Units of Measure for Humans and Machines: Making Units Clear for Machine Learning and Beyond

"If God wanted us to use the metric system, He would have given us ten fingers and ten toes."
— Judith Stone, *Light Elements: Essays in Science from Gravity to Levity*

1: Why are Units Important?

The major challenges that confront human societies are global in reach and complex in nature. They do not have simple, single-discipline, solutions. The intrinsically interdisciplinary and multidisciplinary challenges require trans-sectoral cooperation between academic, commercial and governmental agencies.

For such collaboration to succeed, the essential tools for scientific exchange of information must be fit for purpose if they are to meet the challenges. Quality data is essential, and to be understandable and usable, the data must meet internationally-agreed community-endorsed conventions or standards, a key element being the clear representation of units.

Although the “collaboration imperative” is recognised by the major research funders and international organisations, they often fail to appreciate all the details that are essential to enable the required cooperation. Funding to encourage collaboration, highlighting relevance, pathways to impact, and facilitating international exchanges, are all vital but these are a tower of cards if the foundations for scientific exchange of information are not up to the required challenge.

Providing quality data is essential, but collaborative work will fail unless all those who need to use the data (and the associated information and knowledge) can actually understand it and this requires international, community agreements.

Units of measure are a key part of such agreements. Much of the global output of data lacks clear and unambiguous definitions of the units used. While the units of measure of quantitative values might be conventional and thus unstated for the original application, they are often obscure outside the originating community or discipline. This is particularly problematic when data are analysed at scale using machines. Confusion abounds. Huge efforts are needed when mining data from a neighbouring discipline.

The aim of creating and sharing data has to be to enable problems to be solved, problems that need the active collaboration of other disciplines and sectors; data without well curated, understood, and digitally communicated units is a hindrance rather than an advantage.

Now that we can exchange vast quantities of data in minimal time, there is an even greater need for clarity on all the details, including units. Computers consume numbers, but analysis needs quantities, and quantities are scaled. If the scale is different to that needed by the user, then conversion (re-scaling) is required. It is essential that the major funders of research support the international standards process, especially current efforts on digital representation of units of measure.

As the Royal Society noted in *Science as an Open Enterprise*, a key aspect to the pursuit of science is 'intelligently accessible data'; clear representation of units is essential to all research. However, clarity is not a given: the COVID-19 pandemic has exposed many of the cracks in the international efforts to mount a rapid interdisciplinary effort to tackle a global emergency.[1]

Work on data standards is essential. Without it, the value of much of the huge investment in research capability is lost, or at best, requires major archaeology to retrieve meaningful information from the data, something that can easily be avoided with the appropriate investment. Now!

2: Outline of Important Initiatives around Digital Representation of Units of Measure

The Digital-SI

The International Bureau of Weights and Measures (BIPM: Bureau International des Poids et Mesures) through the International Committee of Weights and Measures (CIPM: Comité International des Poids et Mesures) has initiated a project called the Digital-SI. This project has the following terms of reference:

- To develop and establish a world-wide uniform, unambiguous and secure data exchange format for use in Internet of Things (IoT) networks based on the International System of Units (SI) described in the current SI Brochure.
- To coordinate this effort with all relevant stakeholders by exploring and/or establishing suitable liaisons.
- To propose suitable actions towards making the SI Brochure machine readable.

To start this project, an expert group was convened to discuss what might be the technical infrastructure needed to support the terms of reference. The expert group met at the Physikalisch-Technische Bundesanstalt (PTB, the National Metrology Institute of Germany) on February 4th/5th 2020 and developed the following information:

- Grand Vision for the Digital-SI - high level document focused around "analysis-ready" data that enables AI, ML, automation, reproducible FAIR science, robust simulations, etc.
- Analysis-ready data means that the value, uncertainty and appropriate conditions contextualising the data are presented as a package of information that is reusable

[1] See Barend Mons's Plenary Talk at the virtual Conference on a FAIR Data Infrastructure for Materials Genomics (June 2020) https://www.youtube.com/watch?v=lRciVnP9WIY&feature=youtu.be; and the pre-print at https://osf.io/rtx9m/

The expert group has been defining the scope and participants of a two-day workshop to be held at BPIM in February/March 2021 (originally June 2020) to bring together stakeholders to discuss needs and issues and evaluate proposed developments/tools/services developed by the expert group. The expert group has also been developing ideas for resources and tools needed to support implementation of the Digital-SI.

