Computer Physics and Communication, 2010, 181, 651-662, DOI: 10.1016/j.cpc.2009.11.014

On the Communication of Scientific Data: The Full-Metadata Format

Moritz Riede[a], Rico Schueppel[a], Kristian O. Sylvester-Hvid[b], Martin Kühne[c], Michael C. Röttger[c], Klaus Zimmermann[c], Andreas W. Liehr*[,c]

[a] *Institut für Angewandte Photophysik, Technische Universität Dresden, George-Bähr-Str. 1, 01069 Dresden, Germany*
[b] *Risø National Laboratory, Technical University of Denmark, Frederiksborgvej 399, 4000 Roskilde, Denmark*
[c] *Freiburger Materialforschungszentrum, Universität Freiburg, Stefan-Meier-Str. 21, 79104 Freiburg, Germany*

Abstract

In this paper, we introduce a scientific format for text-based data files, which facilitates storing and communicating tabular data sets. The so-called Full-Metadata Format builds on the widely used INI-standard and is based on four principles: readable self-documentation, flexible structure, fail-safe compatibility, and searchability. As a consequence, all metadata required to interpret the tabular data are stored in the same file, allowing for the automated generation of publication-ready tables and graphs and the semantic searchability of data file collections. The Full-Metadata Format is introduced on the basis of three comprehensive examples. The complete format and syntax is given in the appendix.

Key words: data format, text files, units, physical quantities
PACS: 01.20.+x, 07.05.Hd, 07.05.K

1. Introduction

In the last few years an increasingly sophisticated experimental infrastructure has evolved enabling scientists to share not only knowledge but also primary data via scientific publications [1, 2, 3]. With this increase in sharing primary or processed scientific data the lack of intuitive and well defined data formats for simple tabular data has become increasingly obvious. For complex data sets like those dealt with in the earth sciences, adequate binary formats like the Network Common Data Form (netCDF [4]) or the Hierarchical Data Format (HDF [5]) are well established [6, 7], and the publication of observational geophysical data

*Corresponding author
Email address: liehr@users.sourceforge.net (Andreas W. Liehr)

in World Data Centres has developed into an effective mechanism for data exchange [8]. Another example is the information technology infrastructure for handling the data of the ATLAS experiment [9], where the event data is mainly stored in the ROOT file format [10]. For less complex data structures, like tabular which are typically encountered in many parts of natural and technical sciences, no single standard format has evolved.

The success of the HDF and netCDF relies on the fact that these formats are well defined and integrate smoothly into the workflow of scientists in different laboratories. Although these formats are capable of storing and documenting simple tabular data, the overhead of work needed to process binary files generally poses a barrier to the use of these formats in fields where complex data structures are seldomly dealt with.

A natural requirement of a standardized file format for tabular data is that it allows scientists to add observations, notes, parameter specifications and analysis results by editing in clear text using any given text editor. This constitutes what most of the data formats used in laboratories around the world have in common. However, as text files are easy to handle, every laboratory, working group or even scientist has an individual format of documenting scientific results with text-based formats. While this is completely sufficient in a short term perspective, it becomes intractable with the tendency of research projects to rely on the cooperation of international consortiums involving many different laboratories. Furthermore, in publishing scientific results, there is an increasing demand to also provide processed data as supplementary data or even to publish primary data in OpenData repositories [2]. Thus, there is a need for a common data format for tabular data which is:

Readable and self-documented: The data should be written in the same way the scientist is accustomed to reading it, as e.g. in a laboratory notebook. It should be clear, text based and processable with any word processing tool. The file format should include sections which allow the scientist to document the data and its origin, and this to such an extent that no other source be required to understand the origin of the data. This standard implies that the data files are searchable and individual data sets can be tracked by semantic or keyword based queries.

Flexible but structured: The data format must be flexible enough to allow the individual scientist to structure and classify data in an intuitive and convenient way without compromising the overall structure and readability as stipulated by the format. The overall structure of the format must be such that data files may still be processed with common analysis and visualisation software packages, thereby facilitating the automated processing of data from different measurement sources and measurement series. This further implies that format and syntax specifications are largely decoupled such that annotated data may smoothly cross language zones.

Fail-safe and compatible: A fail-safe data format ensures that the format is robust towards misinterpretation by a parser or deviations from the format

specifications. As the format specification is expected to evolve in time, backwards compatibility must always be retained.

Searchable: Communicating scientific results implies that relevant data sets can be found within a certain data collection by means of simple queries. This requires the documentation of scientific data in the form of self-documenting file formats. Furthermore, a collection of scientific data files must be catalogued not only according to bibliographic items or keywords, but also to physical quantities.

Further, it is of paramount importance that the data format integrates smoothly into the existing workflow of the scientist and supports the established workflows of collecting and structuring information. To become widely accepted, the threshold of annotating primary data with additional information must be as low as possible; scientists should not have to start learning a complex syntax or a sophisticated mark-up language, which for all practical purposes will require specialised software tools.

In the following, we present a syntax for a self-documented scientific data file format for tabular data sets which we call the *Full-Metadata Format* (FMF). It is purely text based and FMF-files consist of two parts: the first contains the metadata describing the data written in the second part of the file. Because most scientific software tools support the skipping of some initial header lines, the data stored in the FMF-file can directly be processed as usual. Yet, the documentation of the data remains always at hand. The proposed file format has evolved from the development of high-throughput experimental setups for the processing and characterisation of organic solar cells [11, 12, 13], and is applicable to all kinds of tabular data encountered in science and engineering. A further demonstration of the capabilities of the Full-Metadata Format is constituted by its incorporation into the scientific analysis software Pyphant [14, 15], which supports the computation with units and the analysis of metadata [16, 17].

In the following sections the Full-Metadata Format is described first by two examples highlighting its principles and potential. Then, a third example sketches the capabilities of searching the metadata of FMF-files for relevant data sets. In the appendix the complete format and syntax definitions are listed.

2. A Basic Example: Communicating Simple Tabular Data

In this example a typical data exchange between two work groups is considered to demonstrate the benefit of readable data formats for the communication between scientists. The goal of the cooperation in this example might be the enhancement of the power conversion efficiency of organic solar cells, or the numerical modelling of the characteristics of solar cells with respect to production parameters. This requires exchanging data between groups. In Fig. 1a the screenshot of a typical data set of a current-voltage characteristic is shown, which is formatted in the most common data format for tabular data: plain columns of numbers. The corresponding graphical representation of the data is

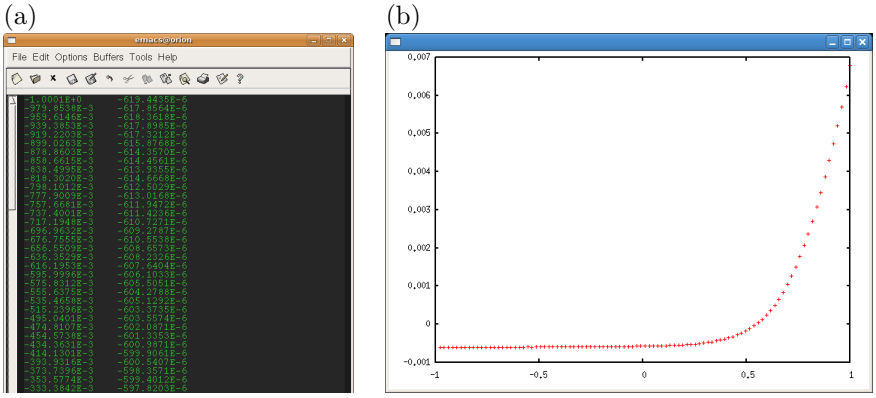

Figure 1: A typical data file exchanged between scientists. (a) Screenshot of a text editor's view of the data file and (b) the corresponding plot with a qualitative relation from the values listed in the data file.

shown in Fig. 1b. The missing axis labels indicate that important information like the name, symbol, and units of the plotted physical quantities are not provided with the data set, thus only allowing for a qualitative assessment of the data.

