Computing and Artificial Intelligence 2025, 3(1), 1884.
https://doi.org/10.59400/cai1884

Article

The potential role of domain vectors in optimizing digital data structure

Wolfgang Orthuber[1,2]

[1] Kiel University, 24118 Kiel, Germany; orthuber@kfo-zmk.uni-kiel.de
[2] University Hospital Schleswig-Holstein, 24105 Kiel, Germany

CITATION

Orthuber W. The potential role of domain vectors in optimizing digital data structure. Computing and Artificial Intelligence. 2025; 3(1): 1884.
https://doi.org/10.59400/cai1884

ARTICLE INFO

Received: 17 October 2024
Accepted: 11 December 2024
Available online: 15 January 2025

Abstract: Each piece of information represents a selection from a set of possibilities. This set is the "domain of information", which must be uniformly known before any information transport. The components of digital information are also sequences of numbers that represent a selection from a domain of information. So far, however, there is no guarantee that this domain is uniformly known. There is still no infrastructure that makes it possible to publish the domain of digital information in a uniform manner. Therefore, a standardized machine-readable online definition of the binary format and the domain of digital number sequences is proposed. These are uniquely identified worldwide as domain vectors (DVs) by an efficient Internet address of the online definition. As a result, optimized, language-independent digital information can be uniformly defined, identified, efficiently exchanged and compared worldwide for more and more applications.

Keywords: definition and domain of information; temporally ordered information; nested domains of information

1. Introduction

To date, there is a wide variety of attempts to define the term "information", typically using other complex terms. The exact definition of the term "information" given below, using only simple set-theoretical and temporal terms, is still isolated [1], but it turns out to be necessary for both the understanding and the efficient technical application of information.

By "information" we mean the superordinate term that encompasses all physical, measurable information. "Digital information" is a subset of this. The title of this paper proposes a new, online defined binary format for optimizing digital data structure. Reason: Although digital information forms the basis of informatics and computer science, its definition is still vague and heterogeneous. This causes considerable difficulties. The well-known interoperability problems [2] are just one example of the consequences. Today, great efforts are still being made to improve interoperability and data compatibility [2–5], as this affects all digital interfaces (including all program interfaces). Nevertheless, the information exchanged is only described and defined individually in very different ways, as there is currently no global infrastructure for defining digital information. In the case of common program interfaces and important file formats, there is partial agreement, which has already proven to be very useful. More specific applications and details such as the interfaces of subroutines are usually redefined, which causes a lot of reprogramming. Missing uniformity (of definition) of digital information leads to redundancy, additional work (also for users) and ultimately to incompatibility and missing interoperability. Another example is the definition of research data [6] and the definition of objectifiable data in general. Common definitions of data including the binary format

would be of great benefit. This applies to all fields of digital communication, for example all types of medical data [7]. Uniform definitions enable comparability for decision-making. It would certainly be very helpful if there would be an appropriate infrastructure, so that, for example, medical experience could be shared worldwide on request and (human and AI-generated) conclusions could be verified, for example by comparing them with patient-specific global, objectifiable statistics. This is not possible today. There is no concept and no infrastructure available for defining digital information globally uniformly. Instead, the number of heterogeneous representations of information continues to increase. Much research has been carried out to overcome the resulting difficulties [8]. Unfortunately, this was not based on a precise definition of digital information, which meant that there was no basis for a systematic structure. Such a structure is necessary to overcome the widespread patchwork step by step.

A well-defined property of digital information was already mentioned by Shannon [9]:

> "The significant aspect is that the actual message is one selected from a set of possible messages."

But this clear aspect was not adequately given focus. Later, there were overviews and numerous attempts to define information more detailed [10–14]. However, the lack of a clear, viable concept is still being criticized [15]:

> "There is a methodological contradiction: The development and application of information technologies requires accuracy and rigor, but at the same time the development is based on a vague, intuitive concept."

It would be necessary to focus on an exact definition, which requires only simple concepts of set theory and also takes into account the temporal, hierarchically nested order of information. This directly can be directly applied by in the same way nestable online definition of information.

The aim of the paper is therefore to explore a possibility of introducing a simple, precise, nestable online definition of information that includes the necessary knowledge about the domain of information as a prerequisite for information transfer. This is followed by a description, demonstration and discussion of its application.

2. Materials and methods: Definition of information and digital application

The most important methods are clear, partly mathematical arguments (using basal set theory):

- Information is defined as a selection from a previously known set or "domain of information".
- It is recalled that numbers also represent such a selection and that the bits of digital information represent sequences of numbers.
- The digital number sequences can be defined online in machine-readable form. This means that the online definitions can be retrieved uniformly worldwide with the help of their unique Internet address, which is provided (in "domain vectors", see 2.3.) before the number sequences.

As there is currently no infrastructure for worldwide use, the principle of online definitions has been implemented in an online prototype in a local database: After logging in, users can enter their own definitions of digital information (**Figure 1**) and associated data. It is also demonstrated how this data can be searched for (**Figures 2–4**). This is done in the local database, but data searches would be possible worldwide if the infrastructure were available. A more detailed description of the online prototype is available, including the algorithms used [16].

After this overview, we begin with the clear definition of information [17,18]. It is the basis for further explanations:

2.1. Defining information

Information is a reproducible selection from its domain.

The domain of information (i.e., its selectable elements and their order) is a set which must be uniformly known by all those who exchange the information. The reproducible, common knowledge of the domain is an essential precondition for the exchange of information. We can say abbreviated: The domain of information must be defined uniformly for everyone beforehand and is inseparable from the information.

Concerning the exact physical definition, we imply that "information" can be transported and copied along increasing common time, as it corresponds to macroscopic everyday experience.

2.2. Digital application of the definition of information

The definition of information (2.1) is universal and generally valid right down to the basics of physics [19], because the result of every physical experiment is information and means a selection from the domain of possible experimental results. In physics, prior knowledge of this domain consequently also requires a common approach to the concept of "reproducible knowledge" or "past" and thus to the contradiction-free common order of our time [20].

Definition 2.1 becomes apparent in the case of digital information, which was designed from its beginning [9] as a selection from a domain. It is presumed that the domain of the digital information is known, e.g., as binary ASCII code. The coding of characters proved to be very useful and was extended by Unicode [21]. It is clear that these digital codes must be known uniformly worldwide, because a common knowledge of the domain of information (2.1) is a prerequisite for the exchange of information.

