ontologien realisiert werden. Um die Mehrdeutigkeiten der natürlichen Sprache zu überwinden, ist es ratsam, die einzelnen Begriffe mit langlebigen Kennzeichen (Persistent Identifiers) zu versehen. Hierdurch wird die Verfügbarkeit der Vokabulare in mehreren (menschliche) Sprachen unter Beibehaltung der logischen Struktur möglich.

Beispiel: Größen und Einheiten

In den quantitativen Wissenschaften (z. B. Chemie, Physik, Materialwissenschaften und andere technische Wissenschaftsgebiete) spielen numerische Faktendaten eine zentrale Rolle. Die grundlegenden Informationen für deren eindeutige Identifizierung sollten Angaben beinhalten über

- Versuchsobjekt: Probentyp, Proben-ID, chemische Identität, Auftraggeber;
- Datengenerierung: experimentelle Methode bzw. Simulationsprozedur,

Identifikation des Messplatzes, Zeitpunkt, Mitarbeiter…;

- Faktendaten-Merkmale: Größe, Maßeinheit, numerischer Wertebereich, numerische Auflösung, Messunsicherheit, Rolle (Messgröße, Variable oder Parameter).

Im vorgestellten Vorgehen wurde zunächst die Darstellung der Datenwerte behandelt. In der Folge wurde ein atomares Metadatenschema für numerische Faktendaten und Vokabulare für physikalische Größen und Einheiten entwickelt.

Ergebnisse: Metadatenschema

Im Rahmen des innerhalb Horizon 2020 europäisch geförderten Projektes *SmartCom* wurde für industrielle Anwendung ein atomares Metadatenschema (**D-SI**), basierend auf internationalen Richtlinien der Messtechnik, für die Beschreibung numerischer Faktendaten definiert. Das Schema sieht Felder für die Angabe eines Messwerts mit Maßeinheit und Messunsicherheit vor. Die Unsicherheit kann entweder als absolute erweiterte Unsicherheit samt Erweiterungsfaktor, oder als Intervallart samt Intervallgrenzen dargestellt werden; in beiden Fällen wird die Überdeckungswahrscheinlichkeit angegeben. Optionale Angaben dazu sind der Größenname, der Zeitstempel sowie die angenommene Wahrscheinlichkeitsdichteverteilung des Messunsicherheitswertes. Die Qualitätskontrolle legt den größten Wert auf die Angabe der Einheiten, wobei SI-Basiseinheiten ohne Präfixe stark empfohlen werden.

Für die Beschreibung von Forschungsdaten für Open Science-Anwendungen wird dieses Schema erweitert. Für jede im Datensatz auftauchende Größe werden sowohl der Wertebereich (Minimum und Maximum) als auch die Unsicherheit angegeben. Die Benennung der Größe ist verpflichtend und soll möglichst aus einem normierten Vokabular stammen. Zusätzlich wird die Rolle der Größe (Messgröße, Messvariable oder Messparameter) sowie eine wörtliche Erläuterung angegeben.

Das vorgeschlagene Schema bildet die hierarchische Anordnung der relevanten Felder ab. Die korrekte Eingabe der Metadatenwerte wird durch Anwendungsregeln unterstützt. Struktur und Regeln sind bewusst

syntaxneutral angelegt und ermöglichen deshalb die äquivalente Auszeichnung bzw. Export der Metadaten in XML, JSON, YAML oder einem anderen beliebigen Auszeichnungsformat bzw. Notation.

Ergebnisse: kontrollierte Vokabulare

Die entwickelten Vokabulare werden in einem „Metrology Terminology Directory“ gesammelt und sollen in maschinenlesbarem Format (JSON, RDF oder OWL) für die Wissenschaft, die Wirtschaft und die Öffentlichkeit zur Verfügung gestellt werden.
Das Vokabular für physikalische Größen besteht derzeit aus ca. 400 Einträge. Für jeden Eintrag werden die folgenden Attribute angeboten:

- Benennung, derzeit in 18 Sprachen.
- Symbol und Definitionsformel.
- Dimensionen nach BIPM Broschüre; Bezug zur vorgegebenen SI-Einheit.
- Kurzer Definitionstext und Anmerkungen.
- Notation aus einer mehrstufigen Klassifikation.

