PROPOSAL FOR A NEW BASIC INFORMATION CARRIER ON THE INTERNET: URL PLUS NUMBER SEQUENCE

Wolfgang Orthuber[1] and Wilhelm Hasselbring[2]
[1]*UKSH, Kiel University, Arnold-Heller-Straße 3, D-24105 Kiel, Germany*
[2]*Software Engineering Group, Kiel University, Christian-Albrechts-Platz 4, D-24118 Kiel, Germany*

ABSTRACT

A new searchable data structure on the Internet is presented, which can be defined with maximized information density for maximal efficiency of information transport. For this purpose, the new basic information carrier "URL (of standardized online definition) plus number sequence" is introduced. It is called "Domain Vector" (DV). The online definition defines a "Domain Space" (DS) in a standardized (machine-readable) way. This DS is a metric space whose elements are the DVs. A DV can precisely represent any definable information, from a simple word to complex multidimensional information, e.g., in science, medicine, industry, or economy. DVs may be defined by anybody, and the definition can be optimized as required by the application. If the length of the URL is minimized (e.g. using a short local substitute), DVs contain only necessary data (numbers) to enable maximal information density. They are internationally uniformly defined and identified (by the URL), and searchable according to user-defined criteria. This way, the proposed data structure fulfills the intended requirements. The online implementation http://numericsearch.com demonstrates examples and shows searchability.

KEYWORDS

Online definition, domain space, domain vector, Internet standard

1. INTRODUCTION

There have been many proposals to improve the data structure resp. information representation (the basic digital coding of information) on the Internet [Auer 2016, Guha 2016]. However, we noticed that e.g. user defined quantitative (numeric) data are not searchable on the Web. This deficit motivated us to study in more detail the existing shortcomings to derive the most important requirements that an approach to data on the Internet should fulfill: The data structure should allow efficient information transport. This can be quantified in several ways, e.g. by the information density D = I / (number of bits used for information transport). At this "I" quantifies the communicated information [Kolmogorov 1968, Section 1]. D should be maximized to minimize cost of information transport. This means that the number of bits used for transport of certain information (the length of information representation) should be minimized.

Information always describes a selection from a set of possibilities or a "domain" [Shannon 2001]. Thus, a well defined domain is precondition for precise digital information. The domain should have an online definition with unique URL to be efficiently accessible and well known by all participants of a conversation. The domain should be ordered and its definition should be standardized, machine readable and complete, such that for information transfer only the URL of the definition and (as information resp. data) a sequence of numbers (for description of the selection within the domain) is necessary. The URL of the definition is also identifier of a certain kind of data. This is important e.g. for search. Due to completeness of the online definition the data (numbers which describe a selection) do not need to contain parts of the definition. The software can download all necessary details for user-friendly representing, editing, comparing and searching the data from the online definition.

After describing the requirements, we show an example for the current data structure on the Internet. Figure 1 shows an RDF example [W3C 2016], which contains 3 variables (patientnumber, date, bodytemperature):

