Worldwide Domain Spaces make quantitative data searchable and prepare these for interoperable exchange

WOLFGANG ORTHUBER, University Medical Center Schleswig-Holstein

Quantitative (numeric) data are often important and decision relevant. By representation as vectors in user defined “Domain Spaces” (DSs) these data become searchable. Every DS represents a nestable metric space with unique “Domain Space Identifier” (DSI), which is the HTTP URL of the DS definition. The elements of a DS are called “Domain Vectors” (DVs). Every DV contains the DSI and a vector which maps or “links” the DV with quantitative data into the DS. This makes the DV accessible to similarity search and further evaluation. The approach is demonstrated in an online database with search engine (http://nummel.com/). The search procedure consists of two systematic steps: 1. Selection of the appropriate DS by conventional word based search within the DSIs or other text parts of the DS definitions. 2. Similarity search of DVs in the selected DS or a part of it. Because DSs can be defined by web users to all their domains of interest, this is a general approach to make quantitative data accessible to similarity search. A web standard for worldwide valid DS definitions and DVs would allow to place quantitative data as identified open data on the web so that they can be searched and exchanged in interoperable way.

Categories and Subject Descriptors: C.2.6 [Internetworking]: Standards; H.3.3 [Information search and retrieval]: Search process – Selection process; E.1 [Data structures]: Distributed data structures; I.5.2 [Pattern recognition]: Design Methodology, Feature evaluation and selection

General Terms: Algorithms, Management, Standardization.

Additional Key Words and Phrases: Metric Space, Domain Space, DS, Domain Vector, DV, Feature Vector, Metric Search, High Resolution Search, Numeric Search, Interoperable quantitative data.

1. INTRODUCTION

Text search is versatile, but also rough, because words usually represent a rough categorization of measurable reality [Black et al. 1963], [Holsgrove et al. 1998]. A finer description is possible using words in combination with numbers (quantitative data) [Nakao et al. 1983]. Example: Compared to the word "cupboard" the term "cupboard, price = 250 Euro, width = 100 cm, height = 200 cm, depth = 50 cm" contains additional information which frequently is decision relevant and therefore important. At this every number represents a quantity [Wikipedia: Quantity 2014] which is a property that can exist as a magnitude or multitude. Quantities can be compared in terms of "more", "less" or "equal", or by assigning a numerical value in terms of a unit of measurement. The above string "width = 100 cm, height = 200 cm, depth = 50 cm" describes quantitative data (the size) of a cupboard. It becomes clear, that searching quantitative data means similarity search of well defined numbers. Because human brain is usually adapted to words of language within their context numbers are seldom in everyday language and the importance of quantitative (numeric) data is usually underestimated. But this is only a subjective impression, from the objective point of view quantitative data are very important as original information. All results of physical measurements are quantitative (numeric) data. These form not only the unbiased basis of our perception, also derived information, e.g. technical data, results of feature extraction (most relevant information about a certain object) [Wikipedia: Feature extraction 2014] can be most efficiently represented in well defined numeric (quantitative) form. Numeric (quantitative) data are often important and decision relevant. But up to now these data are not searchable. This results from the fact that up to now quantitative data (numbers) are on the web usually missing or given in non-uniform way, using heterogeneous units and definitions.
There are approaches to encounter this problem. According to the Linked Data approach [Bizer et al. 2009] of the semantic web [Wikipedia: Semantic Web 2014] OWL [Wikipedia: Web Ontology Language 2014] can be used for numeric definitions. In RDF [Wikipedia: RDF 2014] these definitions can be addressed using HTTP URLs. These approaches are comprehensive and started already before the year 2000. Later more slender syntax proposals for structured data have been introduced. An important initiative is the microdata approach [WHATWG 2014] used in https://schema.org/ [schema.org 2014]. It allows integrating structured data directly into web pages using HTML tags. But all realized approaches together have not been sufficient.

**Despite enormous activities concerning search and the semantic web up to now quantitative data are not searchable on the web.** We try to find the reasons for this clear shortcoming and list some desirable features of a possible solution: An interesting and practicable approach to searchability (and machine readability) of quantitative (numeric) data should have the following features:

a) It is general (usable for all quantitative data).

b) It is slender (avoid unnecessary overhead at the basis).

c) It is hierarchic (to allow high complexity by nesting).

d) Quantitative data are searchable by similarity (user can determine the order of the search result by providing "wished" numbers).

e) Quantitative data are searchable without special knowledge (after providing a keyword or topic search of quantitative data should be possible by filling in a form).

f) Authors of websites can provide quantitative data with the aid of adapted software by filling a form, which is e.g. determined by an online definition (of the used DS as described below).

g) Quantitative data of a web resource are visible after click on it (using adapted web browsers).

[Schema.org 2014] contains a vocabulary for a lot of items and its hierarchic structure implies clearness. But up to now it is no general approach a). For this it is necessary that the users can define the searchable numeric data. Among the above features up to now a) d) e) f) g) are not realized. Condition e) is important for practicability, we cannot expect that users know a priori a large vocabulary or names of searchable variables (e.g. to use these in SPARQL queries).

Here an approach is shown which makes quantitative data searchable and which is designed to fulfill above conditions a) b) c) d) e) f) g). This can be already demonstrated using the online implementation http://nummel.com/ [Orthuber 2012].

An abbreviated description: Quantitative data are represented as "Domain Vectors" (DV) (identified numbers) which are elements of "Domain Spaces" (DS). Every DS represents a metric space [Zezula et al. 2005] with unique "Domain Space Identifier" (DSI), which is the HTTP URL of the DS definition. The DS has a finite count of dimensions (Fig. 1) and it is nestable, i.e. every dimension can represent a number or again a DS (Fig. 2). The DS can be defined by any domain name owner according to a certain domain [Haas 2005] of interest. The DVs are the elements of a DS. Every DV provides identified quantitative data and so well defined similarity relations to other DVs of a DS.
Domain Space (DS):

| Dimension 1 | Dimension 2 | Dimension 3 |
|-------------|-------------|-------------|
|             |             |             |
|             |             |             |
|             |             |             |

Fig. 1. A multidimensional DS. The DS and every of its dimensions have a unique name.