The main advantage of characters as a domain of information is that people already know them uniformly. For this reason, characters are initially suitable as a domain of digital information. Characters can be combined in a variety of ways, for example to represent language vocabulary, again assuming that the chosen language is already known. This already shows details that are important for the coding of information:

- Domains of information can be nested (e.g., bits as letters, letters as words, words as sentences…) by using previously defined domains for defining further

domains. Domains of information (and thus information) can be very deeply nested.

- The more forward in the nesting, the faster the knowledge about the domain must be accessible, right down to the physics. For the transport of digital information, for example, electromagnetic quantities and units must be clear from the beginning in order to encode bits and then use the bits to transport more complex digital information.
- In general, knowledge of the domains of information must be quickly available (largely unconsciously for us humans) for practicable coding of information.
- Language vocabulary is an example of a domain of information that is preferred due to widespread knowledge and universal applicability. However, it is not optimized for the reproducible transport of precise application-specific information.
- The domains of digital information can be defined online as domains of digital number sequences. This enables efficient selection and configurable similarity comparison. Nesting and arbitrary precision and complexity of the definitions are possible.
- Common knowledge of the domains of information is always essential before exchanging information.

So, since the Internet (whose global content can be quickly selected and copied) is available, there is an excellent opportunity: We could define the domains of information for applications of interest in a globally uniform way by defining (domain and binary format of) application-specific optimized number sequences online. This allows the unique Internet address or "UL" (see Section 2.3) of the online definition to link the globally unique definition with all defined data (number sequences).

2.3. Optimized digital data structure: Domain vector (DV)

Digital information consists of number sequences, each of which is a selection from its domain of information (Section 2.1). We can efficiently embed each number sequence in the new "domain vector" or "DV" data structure [17,18]. Each definition of a DV is published online. It is uniform due to its unique Internet address. Every DV has the structure:

$$\mathrm{DV{:}\,UL} \rightarrow \text{plus number sequence} \tag{1}$$

The "UL" is a "Uniform Locator" (an efficient Internet address). It represents a global pointer to the machine-readable, unique online definition of the number sequence. DVs with the same UL are automatically defined globally by the same online definition. The UL is therefore the identifier of a certain uniformly defined type of digital information and also a global pointer to the online definition. In this respect, the UL has a similar function like the Uniform Resource Locator (URL) of a conventional link, but is optimized in terms of efficiency. It is therefore represented directly as a hierarchical number sequence which can very short, for example:

1) Number: If between 0 and 63: Special meaning like "same UL as before" or (if greater than 63) number of the online presence, which allows the standardized

machine-readable (globally uniform) online definition of number sequences (digital information) by registered users.

2) Number: Positive number of the user of the online presence that defines online number sequences.
3) Number: Positive number of the definition of this user.

The UL is directly followed by the online defined number sequence of the DV. The binary encoding of the numbers in the DVs is also optimized for efficiency. We can avoid unnecessary difficulties, energy and efficiency losses from the beginning by using self-extending numbers. **Table 1** shows an example of the binary coding of a self-extending non-negative integer. The first bits always contain the length information about the count of the following bits of the mantissa.

Table 1. Self-extending non-negative integer, example starting with length from 4 bits to 24 bits.

Byte 1				**Byte 2**								**Byte 3**												**Max**
0	0	M	M																					4
0	1	M	M	M	M	M	M																	64
1	0	M	M	M	M	M	M	M	M	M	M													1024
1	1	0	0	M	M	M	M	M	M	M	M	M	M	M	M									16,384
1	1	0	1	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M					262,144
1	1	1	0	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	4,194,304
…																								

If this is not enough, further bits a.re used for the length information (colored blue). M: Bits of the mantissa. Max: Maximum size of the domain.

The concrete coding of DVs first requires the introduction of a standard, which should (in order to be attractive) be optimized in terms of efficiency. As part of the UL, the first number in the DV could have special meanings such as “same UL as before” (if small) or (if large) represent the identification number of an official website with online definitions. Then further numbers of the UL could follow in hierarchically ordered sequence, which address the online definition in this internet presence, followed by the online defined number sequence of the DV.

As soon as an online definition progresses from the draft stage to the final stage, changes are no longer possible. However, it is possible for the owner to mark an online definition, e.g., as “deprecated” with a link (UL) to an updated version. It is easy to extend an online definition by reusing it in a newer version.

Since all digital bits represent sequences of numbers, it is natural for DVs to have the full scope of digital information.

So far, however, there is no infrastructure to enable users to publish the definition of DVs, i.e., the format and domain of digital information, globally uniformly online, optimized for their application. Only then would it be possible to search for this definition globally in a practicable way, compare it and systematically develop it further, efficiently on the Internet without an explicit meeting. The following steps are therefore recommended:

- Foundation of an official international organization, e.g., within ICANN [22], which assigns the first number of the UL (see above) globally uniformly to owners of “Idefsites”, i.e., of websites that enable users to publish online definitions of DVs (digital information) in a standardized machine-readable

manner. This organization should also publish any version of the standard for machine-readable online definitions of DVs.

- The standard for online definitions is constantly evolving. Therefore, the first number of the online definition should reflect the version number of the standard. For efficiency reasons, the first part of the online definition should give the binary format of each number or sequence of numbers in the DV, followed by a link to the corresponding detailed description in the online definition. This description can include standardized abbreviations (numbers) and free text in English, with links to translations into different languages. In this way, a multilingual online definition is possible.
- The standardization and the info sites can be financed among other things by the users who publish their ULs with associated definitions in the Idefsite (see above). However, not every UL should cost a fee, as this would discourage users from publishing their definitions.

Some advantages of DVs can already be demonstrated. Due to their uniform online definition, DVs with the same UL can be compared worldwide. For example, a similarity search of DVs is possible according to criteria that the user can specify in the online definition. The principle of online definition and search of defined information can also be demonstrated within a local database using the existing prototype, see next Section 3.

3. Results

There is a prototype [23] for demonstration purposes (http://numericsearch.com/). It allows users to define DVs resp. number sequences within a local database after logging in (**Figure 1**).