ISBN: 978-989-8533-57-9

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns: cd="http://www.example.ns/cd#">
<rdf:Description
rdf:about="http://www.example.ns/cd/bodytemperature measured">
   <cd:patientnumber>987654</cd:patientnumber>
   <cd:date>2015-02-28</cd:date>
   <cd:bodytemperature_celsius>37.01</cd:bodytemperature_celsius>
</rdf:Description>
</rdf:RDF>
```

Figure 1. Modification from http://www.w3schools.com/Xml/xml_rdf.asp to code the observation of body temperature. There is no Link to a standardized definition of the variables

We did not found in RDF (Figure 1) and not in similar approaches a link to a complete standardized online definition (see e.g. Figure 3) of the variables. However, this is a basic requirement. After detailed study of existing approaches (e.g. schema. org [Khalili 2013], RDF [W3C 2016, Quilitz 2008], HL7 [Benson 2012, Pedersen 2004], FHIR [Smits 2015]) for machine-readable data on the Internet [Orthuber 2015, Section 1.2], we noticed that the above requirements are not yet fulfilled. Before providing a solution, we first define more precisely the used term "URL".

Definition 1 (URL): By "URL" we mean an (internationally) unique resource locator (literally speaking). It can be e.g. the uniform resource locator defined in [W3C 2016]. It can also be a shorter locally defined substitute, e.g. a number (index) which points into a locally stored table with conventional external URLs. If the table contains u external URLs, not more than $\lceil \log_2 u \rceil$ bits are necessary for the local substitute of the (external) URL. In the future, external URLs can be shortened (e.g. replaced by hierarchical number sequences). A short modified start of data (with meaning "SameURLAsBefore") can make repetition of defining URLs superfluous.

2. DESCRIPTION OF THE APPROACH

A new explicit definition of the basic data structure is proposed [Orthuber 2016] that represents digital information as bit sequences. A bit sequence is also a number sequence (like 0 and 1). We prefer the term "number sequence" as carrier of digital information because it can be used more universally to clarify a possible order in the digital information that later can be used for similarity search or as pointer within a set. For example, text can be coded as numbers which each point to a certain character within a coding system such as Unicode, or as numbers that point into an ordered set of words. It depends on the intention whether we see a large data file, a word or only a bit as carrier of information. In every case the digital representation of information is a number sequence (with variable length). It always describes a selection within a set resp. domain. Central requirement of Section 1 is a link to the definition of the numbers. Therefore, additionally to the usual (by context defined) information carrier "**number sequence" (1)** we propose the combination "**URL (of standardized online definition) plus number sequence**" **(2)** as new basic information carrier on the Internet. It is called "domain vector" (DV) [Orthuber 2015] and is element of a set called "domain space" (DS) which is defined by the online definition. Therefore, the online definition can be also called "DS definition". A DS contains all DVs with the same URL and is a metric space to allow similarity search [Kriegel 2010, Zezula 2005]. The URL makes (2) longer than (1). To minimize this disadvantage, the URL is defined according to Definition 1. This means that it can be replaced by a local substitute, e.g. by a short number. If the length of the number sequence is variable, (2) must also contain information (e.g. an additional number) which specifies the actual length.

The DV resp. (2) has substantial advantages compared to (1). For programmers or designers of databases this may be obvious, because any number can be globally optimized for a certain application. A well-defined domain is precondition of precise information transfer. Therefore, the URL of the online definition is important and (2) can be used as general carrier of information on the Internet, which fulfils all requirements listed in Section 1.

Figure 2 shows the structure of a DS definition. The main content is an ordered sequence of dimension definitions. To allow access to dimension definitions, besides the full DS definition every dimension definition is also addressable by a URL. Figure 3 shows a short syntax example of a DS definition. At the given URL it exists *once on the web* and defines all DVs of this DS. Figure 4 shows a possible text form of a DV.

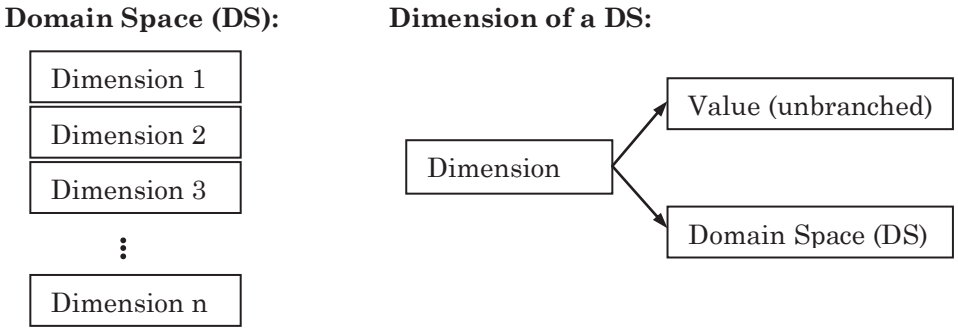

Figure 2. Structure of a DS and DS dimension. The DS and every of its dimensions have a URL. Every dimension of a DS represents an unbranched value (number) or again a DS. This can be used for nesting DS definitions

```
<?xml version="1.0"?>
<DS>
  <v>1</v>
  <kw>BodyTemperature</kw>
  <dim>
    <kw>Patientnumber</kw>
    <co>patient number in data collection xx</co>
    <type>int</type>
  </dim>
  <dim>
    <kw>Date</kw>
    <type>yyyy-mm-dd</type>
  </dim>
  <dim>
    <kw>Body-Temperature-Morning</kw>
    <co>Body temperature at morning directly after stand up</co>
    <unit>degree Celsius</unit>
    <type>float</type>
  </dim>
</DS>
```

Figure 3. XML example of a (not nested) three-dimensional DS definition. Its DVs represent the same data as Figure 1. The keyword (kw) of every dimension must be unique in the DS and can also be used for building a URL of every dimension definition. Thus, every dimension definition can also be accessed as DS definition of a one-dimensional DS