This lack of information could be clarified by a phone call or an email. Typically, the response to such requests depends on how the working group internally documents the primary data. It may either be well documented by means of protocols in the laboratory notebook of the scientist in charge, but the protocol has not been attached to the e-mail. Alternatively, the data file format is standardised by an internal format convention of the working group, but the format has not been documented. In both cases, a useful response would at least communicate the following: the first column is voltage V in units of Volt, the second column is current I in units of Ampere, the current I is measured as a function of V, the device has an active area A_{pv} of 5.3mm^2, and is exposed to a mismatch corrected illumination intensity $I_{AM1.5}$ of 100mW/cm^2 [18]. Having this information at hand, the diagram in Fig. 1b can be labelled correctly as required for further processing, publication, and understanding (Fig. 2). In addition, characteristic properties like the fill factor ($FF = 45.5\%$) and the power conversion efficiency ($\eta = 2.95\%$) can be extracted from the data [18]. However, this is only a temporary solution, as the original data file is unlikely to be annotated accordingly. The next time the data set is used, the same questions will arise. The data set might even become completely useless if the relevant protocol of the laboratory notebook cannot be identified anymore or if the person responsible for the measurement cannot clarify the units [19].

Clearly, it would be better to annotate the data set with the missing information right from the start. Using the proposed format, the FMF-file corresponding to the data depicted in Fig. 2 is shown in Fig. 3. Some metadata

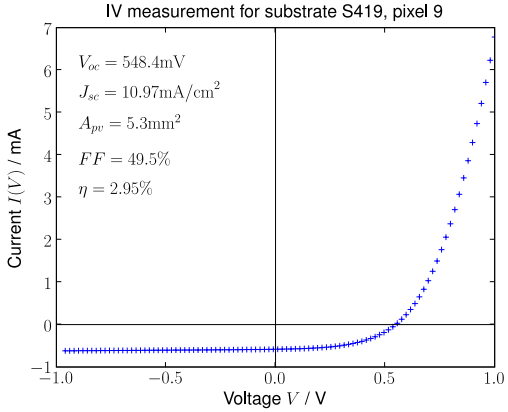

Figure 2: Publication-ready graphics of an IV-characteristic based on an FMF-file (Fig. 3): IV measurement for substrate S419, pixel 9. The solar cell characteristics are measured under illumination with a mismatch corrected intensity of $I_{AM1.5} = 100\text{mW/cm}^2$.

and the first few lines of the raw data are shown. The list of metadata is not exhaustive but shows enough to highlight the possibilities of the file format. Note, that the proposed file format has similarities with the INI file format [20] but goes beyond it in its possibilities due to extra rules. The detailed syntax is given in appendix A.

The file shown in Fig. 3 starts with a single line describing the version of the Full-Metadata Format. The next part contains all the metadata required for understanding the actual data. This metadata is given in a simple and user-friendly way by structuring the file into sections. Bibliographic information resides in section [∗**reference**], column definitions in section [∗**data definitions**], and the corresponding columns of data in section [∗**data**]. The bibliographic information is either used for internal archiving purposes or for publishing the data file in an OpenData repository [2], similar to [21] for example. These three sections are mandatory and comprise the fundamental structure of a FMF-file.

The other sections in the example in Fig. 3: [**setup**], [**parameters**], and [**fingerprints**] are not preceded by an asterisk. These are user defined sections and can contain arbitrary extra metadata. All sections, except the [∗**data**] section, contain items coded as colon separated

$$key : value \tag{1}$$

pairs. The *key* cannot contain a colon, because the first colon per line separates *key* and *value*. A *value* can be boolean, numerical, a quantity, a timestamp, or a string. In section [∗**data definitions**] the value must be a column specifier (A.3).

According to the metadata, the file shown in Fig. 3 was created by *Moritz Riede* on *17th of April 2006* at *18:55:38* local time, which is two hours ahead

```
; -*- fmf-version: 1.0 -*-
[*reference]
creator: Moritz Riede
created: 2006-04-17 18:55:38+02:00
title: IV measurement for substrate S419
substrate name: S419
pixel: 9
place: Materials Research Center Freiburg, Germany
comment: IV illuminated (annealed, 300s, 150C), batch3
[setup]
setup: omm-table
measurement type: IV
setup version: v5.4
[parameters]
pixel area: A_{pv} = 5.3 mm^2
substrate position: p = 3
table position:  x = 43.68 mm
filter: none
illumination intensity: I_{AM1.5} = 100 mW/cm^2
4-wire measurement: true
[fingerprints]
short circuit current density: J_{sc} = 10.97 mA/cm^2
open circuit voltage: V_{oc} = 548.4E-3 V
fill factor: FF = 49.5 %
efficiency: \eta = 2.95 %
[*data definitions]
voltage: V [V]
current: I(V) [A]
[*data]
-1.0001E+0      -619.4435E-6
-979.8538E-3    -617.8564E-6
-959.6146E-3    -618.3618E-6
-939.3853E-3    -617.8985E-6
-919.2203E-3    -617.3212E-6
    ⋮               ⋮
```

Figure 3: An abridged version of the beginning of a data file in the Full-Metadata Format [21].

of UTC (see Tab. 6). It contains data for the solar cell on pixel *9*, located on a substrate with the unique identifier *S419*. This identifier can be used to reference the processing and measurement history of the solar cell [12, 19]. A short comment completes the [**∗reference**] section.

In this example, the section [**setup**] is used to describe the measurement type and the setup used. Many measurements can be carried out with different setups, each with their own distinct features, which are relevant when interpreting the data [12, 13]. A set of important measurement parameters necessary for the interpretation of the data are recorded within the section [**parameters**]. A special mention should be given to key-value pairs which we characterize as *quantities*. For instance, in section [**parameters**] the active area of the solar cell is specified as:

```
pixel area: A_{pv} = 5.3 mm^2
```

It is written like a typical parameter specification and is comprised of a name ("pixel area"), a symbol in LaTeX-notation (A_{pv}) [22], a numerical value and a unit (which may be omitted for unit-less values). Symbols in LaTeX-notation other than characters of the Latin alphabet can easily be included. *Quantities* also support the specification of measurement uncertainties and estimation errors (cf. Tab. 5). The last item shown in section [**parameters**] is boolean, indicating that the measurement was carried out in 4-wire mode.

Initial results derived from the raw data of solar cell on pixel 9 within substrate S419 are listed in the section [**fingerprints**]. As such these data are redundant as they are derived from the IV-data, but can be very helpful for a quick overview and for processing of the recorded data.

The last two sections, [**∗data definitions**] and [**∗data**] differ from the preceding sections. Hence, the n^{th} line of [**∗data definitions**] describes the n^{th} column of the following [**∗data**] section containing tabular measurement data. The format of the column description is chosen to resemble a typical axis label having a name, a symbol, and a unit in brackets. In addition, the functional relationship of the tabulated quantities is given by explicitly denoting current $I(V)$ as being measured with a dependency on voltage V.

3. A More Complex Example: Documenting Experiment And Analysis Together

Applying the basic example of Sec. 2 to other data sets quickly shows that, for general purposes, a more extensive syntax is often needed. For example measurement errors need to be specified or more than one table may be needed for a comprehensive description of the data sets.

An example of an FMF-file with two tables is shown in Fig. 4. It documents the work of two students in measuring the Faraday constant Fa in the course of a practical exercise [23]. The experiment relies on Faraday's second law and uses a coulometer for measuring the volume fractions of hydrogen and oxygen evolving over time due to a constant current I applied to an aqueous solution of sodium

hydroxide. From a progression of measurements, Faraday's constant can be computed by converting the volume fractions to normal conditions (1023mbar and 273K), estimating the evolved gas volume per time interval V' from the time series, and evaluating

$$\mathrm{Fa} = 22.4\frac{\ell}{\mathrm{mol}} \cdot \frac{I}{N_e V'} \tag{2}$$

both for hydrogen and oxygen. Therefore, room temperature and barometric pressure at the time of the experiment comprise important metadata for evaluating the measurement. These physical quantities are specified in section [**measurement**] of the FMF-file shown in Fig. 4 together with their measurement uncertainties:

```
room temperature: T = (292 \pm 1) K
barometric pressure: p = 1.0144 bar \pm 10 mbar
```

This section also gives the current I, applied to the sodium hydroxide solution, and its measurement error. Note that the error specification is very similar to the way in which a scientist would describe the data in a report. Other possibilities for specifying errors are listed in Tab. 5 of the appendix.