**Dimension of a DS:**

```
   Number

   Dimension

   Domain Space (DS)
```

Fig. 2. A dimension of a DS can represent a numeric value or a DS. Therefore DS definitions can be nested and reused within new DS definitions. (A DS can also represent text.)

The paper explains details of the approach. It is organized as follows: Section 2 recounts the metric space concept and its application, especially in case of partially defined vectors. Section 3 describes the Minkowski distance function and its adaptability. Based on this nesting of distance functions is derived which allows integration and combination of metric space definitions. Section 4 expands on the DS concept and its potential for connection of information. Section 5 addresses reusage of DS definitions within new definitions. Section 6 demonstrates the concept by examples using the online implementation. Section 7 provides details about the necessary content of DS definitions and DVs for development of a web standard. Section 8 discusses several important aspects. The conclusion follows in section 9.

2. **METRIC SPACES**

HTTP-URLs can be used for direct connections on the web via hyperlinks. HTTP-URLs can be also used for identification of (online definitions of) metric spaces which define connections (similarity relations) between all their elements. At this the term "similarity" is represented in well defined way by a nonnegative real number called "distance": The smaller the distance between two elements, the greater is their similarity. If the distance between two elements is zero, the compared quantitative data are identical.

Because metric spaces play a central role in this paper, their definition is repeated here: A metric space [Wikipedia: Metric space 2014] is a set $S$ with a distance function $D$ which represents for every two elements (vectors) $X, Y$ in $S$ the distance between $X$ and $Y$ as a nonnegative real number $d(X, Y)$ with
A distance function which fulfills (1)(2)(3) is called a metric. The distance $D$ quantifies the similarity in $S$. Two elements $X,Y$ of $S$ are the more similar, the smaller the distance $D(X,Y)$ is. Note that $D(X,Y)$ is a real number and therefore one-dimensional.

But the set $S$ can be a multidimensional space. In this paper (as content of a DS) it is a $m$-dimensional feature space which is subset of $\mathbb{R}^m$, and the elements $(X,Y,Z$ in the above formulas) are $m$-dimensional feature vectors which are represented by sequences of real numbers ($x_1 \ldots x_m, y_1 \ldots y_m, z_1 \ldots z_m$). The definition of $S$ does not contain a limitation regarding cardinality, so metric spaces can be very large.

2.1 Induced Metric

If $D(X,Y)$ fulfills (1)(2)(3) for all $X,Y \in S \subseteq \mathbb{R}^m$, then it is possible to compare a subset of all dimensions of $\mathbb{R}^m$ using an induced metric:

Let $J=\{x_{j_1}, x_{j_2}, \ldots, x_{j_m}\} \subseteq \{x_1,x_2,\ldots,x_m\}$ denote a selected subset of dimensions. We define the set

$$\mathbb{R}^m_J = \{(x_1, x_2, \ldots, x_m) \vert (x_j \in \mathbb{R}) \text{ for } x_j \in J \text{ and } ((x_j \in \mathbb{R} \text{ or } x_j \text{ is undefined }) \text{ for } x_j \not\in J \}$$

So in contrast to $\mathbb{R}^m$ in $\mathbb{R}^m_J$ values at dimensions outside $J$ can be undefined.

Let $X_m, Y_m \in \mathbb{R}^m$ with $X_m = (x_1, x_2, \ldots, x_m)$ and $Y_m = (y_1, y_2, \ldots, y_m)$. We define the mapping

$$B_J: \mathbb{R}^m_J \rightarrow S_J \subseteq \mathbb{R}^m \text{ with } B_J(X_m)=(b_1, b_2, \ldots, b_m)$$

where $b_j=x_j$ for $x_j \in J$ and $b_j=0$ for $x_j \not\in J$ (4)

So $B_J$ simply replaces possibly undefined values by 0, therefore $B_J(X_m)$ is well defined. Its value set $S_J$ is a subspace of $\mathbb{R}^m$ [Wikipedia: Linear subspace 2014]. So if $D_m: \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ is a metric on $\mathbb{R}^m$, then the restriction $D_J: S_J \times S_J \rightarrow \mathbb{R}$ with

$$D_J(X_m,Y_m)=D_m(B_J(X_m),B_J(Y_m))$$ (5)

is a metric on $S_J$. It is called induced metric. The subspace $S_J$ forms together with $D_J$ a metric space.

2.2 Comparable vectors

In applications we cannot expect that $X_m$ contains values at all dimensions $x_1, x_2,\ldots, x_m$ of $\mathbb{R}^m$. But we can expect, that $X_m$ contains values at an adaptable subset $J$ of dimensions, i.e. we can assume $X_m \in \mathbb{R}^m_J$ and $B_J(X_m) \in S_J$ according to (4). In this case $X_m$ is called comparable in $S_J$.

Two vectors are called comparable, if they have values at an overlapping set of dimensions. The distance between comparable vectors depends on the selected subset $J$ of compared dimensions. These determine the space $S_J$ in which the distance $D_J$ (5) is calculated.

3. SIMILARITY COMPARISON OF QUANTITATIVE DATA

Similarity comparison is done by calculation of the distance between the vectors (Domain Vectors) which represent the compared data in their space (Domain Space).
3.1 The Distance Function

Generally for similarity comparison every metric (see 2) can be used as distance function. If the triangle inequality (2) is not needed, it is even not necessary that the distance function is a metric [Aggarwal et al. 2001]. The optimal distance function depends on the application, and on the definition of "optimal".

3.2 The Minkowski distance

Because it is not possible to discuss every distance function, we need to make a first preselection (which can be expanded later). Because the dimensionality of Domain Spaces can vary, we need a distance function with adaptable dimensionality. Nesting should be possible (see 3.5). Frequently used distance functions like Euclidean and Manhattan distance should be included as special cases. The Minkowski distance [Wikipedia: Minkowski distance 2014] covers these requirements and is established.