CODATA DRUM Task Group

Since 1969, CODATA has convened the Task Group on Fundamental Physical Constants[2]: the group that provides the global scientific community with a self-consistent set of internationally recommended values of the basic constants and conversion factors of physics and chemistry based on all of the relevant data available at a given point in time. From May 2019, all the key units of the SI System are derived from the CODATA recommended values for specific constants.

The Digital Representation of Units of Measure (DRUM) has been identified as a critical tool in facilitating the automated interpretation and use of stored data. Indeed, units are things that are so fundamental and ingrained to scientists that they are often barely noticed, represented in different ways within files and databases, and perhaps even inferred from experience. Unfortunately, software can often struggle with such challenges. Created in 2018, the DRUM Task Group[3] seeks to address this in an interdisciplinary manner by involving scientific union members of the International Science Council, its Committee on Data (CODATA), and working with the BIPM.

Pain Points

UNIT DEFINITIONS: The BIPM, through the SI brochure[4], defines the SI units system. Additionally, concepts in metrology of which units of measure are an important part, are delineated in the International Vocabulary of Metrology (VIM)[5] which characterise how units are related to quantities, quantity kinds, and dimensions. This system works well for the large majority of units, however there are issues with some special cases. One such pain point is with the representation of unitless measurements. While measured values that have no units (i.e., counts of things) are fine, measured values that are ratios of the same unit (i.e. mole fraction - moles of compound A per moles of a mixture of A + B) are normally represented without a unit (as the units cancel out). However, this ratio cannot be used to convert the mass A to the mass of A+B because the units in the calculation do not cancel out. Representation of this scenario in digital systems will require the representation of units to their dimensions so that a computer can make the decision that a mole fraction can only be used to convert between moles of substances.

UNIT REPRESENTATION: As an example of the current fragmented landscape of digital units consider the representation of the meter. The table below shows many of the current digital representations of this unit. Clearly, while these may work in a narrow area (i.e. 'MTR' in international trade), interoperability across these systems is impossible without a mechanism to universally identify them as equivalent.

[2] https://codata.org/initiatives/strategic-programme/fundamental-physical-constants/
[3] https://codata.org/initiatives/task-groups/drum/
[4] https://www.bipm.org/en/publications/si-brochure/
[5] https://www.bipm.org/en/publications/guides/vim.html

Representations

String	Status	Representation System(s)
m (encodings)	preferred	Guide for the Use of the International System of Units (SI) (SP811) IUPAC Quantities, Units and Symbols in Physical Chemistry (IUPAC) (definition) International System of Units (SI) International Virtual Observatory Alliance: Units in the VO (IVOA) Metric Interchange Format (MIXF) Unified Code for Units of Measurement (UCUM)
M	current	Quantities, Units, Dimensions, and Datatypes (v2) (QUDT2) (definition)
MTR	current	United Nations Economic Commission for Europe (UNECE)
Meter	current	Quantities, Units, Dimensions, and Datatypes (v1) (QUDT1) (definition)
NCIT_C41139	current	NCI Thesaurus in OBO (NCIT) (definition)
Q11573	current	Wikidata.org (WDATA) (definition)
UO_0000008	current	Units Ontology (UO) (definition)
meter	current	Semantic Web for Earth and Environmental Technology (SWEET) (definition)

UNIT CONVERSION: A fundamental topic addressed in introductory science classes is how to convert a measured value from one unit to another. Students learn that in order to do this correctly they must apply dimensional analysis and find the correct conversion factor. While this is a trivial process for most students, relying on digital systems to do such a conversion in a systematic and metrologically appropriate way is by no means trivial.