Because the experiment deals with two different gases, namely hydrogen and oxygen, which differ in terms of their number of electrons N_e per reaction, the Faraday constant is retried individually for each time series. Therefore, two tables are needed for adequately describing the experiment: one table specifying the material parameters and the result of the data analysis, and another table listing the time series of measured volume fractions. The names of these tables as well as the associated symbols are defined in section [**∗table definitions**] of the FMF-file in Fig. 4. It tells that the table named *analysis*, A, is followed by the table *primary*, P. Each table then consists of sections [**∗data definitions:** X] and [**∗data:** X] with X referencing the symbol of the table such that each pair can easily be identified.

In this example, two cases of error specifications are needed in the tables: i.e. specifying constant measurement errors valid for elements of a specific column, and assigning special error columns. The specification of constant measurement errors is shown in section [**∗data definitions: P**] of the second table in Fig. 4:

```
time: t [min] \pm 5 [s]
hydrogen volume: V_{H_2}(t) \pm 0.2 [cm^3]
oxygen volume: V_{O_2}(t) \pm 0.2 [cm^3]
```

In this example, time t is measured in units of minutes with an accuracy of 5 seconds and volumes $V_{H_2}(t)$ and $V_{O_2}(t)$ are measured in units of cm^3 with an accuracy of $0.2\mathrm{cm}^3$. With this information at hand, the primary data of section [**∗data: P**] can be plotted as shown in Fig. 5.

The specifications of non-constant errors are shown for V' and Fa in Fig. 4. The errors $\Delta_{V'}$ and Δ_{Fa}, respectively, are defined in section [**∗data definitions: A**] and are explicitly related to the measured quantity as:

```
; -*- fmf-version: 1.0 -*-
[*reference]
creator: Andreas W. Liehr and Andreas J. Holtmann
created: 1995-01-10
title: Measurement of Faraday's constant - An example of documenting ...
place: Physikalisches Institut, Universität Münster
lab excercise manual: Physikalisches Institut (Hrsg.): Anleitung zu ...
[measurement]
room temperature: T = (292 \pm 1) K
barometric pressure: p = 1.0144 bar \pm 10 mbar
current: I = (171 \pm 1) mA
solution: sodium hydroxide
[analysis]
estimation method: line of best fit
[*table definitions]
analysis: A
primary: P
[*data definitions: A]
gas: G
number of electrons: N_e
volume per time interval: V' \pm \Delta_{V'} [cm^3/min]
uncertainty of ratio: \Delta_{V'} [cm^3/min]
Faraday constant: Fa \pm \Delta_{Fa} [C/mol]
error of Faraday constant: \Delta_{Fa} [C/mol]
[*data: A]
;G      N_e     V'      \Delta_{V'} Fa          \Delta_{Fa}
H_2     2       1.256   0.065       91400       5500
O_2     4       0.562   0.04        102200      7800
[*data definitions: P]
time: t [min] \pm 5 [s]
hydrogen volume: V_{H_2}(t) \pm 0.2 [cm^3]
oxygen volume: V_{O_2}(t) \pm 0.2 [cm^3]
[*data: P]
2.5     2.0     2.1
4       4.0     2.4
6       6.6     3.7
9       9.8     4.2
11      13.8    6.0
13      15.0    6.8
15      18.2    8.4
17      20.0    9.4
19      23.4    11.0
21      26.0    12.2
23      28.8    13.8
25      31.6    14.6
27      33.6    15.8
29      36.6    17.2
31      39.0    18.4
```

Figure 4: Measurement of Faraday's constant - An example of documenting experimental data and their analysis within one FMF-file [23].

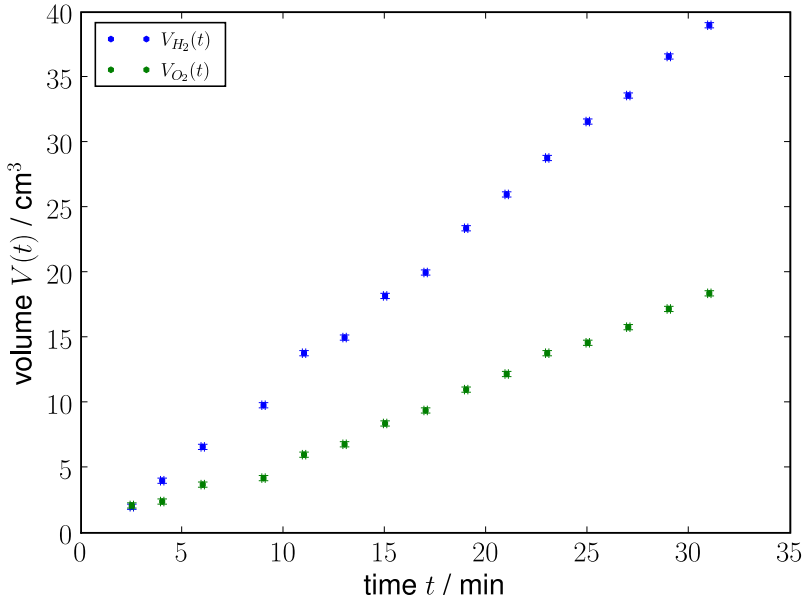

Figure 5: Measurement of Faraday's constant. The diagram visualises the table of data documented in section [**∗data: P**] of Fig. 4 and uses information from section [**∗data definitions: P**] to label the graph accordingly [23].

gas	N_e	$V' \pm \Delta_{V'}$ [cm^3/min]	Fa $\pm \Delta_{Fa}$ [C/mol]
H_2	2	1.256 ± 0.065	91400 ± 5500
O_2	4	0.562 ± 0.04	102200 ± 7800

Table 1: Formatted analysis table A of the FMF-file shown in Fig. 4. The table lists the name of the gas, the number N_e of electrons per reaction, the ratio V' of released gas per time interval and the resulting Faraday constant. The ratio V' has been determined from table P (Fig. 4) by plotting volume fraction against time for each gas (Fig. 5) and estimating the line of best fit.

```
Faraday constant: Fa \pm \Delta_{Fa} [C/mol]
error of Faraday constant: \Delta_{Fa} [C/mol]
```

These data definitions mean that the column listing of the Faraday constant is followed by a column with the corresponding measurement error. Because this table consists of six columns, the creator of the FMF-file decided that the readability would be improved by starting section [**∗data: A**] with a comment repeating the symbols defined in section [**∗data definitions: A**]. The comment is introduced by a leading semicolon.

Altogether, sections [**∗data definitions: A**] and [**∗data: A**] of Fig. 4 give a simple textual representation of Tab. 1, which could be the summary of an experiment. Section [**∗data definitions: A**] lists the name of the gas, the number N_e of electrons per reaction, the ratio V' of released gas per time interval and the resulting Faraday constant. The measured values of Fa depend on the number N_e of electrons per reaction. As can be seen from the last column of Table A in Fig. 4 the measurement of the Faraday constant can deviate up to 6% from the precise value of Fa = 96 485.3399 C/mol [24], but has been correctly determined within the error margins.

In the experiment described, the constant has been determined by means of the line of best fit. Years later however, it occurs to the students (or maybe even their successors in the practical exercise) that moving a ruler around on a piece of paper is not the best way to analyse the data. Since the primary data is available in a form easily understood, they decide to redo the analysis with the more sophisticated means of a least square fit while taking into account that the anode is likely to have an oxide layer, which increases V'_{O_2} in the beginning of the experiment. With $V'_{H_2} = 1.202\text{cm}^3/\text{min} \pm 1.0\%$ and $V'_{O_2} = 0.596\text{cm}^3/\text{min} \pm 2.2\%$ Faraday's constant is now determined as 95600 ± 1500 C/mol for H_2 and 96500 ± 2700 C/mol for O_2. Both these values are far more accurate compared to the original results. This improvement was possible, because the original information was preserved in a way that allowed its interpretation. Furthermore, it could be understood because the method of analysis was indicated.

While this example might seem trivial, it is common that data experimentally gathered by one scientist is useful years later to another scientist. Often, the first scientist has become unavailable and the data can no longer be found, let alone understood. This is a waste of time and resources that should be minimized.