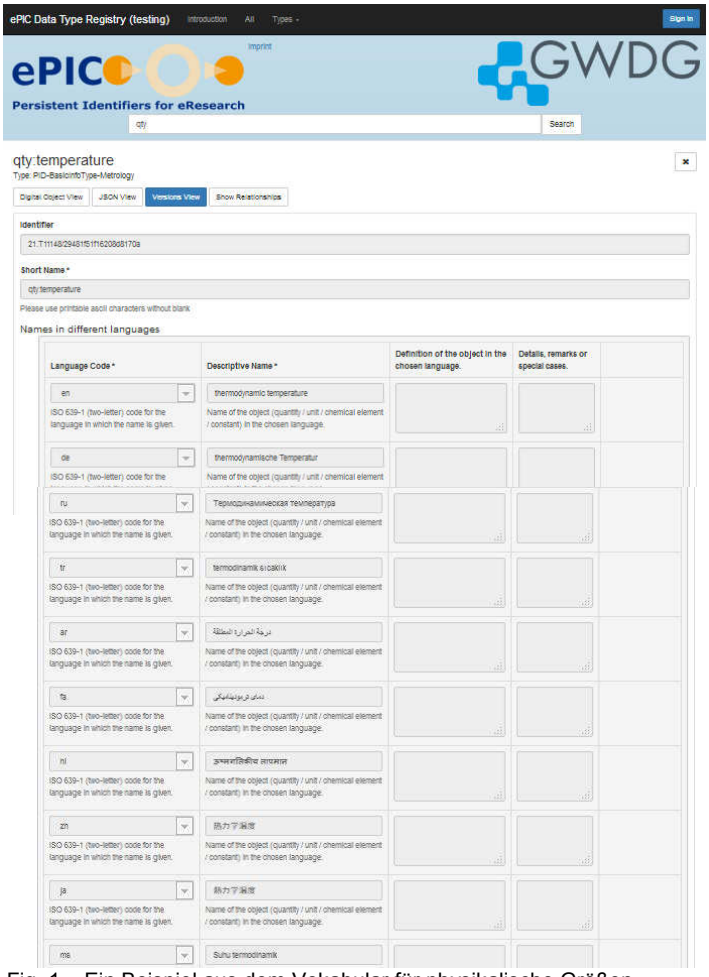

Fig. 1 – Ein Beispiel aus dem Vokabular für physikalische Größen.

Das Vokabular für Naturkonstanten, in Anlehnung an der Liste von CODATA, enthält derzeit 84 Einträge. Für jede Naturkonstante werden die schon bei Größen gelisteten Attribute angegeben, und zusätzlich:

- Der numerische Wert der Naturkonstante mit absoluter und relativer Standardunsicherheit und die Gültigkeitszeitspanne dieses Wertes (basiert auf den CODATA-Ausgaben von 1969, 1973, 1986, 1998, 2002, 2006, 2010, 2014 und 2018).
- Die Korrelationskoeffizienten zu den anderen Naturkonstanten, mit Gültigkeitszeitspanne.

Das Vokabular für Einheiten listet derzeit ca. 100 Einheiten, darunter alle in der BIPM-Broschüre gelisteten (SI-Einheiten und noch akzeptierte nicht-SI-

Einheiten) Einheiten sowie alle bei Messgrößen verwendeten Kombinationen von Einheiten (z. B. J/K, N.m). Für die Einheiten werden folgende Attribute vergeben:

- Bezeichnung, derzeit in 18 Sprachen.
- Symbol in lateinischem, kyrillischen und arabischen Alphabet, sowie LaTeX-Quellcode.
- Definitionsformel mit Umrechnungsfaktor, sofern notwendig.
- Dimensionen nach BIPM-Broschüre und Bezug zu den damit verbundenen Größen und Konstanten.
- Kurzer Definitionstext und Anmerkungen.

