Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210109

Reproducible Transport of Information

Wolfgang ORTHUBER[a,1]
[a] *Department of Orthodontics, UKSH, Kiel University, Germany*

Abstract. Reproducible information is important in science, medicine and other professional fields. Repeating the same experiment with measurement should yield the same information as the result. This original information should also be transported digitally in reproducible form, as a globally well-defined sequence of numbers. The article explains that "Domain Vectors" (DVs) with the structure "UL plus sequence of numbers" are well suited for this purpose. "UL" is an efficient link to the online definition of the sequence of numbers. DVs are globally comparable and searchable and have other important advantages. It is concluded that DVs can fill an important gap in the digital representation of information.

Keywords. Reproducibility, Objectivity, Domain Vector, DV, Online definition

1. Introduction

Global search engines make it possible to find text-based information. However, information is not just text-based. Everything observable generates information. This information is often presented digitally. In general, this digital representation (e.g. as an image) is not globally comparable and searchable. We therefore need a global concept to convert something observable into a globally comparable digital form.
As a prerequisite, the observations (the available source of information) must be reproducible [1][2]. Let us denote by "ORGINFO" the relevant original information from reproducible observations. If we want to share this, we describe preconditions and observations and send this information to other people. This is increasingly done digitally, i.e. as a digital sequence of numbers. But the conversion from ORGINFO to its transportable digital representation DIGINFO is not globally uniform and reproducible. Such non-reproducible digitization of ORGINFO (original information) is quite common. ORGINFO is converted into a digital representation in an individual way, for example into an image or video. This high-dimensional representation can take place in an extremely high number of ways and is therefore not reproducible. When ORGINFO is represented by language (text), the situation is not much better. There is a lot of freedom when combining language vocabulary. The number of possibilities for this grows exponentially with the number of words used. There are many ways to represent the same original information with language. Moreover, the meaning of language vocabulary depends on the context and is often imprecise. It also depends on the viewpoint. For example, waiting 5 minutes may be "short" for one person but "long" for another person, even in the same situation. The language-based conversion of the same original information ORGINFO into its digital representation

[1] Corresponding Author, Wolfgang Orthuber; E-mail: orthuber@kfo-zmk.uni-kiel.de.

DIGINFO (Fig. 1) and the conversion in the opposite direction (Fig. 2) depend on the viewpoint.

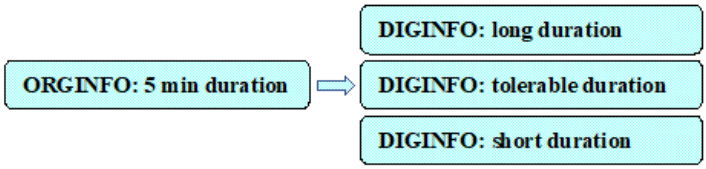

Figure 1. The language-based conversion from ORGINFO to DIGINFO depends on the viewpoint.

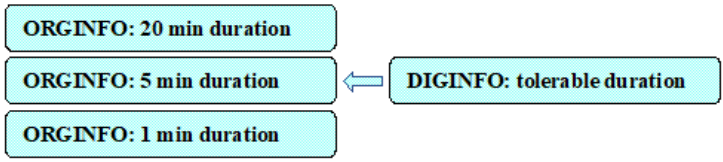

Figure 2. The language-based conversion from DIGINFO to ORGINFO depends on the viewpoint.

It becomes clear that also language-based information transport is not reproducible. This considerably hinders the comparison and the search for precise original information ORGINFO in the case of language-based representation. Fortunately, there are other possibilities that allow a reproducible representation of ORGINFO.

2. Reproducible Digitization of Information

Why are quantitative data commonly used for precise information exchange? This concerns a deeper principle. Quantitative data are *explicitly* transported as numbers. Each number represents a selection from an ordered set, which must be precisely defined for the sender and all recipients of information. This is a matter of course when transporting quantitative information through numbers. The principle can be generalized: Each piece of information is transported as a selection from a common set of possibilities. All digital data are sequences of numbers that represent a selection from a common set that is (or should be) known to all conversation participants. For the sake of brevity, this common set of possibilities is referred to as "domain" here and in other publications [3]-5], and we can summarize:

Information is a selection from a domain. (1)