4. An Advanced Example: Searching Scientific Data In Terms Of Units

How to search data files for certain physical quantities is explained using an example given in Tab. 2, where we consider four energy related quantities pertaining to different experiments: work $W = 23\text{kJ}$, energy $E = 10\text{keV}$, caloric value $H = 10\text{kcal}$, and power $P = 0.01\text{MW}$. A classical full-text search cannot reveal any correlation between their notation *work*, *energy*, *caloric value*, and *energy*. The same holds for the units kJ, keV, kcal, and MW. Therefore, a question like

> "Which measurements involves an energy in the range between one thousand and one billion Joules?"

cannot be formulated as a full-text query. Instead, the values quantifying energies have to be identified and the ones lying outside the desired interval $[1\text{kJ}, 1\text{MJ}]$ have to be sorted out. By aggregating the values to be filtered in set $\mathbb{M} = \{W, E, H, P\}$ the result $\mathbb{M}_E$ is given by

$$\mathbb{M}_E = \{m \in \mathbb{M} | \dim\, m = L^2 T M^{-2}\} \cap [1\text{kJ}, 1\text{MJ}]. \tag{3}$$

Here L, T, M represent the dimensions length, time and mass, respectively.

The computation of this intersection can be carried out by expressing the elements of $\mathbb{M}$ in basic SI-units and decomposing each quantity into an 8-tuple. The elements of the tuple are the numerical value of the quantity and the powers of its SI-units. In terms of pattern recognition, this 8-tuple is denoted the feature vector. Hence, a physical quantity q is uniquely characterised by its feature vector $\vec{q}_F = (q_0, \ldots, q_7)$ with

$$q = q_0 \cdot \text{m}^{q_1} \cdot \text{kg}^{q_2} \cdot \text{s}^{q_3} \cdot \text{A}^{q_4} \cdot \text{K}^{q_5} \cdot \text{mol}^{q_6} \cdot \text{cd}^{q_7} \tag{4}$$

where $q_0 = \{q\}$ is the measure of q and $q_1, \ldots, q_7$ determine its SI-units:

$$[q] = \text{m}^{q_1} \cdot \text{kg}^{q_2} \cdot \text{s}^{q_3} \cdot \text{A}^{q_4} \cdot \text{K}^{q_5} \cdot \text{mol}^{q_6} \cdot \text{cd}^{q_7}. \tag{5}$$

In regard to the example in Tab. 2, all powers except of length, time and mass are zero and only quantities given in units of $\text{m}^2\text{kg}\,\text{s}^{-2}$ $(q_1, \ldots, q_7) = (2, 1, -2, 0, 0, 0, 0)$ are energies and therefore are relevant for determining $10^3 \leq q_0 \leq 10^6$ (Tab. 2). Therefore, only the quantities work $W = 23\text{kJ}$ and caloric value $H = 10\text{kcal}$ pertain to experiments determining energies between 1kJ to 1MJ.

This example illustrates how scientific data sets can be made searchable on the basis of adequate documentation, such that proper documentation of data sets directly enables the re-usability of scientific results.

5. Discussion

We have shown how a text file can be used as a scientific data format enabling storage of tabular data sets in a consistent and self-descriptive fashion. The real

	feature vector $\vec{q}\,(4)$							
meta-information	q_0	q_1	q_2	q_3	q_4	q_5	q_6	q_7
work: $W = 23$ kJ	$23 \cdot 10^3$	2	1	-2	0	0	0	0
energy: $E = 10$ keV	$1,602 \cdot 10^{-15}$	2	1	-2	0	0	0	0
caloric value: $H = 10$ kcal	$41,9 \cdot 10^3$	2	1	-2	0	0	0	0
power: $P = 0.01$ MW	$10 \cdot 10^3$	2	1	-3	0	0	0	0
search interval	$[10^3, 10^6]$	2	1	-2	0	0	0	0

Table 2: Classification of physical quantities by means of feature vectors. The feature vector $\vec{q}\,(4)$ of quantity q is an 8-tuple, which is composed from the measurand $q_0 = \{q\}$ in basic SI units and its dimension coded as powers of units (5). Feature vectors with identical elements $q_1, \ldots q_7$ correspond to the same physical quantity.

novelty of the presented data format is the systematic way in which all relevant metadata required to understand the data can seamlessly be included. In the language of the data-information-knowledge-wisdom hierarchy [25] this means that the data set is upgraded from the data level to the information level. The promotion to the information level has the following significant advantages.

First, it improves the conditions under which scientists can communicate or share scientific data. This may occur within a working group, with external cooperators, or within the scientific community in general. Because unhindered communication is one of the most important prerequisites for a successful collaboration, this feature cannot be overestimated. Second, this approach facilitates the long-term integrity of scientific data; e.g if primary data from an old project must be revisited many years later or if data is passed on to future generations of scientists. At present, it is rather common that a scientist is not able to find all relevant metadata to understand an old data set. Often such data-erosion is simply due to the metadata residing on a different data storage medium than the primary data itself. Working with a data format which includes the relevant metadata avoids this problem altogether.

Using a self-describing format like the one presented here increases the longevity of primary data and thus may improve the quality of science in general. In scientific communities like the geo-sciences or high-energy physics this has especially proven to be the case. In these fields, large data sets and the pressure to communicate them effectively have led to a standardisation of data formats and a culture of sharing such data. Due to the complexity of the data generated in those fields, more sophisticated file formats such as HDF5, netCDF or ROOTS [5, 4, 10] are in use.

In contrast to these complex data formats, the Full-Metadata Format is designed with the needs of so-called *Small Science* [26] in mind. Research by small working groups and individuals producing simple tabular data still occupies a central position in most scientific disciplines. Although the awareness of a systematic management and sharing of data is already rising, an appropriate data format for *Small Science* has yet to fully evolve. One obstacle is that data documentation using the existing eXtensible Markup Languages (XML) like XDF [27] or VOTables [28] adds too much overhead to the content and is

cumbersome to read, edit, and process with existing scientific software tools. On the other hand, the validity of XML files can be checked on the basis of Extensible Stylesheet Language Transformations (XSLT) with respect to a certain syntax [29]. This approach also features the possibility to transform a given set of XML files into another format. Therefore, XSLT could be easily applied for converting XDF or VOTables into the Full-Metadata Format. Concerning the Full-Metadata Format, the same features can be realised on basis of an ANTLR grammar [30], which can be applied for verifying the syntax of FMF-files and converting them into another format including XML.

However, it is much more practical to use a special parser like the Python module fmfile of the information analysis framework Pyphant [14, 15] for checking the syntax of FMF-files, parsing their content and processing it. The most important feature of Pyphant is its ability to compute with physical units. These are realised on basis of the Python module PhysicalQuantities [31], which is also the basis of the SI and non-SI units given in appendix B.

It is important to note, that the Full-Metadata Format accounts for the abbreviation "a.u." in terms of *arbitrary units* and not in terms of atomic units: atomic units form a different system of units in which several physical constants, e.g. Planck's constant and the permittivity of vacuum, are defined as unity [32]. This collides with the searchability of scientific data discussed in Sec. 4.

The approach presented in this paper is simple: describing simple tabular sets of data with simple text files in a way which scientists and engineers are acustomed to and which requires a minimal change in the individual workflow and habits. In general, this means documenting the data in a way one would like to read it in a laboratory journal or in a paper, e.g. within a diagram or a figure caption. The use of plain text files ensures that the scientist can apply this documentation technique instantaneously using a basic information technology infrastructure. Still, these text files can be parsed readily due to their simple and systematic structure [15].

Because of its simplicity and the self-describing character, the Full-Metadata Format offers many possibilities:

- The clear-text documentation of scientific data simplifies its re-use.
- The use of plain text files makes the data ideal for long term preservation [33].
- The communication of data does not require a complex infrastructure; text files can be sent by email or even be printed to analogue media.
- Because the data is connected to the relevant units, special software which is able to process these units during scientific data analysis like Pyphant [14, 15] can be used such that processing and visualisation of the data can be automated.
- Furthermore, the use of the relevant units enables a semantic search within a collection of data sets.

A drawback of the Full-Metadata Format results from the fact that the end-of-line (EOL) character of text files is not uniquely defined for all operation systems, which causes text files to be displayed incorrectly after being transferred to a different type of operation system. However, this problem is generally known and appropriate tools are available [34].

