

The Building Blocks of Information Are Selections - Let's Define Them Globally!

Wolfgang ORTHUBER^{a,1}

^a*Department of Orthodontics, UKSH, Kiel University, Germany*

Abstract. Digital information consists of sequences of numbers that are selections. So far, these are defined by context. We can globalize this by using an efficient global pointer (UL) as "context". The article explains new globally identified and defined "Domain Vectors" (DVs) for transporting digital information. They have the structure "UL plus sequence of numbers", where UL is an efficient identifier and global pointer (link) to the unified online definition of the sequence of numbers. Thus, the format of the number sequence and its meaning is defined online. This opens up far-reaching new possibilities for the efficient exchange, comparison and search of information. It can form the basis for a new global framework that improves the reproducibility, search, and exchange of data across systems, borders, and languages.

Keywords: Online definition, Domain, Domain Vector, DV, Reproducibility, AI

1. Introduction

We need a viable long-term solution to interoperability problems [1] and data silos. Serious problems arise from the variable local definition of digital information. There is a lack of global reproducibility [2] of digital information representation, with all the resulting complications.

This article therefore starts from the basic building blocks of (digital) information, namely selections represented by sequences of numbers. It is shown that an efficient global adaptation of a selection depending on the application (e.g., rough medical diagnosis) is possible. This is done by online definition of a suitable digital number sequence that represents application-relevant information bijectively (one-to-one). A globally defined and identified data structure is derived for the efficient, reproducible, comparable and searchable transport of digital information.

2. Precise Definition of Information

If we want to transport "information" globally and independently of language, we must define it precisely and globally. Set theory provides the precise and very fruitful basis for defining mathematical objects, and mathematical methods are used and applied for dealing with digital information. Therefore, it is only consistent to define the information itself by using the same basis as for other mathematical objects, namely sets and

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

functions on these sets. The selection of an element from a set describes an elementary class of functions on a set. A combination of selections is again a selection - from a more complex set. We note that information consists of building blocks, all of which are selections from sets of possibilities. A combination of information is again a selection - from a more complex set of possibilities. So we observe:

Information is selection (from a common ordered set or "domain"). (1)

This is consistent with reality: "Information" is the (transportable) result of any well-defined physical experiment. This result is a selection from the (common ordered) set of possible experimental results. As a special case, the bits of digital information are also selections. They encode sequences of numbers, all of which are selections.

Now we see why the precise definition of information is important: (Transportable) information selects from a set that must be the same (common) to all participants in the conversation. This set is often an unconscious (early learned) prerequisite for conversation, e.g., language vocabulary. Because of its importance, the ordered common set of possibilities from which information (1) selects is abbreviated called "domain" [2, 3] (as the domain of the definition of the function "information"). The domain is ordered so that we can select (address) its elements, for example by numbers.

2.1. Language Vocabulary as a Domain






Vocabulary of language(s) is a preferred domain for information exchange because our senses and brain are adapted to it. As a result, international character sets such as UNICODE have gained acceptance [4], defining a mapping between characters of multiple languages and numbers for the digital representation of individual language vocabulary. This made it possible to continue the familiar linguistic conversation by digital means. Words from the language vocabulary are small building blocks with broad applicability - if we want to describe a complex issue ORGINFO, we "only" need to combine enough words to get a certain meaning. The number of possible combinations increases exponentially with the number of words. There are many possibilities of word-based representation of the same original meaning ORGINFO. Moreover, words are often open to different interpretations. For example, the same temperature, even in the same environment, may be described as "cold" by one person and "not cold" by another. Therefore, the linguistic representation of information is usually more or less ambiguous and vague [2]. There is no one-to-one mapping or bijection between the original information ORGINFO and its digital representation DIGINFO. This is a serious shortcoming of language-based representation, since bijection would be important if we want to search original information ORGINFO via its digital representation DIGINFO. Strictly speaking, the "language vocabulary" domain is a detour: original information ORGINFO from nature is not language-based, but must be translated into language. This is then transported and translated back at the receiver. Sender and receiver of information are different. This translation into the domain "language vocabulary" and back is not a bijection and thus not reproducible, which causes many problems [2]. Therefore, it is important to remember that besides language vocabulary, it is also possible to use "adapted domains" that are adapted to the application [3] and allow efficient and reproducible communication. We "only" need to ensure that senders and all receivers of information know the adapted domains in order to use them.

2.2. Application-Specific Adapted Domains

Many situations require a more or less precise exchange of information. In such a case we use a quantitative description and adapt the definition of the numbers to the application. This is often necessary even for simple applications, as shown in [Table 1](#), or for professional applications, e.g. in medicine, industry, business, science. There can be many important parameters in a given situation (as original information ORGINFO). Which parameters are important? It depends on the rough situation for which the parameters should provide an accurate description. The parameters are determined from the (total) available information. This is called "feature extraction".

For demonstration purposes, [Table 1](#) shows an example simple feature extraction for the "Ellipse" application. Obviously, this is much more efficient than carrying information about each pixel. In addition, features can be defined to be of direct interest, such as for comparison and search.