Information (in its precise definition as something transportable) can therefore be viewed as a mapping from the domain to an element of the domain. This definition only uses clear mathematical terms from elementary set theory to define "information", which is very useful for its application. First, the relevance of the domain for the information transport is clarified. In the case of free text, the domain consists of words and phrases from the language vocabulary. When sender and receiver speak the same language, this is common and can be used to transport information. However, this is not reproducible, as shown in Section 1. For reproducible digital information transport, we have to define an *adapted domain* [3] for the respective application. All possibilities of the relevant original information ORGINFO must be contained in the adapted domain without redundancy *so that there is a one-to-one correspondence between ORGINFO and its digital mapping DIGINFO*. Like all digital information, DIGINFO

is a sequence of numbers. For its transport the new data structure "Domain Vector" or "DV" is introduced [3][6]:

DV: UL plus sequence of numbers (2)

"UL" is a "Uniform Locator", a global pointer to the online definition of the sequence of numbers. It has a similar function as a "URL", but the UL can be optimized for efficiency [4], for example as hierarchical sequence of numbers (Table 1).

Table 1. Proposal for numbers in the UL (The UL is also a sequence of numbers). Each number is a self-expanding positive integer [4] with initially 2 bits, which specify the length in half bytes.

Number N1	Number N2	Number N3	Further numbers
N1 is pointer within official table: The first entries 0..31 of the table have special meanings like "same UL as before". The following entries contain links to online presences with collections of online definitions.	If N1 points to (link to) online presence: N2 is number of the user of the online presence who creates online definitions.	Number of definition created by this user.	Hierarchical sequence of D numbers as pointers in D-fold nested subgroups of this definition. 0 means "end of UL".

Due to the one-to-one correspondence between the reproducible ORGINFO and its digital mapping DIGINFO, *the DV transports information (DIGINFO) in a globally reproducible way*. It is globally uniformly defined and identified. This means that it is also findable and thus solves the problems mentioned at the beginning of Section 1. The complete DV can have a very efficient binary form based on self-expanding numbers [4]. Comfortable editing of DVs is possible because all details about the format and meaning of the numbers and further description of the DV structure can be downloaded from the online definition via UL. The count of the numbers after the UL or the dimensionality of the DV depends on the count of the variables that we want to transport. The fewer dimensions we compare, the easier is the comparison. Therefore, we try to describe ORGINFO by as few variables as possible. Since these should be as meaningful as possible, we first think about the most important independent features (parts) of ORGINFO and sort each feature from "small" to "large" for the mapping to a number. In the case of a measurable quantitative feature, we can use the measured value directly with the appropriate unit. Each (possibility of the) ordered feature is mapped bijectively or one to one to a (possibility of the) number. Finally, all relevant features [3] are globally reproducibly mapped to a sequence of numbers (DIGINFO). It's the part of the DV after the UL. Since the online definition of the DVs is machine-readable and unique for all DVs with the same UL, these DVs (which transport DIGINFO) are calculated from ORGINFO in the same globally reproducible way.

The following example should clarify the principle. Figure 3 shows a medical situation that is difficult to quantify. It is an infected surgical suture. We want to follow up on this finding. Without laboratory data, quantitative progress monitoring should be possible immediately from the image. For this we need a reproducible feature extraction. This can be done by adapted open source software that is uniformly available from the online definition (directly or via link). From the features of the color gradient estimates can be made in a reproducible manner[2] about the central and surrounding areas of inflammation, as shown in Figure 3. The chronological sequence of the relative areas can provide initial information about the progression of inflammation. The basic DV for this finding may include date and time, location

[2] Globally, we need a good (open source) feature extraction software only once for each application (such as rough diagnosis). The unique online definition could provide a link to the software.

coordinates (from a common anatomical reference coordinate system of the skin), area A and area B from Figure 3 and other important and immediate measurements (Table 2). This basic DV can be a "sub-DV" or part of a larger DV for chronological progress documentation, which is calculated from a sequence of these basic DVs and contains additional data from other DVs, e.g. a sequence of results from blood tests.

Figure 3. Left: Infected surgical suture. Middle: The skin area A of the central inflammation is colored blue. Right: The skin area B of the surrounding inflammation is additionally colored red.

Table 2. The basic DV for the "Immediately measurable state of an infected surgical suture" consists of sub-DVs. This nesting is done in the online definition, therefore only the UL of this online definition (not of sub-DVs) is necessary in the transported DV. Numbers "n.." are positive, self-expanding integers [4]. Floating point numbers "n..f" consist of 2 such numbers with sign bits. If needed (e.g., for "datetime" or "location"), the online definition provides links to open-source software for editing, displaying, and processing the numbers. Customized editors for editing DVs automatically use the online definition.