The Minkowski distance \( D(X,Y) \) of order \( k \geq 1 \) between two vectors \( X = (x_1, x_2, \ldots, x_n) \) and \( Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n \) is

\[
D(X,Y) = \left( \sum_{j=1}^{n} |x_j - y_j|^k \right)^{\frac{1}{k}}; \quad (k \geq 1)
\]

(6)

Here we presuppose \( k \geq 1 \), because in this case the Minkowski distance \( D \) fulfills besides (1)(3) also (2) and is a metric. In (6) there is freedom regarding the unit or scale of \( x_i, y_i \), therefore we can multiply every dimension with a constant \( r_j > 0 \) and get the weighted Minkowski distance.

3.3 The weighted Minkowski distance

The (with constants \( r_j > 0 \)) weighted Minkowski distance \( D(X,Y) \) of order \( k \geq 1 \) between two vectors \( X = (x_1, x_2, \ldots, x_n) \) and \( Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n \) is

\[
D(X,Y) = \left( \sum_{j=1}^{n} (r_j |x_j - y_j|^k) \right)^{\frac{1}{k}}
\]

(7)

\( D \) is a metric for \( k \geq 1 \) and we can search given quantitative data \( x_1, x_2, \ldots, x_n \) by inserting the \( x_j \) with \( r_j > 0 \) as coordinates of a searched feature vector \( X_0 \) as described in 2.2 and calculating and sorting the distances to \( X_0 \). At this the factor \( r_j \) determines the relative weight of dimension \( j \). The larger \( r_j \), the more dimension \( j \) influences the distance and derived search results. In case of \( r_j=0 \) the dimension \( x_j \) is ignored. If we calculate (7) within a subspace \( S_J \) (4), it is sufficient to sum up only over dimensions in \( J \) (therefore an index should require disk access only for the searched dimensions, see 6.5.).

3.4 Selection of the exponent \( k \)

Due to their importance we explicitly write weighted Minkowski distances with certain \( k \). Special cases are the weighted Manhattan distance with \( k=1 \), the weighted Euclidean distance with \( k=2 \) and the weighted Maximum distance with \( k \rightarrow \infty \);
D_1(X,Y) = \sum_{j=1}^{n} |x_j - y_j| \quad \text{Manhattan dist.} \quad (8)

D_2(X,Y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2} \quad \text{Euclidean dist.} \quad (9)

D_\infty(X,Y) = \max_{j=1}^{n} |x_j - y_j| \quad \text{Maximum dist.} \quad (10)

Among these D_1 provides the best contrast between different vectors [Aggarwal et al. 2001]. D_2 can be used e.g. for calculating distances with direct geometrical meaning. D_\infty can be e.g. used for limiting the range of dimensions.

3.5 Nested distance functions

It is efficient to use definitions of established metric spaces within other new definitions. For this combination of distance functions to one nested distance function is required. We now show, that analogously to (7) instead of differences |x_j - y_j| also metrics can be nested to a superordinated metric:

**Proposition.** Let V denote a vector space whose dimensions are a concatenation of the dimensions of vector spaces V_1, V_2,..., V_n and X,Y,Z \in V. We presuppose w_j > 0, k \geq 1 and that for j \in \{1,2,...,n\} D_j is a metric on V_j and X_j,Y_j,Z_j \in V_j. Then the following nested distance function is a metric:

\[
DC(X,Y) = \left( \sum_{j=1}^{n} (w_j D_j(X_j,Y_j))^{\frac{1}{k}} \right)^{\frac{k}{1}}
\]  

**Proof.** We have to show (1)(2)(3) of section 2. Due to nonnegativity of D_j(X_j,Y_j) from DC(X,Y)=0 follows D_j(X_j,Y_j)=0 and therefore X_j=Y_j for j \in \{1,2,...,n\} and X=Y. Reversely from X=Y follows X_j=Y_j and D_j(X_j,Y_j)=0 for j \in \{1,2,...,n\} and so DC(X,Y)=0, therefore (1) is true. Symmetry (3) of DC(X,Y) follows from symmetry of D_j(X_j,Y_j).

We now prove (2). For k\geq1 and u_j, v_j \in R we have due to the Minkowski inequality [MATH41002 2014] :

\[
\left( \sum_{j=1}^{n} |u_j + v_j|^k \right)^{\frac{1}{k}} \leq \left( \sum_{j=1}^{n} |u_j|^k \right)^{\frac{1}{k}} + \left( \sum_{j=1}^{n} |v_j|^k \right)^{\frac{1}{k}}
\]

we set u_j=w_jD_j(X_j,Y_j)\geq0 and v_j=w_jD_j(Y_j,Z_j)\geq0 and get

\[
\left( \sum_{j=1}^{n} (w_j(D_j(X_j,Y_j) + D_j(Y_j,Z_j)))^{\frac{1}{k}} \right)^{\frac{k}{1}} \leq
\]

\[
\left( \sum_{j=1}^{n} (w_jD_j(X_j,Y_j))^{\frac{1}{k}} \right)^{\frac{k}{1}} + \left( \sum_{j=1}^{n} (w_jD_j(Y_j,Z_j))^{\frac{1}{k}} \right)^{\frac{k}{1}}
\]

Due to D_j(X_j,Z_j) \leq D_j(X_j,Y_j)+D_j(Y_j,Z_j) (triangle inequality)
\[ \left( \sum_{j=1}^{n} (w_j D_j(X_j, Z_j))^k \right)^{\frac{1}{k}} \leq \left( \sum_{j=1}^{n} (w_j D_j(X_j, Y_j))^k \right)^{\frac{1}{k}} + \left( \sum_{j=1}^{n} (w_j D_j(Y_j, Z_j))^k \right)^{\frac{1}{k}} \]

this is just the triangle inequality for DC in (11):

\[ DC(X,Z) \leq DC(X,Y) + DC(Y,Z) \]

So we have shown also (2) and so proven that (11) is a metric. (11) has a similar structure like the Minkowski metric (7), only the \(|x_i-y_i|\) are replaced by \(D_j(X_j,Y_j)\). Because the \(|x_i-y_i|\) are special cases of a metric, (11) is a generalization of (7). So in a (weighted) Minkowski distance function (7) for every \(j\) the absolute difference \(|x_j-y_j|\) can be replaced by a metric \(D_j(X_j,Y_j)\). The result (11) remains a metric.