Table 1. Application "Ellipse", example feature extraction: angle in 0..45 degrees, height, width, red, green, blue in relative units 0..255 . The domain of these 6 numbers is sufficient, but not the domain of "language vocabulary". Phrases of language like "deep blue, very wide and flat ellipse" or "dark, thin and long ellipse" etc. are not reproducible and not precise enough for comparing and searching such objects.

Ellipse	Angle	Height	Width	Red	Green	Blue
	0	3	3	255	0	0
	0	3	10	255	0	255
	0	3	20	0	255	0
	0	2	20	0	0	255
	15	2	5	0	0	255

2.3. Online Definition of Digital Information for Global Communication

We want to exchange the parameters of an adapted domain (as shown in [Table 1](#)) in a globally reproducible and comparable way. As mentioned earlier ([2.1](#)), for this we not only need a one-to-one mapping (bijection) between these parameters and their digital representation DIGINFO (typically by "quantification"), we also need to ensure that this mapping is globally consistent. This is efficiently achievable through an online definition. The online definition is unique and can be localized uniformly by an efficient global pointer. This global pointer is called "UL" below ([2](#)).

Selection within the adapted domain is done by a sequence of numbers. We can consider sequences of numbers as building blocks of digital information. Since the existence of the Internet, it would have been possible to provide a standard and convenient software for online definition of number sequences and their domains [[3](#)]. These domains are language independent and globally valid! Their size grows exponentially with the count of defined numbers (dimensions). Existing online definitions can be reused and nested in new definitions. Thus, we can globally define huge multidimensional domains (huge spaces as "sets of possible messages"). The online definition can be sharp and very detailed (multilingual, even with multimedia parts). It is automatically globally unique. Each online definition can contain links and additions, e.g., the definition of a distance function for similarity comparison and similarity search [[5](#)]. The data structure for the globally defined information transport is called "DV" [[2](#), [3](#)] as an abbreviation for "Domain Vector" and has the form:

$$\text{DV:} \qquad \text{UL} \quad \text{plus} \quad \text{sequence of numbers} \qquad (2)$$

"UL" is a "Uniform Locator" that refers to the (unique) online definition of the sequence of numbers and is also (directly or indirectly) a unique identifier. The online definition contains machine-readable information about the format of the number sequence and about the domain from which it selects. The UL has similar tasks as a URL, but the UL can be optimized to meet the requirements with maximum efficiency.

Notes on the UL format: as shown earlier [2], the UL can be represented by a hierarchical sequence of numbers. The binary format of the numbers can be based on self-expanding positive integers starting with a half-byte [6]. The compact UL refers to a unique online definition that can be very detailed and explanatory. There is no reason to waste bits in a DV which is designed for efficient information transfer. For example, consider the following DV:

$$\text{UL} \quad \text{plus} \quad 0, 3, 3, 255, 0, 0 \qquad (3)$$

As prior knowledge, we only know the format of the UL in the DV, for example, a hierarchical sequence of numbers N1, N2, N3, 0, as proposed in [2]. Here, N1 is an integer pointer within an official table of web addresses of online presences with collections of online definitions, N2 is the number of the user within that online presence, and N3 is the number of the definition of this user. In this way, we know the online definition of the following sequence of numbers. In this case, it contains the information about the first ellipse given in **Table 1**.

In the DV (3), the quantified features (as identified numbers) are directly comparable and searchable. Such number sequences can be defined for any application! Further examples have been published [2][3][6][7][8] and search of information has been demonstrated [9]. When defining the numbers, care must be taken to ensure that they represent the relevant information bijectively (one-to-one). Properly defined DVs with adapted domains change reproducibly if and only if the relevant original features change. More and more adapted domains can be defined online. For example, adapted domains can be defined for diagnoses from ICD-10 [10] or SNOMED-CT [11] as well as for diagnostic devices and diagnostic procedures. In medical reports, DVs with such adapted domains can be used to reduce non-reproducible parts step by step. The more adapted domains are defined online, the more tools are available for a globally reproducible, searchable *and* precise individual digital description.

Correctly (with adapted domain) defined DVs realize:

- globally the same (online) definition for (bijective) one-to-one conversion of original information (application-specific relevant features) into digital information (number sequence).
- globally the same uniform labeling ("UL" = efficient link to the unique machine-readable online definition) in front of the transported digital information (number sequence).

DVs are deliberately designed to transport the online defined digital information (sequence of numbers) as efficiently as possible. Existing online definitions can be reused and nested in new online definitions (e.g., the 3 components of color in **Table 1**).

3. Discussion

It has already been recognized that global definitions are important for certain quantitative data. Therefore, LOINC [12] was introduced. However, the generalization has not yet been realized. So far, it is not common to consider information in general as a selection from a domain (see (1)) and conclude that the global definition of domains of information is important. Technically, this can be realized very efficiently by online definitions of number sequences (for the selection in their domain). DVs (2) are optimized for transport of such globally defined number sequences and universally applicable. Existing work can be continued in online definitions. For example, any ontology, nomenclature, and terminology such as ICD-10 [10], SNOMED-CT [11], and LOINC [12] could be efficiently implemented in a single online definition. In the case of LOINC, for example, the first defined numbers could carry the LOINC code followed by one or more numbers containing measurement data. All DVs transporting this information (from the same online definition) would have the same UL. The online definition is machine-readable, which can also be used for convenient editing of DVs.