6. Conclusion

The advantage of the suggested file format is its ease of use and its user-friendly syntax, which is in contrast to the computer-friendly syntax of markup-languages. The purpose of the Full-Metadata Format is to document small tabular data sets, mainly produced in fields generalised as *Small Science*. For these scientific communities, the use of the Full-Metadata Format can be the starting point to a systematic management for scientific data in the form of information, and thus the starting point for participation in the growing culture of data sharing.

The authors would like to encourage the reader to engage in the application of the Full-Metadata Format and to actively participate in the improvement of the proposed format.

A. The Syntax Of The Full-Metadata Format

The appendix comprises a technical description of the syntax characterising the Full-Metadata Format in version 1.1. It is meant as a guide to the format and shows comprehensive tables of coding examples. Therefore the appendix intentionally repeats certain parts of the format in order to minimise browsing for a specific piece of information.

Data files written in the Full-Metadata Format always consist of three parts:

Headline (A.1),

Metadata (A.2),

Tables (A.3).

The headline is a comment indicating how to interpret the file on a formal level. Following the headline is the main body of the file, which is structured in sections. While the file body can contain arbitrarily many sections with metadata and measurement data, at least three mandatory sections are needed for a meaningful FMF-file. These sections are named **[*reference]**, **[*data definitions]** and **[*data]**. The **[*reference]** section contains the metadata necessary for referencing the data set and the **[*data definitions]** and **[*data]** sections represent a table of data. This minimal structure is shown in Fig. 6, while the general structure is summarized in Fig. 7.

Headline	; −∗− fmf−version: 1.0 −∗−
Metadata	**[∗reference]** *title*: A concise description of the data set *creator*: The persons in charge *created*: Timestamp *place*: The location, where the data have been collected
Tables	**[∗data definitions]** ;One *key:column* item per column of data tabulated in **[∗data]** ⋮ **[∗data]** ;One column for each *key:column* item specified in **[∗data definitions]** ⋮

Figure 6: Minimal structure of a FMF-file.

From a grammatical point of view, the Full-Metadata Format consists only of three different types of lines, defined as follows:

Comments are indicated by a leading semicolon (;) or a leading hash (#). The comment character used for the headline (A.1) has to be used consistently for all other comments in the same file. A comment character in a key or value is treated as a normal character.

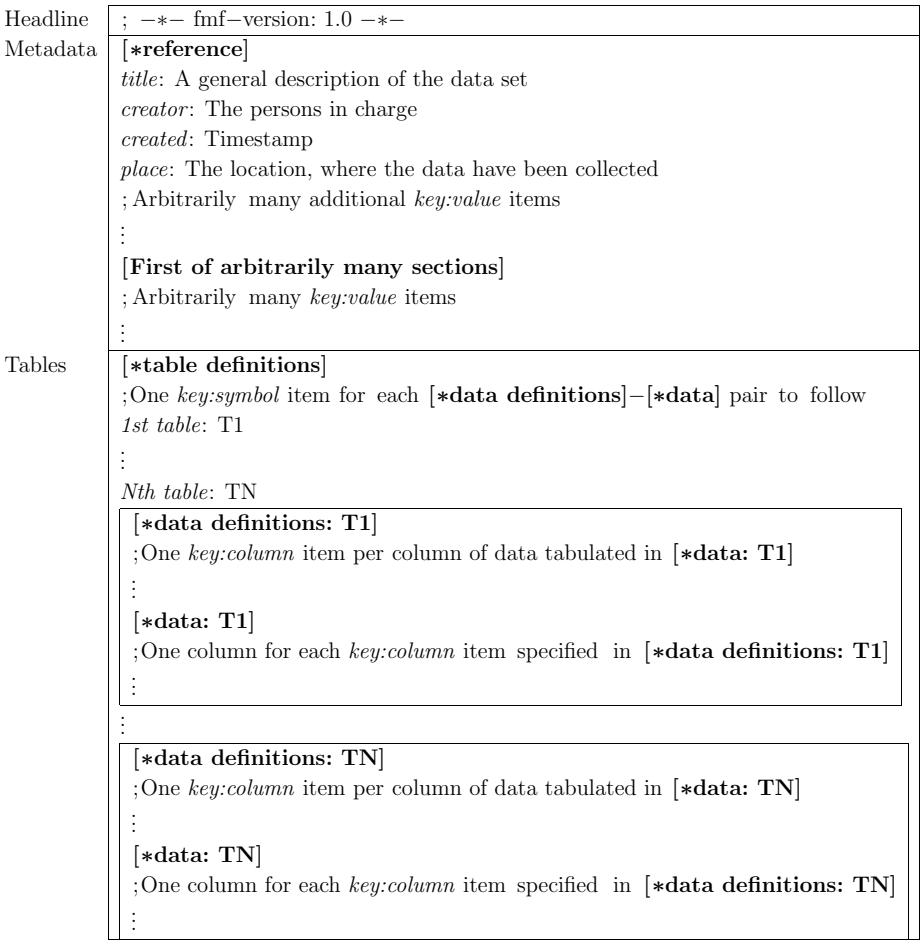

Figure 7: The general structure of an FMF-file. The headline is used to define the file coding and delimiter in the tables, of which several are present in the file.

Section headers are embraced by square brackets [ ] and have to be unique throughout the file. In this and future versions of the FMF format specification only reserved sections may start with an asterix (*). Any other allowed character sequence can be used for arbitrary sections. In this version, the following reserved sections are put to use:

- **[*reference]**,
- **[*table definitions]**,
- **[*data definitions]**, and
- **[*data]**.

***Key:value* items** are used in all but the **[*data]** section. A *key* can consist of

all characters except the colon (:), which is used to separate *key* and *value*. Each *key* has to be unique within its section. The different types of *values* are discussed in A.2. In the [***reference**] and all user defined sections, arbitrary value types may be used. The [***table definitions**] section may only contain *symbols* as values and the [***data definitions**] sections only *column specifications*, both for reasons that will become clear later on.

Rows of data are collected in [***data**] sections. They represent classical tabulated data sets. Other column separators than tab stop can be specified in the headline (A.1).

A.1. Headline

The headline is a special comment, which indicates how the content of the file is to be interpreted. Foremost, this includes foremost the encoding, which tells the computer how to translate the bytes of the file into characters, and the separator which splits the table rows of the [***data**] section into the appropriate cells. The headline also mandatorily specifies the version of the Full-Metadata Format employed in the file: It uses the Emacs style file syntax [35] and thus looks like

```
; -*- fmf-version: 1.1 -*-
```

In addition, *coding* (default = utf-8) and *delimiter* (default = tab) can be specified (Tab. 3). The *key:value* items have to be separated by a semicolon. Although the semicolon (;) is the default comment character, comments can alternatively be introduced by a hash (#). The comment character used in the headline has to be used throughout the file.

A.2. Metadata

Metadata is an essential part of the data file, because it describes the context from which a data set has been collected. It is structured by sections, which start with a unique section header consisting of a section identifier enclosed in square brackets. Section identifiers starting with an asterisk are reserved for this or any future version of this specification. All section headers except the [***data**]-header are followed by lines of

```
key: value
```

pairs. The key can contain any valid character except a colon, which separates *key* and *value*. The value is always a textual representation of some information. However, in order to allow for an automated interpretation of the information the Full-Metadata Format defines some conventions for the representation of numerical and boolean values, quantities and complex strings:

Boolean values are given by the words “true” or “false”. They can be written in lower case letters, capital letters, or with a starting capital letter. A list of boolean values is defined by separating the individual values by commas.

Numerical values are textual representations of integer, real or complex scalars. Due to the restrictions of floating point arithmetics, the accuracy of real and complex scalars is restricted by the number of bits used for encoding the scalar. Optionally, a numerical value can be complemented by an uncertainty specified in common scientific notation. Furthermore a numerical value can be annotated by a symbol in LaTeX-notation, which is prefixed to the number and is related to the latter by an equal sign. A list of numerical values is defined by separating the individual values by commas. A comprehensive list of possible numerical formats is given in Tab. 4.

Quantities are measurands, estimations, or control parameters of an experiment or simulation. They are characterised by a numerical value and a unit. Units are extensively described in appendix B. A list of quantities is defined by separating the individual quantities by commas. A comprehensive list of examples of quantities is given in Tab. 5.