Important special cases of (11) are the following nested distance functions

\[ DC_1(X,Y) = \sum_{j=1}^{n} (w_j D_j(X_j, Y_j))^k \quad \text{nested Manh. dist.} \quad (12) \]

\[ DC_2(X,Y) = \left( \sum_{j=1}^{n} (w_j D_j(X_j, Y_j))^k \right)^{\frac{1}{k}} \quad \text{nested Eucl. dist.} \quad (13) \]

\[ DC_{\infty}(X,Y) = \max_{j=1}^{n} (w_j D_j(X_j, Y_j)) \quad \text{nested Max. dist.} \quad (14) \]

### 3.6 Estimation of the weights

We assume that the functions \(D_j(X,Y)\) in (11) are Minkowski distances of the form (7). The weights \(r_j\) and \(w_j\) are free parameters and we need an estimation. If there is no further information available, the initial values are \(r_j=1\) in (7) and \(w_j=1\) in (11). The \(w_j\) can be left unchanged, because all necessary modifications can be done by adjusting the \(r_j\) in (7). They are important for multidimensional similarity search, because they determine the relative weight of dimensions for calculation of the overall distance.

If there are no individual preferences, at least the influence of a dimension's unit should be eliminated. The smaller the unit of a dimension, the larger is its numerical variation. It is possible to take into consideration the numerical variation of a dimension \(j\) by setting \(r_j = 1/s_j\), where \(s_j\) can be e.g. the standard deviation (or a difference between two given percentiles) of dimension \(j\). Calculation of the \(r_j\) can be done also retroactively.

We now come to the application of the above theoretical background.

### 4. DOMAIN SPACES

The metric space concept can be used systematically on the web to make multidimensional quantitative data searchable. When defining a metric space about a certain domain of interest, regularly it is desirable to include all possibly interesting dimensions. This leads to a domain specific metric space \(DS_m\) with maximal dimensionality \(m\). Because we cannot expect that users always provide values for all \(m\) dimensions we introduce the following convention:
A Domain Space (short "DS") is defined by the domain specific metric space $DS_m$ with maximal dimensionality $m$. The elements of a DS are called Domain Vectors. Every Domain Vector (short "DV") has a feature vector which has values at all or a subset of all $m$ dimensions. Partially defined feature vectors can be mapped by (4) into subspaces of $DS_m$. Subsequently we abbreviate, where clear, the term "$DS_m$ of DS" by "DS". For example we call "subspaces of $DS_m$ of the DS" simply "subspaces of the DS". Analogously we abbreviate, where clear, "feature vector of the DV" simply by "DV".

DVs are called "comparable" if their feature vectors are comparable (see 2.2), i.e. if they have values at an overlapping set of dimensions. It is possible to select a subset $J$ of these dimensions for similarity comparison and to calculate distances $D_J$ (5) between all DVs which are comparable in $S_J$ (4). The smaller the distance is, the greater is the similarity in $S_J$.

The subspaces $S_J$ of a DS form a set of domain specific metric spaces [Kriegel et al. 2010], [Wikipedia: Vector space model 2014] or conceptual spaces [Gaerdenfors 2000; 2004].

4.1 The addressable web extended by Domain Spaces

Every DS has a HTTP URL [Berners-Lee et al. 1998]. It is the HTTP URL of its definition on the web and it is the worldwide unique Domain Space Identifier (DSI). The DS consists of its definition and of its elements which are the DVs. Every DV contains at least the following information:

- Its "Vector Location" (short "VL"). This is the HTTP URL of the DV on the web.
- The DSI (HTTP URL of the DS definition)
- The numeric representation of its feature vector, given by (hyperlinks to) identified numeric values for all or a part $x_{j1}, x_{j2}, ..., x_{jn}$ of all dimensions $x_1 ... x_m$ of the DS (see 2.1).

Domain Vectors (DVs) can be regarded as links (or mappings) into Domain Spaces. Fig. 3 compares the most important characteristics of Hyperlinks and Domain Vectors. It shows that the destination of a hyperlink can be every HTTP URL of the web, while the destination of a DV can be only the HTTP URL of a DS (which represents all contained DVs). The DS definition is an intermediate station of bidirectional links (Fig. 4). DVs are especially useful when multiple resources with quantitative features of the same domain of interest should be connected by "similarity". The similarity depends on the (by the user) selected subset $J$ of compared dimensions. These determine the subspace $S_J$ in which comparison is done, and these determine the distance function $D_J$ in (5). The distance is defined to all DVs which have values at least at the selected dimensions. The similarity (of the compared dimensions) is the greater, the smaller the distance is.
|                                | Hyperlink                                         | Domain Vector (DV)                                                                 |
|--------------------------------|--------------------------------------------------|-----------------------------------------------------------------------------------|
| **Usual location**             | resource with HTTP URL                           | resource with HTTP URL                                                            |
| **Connections from location**  | unidirectional to a HTTP URL on the web           | via HTTP URL of the DS bidirectional similarity relations to all (locations of) comparable DVs |
| **Connections back to location** |                                                 | given from (the locations of) all comparable DVs of the same DS                   |
| **Essential information (besides location)** | HTTP URL of destination | HTTP URL of DS and values $x_i$ of a subset of dimensions of the vector          |
| **Distances**                  |                                                 | available to all (locations of) comparable DVs of this DS and included DSs, see section 5 |
| **Additional purpose**         |                                                 | searchable quantitative description                                               |

Fig. 3. Comparison Hyperlink / Domain Vector. A Domain Vector can be regarded as a link into a Domain Space.