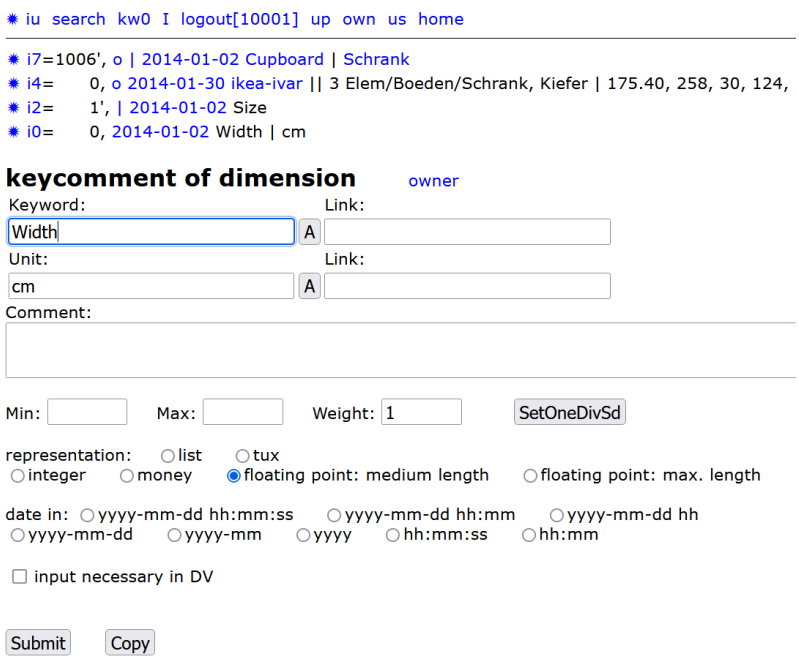

Figure 1. Definition of a DV using the online prototype [23].
In this example, the DV contains 4 dimensions (numbers) that provide a simplified description of a cupboard. Here, the dimension with the name “width” and the unit “cm” is edited.

The definition of number sequences also determines the meaning, dimensionality and structure of the domains. It is therefore explicitly demonstrated that number sequences (and thus domains of information according to definition 2.1) can be defined online. These online definitions are currently only valid within the local database of the prototype, as there is no global standard yet. Nevertheless, important applications can be shown: The comparison and search of digital data or number sequences defined online. The principle of this numeric search can be quickly demonstrated using an example (**Figure 2**).

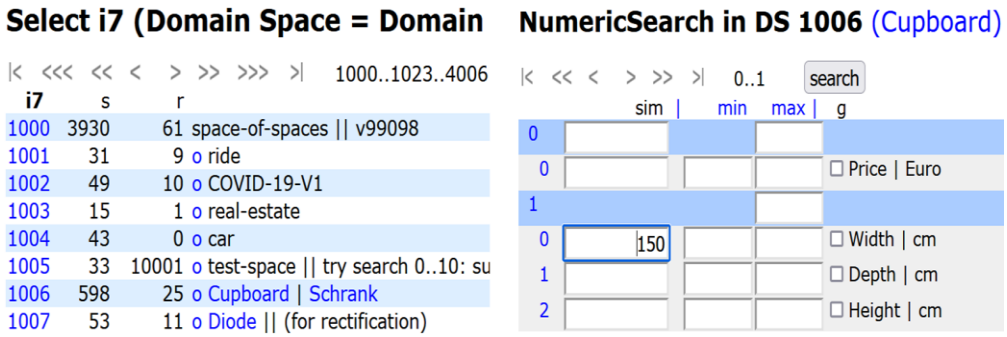

Figure 2. Demonstration of Numeric Search.

Number sequences (with domains) have been defined in a local database. Some of them are listed on the left and indexed by an index i7. After clicking on i7 = 1006, a window opens with the definition of the domain "Cupboard" shown on the right. Here we are searching for cupboards measuring 150 cm in width.

Figure 2 contains the titles of some exemplary domains of Information listed on the left. The domains here are also metric spaces and were therefore called "Domain Spaces" or "DSs". The exact metric for the similarity search can be specified by the user in the definition. After clicking on the domain "Cupboard" with index i7 = 1006 on the left, the search field appears on the right. It shows that the numbers "Price", "Width", "Depth" and "Height" have been defined as number sequence for this domain of information. In this case, the system searches for number sequences with the 2nd number "Width" = 150, i.e., for cupboards whose width is as close as possible to 150 cm.

In the case of a worldwide standard of online definition and DVs, we could search worldwide for DVs with this definition. This does not yet exist, so we first searched a local database for demonstration purposes. **Figure 3** shows the search result. Each line contains data from a DV, which describes a "cupboard" with the 4 numbers "Price", "Width", "Depth" and "Height". In this case, similarity search was performed for DVs with the 2nd number "Width" close to 150. Therefore, the entries are listed the higher the closer the 2nd number is to 150.

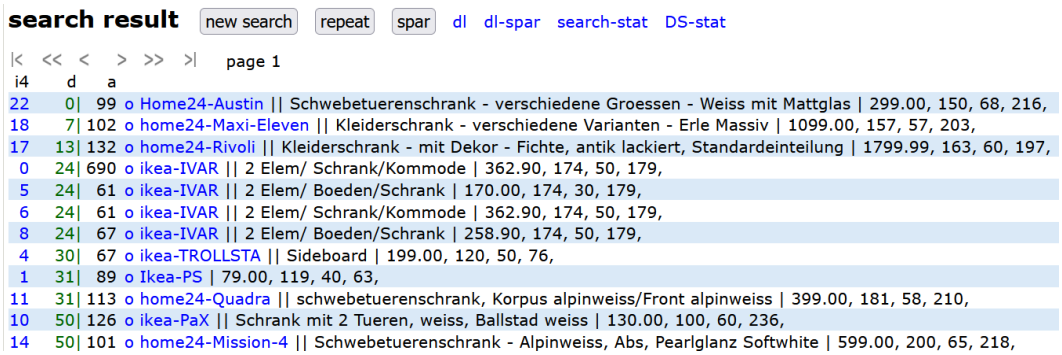

Figure 3. Search result after the similarity search shown in **Figure 2**.

The DVs are listed the higher, the closer their 2nd number "Width" is to 150. The search criterion is therefore the smallest possible absolute value of the difference between the 2nd number and 150 and is listed by the green numbers in column "d".

This is a simple example with one search criterion. Multi-dimensional searches are also possible according to the criteria selected by the user.