Timestamps are ISO formatted date-time strings [36], for example "2006-04-17 18:55:38+02:00" for the 17th of April 2006 at 18:55:38 local time, which is 2 hours ahead of UTC (see Tab. 6). If the time zone information is omitted, the local time zone is assumed. However, in view of international cooperations the reference to UTC should always be included. A timestamp can also be amended by an uncertainty, which is indicated by $+-$ and a temporal quantity. This is useful for applications in forensics [37]. A list of timestamps is defined by separating the individual timestamps by commas.

Strings are the most flexible type of *values* to be returned, because a string of characters can map any textual information. In particular this applies if the mapping to boolean values, numerical values, quantities, or timestamps does not match. In order to prevent the interpretation of a textual *value* in terms of numerical values or quantities the information can always be enclosed in quotation marks. However, for more complex strings like multi-line strings, lists of strings or strings containing quotation marks, some conventions have to be met, which are listed in Tab. 7.

A.3. Tables

The [***tables**] section is a means to include more than one table in a single FMF-file. This creates the need to identify corresponding [***data definitions**] and [***data**] sections. To this end, each table is assigned a name and a symbol, which in turn is used for identifying the table throughout the file. This information is found in the [***tables**] section. The relevant sections for multi-table files are:

[***table definitions**] This section has one *key:symbol* item per table. While the *key* acts as a descriptive name for the table, the *symbol* is used to relate the [***data definitions**] and the [***data**] sections to each other. Therefore

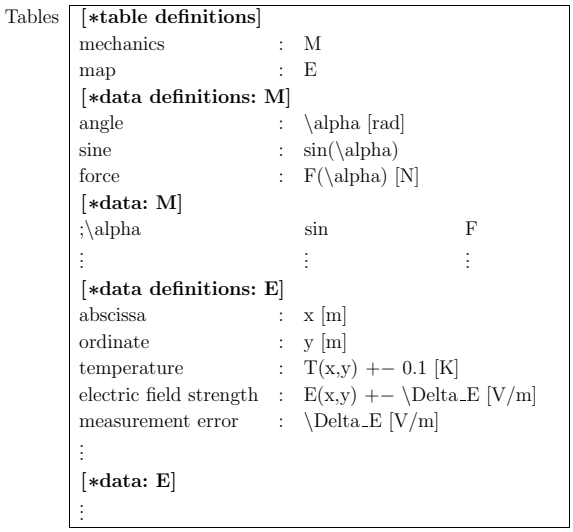

Figure 8: Structure of the tables part comprising two tables.

these sections reference the table *symbol* within their section header as **[*data definitions: symbol]** and **[*data: symbol]**. The **[*table definitions]** section can be skipped, if only one table is given within the FMF-file. In this case, the data definitions and the data sections do not reference a symbol and thus are captioned by **[*data definitions]** and **[*data]** (see Fig. 6). In general, LaTeX-notation for symbols is allowed.

[*data definitions] These sections describe the columns of data given in the respective **[*data]** sections by means of *key:column* items. The n^{th} item of a **[*data definitions]** section describes the n^{th} column in the **[*data]** section. A *column* value specifies a *symbol* referencing the tabulated quantity. Optionally, it can also define a unit, an uncertainty and the functional dependency on another quantity. The uncertainty is either constant or might be tabulated in another column. For the details refer to Fig. 8.

[*data] These sections are tables of data as shown in Figs. 3 and 4. The columns can contain strings, numerical values, and quantities, whose symbols and names are defined in section **[*data definitions]**. The same holds for uncertainties and units. By default, columns are separated with tabs. Other delimiter like whitespace can be explicitly defined in the header line of the FMF-file (Table 3).

Variable	: Value	Status
fmf-version	: 1.1	Mandatory. Version presented in this paper is 1.1.
coding	: utf-8	Default character encoding [38].
	: cp1252	Example for character encoding with WinLatin1 code page [39].
delimiter	: \t	Default delimiter is tab.
	: whitespace	Example for column separation by whitespaces.
	: semicolon	Example for column separation by semicolons (;).
	: ,	Example for column separation by commas.

Table 3: Variables defined in the headline. Comprehensive information on alternative code pages can be found at [40].

Explaining key	: Value
Integer	: 1
Negative integer	: -2
Floating point number	: 1.0
Floating point number with leading decimal dot	: .1
Floating point number with exponential	: 1e-10
Another floating point number with exponential	: -1.1E10
Complex number	: 1+2j
Another complex number	: 1.1+2j
Complex number with zero real part	: 2j
Complex number with zero imaginary part	: 1+0j
List of floats	: 1.0, .1, 1e-10, -1.1E10
Parameter	: P = 42.0
Parameter with uncertainty	: Q = 42.1 +− 0.2
Parameter with relative uncertainty	: Q' = 42.1 +− 0.48%

Table 4: Examples for textual representations of scalars. A value is interpreted as integer if the respective string contains only digits and an optional leading sign. A string is interpreted as floating point number if it contains a decimal dot or an exponent indicated by an embedded 'e' or 'E'. Complex numbers are coded as a sum of real and imaginary parts in integer or floating point notation. The imaginary part is indicated by a trailing 'j'. Lists of numbers are built from comma separated numbers. Special values like NaN (not a number) or +INF and −INF for $\pm\infty$ are also allowed (IEEE 754) [41]. Optionally numerical values can be complemented by uncertainties and a symbol in LaTeX-notation. Note that the uncertainty sign can also be given by \pm.

Explaining key	: Quantity
Physical quantity	: 2.0 ohm
	: 2.0 kg*m**2/A**2/s**3
	: 2.0 kg*mˆ2/Aˆ2/sˆ3
	: 2.0 kg*mˆ2*Aˆ-2*sˆ-3
Physical quantity with uncertainty	: 2.0 ohm +− 0.02 ohm
	: 2.0 ohm +− 20 mohm
	: (2.0 +− 0.02) ohm
	: (2.0 +− 1 %) ohm
	: (1.0 +− 0.01) 2.0 ohm
	: (1.0 +− 1%) 2.0 ohm
Monetary quantity	: 19.99 EUR/m**2
List of quantities	: 2.0 ohm, 2.0 ohm +− 0.02 ohm, 19.99 EUR/m**2
Resistance	: R = 2.0 ohm
Temperature	: \theta = 32.0 K
Measured resistance	: R = 2.0 ohm +− 0.02 ohm

Table 5: Examples for textual representations of *quantities*. They are specified by a numerical value (Tab. 4) and a unit (appendix B). Optionally, quantities can be complemented by uncertainties and a symbol in LaTeX-notation. Note that the uncertainty sign can also be given by \pm.

Explaining key	: Value
date	: 2008-12-16
week date	: 2008-W47-1
date-time	: 2008-12-16T16:51
another date-time	: 2008-12-16 16:51
date-time with seconds	: 2008-12-16T16:51:05
date-time UTC	: 2008-12-16T16:51Z
date-time+2h	: 2006-04-23 14:25:51+02:00
date-time with uncertainty	: 2008-12-16 16:30+−2 hr
list of dates	: 2008-11-17,2008-1-3,2006-2-17,2008-W47-1

Table 6: Examples for ISO formatted date-time strings [36], which can be gathered in lists and can be supplemented by temporal uncertainties.

Explaining key	: Value
Text	: Demonstrating the flexibility of the Full-Metadata Format
Comma separated list	: Freiburger Materialforschungszentrum, Universität Freiburg
Quoted text	: "Freiburger Materialforschungszentrum, Universität Freiburg"
Single quotes	: 'Freiburger Materialforschungszentrum, Universität Freiburg'
Inside quotation	: Arthur C. Clarke's "The Sentinel"
Multi-line	: ''' A multi-line value, that spans more than one line: The line breaks are included in the value. '''
Another multi-line	: """A multi-line value, that spans more than one line: line breaks are included in the value."""
Enclosed quotation marks	: """ "Don't visualise data, document it!" """

Table 7: Examples for textual representations of information, which are mapped to strings of characters. Text values can be quoted by single quotes, a single quotation mark (') and by double quotation marks (") in order to prevent the interpretation of the text value by the parser. Triple quotes are used for multi-line text values or in cases for which the text value starts and ends with quotation marks.