![Diagram](image)

Fig. 4. Bidirectional weighted connections (similarity relations) defined by a group of DVs to a (HTTP URL of a) DS. Shown are connections within a two-dimensional subspace. They connect the (locations of the) DVs. The length of the blue lines illustrates the distance. The smaller the distance, the greater is the similarity.
As example we assume that the user has selected 2 dimensions for comparison in a DS. We can use their values as coordinates and represent every DV as point in the 2D plain. Fig. 4 shows an example with 7 comparable DVs in the subspace determined by the selected dimensions. If we assume Euclidean distance (9), the lengths of the blue lines can be used to represent the distances between the DVs.

Fig. 5 shows the connections defined by a group of hyperlinks to a HTTP URL. Hyperlinks generate no implicit connections, so k hyperlinks generate also k (unidirectional) connections. When this is wished, the hyperlink is appropriate. But if a group of resources (with interesting quantitative descriptions) of the same domain should be (described and) connected, DVs are efficient. If we assume that there are k DVs in a m-dimensional DS, each with values at all m dimensions, then there are $2^m$ different subsets of dimensions which each define $k(k-1)/2$ distances (weighted connections). These can be evaluated, e.g. for similarity search. The quantitative data are also available for further calculations, e.g. statistics.

4.2 Similarity search in a DS
In 2.2 is described how to calculate distances between DVs after selection of the dimensions for comparison. Analogously it is possible to select a set $J$ of dimensions for similarity search. If $S_J$ is the subspace (4) which is determined by these dimensions, distances $D_J$ (5) can be calculated to all DVs which are comparable in $S_J$. If there are no further restrictions, these DVs form the search result. The smaller the distance of a DV, the higher is its rank in the search result. Further restrictions of the search result are possible, e.g. ranges (minima and maxima) of independently selected dimensions.

4.3 Storing relevant dimensions
When storing information about the world on the web, usually most measurable quantitative features are omitted. Macroscopic world has so many quantitative features, that it is necessary to select in dependence of the domain few most interesting (decision relevant) features for further information processing. These need not be "direct" physical measurement results, they can be completely derived, e.g. a result after feature extraction and/or prepared for usage with special software. The main issue is that they are interesting (preferably decision relevant) for future readers or users. So DSs typically contain those dimensions which describe the part
of the world, which is interesting in the chosen domain. In the course of time new features can become interesting, other features can become deprecated. The distance function is alterable, dimensions can become unused, new dimensions can be added. This can lead to high dimensional DSs. But this does not mean that search becomes high dimensional.

4.4 Searching in a subset of dimensions

Due to the curse of dimensionality [Aggarwal et al. 2001] high dimensional similarity search in not clustered data tends to become inefficient. Moreover DSs often have dimensions which describe incommensurable data, e.g. data with incommensurable units. In this case the relative weight of a dimension depends on the intention of the user at search time and it is not possible to anticipate it as described in 3.6. Therefore it is recommendable to select only a small subset J of dimensions with meaningful common distance function for similarity search (2.2) to get a well interpretable ranking of the search result. If only one dimension is included into similarity search, the search result is simply ordered by the absolute difference of the searched dimension. The smaller it is, the higher is the ranking of a DV in the search result.

Additionally the search result can be restricted by determining minima and maxima of dimensions. These dimensions are also called "searched", together with the dimensions in J.

Search is done over DVs which contain values at all searched dimensions. This implies, that the probability to be found is the greater, the more numerical values (dimensions) are given in a DV.

5. COMBINING DSS

5.1 Grouping DSs

A DS group is a set which contains as elements DSs and/or DS groups. Grouping of DSs or DS groups can be for example useful if they have a common topic. So a DS group with topic "clothes" may contain the DSs or DS groups "trousers", "shirts", "coats" etc.

Grouping of definitions can be appropriate to (simply) build a thematic structure of DSs and/or DS groups without involvement of dimension definitions.

5.2 Nesting DS definitions

Every DS combines an expandable set of named dimensions, and every dimension can represent:

- a number (with selectable precision, date included, accessible to similarity search and min max conditions)
- a DS (DVs are also accessible to similarity search and min max conditions, an additional condition can be a maximal distance of its DVs. Because a DS can represent text, a dimension can represent text.)

So a dimension can represent not only a numeric value, it can also represent another DS. It can be efficient to use DS definitions within other new definitions. For example a DS with the DSI "hemogram-1" can be used within many DSs which describe medical findings. It is not necessary to reinvent it. We can use its metric in a
nested distance function (11). This is important due to the following reasons:

- Established DS-definitions can be reused and included as "Sub-DS definitions" into new (higher dimensional) DS-definitions with weighted Minkowski metric (7), in which coordinate differences \(|x_j - y_j|\) are replaced by the distances \(D_j(X_j, Y_j)\) of the Sub-DSs. It is sufficient to use their HTTP URLs as reference. So also updates are automatically forwarded.

- The included DS can have any metric, also a non-Minkowski metric, or again a nested metric (11).

There are well defined similarity relations (Fig. 4) between all DVs which have a common HTTP URL of (their DS or) an included DS. One application of nested DS definitions is realization of ontology based structures [Wikipedia: Domain Ontologies 2014] in quantitative data. The nesting level of a DS is the maximal count of nested layers. Fig. 6 shows an example of a nested DS definition with 2 layers (nesting level 2).

Nesting of DS definitions can quickly lead to high dimensional DSs. But according to 4.4 it is recommendable to select only a small subset of dimensions for search.

![Fig. 6. Exemplary nested DS definition in the structure of Fig. 8 (nesting level is 2).](image)

6. LOCAL IMPLEMENTATION

Up to now there is no web standard for worldwide valid DSs on the web. Nevertheless it is possible to implement the search principle locally. We have done this in the online implementation [http://nummel.com/](http://nummel.com/). It contains a local database with DSs definitions and to every DS definition a local database with the DVs of this DS.