For this purpose, each domain is additionally equipped with a metric or distance function [24–26] in the prototype, so that it forms a metric space. Each domain is therefore called "Domain Space" (DS) in the prototype. For demonstration purposes, a DS was generated with more than 100 dimensions and 100,000 DVs, which were filled with evenly distributed floating-point pseudo-random numbers between 0 and 10. Selectable dimension groups can form subspaces with different metrics (distance functions). For example, 2 dimensions were selected from a subspace with a Euclidean metric (distance function) and a 2-dimensional similarity search was carried out for the point (x, y) = (5.0, 3.0). As a result of the similarity search, the left half of **Figure 4** graphically shows the 1000 points that represent the coordinates (of the selected dimensions) of DVs that are closest to the point (5.0, 3.0). Since Euclidean metric was chosen as the distance function, the points with minimum distance are located within an ellipse (circular area, which was widened due to the different vertical and horizontal scale). The right half of **Figure 4** shows the analogous result when 2 other dimensions are chosen in a subspace with Manhattan metric. The 1000 points nearest to (5.0, 3.0) now lie within a rhombus, as this contains the points with the smallest distance in the case of the Manhattan metric as distance function.

The use of metrics (distance functions) is not only important for the direct search of digital information, metrics are also decisive for the training of neural networks (AI) [27–29].

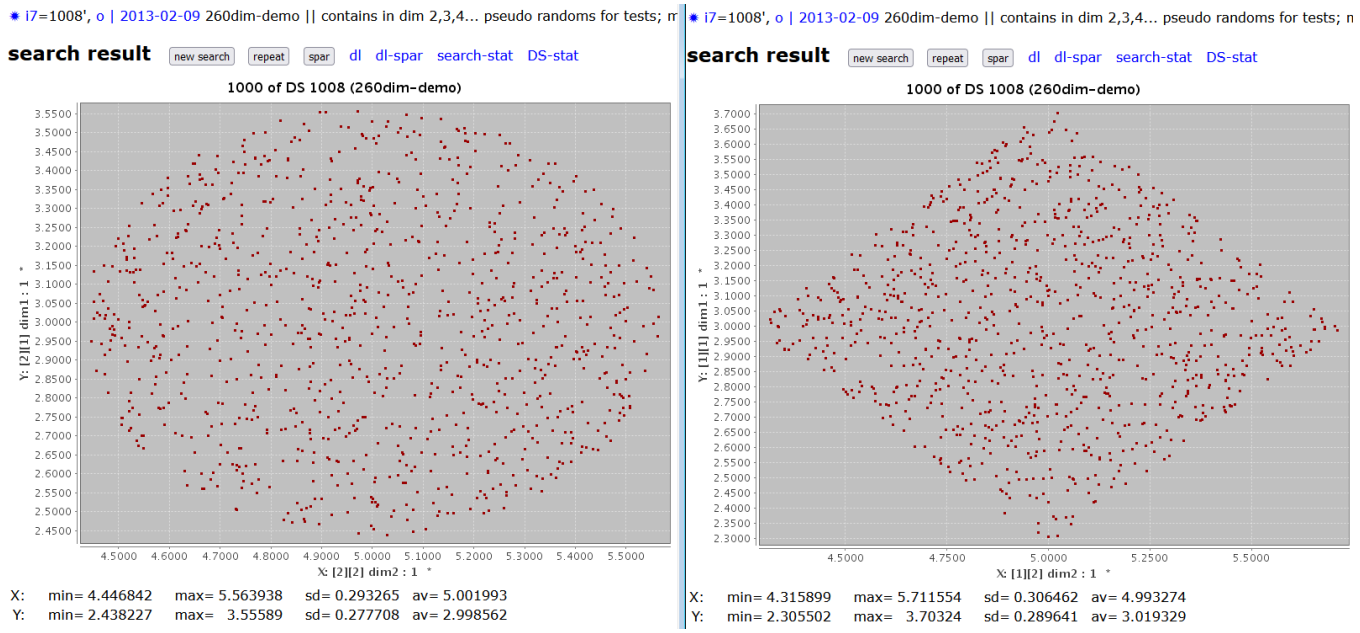

Figure 4. Graphical representation of the search result after a 2-dimensional similarity search in a domain with 100,000 DVs, which contain pseudo-random numbers between 0 and 10 as coordinates in these 2 dimensions.

Each red dot represents these coordinates of a DV. The coordinates of the 1000 DVs closest to the point (3, 5) are displayed. The left graphic shows the result of similarity search using the Euclidean metric, the right graphic shows the result using the Manhattan metric.

Due to the lack of a global standard, the prototype can only demonstrate some exemplary applications using a local database. The scope of online (and thus globally) defined DVs covers the entire spectrum of digital information (number sequences), as online definitions of DVs (number sequences) on any selectable topic are possible. **Figure 2** shows a few examples on the left. It is advisable to define reproducible number sequences in the online definition which contain information of user interest on the selected topic. The fewer numbers are sufficient for the search, the more targeted the search can be. Feature extraction according to application-specific criteria can be useful for this. This can also be automated using AI, e.g., Deep Learning [30,31].

4. Discussion

The exact definition 2.1 of information contains important details about time and the prerequisites for general information exchange, which can be used systematically since the introduction of digital information: Before information can be exchanged, there must be a domain of information [17,18], i.e., a common set of possibilities that all participants of a communication must know (e.g., language vocabulary). Only then is it possible to exchange information as a selection from this domain of information.

This clear connection between time and information could be more strongly addressed in information science and systematically applied in computer science [16–18,32–35], because the components of digital information are sequences of numbers that represent a selection from a domain of digital information. This domain must be known before digital information is exchanged.

Language vocabulary is often already known and therefore an example of a useful (quickly and unconsciously used) domain of information, but it is increasingly reaching its limits. The limits of language communication become obvious, for example (among nations or groups with different languages or) in reproducible, precise, technical and fast machine-readable communication. Since the existence of the Internet, however, we could define optimized language-independent domains of information online and thus uniformly worldwide for any application (e.g., for medical information exchange). Based on the definition of Information 2.1 the online defined DV data structure presented in Section 2.3 almost inevitably results in case of consequent optimization of efficiency:

- Since information can only be understood if the domain of information is known beforehand (2.1), the UL is placed before the number sequence in the DV data structure.
- The Internet with the machine-readable online definition (efficiently referenced by unique UL) ensures that this definition of the number sequence is available uniformly worldwide as quickly as possible.