As an example, take the conversion of length to the unit of Ångström (symbol Å). The conversion factor for the Ångstrom (a non-SI unit but one that uses the SI dimension of length) is 1 Å = 1 x 10^{-10} m, and thus it can be processed the same way as an SI prefix would be processed (a scaling factor). For digital systems this unit and the associated factor must be represented such that:

- It is aligned with the SI system of units (and not another system of units)
- It can be identified as a unit with the same dimensionality as the meter
- The conversion factor is identified as ‘exact’ to allowing the processing of uncertainty
- The magnitude of the conversion factor is certified as correct

Currently, none of the above criteria can be evaluated/enforced by a computer system and therefore the reliability of unit conversions would be very low.

3: Request to Name ‘Ambassador(s)’ for Digital Representation of Units of Measure

With the support of the CODATA Secretariat and in liaison with the ISC, the DRUM TG is approaching International Unions (IUs) and International Associations (IAs) on a number of related matters.

First, we invite each IU/IA to name an ‘ambassador’, who would be the point of contact for DRUM and engage with the TG. This ‘ambassador’ will act as the point of contact for DRUM on matters related to units of measure. This will include participation in activities and events to make the case for the wider, systematic development and application of the Digital Representation of Units of Measure.

Second, we would like to propose the same person as a liaison to BIPM for the Digital SI for the purpose of units and measure. If the IU/IA in question already has a liaison with BIPM, we would request that that person is also named as the ‘ambassador’ to engage with DRUM.

Third, we propose to liaise with BIPM and nominate these ambassadors to participate in the 2021 workshop. By this step we seek to maximise IU/IA involvement in this process. This will ensure, both that the scientific community is fully represented in the Digital-SI, and that the importance of the Digital-SI and DRUM is disseminated fully in the scientific communities represented by the IUs and IAs.

Finally, in the next section, below, we invite the IUs and IAs to share with us use cases that demonstrate the utility and importance of the digital representation of units of measure, or illustrate pain points created by their absence. If the IU/IA has a solution in place for Units of Measure, agreed by the community, then we would like to build on that to convert between domains.

By engaging with this process, ISC, CODATA and DRUM aim to help mobilise the IUs and IAs, to present a united front and to develop a 'manifesto' that can be presented to funders to obtain greater resources for the initiative of develop digital and standardised representations of units of measure, both through the Digital-SI and reaching beyond those units directly encompassed by the SI System.

4: Request for Use Cases

Use Cases for the Development of Fit-for-Purpose Analysis-Ready Data

The application of units of measure across the International Unions (IUs) and International Associations (IAs) is broad, variable, and largely defined at the user level. Given the movement toward FAIR data, in particular the interoperability of data, this area clearly needs attention and we need the help of the IUs/IAs.

The BIPM Digital-SI project aims to revolutionise units of measure such that every discipline can easily and consistently implement SI units for their needs. In order for the Digital-SI project to fully understand the complexity of usage of units of measure, we encourage the IUs/IAs to document and share use cases that are critical to their disciplinary needs. In this way the project will be able to develop a framework for digital units that can provide every discipline with analysis-ready data, that is, data that is accurately described in measured value, uncertainty, and units and provides the contextual information that is critical to use of the data in subsequent analysis and metaanalysis, both within and between disciplines.

At this point we are seeking from each IU/IA the single most impactful use case of SI units of measure. We encourage each IU/IA to put together a small, high level group (coordinated through the IU/IA ambassador) to develop/articulate such a use case, taking into account (but not limited to):

- What is most impactful data that your discipline needs to accurately represent internally and/or with other IUs/IAs?
- Where are the pain points in using units of measure (and/or metrology in general) in your discipline?
- What topics do students have the most problem with when reporting units of measure?