B. Units

Units are defined on the basis of the SI units Metre (m), Kilogram (kg), Second (s), Ampere (A), Kelvin (K), Mol (mol), Candela (cd), and the derived units Newton (N), Pascal (Pa), Joule (J), Watt (W), Coulomb (C), Volt (V), Farad (F), Ohm (ohm), Siemens (S), Weber (Wb), Tesla (T), Henry (H), Lumen (lm), Lux (lx), Becquerel (Bq), Gray (Gy), Sievert (Sv), Radian (rad), and Steradian (Sr). Moreover, monetary values can be defined on the basis of the Euro (EUR) exchange rates as published by the European Central Bank [42]. The units Bit (bit) and Byte (B, 1B = 8bit) are used for documenting topics of information technology [43]. The order of magnitude for all SI units and units derived thereof can be specified by metric prefixes (Tab. 8). For currencies and units of information technology only prefixes with positive exponents are valid. Additionally binary prefixes as defined by norm IEC 60027–2 [43] are used in combination with Bit and Byte (Tab. 9). Constants and additional non-SI units are listed as follows:

Tab. 10 Mathematical and physical constants,

Tab. 11 time units,

Tab. 12 length and area units,

Tab. 13 volume and concentration units,

Tab. 14 mass and force units,

Tab. 15 energy and power units,

Tab. 16 pressure units,

Tab. 17 geometrical, and thermo-dynamical degrees.

This comprehensive collection of SI units and derived SI-units is based on the Python module PhysicalQuantities developed by K. Hinsen [31].

Note, that the Full-Metadata Format accounts for the abbreviation “a.u.” in terms of *arbitrary units* and not in terms of atomic units. This is due to the fact that atomic units form a system of units in which several physical constants are defined as unity [32]. E.g. for Hartree atomic units the mass and charge of the electron, the Bohr radius, the absolute value of the electric potential energy of the Hydrogen atom in its ground state, Planck’s constant and the permittivity of vacuum are unity by definition, which of course collides with the searchability of scientific data discussed in Sec. 4.

Symbol	Prefix	Order of magnitude
Y	yotta-	10^{24}
Z	zetta-	10^{21}
E	exa-	10^{18}
P	peta-	10^{15}
T	tera-	10^{12}
G	giga-	10^{9}
M	mega-	10^{6}
k	kilo-	10^{3}
h	hecto-	10^{2}
da	deca-	10^{1}
d	deci-	10^{-1}
c	centi-	10^{-2}
m	milli-	10^{-3}
mu	micro-	10^{-6}
n	nano-	10^{-9}
p	pico-	10^{-12}
f	femto-	10^{-15}
a	atto-	10^{-18}
z	zepto-	10^{-21}
y	yocto-	10^{-24}

Table 8: Prefixes that can be used for base and derived SI units only.

Symbol	Prefix	Order of magnitude
Yi	yobi-	2^{80}
Zi	zebi-	2^{70}
Ei	exbi-	2^{60}
Pi	pebi-	2^{50}
Ti	tebi-	2^{40}
Gi	gibi-	2^{30}
Mi	mebi-	2^{20}
Ki	kibi-	2^{10}

Table 9: Binary prefixes as defined by norm IEC 60027–2 [43].

Symbol	Value	Description
pi	3.141592653589793	Area of unit circle
c	299792458.*m/s	Speed of Light
mu0	4.e-7*pi*N/A**2	Permeability of vacuum
eps0	1/mu0/c**2	Permittivity of vacuum
G	6.67428e-11*m**3/kg/s**2	Gravitational constant
h	6.62606896e-34*J*s	Planck constant
hbar	h/(2*pi)	Planck constant / 2pi
e	1.602176487e-19*C	Elementary charge
me	9.10938215e-31*kg	Electron mass
mp	1.672621637e-27*kg	Proton mass
Ryd	10973731.568527/m	Rydberg constant
Fa	96485.3399 C/mol	Faraday constant
NA	6.02214179e23/mol	Avogadro number
k	1.3806504e-23*J/K	Boltzmann constant
u	1.660538782e-27*kg	Atomic mass unit

Table 10: Mathematical and physical constants. The physical constants reflect the CODATA-2006 recommendations [24].

Symbol	Value	Description
min	60*s	Minute
hr	60*min	Hour
d	24*h	Day
wk	7*d	Week
yr	365.25*d	Year

Table 11: Time units.

Symbol	Value	Description
AU	149597870691m	Astronomical unit
Ang	1.e-10*m	Angstrom
Bohr	4*pi*eps0*hbar**2/me/e**2	Bohr radius
ft	12*inch	Foot
inch	2.54*cm	Inch
lyr	c*yr	Light year
mi	5280.*ft	(British) mile
nmi	1852.*m	Nautical mile
pc	3.0856776e16m	Parsec
yd	3*ft	Yard
acres	mi**2/640	Acre
b	1.e-28*m**2	Barn
ha	10000*m**2	Hectare

Table 12: Length and area units.

Symbol	Value	Description
l	dm**3	Litre
dl	0.1*l	Decilitre
cl	0.01*l	Centilitre
ml	0.001*l	Millilitre
tsp	4.92892159375*ml	Teaspoon
tbsp	3*tsp	Tablespoon
floz	2*tbsp	Fluid ounce
cup	8*floz	Cup
pt	16*floz	Pint
qt	2*pt	Quart
galUS	231*inch**3	US gallon
galUK	4.54609*l	British gallon
M	mol/m**3	Molar concentration
mM	mol/l	Millimolar
muM	0.001*mol/l	Micromolar

Table 13: Volume and concentration units. Note, that the unit *molar concentration* represented by symbol M does not collide with prefix Mega, which also is represented by symbol M. This is due to the fact, that units are combined by multiplication or division while prefixes are directly prepended to a SI-unit or a derived SI-unit. Thus MJ denotes Mega-Joule, and M*J represents the hypothetical unit *molar Joule*.

Symbol	Value	Description
oz	28.349523125*g	Ounce
lb	16*oz	Pound
ton	2000*lb	Ton
dyn	1.e-5*N	Dyne (cgs unit)

Table 14: Mass and force units.

Symbol	Value	Description
erg	1.e-7*J	Erg (cgs unit)
eV	e*V	Electron volt
Hartree	me*e**4/eps0**2/h**2/4	Hartree
invcm	h*c/cm	Wave-numbers/inverse cm
Ken	k*K	Kelvin as energy unit
cal	4.184*J	Thermo-chemical calorie
kcal	1000*cal	Thermo-chemical kilo-calorie
cali	4.1868*J	International calorie
kcali	1000*cali	International kilo-calorie
Btu	1055.05585262*J	British thermal unit
hp	745.7*W	Horsepower

Table 15: Energy and power units.

Symbol	Value	Description
bar	1.e5*Pa	Bar (cgs unit)
dbar	1.e4*Pa	Decibar (cgs unit)
mbar	1.e2*Pa	Millibar (cgs unit)
atm	101325.*Pa	Standard atmosphere
torr	atm/760	Torr = mm of mercury
psi	6894.75729317*Pa	Pounds per square inch

Table 16: Pressure units.

Symbol	Value	Description
deg	pi*rad/180	Degrees
degR	(5./9.)*[K]	Degrees Rankine
degC	[K]-273.15	Degrees Celsius
degF	5./9.*[K]-459.67	Degrees Fahrenheit

Table 17: Geometrical and thermo-dynamical degrees.

Acknowledgements

The authors would like to thank H. H. Winter, M. Walter, and K. Kaiminsky for fruitful discussions on the topic and Michael Machala for proof-reading the manuscript. A. W. Liehr gratefully acknowledges the Apple Research & Technology Support (ARTS).

References

[1] J. Klump, R. Bertelmann, J. Brase, M. Diepenbroek, H. Grobe, H. Höck, M. Lautenschlager, U. Schindler, I. Sens, J. Wächter, Data publication in the open access initiative, Data Science Journal 5 (2006) 79–83.

[2] P. F. Uhlir, Open Data For Global Science: A Review of Recent Developments in National and International Scientific Data Policies and Related Proposals, Data Science Journal 6 (2007) OD1–70.

[3] Earth system science data (ESSD), `http://www.earth-system-science-data.net`, [Online; accessed 2009-02-13] (2008).

[4] Unidata, NetCDF (network Common Data Form), `http://www.unidata.ucar.edu/software/netcdf/`, [Online; accessed 2009-02-13] (2009).