As an efficient sequence of numbers, the DV data structure can be used universally for all types of digital information. It is therefore a primary stage of digital information where bitwise efficiency is particularly important to save time, energy and hardware (which is also a basis for long-term financial attractiveness). Therefore, **Table 1** gives an example of self-extending numbers and reminds us that efficiency should be further considered when agreeing on the concrete common standard of DVs. Thus, the DV data structure is as short as possible: It contains only the UL as an efficient (globally unique identifier and) link to the globally standardized online definition and the (online defined) sequence of numbers. This automatically ensures that the number sequence is defined uniformly worldwide. Furthermore, the structure of the DV (2.3) offers maximum freedom for the online definition.

It allows application-optimized, language-independent digital information to be uniformly defined, identified, efficiently exchanged and compared worldwide. This applies to all types of digital information, including "quantitative" information, for which there is not yet a uniform global definition. Software interfaces, e.g., the content of reproducible layers of artificial neural networks [31,36], can also be defined online in this way. As a result, AI can be connected and trained with a maximum global data set and its results can be compared globally. In this way, the results and benefits of AI for humans can be objectively evaluated by humans.

Despite their clear advantages, the global definition of digital information and the DV data structure have been ignored for many years. The next section tries to find some reasons for this.

4.1. Obstacles and concerns regarding the DV data structure

It seems that the exact set-theoretical approach to "information" is no more familiar today.

- **A new approach, which starts at the very beginning** of the definition of (digital) information, needs the creation of new software for a practicable

implementation. The effort required for this has so far been avoided. Several publications (e.g., [16–18,32–35]) have been ignored for years. It seems that the relevance of the definition of information (Section 2.1) is underestimated. It is also not recognized that the DV data structure (see Section 2.3), which is optimized in terms of efficiency, almost inevitably results as a conclusion. So far, none of this has been a topic in computer science and information science. The unambiguous definition of information with the help of (globally uniform) defined domains (of number sequences) is a wide, open field that is not yet the focus of university education and scientific research, although it forms a well-defined basis of digital information and makes important mathematical tools directly applicable. In the prototype (http://numericsearch.com/, see Section 3), various functions or "metrics" can be applied and compared to search for similarities in digital information.

- **This technical solution** (for a systematic, globally standardized online definition of digital information) **has so far been ignored**. In the meantime, a lot of patchwork has been established. However, this cannot replace the missing (common domain and definition of information as) basis. There is a lack of comparability or verifiability of digital information. This can cause an increase in misinformation and meanwhile also uncertainty due to "fake news".
- **This approach should not be confused with the semantic web concepts** [37–40], which attempt to make digital information readable by combining it with variable metadata. Global uniformity is therefore not guaranteed, and we are back to the old problems. In contrast, the online definition of DVs is automatically globally unique (due to the unique UL), and the data (DVs) can therefore be defined globally in a globally uniform, reproducible and much more efficient way. This can be made user-friendly right from the beginning, as the online definition (of number sequences) requires only little prior knowledge and is simple if supported by suitable software. Once online definitions are available, it can even be easier to use them than to create new definitions. It is possible to generate convenient online sites where each user can search the existing online definitions and select the most frequently used definitions (alone, without meeting other users). Users can also create a new definition on such an online site. The frequency of use can be displayed immediately later and indicate whether this is well done and attractive. The existing prototype [23] (see Section 3) already shows that it is possible in principle to program online presences in which users can create their definitions of digital information (number sequences) and search in existing definitions.
- There may be concerns that **privacy** [41–44] could be reduced in the future as a result of globally comparable data. In fact, the global definition is a powerful tool, and it depends on how you use it. One can also use this tool to improve individual privacy, as it enables the creation of anonymized statistics from global data, which serve their purpose better than individual data and can be very helpful, e.g., for medical decisions. The online definitions of DVs help to reduce the unnecessary variety of interfaces, which makes them less vulnerable to hackers. It is also possible to define securely encrypted DVs.

- The **publication** of a definition may not be desirable. In any case, everyone should know that online definitions do not have to be perfect. They may also contain information about a draft stage or preliminary stage of development.
- The **organizational** [45] effort is not trivial. However, this must be compared with the organizational and regulatory effort that is necessary today as a result of incompatible data and a lack of interoperability. Instead of regulations, it is better to define the DVs in such an attractive, efficient and economically advantageous way that people are happy to use them.
- **Imprecise definitions can also be created online**. This is evident when the same primary data source can lead to significantly different digital information (as DV), i.e., when the conversion from original to digital information is not sufficiently reproducible. For example, the online definition may simply allow the DV to reproduce a medical patient report as free text, without any further specifications. The DV is therefore not better comparable than medical findings in free text. Such free text can be used as an introduction to a specific application (e.g., rough medical diagnosis). Further precise data should then be digitized in a suitable DV whose online definition has been optimized for this application (example: After rough medical diagnosis, a DV should be selected whose online definition contains reproducible results of precise patient findings relevant for this diagnosis). To ensure sufficient reproducibility, optimized online definitions must therefore be written for the specific applications.
- **Poor quality of the primary data source** is also problematic if the online definition (of the conversion into the digital representation as DV or number sequence) is well-defined and unambiguous (reproducible). In particular, the primary data source may contain systematic errors that are not easily recognizable. After all, globalized data collection can make the data volume so large that the systematic errors that would otherwise result from local data collection can be avoided. Random errors can also be better corrected by global averaging. Statistical parameters (size of the data collection, standard deviation) make it possible to estimate this.
- **Redundant definitions** arise when online definitions already exist for an application and users nevertheless post new definitions online. Until now, however, the redundancy is much greater: Due to the lack of infrastructure for online definitions, users and programmers cannot systematically search for existing definitions and are forced to define them as new digital information (number sequences) locally and redundantly.
- **Investments are needed** to support online definitions and handle the new DV data structure (2.3) with comfortable software. The global machine readable standard introduced should pay attention to efficiency so that the format of globally defined data (DVs) is generally attractive, even for software interfaces and programmers. The version number can be specified at the beginning of each online definition to enable continuous updates of the standard.
- The **UL** in the DV data structure (2.3) requires some bits. However, this is only noticeable if the UL is long compared to the following number sequence defined online and if the UL is repeated frequently. The UL should therefore be designed as a short (hierarchical) number sequence that is optimized for

efficiency. As shown in Section 2.3, it is sufficient that the UL contains 3 self-extending numbers (**Table 1**). This makes an optimized standard possible, so that DVs are also superior in terms of efficiency.