6.1 Keycomments

For description of DSs, DS dimensions and DVs we used "keycomments" which are structured comments: Every keycomment starts with an ordered list of one or more (optionally linked) keywords, followed by a comment which is a string.

| Keyword | Link                                      |
|---------|-------------------------------------------|
| kw0     | ![http://www.optional.hyperlink0.com](http://www.optional.hyperlink0.com) |
| kw1     | ![http://www.optional.hyperlink1.com](http://www.optional.hyperlink1.com) |
| kw2     | ![http://www.optional.hyperlink2.com](http://www.optional.hyperlink2.com) |

Comment: A summary comment

![Fig. 7. Input mask of a keycomment with 3 keywords, in which the second is also a hyperlink](image)
The advantage of the structured keycomment is that the meaning of the ordered keywords or hyperlinks can be defined a posteriori, depending on application. So for example the first keyword "kw0" can be a unique identifier, the second keyword can be a unit etc..

6.2 Implemented DS and DV structure

Every DS has a distance function of the form (12) or (13) or (14) where every $D_j$ has the form (8) or (9) or (10) or GPS distance (see e.g. [Movable Type Scripts 2014]). So there are exactly 2 nested layers (nesting level 2): Every DV has a vector which contains one or several "subvectors" (DVs of "Sub-DSs") with numeric dimensions. Fig. 8 shows an example of a DS with 2 subvectors. Fig. 9 shows the definition of a subvector, Fig. 10 the definition of a dimension.

**Definition of DS 1006 (Cupboard)**

| < | <<< | < | > | >> >| 0..1 |
|---|---|---|---|---|---|
| 0 | Finances | 0 | Price | Euro |
| 1 | Size | 0 | Width | cm |
|   |   | 1 | Depth | cm |
|   |   | 2 | Height | cm |

Keyword: Cupboard Link: http://en.wikipedia.org/wiki/cupboard
Keyword: Schrank Link: http://de.wikipedia.org/wiki/schrank
Comment: 

This is: @ draft @ ok @ deprecated

conn: @ Manhattan @ Euclidean @ Maximum

Fig. 8. Exemplary definition of a DS. The DSI is "Cupboard" (in a web standard it would be a HTTP URL). The first column shows the internal index of the subvectors and the second column the internal index of the numeric dimensions (2 layers, see Fig. 6).
It is possible to define minimum, maximum and weight (r; default in (7)) of every dimension. Also the representation can be adapted to the user's needs. The internal representation is a 64 bit double, the external representation can be:

**List**: Appropriate if the dimension represents the position in a list of items, e.g. a selection between two possibilities like "yes" and "no". By default the items are internally represented by integers (0,1,2...) in this implementation. For a later implementation the following option is possible: Every item can be considered as interval, so that all items represent an ordered list of labeled intervals (a partition) of the set of real numbers \( \mathbb{R} \). The first interval may have no lower border, else by default the lower border of an interval is the upper border of the previous interval, the last interval may have no upper border. If these intervals are defined for a dimension and the user selects the input option "intervals", the ordered list of interval names is opened and the user can select a name. If the name is given
as sort criterion, the mean of the interval (or the border, if only one border exists) is internally inserted into the similarity field. If the name is given as condition, the lower bound of the interval is (if existing) internally inserted into the min field and the upper bound (if existing) into the max field.

**tux**: Appropriate if the dimension represents a short alphanumeric text which can contain up to 8 lowercase letters a..z or digits 0..9. It is introduced because it can be easily remembered. Similarity search is defined so that all DVs are found whose initial letters are at this dimension identical to the searched tux. So the initial letters should be most significant.

Compared to tux a list of intervals or the following ordered representations have the advantage that they allow similarity comparisons and further algebraic evaluation:

- **date**: for representation dates in variable accuracy (highest significant numbers first, e.g. yyyy-mm-dd hh:mm:ss)
- **floating point**: floating point number in variable accuracy, e.g. for measurement results
- **integer**: integer number, e.g. for counts

An interesting but in the current version not implemented possibility is the definition of dimensions as computational results of other dimensions.

Fig. 8 shows that Manhattan metric is chosen to connect the distances of 2 subvectors according to (12). In every subvector (e.g. Fig. 9) again Manhattan metric is chosen (to connect dimensions) according to (8) with all weights \( r_j = 1, w_j = 1 \) (1 is default value for weights). Therefore the distance function of DS\(_m\) (with maximal dimensionality) is

\[
D(X,Y) = |x_{\text{Price}} - y_{\text{Price}}| + |x_{\text{Width}} - y_{\text{Width}}| + |x_{\text{Depth}} - y_{\text{Depth}}| + |x_{\text{Height}} - y_{\text{Height}}| \quad (15)
\]

According to 4.4 a subset \( J \) of these dimensions can be searched and so used for sorting the search result. Then the sum includes only the searched dimensions. If for example only a value of dimension "Price" is given for similarity search, then the distance function reduces to

\[
D(X,Y) = |x_{\text{Price}} - y_{\text{Price}}| \quad (16)
\]

This distance function is used in the search example (Fig. 13) of section 6.3.

### 6.3 Search

DV Search consists of 2 systematic steps:

1 of 2: In the first step the appropriate DS is selected. This can be done by clicking on its index number directly in the list of all DSs (Fig. 11) or after word based search within the DSIs (Fig. 12) which are here (in the local database) the first keywords (kw0) of the space definitions. After selection of the appropriate DS its specific search mask appears (Fig. 13).

2 of 2: The second step is metric similarity search in the selected DS or a part of it. All data for this are provided in the search mask of the DS.