5: Example Use Cases

Mars Orbiter

The relatively wide adoption by engineers of two different unit systems, SI in Europe and British or Imperial units frequently (but not exclusively) used in the USA is the cause of major worries and required additional human resources to assure that correct conversions have been done where necessary.

In the case of the Mars Orbiter[6] this had a major impact and contributed to the loss of this space mission. The enquiries revealed that the navigation software assumed that the SI units were being used e.g. force measurements were in newtons and thrust in N s, while the software that instructed the thrusters assume that units of pounds (force) seconds (lbf s) this led to underestimating the effect of the thruster firings on the spacecraft trajectory by a factor of 4.45 (1 lbf s = 4.45 N s).

Often in the case of unit conversions the errors induced by incorrect conversions are large and relatively easy to spot. In this case the factor was relatively small, and the origin was much harder to spot: the small deviations were sufficient to send the craft well off the correct trajectory.

The reports of the Mars Orbiter investigation vary but clearly show that the unit conversion problem contributed to wider analysis and reporting issues. There is speculation that the software engineers were well aware of the need for conversion, but each side of the software exchange did the conversion (i.e. double conversion was undertaken), resulting in the error.

Systematic representation of units in a digital framework would facilitate the consideration of quantities (represented by number and unit) rather than just the number and therefore the checking of software compatibility at compile and runtime. The concept of exchanging and communication quantities, rather than just numbers, is one that is clearly found to be difficult by those entering studies and all the way through to experienced professionals.

Microstructures

A pressing challenge in materials science concerns microstructures, i.e., the arrangement of the constituents of a material (alloy, polymer, etc.) at the atomic scale. Microstructure properties such as grain size distribution and grain boundaries have a profound impact on the macro-scale properties of the material (stiffness, flexibility, brittleness).

In order to study microstructures and how they correlate with overall material properties (impactful data), it is important to aggregate images (high resolution pictures of a materials surface) from multiple experiments and process them using machine learning algorithms.

[6] See https://www.sciencemag.org/news/1999/09/english-metric-miscue-doomed-mars-mission; https://spectrum.ieee.org/aerospace/robotic-exploration/why-the-mars-probe-went-off-course; http://web.mit.edu/16.070/www/readings/Failures_MCO_MPL.pdf

Microstructure images are collected with a variety of electron microscopes from different vendors, unfortunately the pixel sizes and electron beam energies vary so there is no guarantee they are in the same units (pain point).

Being able to dynamically parse a digital representation of units (for both the grain size and electron beam energies) and convert them to a common framework would demonstrate the power of a widely agreed Digital-SI and its importance for unambiguous and reproducible science.

DRUM, ISC and CODATA Group

This document was prepared by:

- Geoffrey Boulton, OBE FRS FRSE; Regius Professor of Geology Emeritus; member of the ISC Governing Board; CODATA Past President.
- Stuart Chalk, DRUM Task Group member; BIPM Digital-SI Expert Group member.
- Simon Cox, Research Scientist CSIRO Land and Water; member of the CODATA Executive Committee
- Jeremy Frey, Professor of Physical Chemistry, University of Southampton, DRUM Task Group Member, Member of IUPAC Div I & CPCDS committees and Chair of the IUPAC Green Book 5th Edition project.
- Robert Hanisch, NIST Office of Data and Informatics; Chair, DRUM Task Group; member of BIPM Digital SI Expert Group.
- Richard Hartshorn, Professor, School of Physical and Chemical Sciences, University of Canterbury; Secretary General of the International Union of Pure and Applied Chemistry; member of the CODATA Executive Committee.
- Simon Hodson, CODATA Executive Director.
- Barend Mons, Professor in Biosemantics, Leiden University Medical Center; President of CODATA; Director GO FAIR International Support and Coordination Office.
- Hana Pergl, Operations Manager, CODATA.
- Anne Thieme, ISC Membership Liaison Officer.