- Today, there seems to be little interest from large Internet companies in the new DV data structure. Perhaps they are currently satisfied with the **established structures**, or the potential of DVs has not yet been recognized by companies. The globally defined DV data format (2.3) can actually set a new standard due to its superior efficiency, universal applicability and many other advantages (see below). This could actually be seen as an opportunity for companies to enter a new important market.
- Today, the introduction of online definitions is **only possible gradually** and is not as natural as their systematic use from the beginning. However, it is also possible today to introduce the DV (2.3) data structure and to define more and more attractive domains of digital information (systematically building on each other) in a standardized way worldwide.
- Temporary variables whose definition is only required internally for a short time do not need to be defined online. But anything that could be of repeated interest to other users is suitable for an online definition. In particular, this includes all reproducible number sequences (data) that are needed for a specific application.
- Creating an online definition first **requires reflection**: Which features are interesting for this application, and how can these best be represented as numbers? However, such questions also arise when software is created for the same application. In particular, without an online definition, these questions arise redundantly again and again and are then answered (incompletely or) in different ways. This ultimately means extra work with data gaps, and the resulting data (number sequences) are not comparable.
- **Online definitions in the "final" stage can no longer be removed**. However, subsequent comments are possible, which can also contain a link to a more up-to-date version.

4.2. Advantages of the DV data structure

After their introduction, the advantages of DVs can become increasingly apparent. Since DVs are optimized bit by bit as universal digital data carriers, their data structure (2.3) is also attractive for reasons of saving resources and energy (especially in view of the huge amount of digital data). However, the most important advantages of the DV data structure result from its globally uniform definition and identification (by unique UL, see 2.3). This makes it possible to gradually reduce significant current problems of digital information, which result from a lack of knowledge of the domain of information (2.1). Without this knowledge digital information is simply not readable - not usable. We have already learned that this leads to inefficiency, redundancy, interoperability problems, unnecessary redefinition with reprogramming with inconsistent operation, lack of comparability and therefore lack of objectivity in the evaluation of digital information (including AI).

Apart from reducing such problems, the DV data structure also opens up many new possibilities and advantages. The UL at the head of the DV data structure identifies a specific type of globally standardized digital information. The user is free to decide what the sequence of numbers defined online represents in the DV. The number sequence can universally represent all types of digital information. Due to the globally uniform identification by UL, the DVs can be extracted from the entire open web and are thus accessible for various evaluations and applications that were previously not possible. Some of the advantages of online definitions and the DV data structure (2.3) are listed below:

- **Specific professional support:** As in medicine, the online definitions can also be optimized for other applications in order to improve the precise professional exchange of information.
- **Reduction of redundancy:** To this day, the wheel is reinvented again and again. Digital data is also redefined again and again for the same application. Without an online definition, new definitions are usually different and the resulting data are no longer comparable. Data for special applications are often incompatible if the generating software comes from different providers. However, if the providers would define their digital data online as DVs, they would all be able to reuse exactly the same definition.
- **Global program interfaces:** Digital program and subprogram interfaces can be defined online as DVs to facilitate data exchange. (Parts of) software architectures can also be shown online to achieve greater uniformity in programming and globalized digital collaboration. This enables the step-by-step creation of even very large software projects and can be combined with open source if desired.
- **Expanding international communication with an increasing common vocabulary:** The total vocabulary of all human languages is too large for our mind. Nevertheless, this is only a tiny part of the possible domains of information that can be represented as domains of DVs. The domains of DVs can be optimized language-independently for the respective application and thus enable language-independent international communication. The definitions of existing DVs can be reused in new online definitions via UL. The domains of DVs form an internationally standardized vocabulary for more and more topics and help to reduce linguistic and technical separation. Vocabulary from different languages can be used in online definitions of DVs. Since not everyone can know every language and the English language has become established, it is recommendable that each online definition (also) contains a complete definition in English as a reference.
- **Reuse of existing definitions:** Online definitions of DVs are nestable, i.e., existing online definitions can be integrated (via UL) and reused in later definitions. The reuse of efficient online definitions is recommendable. For example, the reuse of an online definition “age” (**Figure 5**) enables a comprehensive search for (e.g., patients with) a similar age in the entire group of DVs that also use this online definition. The systematic reuse of online definitions is beneficial for all types of digital information.

- **Connecting and combining application-specific data:** Separate data on different applications become internationally visible by means of online definitions. Their reuse connects the data of more and more applications and thus opens up new possibilities for data evaluation and data exchange, for example in science.
- **Connecting and combining scientific data:** To this day, the reuse of research data is not a given. This results in unnecessary redundancy and inefficiency in scientific work. If, on the other hand, the data used in scientific work would be defined online as DVs, both research data and their definitions would be directly reusable. This would not only increase efficiency and lower redundancy, it would also open up new comprehensive possibilities for the evaluation and collaboration of scientific research. This should actually be of great interest to the scientific community and also to scientific organizations such as RDA [6].
- **The author** of data can also be efficiently shown by an additional DV.
- **Precise global data search:** A universal and global similarity search is immediately possible (**Figures 2** and **3**) if the DV for the selected topic quantifies the features so that similar numbers also have similar meaning.
- **Sophisticated global data search:** Even if the DVs only indirectly reflect features of interest, their similarity search and comparison is possible after intermediate steps. Suitable intermediate steps include dimension reduction and feature extraction, which can also be automated with the help of AI [30].
- **New globalized medicine:** Due to the increasingly detailed (high-dimensional) diagnostic data provided by digital medical diagnostics, it is no longer possible for the human mind to keep track of all the details. It is therefore possible that the importance of details is overlooked, e.g., the combination of certain genetic data with certain blood findings. Much more would be recognizable by precise data comparison of these findings with the globalized data stock of medical data processing. The obvious first step is a similarity search (see above) to localize the current finding in the global database. Particularly sophisticated evaluation is possible by applying AI to the DV database: In medicine, DVs can transport the AI input data, including certain findings and medical images. Additionally defined DVs can be suitable for data exchange after various AI processing steps such as convolution, dimension reduction and specific feature extraction. The DVs resulting from AI processing can directly reflect decision-relevant features of the current global data set, e.g., probabilities of possible diagnoses and treatments (**Figure 5**).