Fig. 13 shows an example of a search mask with exemplary input. It shows the search of the cheapest cupboards (those nearest to price=0). Two checkboxes in column "g" are checked to signal the wish for graphical and statistical output of Price and Width. Fig. 14 shows the resulting graph over the checked dimensions together with the search result.
Fig. 11: Excerpt of the start screen. The first column "i7" shows index numbers of the DSs. Clicking e.g. on "1006" opens the search mask (Fig. 13) of the DS with DSI "Cupboard". The second column "s" shows the search count, the third column "r" the count of resources in a DS. Clicking on "o" in the next column shows the owner of a DS. Then follows the first obligatory keyword kw0 which here is the DSI (blue if HTTP-link), after "|" further optional keywords, after "||" a comment. After clicking on "kw0" text search is done over the DSIs of the DSs, the result is shown in Fig. 12.

search KW0

|< |< |< |< |< |< |> |> |> |> |0..0 |
|---|
|17 |
|1006 |
|0 |
|o Cupboard | Schrank |

"cup" searched in 54 first keywords along i7 1 found which begins with the searched string

Fig. 12: Text search result after entering the first letters of the DSI in Fig. 11 and clicking on "kw0" (Keyword 0).

i7=1006, o | 13-02-09 Cupboard | Schrank

DL search in DS 1006 (Cupboard)

|< |< |< |< |< |< |> |> |> |> |0..1 | search |
|---|---|---|---|---|---|---|---|---|---|---|---|
|0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |0 |
|1 |
|0 |1 |2 |3 |4 |
|Price | Euro |
|Size |
|Width | cm |
|Depth | cm |
|Height | cm |

Fig. 13: Similarity search mask. It appears after selection of the DS (click on 1006 in Fig. 11 or Fig. 9). Similarity search of Price "0" is selected with graphic output of price in dependence of width.
Fig. 14. Search result with preceding graphic output. Click on the index in the 1. (left) column opens the DV (resource). The 2. column "d" shows the distance, here \(d = |\text{price} - 0|\), the 3. column "a" shows the access count, click on "o" in the 4. column shows data of the owner. Then follows the first keyword with optional link for description of the resource, after "\|\|" an optional comment, after "|" the quantitative data in order of Fig. 13 (Price, Width, Depth, Height).

Together with the graph in Fig. 14 statistical data of the checked dimensions are
shown (average, standard deviation, minimum, maximum of price and width). This allows to check dependencies.

The min and max fields (Fig. 13) can be used to restrict the range of certain dimensions in the search result. (The checkboxes "o" and "w" can be used to restrict the search result to offered or wanted resources, the "pcnt" field allows to enter the maximal count of shown resources in the search result, pcnt=1000 is default and maximum.)

So the search result can be restricted to a certain part of all DVs. Using the checkboxes of column g allows to get statistical data of selected dimensions in the search result.

According to Fig. 13 the searched value of dimension "Price" is 0. This is inserted into (16) and leads to the distance \( d = D(X,Y) = |Price - 0| = |Price| \). Therefore in Fig. 14 the distances in the second column "d" are equivalent to the Price of the resources which is the first number after "|".

Fig. 14 shows most important data of the search result in compressed form. Clicking on the index of a DV (left column) shows its data in more detailed form in a new window (Fig. 15).

6.4 Search results in case of equally distributed pseudo random numbers

To illustrate the effect of the distance function (7) we generated a high dimensional Domain Space with 1500001 DVs whose dimensions have been filled with equally distributed pseudo random numbers between 0 and 10.

Fig. 16 shows the search result of (7) with default \( r_j = 1 \) in case of \( k=2 \) (Euclidean distance (9)), Fig. 17 shows it in case of \( k=1 \) (Manhattan distance (8)).
Fig. 16. Elliptic shape of output after searching within 1500001 DVs the 1000 nearest around point \((x,y)=(7,2)\) in case of Euclidean distance (9).

Fig. 17: Shape of output after searching within 1500001 DVs the 1000 nearest around point \((x,y)=(7,2)\) in case of Manhattan distance (8).

While both graphs are identical near the center, in case of large deviation of one dimension (near to the border of the graph) Manhattan distance restricts the other dimension sharper than Euclidean distance. Also research [Aggarwal et al. 2001] shows that Manhattan distance provides more contrast. Therefore we selected Manhattan distance (8) as default metric. Independently of this Euclidean metric can be recommendable e.g. if the dimensions represent Cartesian coordinates and the distance should have a geometric meaning.
6.5 Synchronized index

In the implementation we used for every DS a synchronized index. Subsequently we give a short description of it in a version which can be used also for DV-groups (see 7.3) on the web.

The dimensionality of a DS (or even all DSs on the web) can become large, nevertheless meaningful similarity search includes only a few dimensions. There are $2^n$ possibilities to select a subset of dimensions within a n-dimensional space. This selection is done before search, not before index calculation. So the synchronized index does not anticipate a certain combination of dimensions, i.e. every dimension has an own dimension database (three in Fig. 18) in the index which is optimized for quick access. During index creation the original DVs are scanned on the web and the count of scanned DVs (resp. DV-groups, if DVs are grouped, see 7.3) is increasing. We will call this increasing count "c". As long as dimensions belong to the same DV, the count c is constant. After a DV is fully scanned, c is stored in a separate database (A) together with the HTTP-URL of the DV and all further information which should be available in the search result. All (short) data records of the dimension databases get c as identifier ("c" in Fig. 18) of the DV and the numerical value ("x" in Fig. 18) of the dimension. Data records with the same c belong to the same (multidimensional) vector of a DV. After all data of a DV are stored in the index, c is incremented by 1. We can say that the increasing c "synchronizes" the dimension databases (B). Therefore we call this ((A) and (B)) a synchronized index.

So in case of multidimensional similarity search only the (few) dimension databases of the searched dimensions are serially scanned along increasing identifier c (linear performance, or better if c makes large jumps). As soon as an identifier c is found for which all searched dimension databases contain values (in Fig. 18 for $c \in \{9, 21, 29, 42\}$), it is checked whether the values fulfill the requirements (especially the min-max conditions). If yes, the distance is calculated ((12) can be used for combining multiple DSs). It is appended to the preliminary search result together with the dimension values and the most important information (A) about the DV.

This can be used for scanning all combinations of synchronized dimensions in linear performance or better. As soon as the searched dimensions are fully scanned, the preliminary search result is complete. After sorting it according to distance (the smallest distance first) we get the final search result.