[5] The National Center for Supercomputing Applications (NCSA), HDF5, `http://hdf.ncsa.uiuc.edu/HDF5/`, [Online; accessed 2009-02-13] (2009).

[6] Unidata, Where is NetCDF used?, `http://www.unidata.ucar.edu/software/netcdf/usage.html`, [Online; accessed 2009-02-13] (2009).

[7] The National Center for Supercomputing Applications (NCSA), HDF5 users, `http://hdf.ncsa.uiuc.edu/HDF5/users.html`, [Online; accessed 2009-02-13] (2009).

[8] National Research Council / Committee on Issues in the Transborder Flow of Scientific Data, Bits of power : issues in global access to scientific data, National Academy Press, 1997, [Online; accessed 2009-02-13].

[9] M. Nowak, D. Malon, P. van Gemmeren, A. Schaffer, S. Snyder, S. B. and. K Cranmer, Explicit state representation and the atlas event data model: theory and practice, Journal of Physics: Conference Series 119 (4) (2008) 042024.

[10] I. Antcheva, O. Couet, ROOT. an object-oriented data analysis framework, Users Guide 5.21, CERN, [Online; accessed 2009-02-13] (2008).

[11] M. K. Riede, A. W. Liehr, M. Glatthaar, M. Niggemann, B. Zimmermann, T. Ziegler, A. Gombert, G. Willeke, Datamining and analysis of the key parameters in organic solar cells, in: A. Gombert (Ed.), Photonics for Solar Energy Systems, Vol. 6197 of Proceedings of SPIE, 2006, p. 61970H, conference Location and Date: Strasbourg, France, 2006.

[12] M. Riede, Identification and analysis of key parameters in organic solar cells, Ph.D. thesis, Universität Konstanz, Fachbereich Physik (2006).

[13] M. K. Riede, K. O. Sylvester-Hvid, M. Glatthaar, N. Keegan, T. Ziegler, B. Zimmermann, M. Niggemann, A. W. Liehr, G. Willeke, A. Gombert, High throughput testing platform for organic solar cells, Progress in Photovoltaics: Research and Applications 16 (7) (2008) 561–576.

[14] K. Zimmermann, L. Quack, A. W. Liehr, Pyphant - a Python framework for modelling reusable information processing tasks, The Python Papers 2 (3) (2007) 28–43.

[15] K. Zimmermann, A. W. Liehr, fmfile: A Python parser for the Full Metadata Format, `https://pyphant.svn.sourceforge.net/svnroot/pyphant/trunk/src/workers/fmfile/`, [Online; accessed 2009-02-13] (2008).

[16] M. Hanko, Entwicklung eines neuen optochemischen Gassensors zur Bestimmung von SO_2, eines vorteilhaften Messsystems für die Gasanalytik und einer vielseitigen Matrix für die einfache Herstellung chemischer und biochemischer Sensoren, Ph.D. thesis, Fakultät für Chemie, Pharmazie und Geowissenschaften, University of Freiburg (2006).

[17] N. Bruns, W. Bannwarth, J. T. Tiller, Amphiphilic conetworks as activating carriers for the enhancement of enzymatic activity in supercritical CO_2, Biotechnology and Bioengineering 101 (2008) 19–26.

[18] V. Shrotriya, G. Li, Y. Yao, T. Moriarty, K. Emery, Y. Yang, Accurate measurement and characterization of organic solar cells, Advanced Funtional Materials 16 (15) (2006) 2016–2023.

[19] M. Kühne, A. W. Liehr, Improving the traditional information management in natural sciences, Data Science Journal 8 (2008) 18–26.
URL `http://www.jstage.jst.go.jp/article/dsj/8/0/18/_pdf`

[20] Microsoft Support, WININI.WRI from Windows for Workgroups 3.11, `http://support.microsoft.com/kb/109496/en`, [Online; accessed 2009-11-13] (1999).

[21] M. K. Riede, Characteristic of an organic solar cell - an example of documenting experimental data with the Full-Metadata Format (FMF), Scientific Information SI20090303a, Freiburg Materials Research Center, doi:10.1594/fmf.SI20090303a (2009).

[22] L. Lamport, LaTeX: User's Guide and Reference Manual, 2nd Edition, Addison-Wesley Longman, Amsterdam, 1994.

[23] A. W. Liehr, A. J. Holtmann, Measuring of Faraday's constant - an example of documenting experimental data and their analysis with the Full-Metadata Format (FMF), Scientific Information SI20090303b, Freiburg Materials Research Center, doi:10.1594/fmf.SI20090303b (2009).

[24] P. J. Mohr, B. N. Taylor, D. B. Newell, CODATA recommended valued of the fundamental physical constants: 2006, Review Modern Physics 80 (2008) 633–730.

[25] R. L. Ackoff, From data to wisdom, Journal of Applied Systems Analysis 16 (1989) 3–9.

[26] H. Onsrud, J. Campbell, Big Opportunities in Access to "Small Science" Data, Data Science Journal 6 (2007) OD58–OD66.

[27] E. Shaya, B. Thomas, C. Cheung, Specifics on a XML data format for scientific data, in: J. F. R. Harnden, F. A. Primini, H. Payne (Eds.), Astronomical Data Analysis Software and Systems X, Vol. 238 of ASP Conf. Ser., San Francisco, 2001, p. 217.
URL http://www.adass.org/adass/proceedings/adass00/O6-02/

[28] F. Ochsenbein, R. Williams, C. Davenhall, D. Durand, P. Fernique, D. Giaretta, R. Hanisch, T. McGlynn, A. Szalay, M. B. Taylor, A. Wicenec, Votable format definition version 1.1, IVOA Recommendation 11 August 2004, International Virtual Observatory Alliance (2004).
URL http://www.ivoa.net/Documents/latest/VOT.html

[29] S. Mangano, XSLT Cookbook, o'Reilley, Köln, 2006.

[30] T. Parr, The Definitive ANTLR Reference. Building Domain Specific-Languages, The Pragmatic Bookshelf, Raleigh, 2007.

[31] K. Hinsen, Module physicalquantities, http://dirac.cnrs-orleans.fr/ScientificPython/ScientificPythonManual/Scientific.Physics.PhysicalQuantities-module.html, [Online; accessed 2009-02-13] (2006).

[32] H. Shull, G. G. Hall, Atomic units, Nature 184 (1959) 1559–1560.

[33] J. Rog, C. van Wijk, Evaluating file formats for long-term preservation, http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf, [Online; accessed 2009-02-13] (February 2008).

[34] Wikipedia, Newline — Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Newline&oldid=270258348, [Online; accessed 2009-02-19] (2009).

[35] M.-A. Lemburg, M. von Löwis, Defining Python Source Code Encodings, http://www.python.org/dev/peps/pep-0263/, [Online; accessed 2009-02-13] (2007-06-28).

[36] M. Kuhn, A summary of the international standard date and time notation, http://www.cl.cam.ac.uk/~mgk25/iso-time.html, [Online; accessed 2009-02-13] (2004).

[37] M. Bohnert, K. Schulz, L. Belenkaia, A. W. Liehr, Re-oxygenation of hemoglobin in livores after postmortem exposure to a cold environment, International Journal of Legal Medicine 122 (2) (2008) 91–96.

[38] F. Yergeau, UTF-8, a transformation format of ISO 10646, Request for Comment RFC 3629, Network Working Group, [Online; accessed 2009-02-13] (2003).

[39] Windows 1252, http://www.microsoft.com/globaldev/reference/sbcs/1252.mspx, [Online; accessed 2009-02-13] (May 2005).

[40] T. Texin, Character Sets And Code Pages At The Push Of A Button, http://www.i18nguy.com/unicode/codepages.html, [Online; accessed 2009-02-13] (2005).

[41] Wikipedia, IEEE 754-1985 — Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=IEEE_754-1985&oldid=221507406, [Online; accessed 2008.06.28] (2008).

[42] E. C. Bank, Euro foreign exchange reference rates, http://www.ecb.europa.eu/stats/exchange/eurofxref/html/index.en.html, [Online; accessed 2009-02-13].

[43] IEC 60027-2, Letter symbols to be used in electrical technology – part 2: Telecommunications and electronics, Tech. rep., 3. Edition (2005).