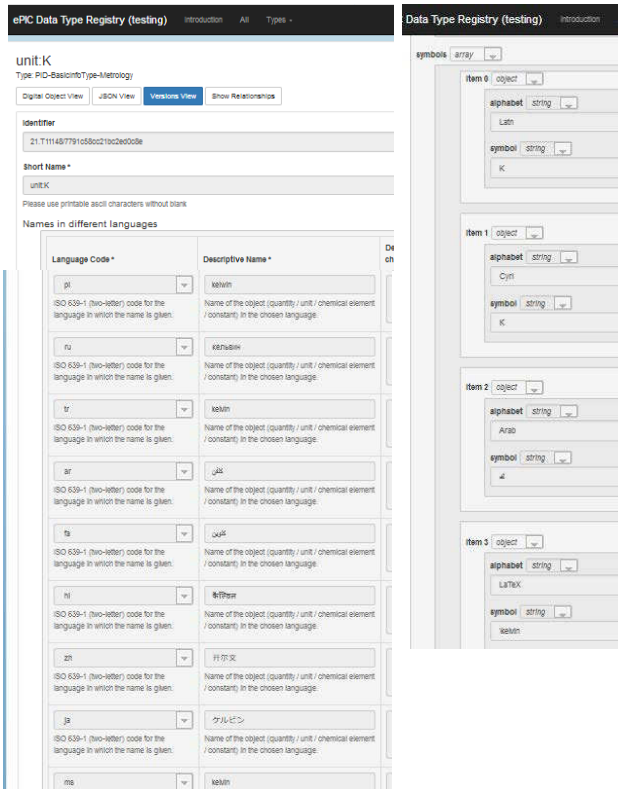

Fig. 2 – Ein Beispiel aus dem Vokabular für Einheiten.

Geplant ist ein viertes Vokabular, das auf Basis des VIM (Internationales Wörterbuch der Metrologie) und des GUM (Guide to the expression of uncertainty in measurememement) die grundlegenden metrologischen Begriffe mehrsprachig und maschinenlesbar darstellen wird. Unter anderen werden

die Grundbegriffe in Bezug auf Messverfahren, Messunsicherheit sowie die Rolle einer Größe aufgeführt.

Beispielanwendung

Das hier eingeführte Schema samt Vokabularen stellt eine mögliche Grundlage dar für die interoperable Erfassung von Forschungsdaten in einem Datenportal oder Repositorium, die eine feingranulare Suche ermöglicht. Dafür soll eine spezialisierte Suchmaschine eingerichtet werden, die nach Größennamen und Wertebereichen indexieren und filtern kann.

Als Beispiel könnte eine Recherche nach aktuellen Erkenntnissen über die Temperaturabhängigkeit verschiedener Materialeigenschaften eines Stoffs unter gewissen Umweltbedingungen ausgeführt werden. In diesem Beispiel sollen alle Datensätze relevant sein, in denen „Graphen“ als Substanz und „Temperatur“ als Größe vorkommen und bei denen Temperaturwerte zwischen vorgegebenen Grenzwerten (z. B. zwischen 200 K und 1000 K) vorliegen.