```
<v http://numericsearch.com/patbt.xml; 987654; 2015-02-28; 37.01>clickable text</v>
```

Figure 4. Exemplary syntax of a DV in text form (e.g. for insertion into HTML) which represents the same data as Figure 1, using the DS definition of Figure 3. Still more abbreviation is possible by using a binary form

The DV in Figure 4 contains the same numeric data as the RDF example in Figure 1. In the DV it is not even tried to provide a (partial) definition together with the data. The DV contains only the (numeric) data and the URL of its definition, this way it is much shorter. Nevertheless, it contains useful additional information: the clickable text is e.g. necessary for integration into HTML and the URL is necessary for identification of the data *and* for providing a link to the definition (Figure 3), which can contain much more information than Figure 1. According to Definition 1, the URL can be abbreviated as a number. Still more abbreviation is possible by direct binary coding of the DV, and of the DS definition (Figure 3). Using the DS definition DVs can be user friendly edited using e.g. an appropriately designed table (Figure 5).

0	987654	Patientnumber
1	2015-02-28	Date | yyyy-mm-dd
2	37.01	Body-Temperature-Morning | degree Celsius

Figure 5. Tabular representation of a DV in our implementation [Orthuber 2012], which can be used for editing the DV

Such a table (Figure 5) is a basic possibility for viewing and editing data. In the DS definition, more sophisticated (e.g. graphical) possibilities can also be defined. As already introduced [Orthuber 2015], search

of DVs is possible according to criteria defined by users in the DS definition. This includes similarity search [Zezula 2005] and is demonstrated [Orthuber 2012]. A detailed description is available [Orthuber 2016].

3. EVALUATION, EXAMPLES, SCOPE

The approach includes several new aspects for possible evaluation, which depend on the application, for instance, the exact increase of information density D (Section 1). Because users can adapt the DS definition to the application and repetition of defining context is no longer necessary, relevant improvement of information density can be expected (compare e.g. Figure 1 and Figure 4).

Evaluation of searchability is of particular interest. We programmed an implementation [Orthuber 2012] with a new scalable "synchronized index". The evaluation [Orthuber 2012, Section 4] showed that this index allows a posteriori combination of DS dimensions for search. This is important especially in case of high dimensional DSs or if later dimensions from different DSs need to be combined for multidimensional search.

We now provide an example of a typical application: Combination of searchable data with a resource (e.g. a picture) or its URL. Subsequently the name of a DS starts with "DS..". Let DS_gps contain (two-dimensional DVs with) GPS coordinates (latitude, longitude) and DS_map contain (rectangular, west east oriented) maps in jpg form. Let DS_gps_midpoint contain (DVs with) GPS coordinates of the map's midpoint, let DS_date contain a dimension with dates, let DS_kwmap contain DVs which each represent an index within a table of appropriate keywords (e.g. city, satellite, air-pollution, weather, ...), let DS_km_westeast and DS_km_northsouth contain dimensions of the map in km. According to Figure 2, we can combine these definitions in the definition of a new DS_map_searchable in the order DS_gps_midpoint, DS_date, DS_kwmap, DS_km_westeast, DS_km_northsouth, DS_map. Figure 6 shows the definition in our implementation [Orthuber_2012].

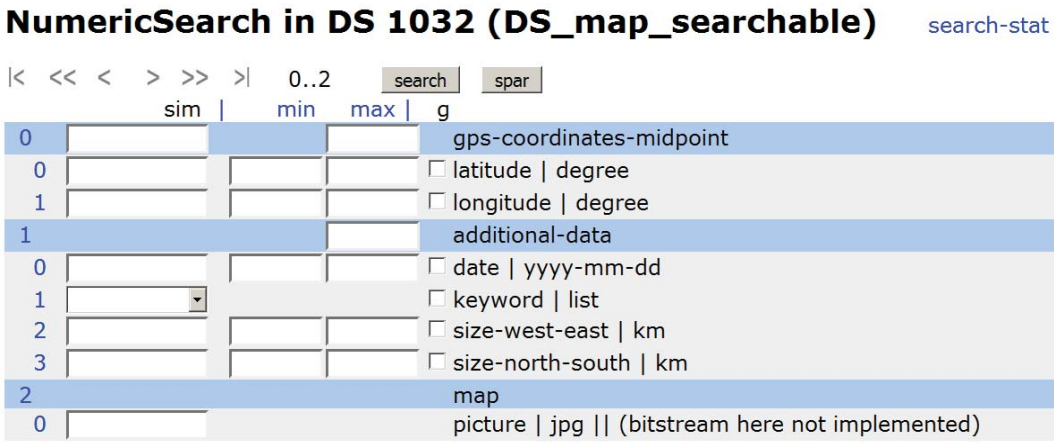

Figure 6. Definition of DS_map_searchable, except DS_map (long bitstream)

Because long bitstream (DS_map) is not defined precisely, it is not suitable for meaningful similarity search. But the other data (numbers) of DS_map_searchable are precisely searchable.

DVs with partial content are possible, e.g. with placeholders instead of numbers at omitted positions. We can also allow definitions of DSs with variable length. A DV with variable length can contain any kind of binary data ("bitstream"), e.g. those which are currently stored as files with certain ending (zip, pdf, jpg, doc, docx, mp3, mp4 ...) or also text with certain meaning (xml, txt, cpp, jsp, ...). This includes URLs. One advantage is that such data can be identified more precisely than with current approaches. For example, a DV can be defined as bitstream in jpg form with additional comment "passport photo" and numbers with specific measurements for such a photo. Further applications are nesting (combination of DS definitions, see Figure 2). In healthcare grouping of dimensions about anamnesis, therapy and result is useful for decision support [Orthuber 2016 August, Pedersen 2004, Haas 2005]. The global validity of DS definitions facilitates a large data collection and is important for interoperability [Orthuber 2012]. Global validity is even useful for zero dimensional DSs. A zero-dimensional DS contains only one DV with the URL of the DS definition. Such a