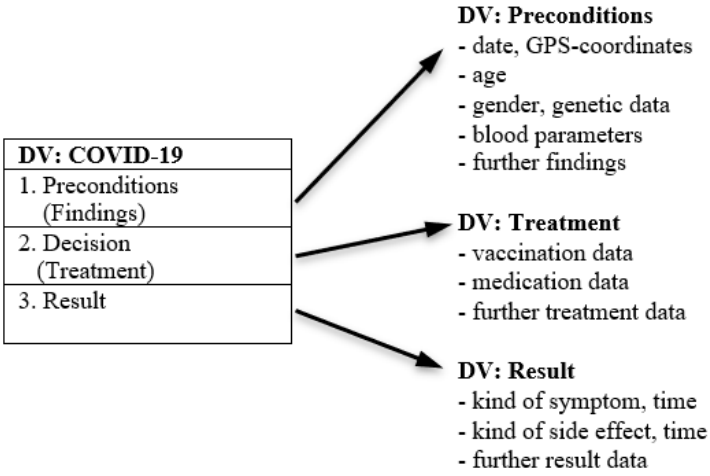

Figure 5. Examples of DVs defined online in medicine.

Online definitions can be nested. It is attractive to reuse efficient online definitions by incorporating them into new definitions.

- **Application of Large Language Models (LLM)** [46–48] within the new, globally defined framework: AI architectures that use LLM (e.g., ChatGPT [49]) can be adapted and trained (using online definitions e.g., of medical treatment data) to translate precise questions (e.g., with medical findings) into corresponding online defined data and to translate interesting results (e.g., related precise online defined medical treatment results) back into human language. However, language vocabulary is only a very small subset of the possible domains that can be represented by online defined number sequences, as these have a much higher cardinality and resolution. It is therefore advisable to extend LLMs. As a first step, such extensions could provide graphical output in addition to text in order to better visualize the exact original numerical data.
- **Global AI training and global control of results is made possible:** The importance of this is illustrated by a short citation [50]: "…with AI models, the essential need is in data".

It makes sense to use DVs as training data for AI because of their globally uniform online definition. This allows the globalization of AI with a maximum amount of training data. The online definition can be adapted to this application. DVs can contain input and output data of AI (**Figure 6**) and also the data of reproducible intermediate steps, for example intermediate steps of deep learning [30]. For new "global open AI" and also to support existing collaboration initiatives [50], the definition of DVs can also be provided with (links to) information about the AI architecture. Globalization of the data for the AI models helps to avoid hidden local bias. It is an important step towards the comparability and verifiability of AI results and towards improving their quality.

It seems appropriate to take a closer look on the application of AI to globally (online) defined digital data. Input layer, output layer (**Figure 6**) and even intermediate layers of artificial neural networks can be globalized by online definitions if they are sufficiently reproducible.

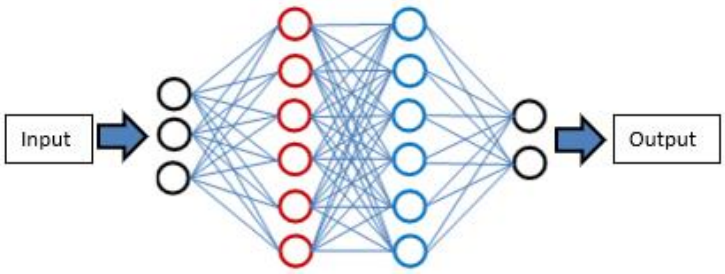

Figure 6. Simplified representation of an artificial neural network.

If the numbers representing the neurons of the input layer and output layer are well-defined, they can also be defined online (globally). This allows training data to be globalized and the quality of AI results to be checked globally and uniformly by humans.

Training of AI on online defined DVs would be a relevant topic for further research, because this allows the training and result data for AI to be globalized and maximized. By training on a globally defined data set, AI bias can be avoided, which would otherwise have occurred when training on a locally defined subset of the data. In addition, the quality of the results can be checked globally, uniformly and efficiently according to objectifiable standards—by comparing them with global reality.

In medicine, for example, such globalization of data would be helpful for the efficient automation of learning from real experiences. We could use the global data pool for research purposes and directly for numerical searches of "similar" cases in domains optimized for the diagnoses of interest (e.g., "COVID-19" in **Figure 5**). The globalized data pool could also be used automatically for machine learning (**Figure 6**) or deep learning. Not only the data, but also the online definitions could be expanded to include additional dimensions (numbers) about interesting details on findings, diagnoses, treatments and treatment outcomes (**Figure 5**). Reusing existing terminologies can be helpful, e.g., integrating ICD-10 codes [51] into DVs with health data. The reuse of efficient online definitions is explicitly recommended. The more existing definitions are reused, the more connected the resulting data space can be.

Since DVs (2.3) are universally applicable to all types of digital information, it is not possible to list all the advantages resulting from their globally uniform identification and definition. Existing publications [17,18] contain additional details.

5. Conclusion

The title of the paper is meant literally: Domain vectors are a new fundamental approach that can play a central role in optimizing the structure of all digital information. The pieces of digital information (sequences of numbers) can be defined globally uniformly online as domain vectors, in efficient structure down to the bit level. As the online definition can be used universally, its technical potential is far-reaching. However, there is still no standard and no infrastructure for defining digital information online. The correction would begin with publication (e.g., by ICANN [22]) and investment in a convenient online presence where users can define standardized DVs (2.3) for their own needs. It also makes sense to invest in software for convenient handling of DVs. Then more and more machine-readable online

definitions of DVs can be created and optimized for more and more applications. This would enable language-independent, uniform global information exchange, which is superior in terms of efficiency, precision and speed. The machine readability of DVs can also be used to train AI on a maximum global amount of data and then for objectifiable, standardized comparison of the quality of AI results by humans.

The introduction of online defined digital information or DVs (2.3) would mean a significant improvement in the digital exchange of information and is therefore highly recommended.

Conflict of interest: The author declares no conflict of interest.