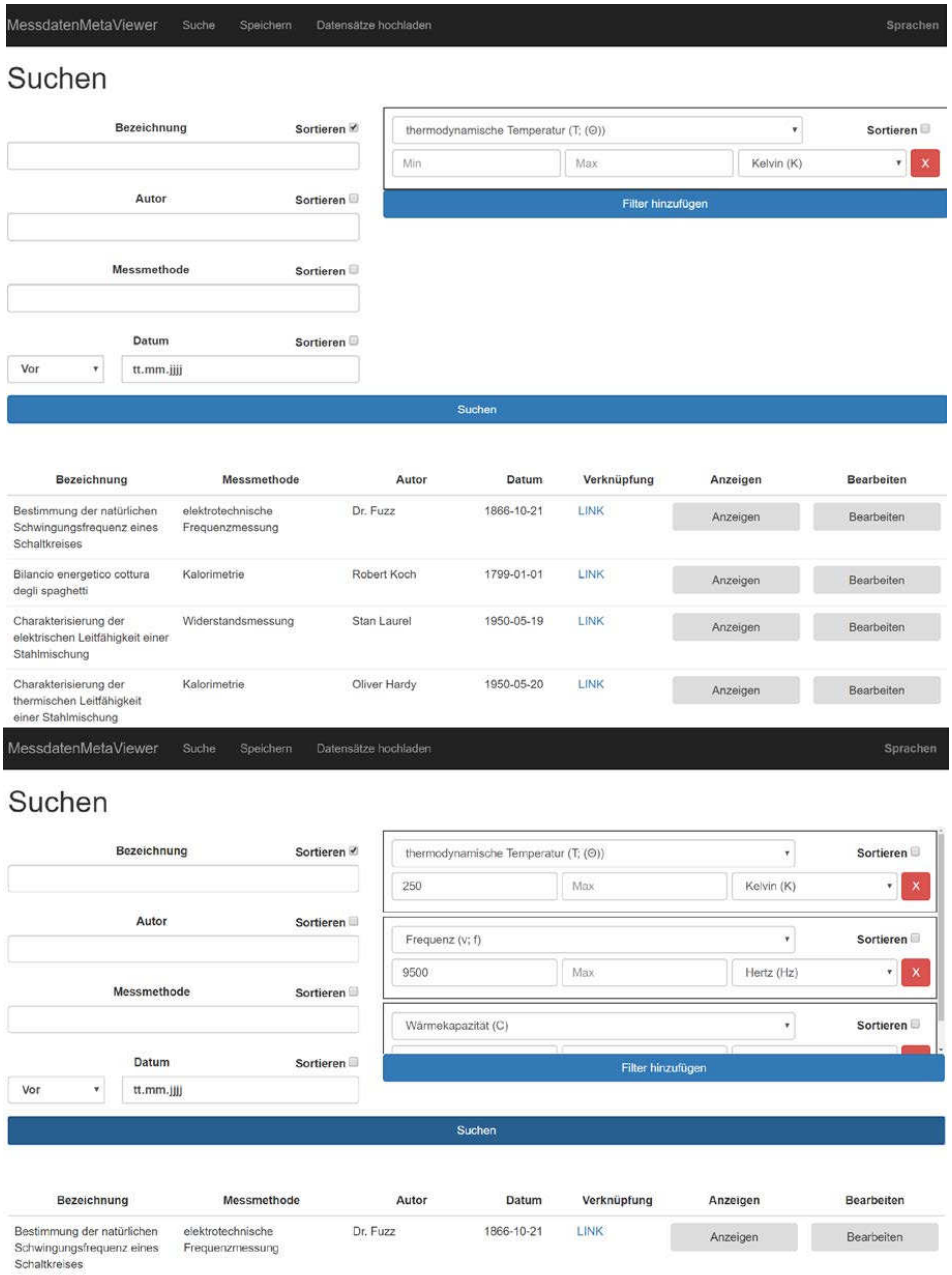

Fig. 3 – Beispielanwendung: eine Suchplattform für Forschungsdaten, die eine erweiterte Suche auf Basis der von uns definierten Felder und der festgelegten Vokabulare ausführt und nach Größen und Wertebereichen filtern kann.

Ausblick

Die Erfassung der Metadaten kann manuell oder automatisch erfolgen. Schon heute werden bei einer softwaregesteuerten Messung wichtige Parameter über den Messablauf von den Messgeräten selbst in den Kopfzeilen der erzeugten Rohdaten oder in zusätzlichen begleitenden Dateien protokolliert. Auch erlaubt bestimmte Software für die wissenschaftliche Datenauswertung die Speicherung von Informationen über enthaltene Variablen und deren numerischen Werte. Die Lesbarkeit und Nachnutzbarkeit dieser Metadaten hängt vom Format und von der verwendeten Syntax ab. Wären sie in einem universellen Format abgespeichert, so könnten sie leicht von anderer Software gelesen, interpretiert und genutzt werden. Somit würde die Metadatenbeschreibung eine Messobjektes entlang der Datenbearbeitungskette stetig angereichert werden und am Ende würde die gesamte Dokumentation über den Messprozess und seine Ergebnisse vorliegen. Metadatenerfassung durch einen Menschen wird so weitgehend vermieden und beschränkt sich daher auf diejenigen Felder, die nicht im Verlauf des Prozesses automatisch beschrieben werden können. Weitere wesentliche Vorteile dieser Vorgehensweise sind, dass fehleranfällige nachträgliche „händische“ Erfassung von Metadaten minimiert wird und die Kooperationsbereitschaft für Forschungsdatenmanagement mit Zielrichtung Open Science beim für die Datengewinnung zuständigen Personal wächst.
Um die Interoperabilität, eine breite Akzeptanz und die Praxistauglichkeit sicherzustellen, wird die Einbeziehung von Messgeräteherstellern und Herstellern wissenschaftlicher Software bei der Festlegung des Formats angestrebt.