DV can e.g. represent a globally well-defined (partial) meaning of a word or a phrase. A multilingual DS definition can contain many translations of this representation. If needed, dimensions with additional numeric data can be added.

DSs can be freely defined on the web. Our implementation [Orthuber 2012] for demonstration purposes shows some exemplary definitions of DSs to certain topics, e.g. ride, my-location, real-estate, car, cupboard, diode, screw, opinion-about-xx, climate-fluctuations, meeting, traffic-accident etc.. On the Internet much more data are available and all DS definitions can be reused in other DS definitions. Therefore, published DS definitions should be freely usable and restrictions e.g. by patents or copyright are counterproductive. Combination of existing definitions is possible also a posteriori by using algebraic equations. For example, a simple proportionality factor could be used to connect definitions with different units. Hereby, worldwide search and comparison over multiple online defined (dimensions of) DSs becomes feasible.

Complex applications are feasible also by integration into programming systems, e.g. by automatic creation of DS definitions from classes of well known programming languages like Java or C++. This can be useful (especially in open source applications) if a certain task can be only solved by automatic exchange of (a limited amount of) specific data between many local systems. If the data are only temporarily used and not published on the Internet, local usage of temporary DS definitions can be an option.

4. DISCUSSION

After considering existing well-known concepts [W3C 2016, Quilitz 2008, Smits 2015, Benson 2012] for machine-readable data on the Internet, we decided for our approach - to achieve the requirements mentioned in Section 1. We list some possible issues and answer to them:

- Without (online) DS definition the data of the DS are no longer readable.
  Answer: This can be solved by periodical reliable data storage (backup) of published DS definitions.
- How is separation of subject, predicate and object (as in RDF) possible?
  This separation is only a special case. By nesting DS definitions (see Figure 2) groups of dimensions can be separated like subject, predicate and object or used for building more complex sentences, or grouped for other purposes (Section 3).
- Data are not self-contained.
  Answer: To get self-contained data, DS definitions may automatically be downloaded together with DVs. This has to be done *only once* for the same DS (kind of data). If definitions are mixed with data (Figure 1), these are repeatedly downloaded also during repetition of the same kind of data. (Remark: currently many data are not self-contained due to not standardized or not locatable or missing definition by context. This approach provides a systematic solution.)
- There will be many DS definitions for the same thing, such that an unclear situation can result.
  Answer: Up to now **(1)** is defined by context (see Section 2) which (e.g. free text) is even for the same thing *very* variable. By **(2)** we get a systematic approach to solve resulting problems. Existing DSs definitions can be searched before and during creation of a new DS definition. New definitions can be connected (e.g. by "sameAs") with or derived from existing definitions (by using algebraic expressions).
- The length of the URL makes **(2)** longer than **(1)**.
  Answer: Usually the length of a URL is much shorter than the length of the information provided at a URL. Furthermore, the URL can be abbreviated, see Definition 1. Additionally, it is an identifier.

We made the experience that even experts initially often underestimate the range and potential of **(2)**. It is important to recall that also today numbers are basic carriers of all information on the Internet and despite of a short URL, the online DS definition can be very sophisticated and optimized for individual applications.

5. CONCLUSION AND FUTURE WORK