References

1. Exact Definition of Information and digital application. Available online: https://en.wikipedia.org/wiki/Information (accessed on 29 September 2024).
2. Hodapp D, Hanelt A. Interoperability in the era of digital innovation: An information systems research agenda. Journal of Information Technology. 2022; 37(4): 407–427.
3. Brownsword LL, Carney DJ, Fisher D, et al. Current perspectives on interoperability. Rapport Technique. 2004; 13.
4. Rubinfeld D. Data portability and interoperability: An EU-US comparison. European Journal of Law and Economics. 2024; 57(1): 163–179.
5. Barbu M, Vevera AV, Barbu DC. Standardization and Interoperability—Key Elements of Digital Transformation. Springer; 2024. pp. 87–94.
6. Treloar A. The Research Data Alliance: Globally co-ordinated action against barriers to data publishing and sharing. Learned Publishing. 2014; 27(5): 9–13.
7. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. Springer; 2016.
8. Rubin RE, Rubin RG. Foundations of library and information science. American Library Association. 2020.
9. Shannon CE. A mathematical theory of communication. The Bell System Technical Journal. 1948; 27(3): 379–423.
10. Kolmogorov AN. Three approaches to the quantitative definition of information. International Journal of Computer Mathematics. 1968; 2(1–4): 157–168.
11. Yuexiao Z. Definitions and sciences of information. Information Processing & Management. 1988; 24(4): 479–491.
12. Losee RM. A discipline independent definition of information. Journal of the American Society for information Science. 1997; 48(3): 254–269.
13. Madden AD. A definition of information. Aslib Proceedings. 2000; 52(9): 343–349.
14. Hlaváčková-Schindler K, Paluš M, Vejmelka M, Bhattacharya J. Causality detection based on information-theoretic approaches in time series analysis. Physics Reports. 2007; 441(1): 1–46.
15. Kuzenkov OA. Definition of information in computer science. Izvestiya VUZ. Applied Nonlinear Dynamics. 2024; 32(4): 541–562.
16. Orthuber W. Uniform definition of comparable and searchable information on the web. Available online: https://arxiv.org/pdf/1406.1065 (accessed on 29 September 2024).
17. 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. doi: 10.3390/ijerph17082975
18. Orthuber W. We Can Define the Domain of Information Online and Thus Globally Uniformly. Information. 2022; 13(5): 256. doi: 10.3390/info13050256
19. Dirac PAM. The principles of quantum mechanics. Oxford University Press; 1981.
20. Orthuber W. All physical information is discretely connected from the beginning and all geometrical appearance is a delayed statistical consequence. Ann. Math. Phys. 2023; 6(2): 159–172. doi: 10.17352/amp.000097
21. Aliprand JM. The Unicode standard. Libr. Resour. Tech. Serv. 2011; 44: 160–167.
22. ICANN: Internet Corporation for Assigned Names and Numbers. Available online: https://www.icann.org/ (accessed on 3 September 2024).

23. Orthuber W. Demonstration of Numeric Search in User Defined Data. Available online: http://numericsearch.com/ (accessed on 3 September 2024).
24. Chen S, Ma B, Zhang K. On the similarity metric and the distance metric. Theoretical Computer Science. 2009; 410(24–25): 2365–2376.
25. Chauhan SS, Garg P, Thakur K. Study of metric space and its variants. Journal of Mathematics. 2022; 1: 7142651.
26. Ghazal TM. Performances of k-means clustering algorithm with different distance metrics. Intelligent Automation & Soft Computing. 2021; 30(2): 735–742.
27. Kulis B. Metric learning: A survey. Foundations and Trends® in Machine Learning. 2013; 5(4): 287–364.
28. Moutafis P, Leng M, Kakadiaris IA. An overview and empirical comparison of distance metric learning methods. IEEE Transactions on Cybernetics. 2016; 47(3): 612–625.
29. Roth K, Milbich T, Sinha S, et al. Revisiting training strategies and generalization performance in deep metric learning. PMLR; 2020. pp. 8242–8252.
30. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553): 436–444.
31. Kukreja, H., Bharath, N., Siddesh, C. S., Kuldeep, S. An introduction to artificial neural network. 2016, Int J Adv Res Innov Ideas Educ, 1(5), 27-30.
32. Orthuber W. Reproducible Transport of Information. IOS Press; 2021. Volume 281. pp. 3–7.
33. Orthuber W. Online definition of comparable and searchable medical information. Digital Medicine. 2018; 4(2): 77–83.
34. Orthuber W. Global predefinition of digital information. Digital Medicine. 2018; 4(4): 148–156.
35. Orthuber W, Hasselbring W. Proposal for a new basic information carrier on the Internet: URL plus number sequence. IADIS; 2016. pp. 279–284.
36. Uzair M, Jamil N. Effects of hidden layers on the efficiency of neural networks. In: Proceedings of the 2020 IEEE 23rd international multitopic conference (INMIC); 5–7 November 2020; Bahawalpur, Pakistan. pp. 1–6.
37. Lassila O, Hendler J, Berners-Lee T. The semantic web. Scientific American. 2001; 284(5): 34–43.
38. Hitzler P, Krötzsch M, Rudolph S, Sure, Y. Semantic Web: Grundlagen. Springer-Verlag; 2007.
39. Patel A, Jain S. Present and future of semantic web technologies: A research statement. International Journal of Computers and Applications. 2021; 43(5): 413–422.
40. Berners-Lee T, Hendler J, Lassila O. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Association for Computing Machinery; 2023. pp. 91–103.
41. Quach S, Thaichon P, Martin KD, et al. Digital technologies: Tensions in privacy and data. Journal of the Academy of Marketing Science. 2022; 50(6): 1299–1323.
42. Tawalbeh LA, Muheidat F, Tawalbeh M, Quwaider M. IoT Privacy and security: Challenges and solutions. Applied Sciences. 2020; 10(12): 4102.
43. Liu B, Ding M, Shaham S, et al. When machine learning meets privacy: A survey and outlook. ACM Computing Surveys. 2021; 54(2): 1–36.
44. Letafati M, Otoum S. On the privacy and security for e-health services in the metaverse: An overview. Ad. Hoc. Networks. 2023; 103262.
45. Chen L, Tong TW, Tang S, Han N. Governance and design of digital platforms: A review and future research directions on a meta-organization. Journal of Management. 2022; 48(1): 147–184.
46. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology. 2024; 15(3): 1–45.
47. Minaee S, Mikolov T, Nikzad N, et al. Large language models: A survey. Computer Science. 2024.
48. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nature medicine. 2023; 29(8): 1930–1940.
49. Deng J, Lin Y. The benefits and challenges of ChatGPT: An overview. Frontiers in Computing and Intelligent Systems. 2022; 2(2): 81–83.
50. Lin Z, Ma W, Lin T, et al. Open-Source AI-based SE Tools: Opportunities and Challenges of Collaborative Software Learning. Computer Science. 2024.
51. Fung KW, Xu J, Bodenreider O. The new International Classification of Diseases 11th edition: A comparative analysis with ICD-10 and ICD-10-CM. Journal of the American Medical Informatics Association. 2020; 27(5): 738–746.