The domain "language vocabulary" (see Section 2.1) is widespread because it is obvious that it can be used as a domain of information (as a common set of possible messages). Thus, standards such as Unicode [4] have been introduced for (indirect) selection of elements (words) of the language vocabulary. Therefore, language-based information is globally searchable. This may suffice if we are searching for words in a particular language. But the above considerations about the domain "language vocabulary" make it clear that language-based information is not reproducible and often inadequate, even for transport of simple natural information or for applications, as shown in Table 1. It is not possible to correct such deficiencies subsequently. Even sophisticated linguistic and semantic analysis or AI cannot. In this respect, we need to rethink the acquisition and storage of more or less complex original data from nature, e.g., data from medical findings. If we convert them early into a language-based representation of information for communication, we will lack the required accuracy (Table 1) and reproducibility. But what can we do if the human brain is not adapted to accurately handle high-dimensional original data ORGINF?

We can first focus on the most important features and build on them in a targeted way: We can search and select an online definition of DVs with adapted domain that is suitable for the application or situation (e.g. ICD-10 diagnosis, see below). Optimally, the online definition would also provide appropriate software for reproducible feature extraction and automatic conversion of the original data relevant in this situation to their digital representation as DVs. Thereafter, the reproducible collection, comparison and search of such precise data could become everyday routine. Instant and individual global statistics for decision support would also be possible, more than an elaborate scientific study today. The DVs are also adaptable as *precise* source information for *global* machine learning and AI. The exemplary DV (3) clarifies also the efficiency of the DV data structure. Using directly editable formats for transport of these data (3), we would need several lines of code [8]. This is not only less efficient. In particular, the combinatorial freedom of directly editable formats (like XML, JSON, Turtle) contradicts the reproducible digitization of information. To improve the situation, many rules were published (e.g., in FHIR [13], LOD [14]) to make the transformation of the original information ORGINFO into its digital representation DIGINFO more reproducible. In the process, many problems came to light. Obviously, the optimal conversion depends on the application and the situation, e.g., medical diagnosis.

This could be one of the first applications of DVs. The online definitions of adapted domains (2.2) can be adapted to more and more applications and situations and provide rules and software for reproducible digitization. These are immediately accessible, globally valid, and can be created for more and more topics (e.g., given by a medical term from ICD-10 [10] or SNOMED-CT [11]). Using RFC4648 [15] and suitable editors, we can integrate DVs into existing editable formats, automatically incorporating FAIR principles [16]: Each DV is globally identified via UL and is thus "findable". The online definition clarifies the content of the DV (this can be done descriptively and extensively) and makes it editable and "accessible". The identification of the DVs (by UL) makes them "interoperable" and "reusable." The online definitions themselves are also reusable and can be embedded and nested within other online definitions via link. Thus, DVs can be online defined as globally consistent and comprehensive building blocks of digital information.

4. Conclusion

The proposed DVs can be viewed as globally defined building blocks of digital information. Online definitions of DVs can be adapted to the application or situation to globally determine a one-to-one mapping from the important features of the original information (ORGINFO) to the digital representation (DIGINFO) as DV. The DV data structure is well-defined and has further objectifiable advantages, e.g. in terms of global reproducibility and efficiency of information transport. Its introduction is recommended.

References

- [1] Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. *NPJ Digital Medicine* 2019; 2(1):1-5.
- [2] Orthuber W. Reproducible Transport of Information. *Studies in Health Technology and Informatics* Vol. 281; IOS Press 2021: 3-7. <https://ebooks.iospress.nl/doi/10.3233/SHTI210109>
- [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] Allen JD, Anderson D, Becker J, Cook R, Davis M, Edberg P, et al. The unicode standard. Mountain view, CA, 2012.
- [5] Zezula P, Amato G, Dohnal V, Batko M. Similarity Search: The Metric Space Approach. Vol. 32. New York, USA Springer Science & Business Media, 2006.
- [6] Orthuber W. How to make medical information comparable and searchable. *Digit Med* 2020;6:1-8.
- [7] Orthuber W. Global predefinition of digital information. *Digit Med* 2018;4:148-56.
- [8] Orthuber W. Online definition of comparable and searchable medical information. *Digit Med* 2018;4:77-83.
- [9] Orthuber W. Demonstration of Numeric Search in User Defined Data. Available from: <http://www.numericsearch.com> [Accessed on 2021 Jun 08].
- [10] World Health Organization. ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision; 2016.
- [11] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 2006; 121: 279.
- [12] 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.
- [13] Benson T, Grieve G. Principles of health interoperability: SNOMED CT, HL7 and FHIR. Springer, 2016.
- [14] Bauer F, Kaltenböck M. Linked open data: The essentials. Edition mono/monochrom, Vienna, 2011.
- [15] Josefsson S. The base16, base32, and base64 data Encodings. RFC 4648, October 2006; 1-18.
- [16] 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.