The DV data structure "URL (of standardized online definition) plus number sequence" is searchable and enables the combination of maximal competence (online definition by all Internet users) and maximal efficiency (number sequence). Therefore, it has high potential as general carrier of digital information.

As future work, we intend to investigate and optimize binary coding of data and definitions, as well as the process of defining interoperability standards in our setting [Grimson 2000, Hasselbring 2000, Hasselbring 2002, Niemann 2002].

REFERENCES

[Auer 2016] Auer, S., Heath, T., Bizer, C., Berners-Lee, T. 2016. LDOW2016: 9th Workshop on Linked Data on the Web. In Proceedings of the 25th International Conference Companion on World Wide Web (pp. 1039-1040).

[Benson 2012] Benson, T. 2012. *Principles of health interoperability HL7 and SNOMED.* Springer-Verlag.

[Grimson 2000] Grimson, J., Grimson, W., Hasselbring. W. 2000. The SI challenge in health care. *Communications of the ACM* 43 (6), 48-55.

[Guha 2016] Guha, R. V., Brickley, D., Macbeth, S. 2016. Schema.org: Evolution of structured data on the web. *Communications of the ACM*, 59(2), 44-51.

[Guyon 2006] Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. 2006. *Feature Extraction, Foundations and Applications. Studies in Fuzziness and Soft Computing*. Springer, Berlin, Heidelberg.

[Haas 2005] Haas, P. 2005. Medizinische Informationssysteme und Elektronische Krankenakten. Springer-Verlag.

[Hasselbring 2000] Hasselbring, W. 2000, The Role of Standards for Interoperating Information Systems. *In Information technology standards and standardization: A global perspective,* IGI Global. Pp. 116-130.

[Hasselbring 2002] Hasselbring, W. 2002. Web data integration for e-commerce applications. *IEEE Multimedia* 9 (1), 16-25

[Khalili 2013] Khalili, A., Auer, S. 2013. Wysiwym authoring of structured content based on schema. org. *In International Conference on Web Information Systems Engineering.* pp. 425-438. Springer Berlin Heidelberg.

[Kolmogorov 1968] Kolmogorov, A. N. 1968. Three approaches to the quantitative definition of information. *Problems of information transmission,* 1(1), 1-7.

[Kriegel 2010] Kriegel, H.P., Kröger, P., Renz, M., Schubert, M. 2010. Metric spaces in data mining: applications to clustering. SIGSPATIAL Special Volume 2 Issue 2, 36-39.

[Niemann 2002] Niemann, H., Hasselbring, W., Wendt, T., Winter, A., Meierhofer, M. 2002. Kopplungsstrategien für Anwendungssysteme im Krankenhaus, *Wirtschaftsinformatik* 44 (5), 425-434

[Orthuber 2010] Orthuber, W., Dietze, S. 2010. Towards Standardized Vectorial Resource Descriptors on the Web. *In Lecture Notes in Informatics, Service Science - Neue Perspektiven für die Informatik*. Band 2 pp. 453-458.

[Orthuber 2012] Orthuber, W. 2012. Online implementation of vectorial description and numeric search. *http://numericsearch.com*. and http://numericsearch.com/intro.pdf

[Orthuber 2015] Orthuber, W. 2015. How to make quantitative data on the web searchable and interoperable part of the common vocabulary. *Lecture Notes in Informatics (LNI), Gesellschaft für Informatik*, Bonn

[Orthuber 2016] Orthuber, W. 2016. Uniform definition of comparable and searchable information on the web. *arXiv preprint arXiv:1406.1065*. http://arxiv.org/abs/1406.1065.

[Orthuber 2016 August] Orthuber, W. 2016. August. Collection of Medical Original Data with Search Engine for Decision Support. *Studies in health technology and informatics 228,* 257.

[Pedersen 2003] Pedersen, S. und Hasselbring, W. 2003. Structuring and combining domain-specific standards for interoperability in health care. *In: Proceedings of EFIS 2003*, Coventry, UK.

[Pedersen 2004] Pedersen, S. und Hasselbring, W. 2004. Interoperabilität für Informationssysteme im Gesundheitswesen. In *Informatik Forschung und Entwicklung*. pp. 174-188. DOI 10.1007/s00450-004-0146-8.

[Shannon 2001] Shannon, C. E. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1), 3-55.

[Quilitz 2008] Quilitz, B., Leser, U. 2008. Querying distributed RDF data sources with SPARQL. In *European Semantic Web Conference* (pp. 524-538). Springer-Verlag

[Smits 2015] Smits, M. et al, 2015. A comparison of two Detailed Clinical Model representations: FHIR and CDA. *EJBI*, 11(2), 2015.

[W3C 2016] W3C RDF Examples, http://www.w3schools.com/xml/xml_rdf.asp , viewed January 2016.

[Zezula 2005] Zezula, P. et al. 2005. The Metric Space Approach. *Series: Advances in Database Systems*, Vol. 32., Springer, Berlin, Heidelberg.