UL	**sub-DV 1**	**sub-DV 2**	**sub-DV 3**
See	datetime:	n3f: Area A in mm^2	If n6==0 "same location as before", else
Table 1	n1: Months since Jan. 2048=2^{11}(with sign)	n4f: Area B in mm^2	n6: Number of body coordinate system
		n5f: Body temperature in Degree Celsius minus 37	n7: Number of organ, for example "skin"
	n2f: Days since 01.		n8f, n9f,n10f: Location coordinates in mm

3. Discussion

Reproducibility is a basic requirement in science and other professional fields. Original scientific information ORGINFO (Section 1) must be reproducible [1][2]. Section 2 explains the reason why this information is usually exchanged as quantitative data, i.e. as explicit numbers: The domain (range) of these numbers is well-defined for the conversation participants (to give meaning to the numbers). For this purpose, the numbers, their units and names are transported together. In addition, a more detailed definition by context is also available. Usually names and units are known so that an expert with sufficient prior knowledge, who also reads the more detailed textual description, knows exactly the domain and the meaning of the numbers. By definition, these numbers directly (bijectively) represent the relevant original information ORGINFO. If ORGINFO is reproducible, the transporting numbers (DIGINFO, Section 1) are *locally* also reproducible, since the same bijective local definition of the numbers is used. Such a restriction to a local definition and collection of data (e.g. database) is common in science. A global collection (across all languages, countries and institutions concerned) did not seem feasible, as the situation is more complicated globally: The sequences of numbers that transport quantitative (and other, strictly speaking all digital) data are defined locally again and again in different ways and transported in flexible formats. This is *not reproducible from a global viewpoint* and led to today's split into countless databases and data silos, despite decades of considerable efforts [7][8][9][10][11] to achieve interoperability and standardization by providing specialized vocabularies, codes, and formats such as XML, JSON, Turtle.

Unfortunately, a general global connection of digital information (sequences of numbers) via global identification and online definition has not been introduced yet. It is this gap that is filled by the Domain Vector or DV (2) structure described in Section 2. It starts with the UL, which is both a global identifier and an efficient pointer to the (globally unique) online definition. We must remember that all digital data are sequences of numbers. Therefore, DVs (2) are a general approach to globally defined and efficiently identified digital information. When the data are DVs (2), global numerical search and statistics (defined by the sequence of numbers in the DV) is not only technically feasible, it is possible within a single search command [6][12]. DVs are globally reproducible, as described in Section 2, if the original information ORGINFO is reproducible. For example, if the data of one or many scientific studies are published in the structure of a DV (2), they are searchable as a collection of data identified by a uniform UL. As shown and demonstrated [5][6], a universal numerical search engine could flexibly extract subsets from this collection and immediately compute statistics (depending on the definition of the DV). Since the numerical search is reproducibly defined, the results of the local search can be exported (as privacy-compliant statistics) and combined into global statistics for decision support, for example.

4. Conclusion

Original information is represented digitally (i.e. as a sequence of numbers) in a very variable and globally non-reproducible way. Therefore, "Domain Vectors" or "DVs" with the structure "UL plus sequence of numbers" are proposed, where "UL" is an efficient pointer to the (globally unique) online definition of the sequence of numbers. DVs are well suited for globally reproducible digital representation of information. They are globally comparable and searchable, and have other important advantages. Thus, the introduction of DVs (Domain Vectors) would fill an important gap.

References

[1] Baker, M. Is there a reproducibility crisis? Nature. 2016;533:452–454.
[2] McNutt, M. Reproducibility. Science. 2014;343:229-229
[3] Orthuber W. Information is Selection-a Review of Basics Shows Substantial Potential for Improvement of Digital Information Representation. Int. J. Environ. Res. Public Health. 2020;17(8): 2975.
[4] Orthuber W. How to make medical information comparable and searchable. Digit Med. 2020;6:1-8.
[5] Orthuber W. Global predefinition of digital information. Digit Med. 2018;4:148-56.
[6] Orthuber W. Demonstration of Numeric Search in User Defined Data. Numeric Search viewed March 2021. Available from: http://www.numericsearch.com.
[7] Guha RV, Brickley D. Macbeth S. Schema.org: evolution of structured data on the web. Communications of the ACM. 2016;59(2): 44-51.
[8] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3(1):1-9.
[9] Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digital Medicine. 2019;2(1):1-5.
[10] Benson T, Grieve G. Principles of health interoperability: SNOMED CT, HL7 and FHIR. Springer 2016.
[11] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: A 5-year update. Clin Chem. 2003;49:624-33.
[12] Zezula P, Amato G, Dohnal V, Batko M. Similarity Search: The Metric Space Approach. Vol. 32. New York, USA Springer Science & Business Media, 2006.