Additionally special indices can be calculated for certain (combinations of) dimensions. If for example a priori is known that frequently only a certain dimension is searched, it can be efficient to calculate additionally an index which is sorted along this dimension so that similarity search of this dimension is possible in logarithmic time (e.g. using binary search). Calculation of additional indices for selected (combinations of) dimensions, however, needs additional computational cost. Future research can show, under which conditions it is efficient to calculate additionally indices which are a priori optimized (e.g. by adapted sorting) for certain (combinations of) searched dimensions.
Fig. 18. Synchronized index of a DS with 3 dimension databases. Every (short) database record contains 2 numbers: "c" represents during index creation the current count of scanned DVs, "x" represents the numeric value of this dimension in this DV. The blue lines connect records with values from the same DV, 4 DVs are found which contain values at all 3 dimensions. Because \( c \) is increasing, the connecting blue lines cannot cross and it is possible to scan all dimensions in one pass without redundancy.

### 6.6 Index performance

The synchronized index was realized for our local database and we measured the search time. As expected the time of multidimensional search depends not on the total dimensionality of the DS but on the dimensionality of the search (the count of simultaneously sought dimensions). Fig. 19 displays the similarity search time (for searching the most similar 1000 DVs, sorted along distance) within a 260 dimensional DS with 1500001 DVs (see 6.4). For each dimensionality 1..10 the average search time of 20 searches is shown.

Fig. 19. Average search time in milliseconds (vertical axis) in dependence of the dimensionality of the search (horizontal axis) within a DS with 1500001 DVs.
The time for every pass of index calculation for this DS was between 23 and 24 minutes. The implementation was programmed using java jdk-7u4-linux-x64.rpm, apache-tomcat-7.0.27.zip, performance was measured on a dedicated server with Intel Core i7 3930K (3,20 GHz), 64 GB RAM, 256 GB SSD, with Linux Cent OS 6.3.

7. TOWARDS STANDARDIZED DSS AND DVS

First recommendations to a web standard for DSs and DVs can be given already in this paper. DS definitions can be expanded a posteriori. Already defined content cannot be changed, but commented. To meet these requirements, keycomment-pairs can be used. Two keycomments (Fig. 7) are joined together in a keycomment-pair. The first keycomment should already initially contain a precise and complete definition as text, the second keycomment can provide structured changeable information. A keycomment-pair has 3 possible states: "draft", "ok", and "deprecated". Once the initial (default) state "draft" is left, the first keycomment is fixed. In an appropriate environment this should ensure, that already defined essential content of a DS definition is stable.

We now summarize the content of DS definitions and DVs:

7.1 Content of a DS definition

- String: Domain Space Identifier (DSI): The DSI is given implicitly by the HTTP URL of the DS definition. The DSI must be stable
- keycomment-pair
- String: Information about the distance function
  (e.g. "M" for usage of (11), followed by the exponent k)

Sequence of dimension definitions:

7.1.1 Content of a dimension definition

- String: Dimension Identifier (DI):
  The DI is a string which is distinct to the DIs of other dimensions of the same DS. The dimension definition can be regarded as definition of a one-dimensional DS. The DSI of this DS can be formed by the HTTP URL of the DS definition together with the DI.
- keycomment-pair
- floating point number: weight ($w_j$ in (11) if nested, else $r_j$ in (7))
- String for content description:
  - if nested (according to section 5): DSI (HTTP URL) of the integrated DS definition. A dimension definition can be integrated using the DSI plus the Dimension Identifier (DI).
  - if not nested: String which describes the content e.g.
    - "integer", "floating-point-precision-8",
    - "date-YYYY-MM-DD", "tux", list etc. (see section 6.2)
    - borders and names of intervals, if given (option "list" in section 6.2)
    - optionally an expression which describes this dimension as algebraic result of other dimensions.
Apart from the above content a DS definition can contain additional information, e.g. default preferences for representation and formatting of DVs in the Web browser.

7.2 Content of a DV
- String: HTTP URL of the Domain Space definition (DSI)
- keycomment (optional)

content of given dimensions as semicolon separated list of numbers (this short form is possible if the order of DIs in the DS definition is used) or a sequence of:

7.2.1 Content of a dimension
- String: Dimension Identifier (DI)
- content:
  - if dimension definition nested: integrated DV or HTTP URL of integrated DV
  - if dimension definition not nested: Number, depending on definition

Additional formatting information can overwrite the (in the DS definition given) default preferences for representation of the DV in the Web browser.

7.3 DV-groups
DV-s can be grouped together, so that one group describes the same resource. Later it is possible to combine the dimensions of these DV-groups (i.e. dimensions of different DSs) also for search if the search engine uses a synchronized index (see 6.5) for all dimensions. This index is not restricted to one DS.

7.4 Adaptable syntax: DVs as identified numbers
The concept can be realized using variable syntax. For example OWL [Wikipedia: Web Ontology Language 2014] and RDF [Wikipedia: RDF 2014] can be used to realize DS definitions and DVs. In the long run conciseness is relevant for feasibility and efficiency. Therefore we recommend to start with concise and minimal syntax with minimal overhead. Minimal precondition for DVs is that the numbers are identified, using the DSI of the DS and the DI of the dimension (see 7.1.1). e.g.: http://example.org/DS-Cupboard.htm#Size/Width=100

Because at this the dimensions of the same DV should be grouped together, it is sufficient to specify the DSI of the DV (http://example.org/DS-Cupboard.htm) only once at the beginning of the group. If the default order of dimensions of a DSs is inherited, it is even sufficient to use one special character as separation between dimension values in DVs (e.g. use a sequence of semicolon separated numbers).

8. DISCUSSION
Similarity search of quantitative data in metric spaces is well investigated. Due to the potential of this technique it is desirable to enhance the web by metric spaces. A DS represents a nestable metric space with unique identifier (HTTP-URL) on the web. Because DSs can be defined by web users according to their domain of interest, and because of their applicability in Domain Ontologies (see [Haas 2005]) we called these "Domain Spaces". The following subsections discuss some important aspects (partially derived from [Orthuber 2013]) of the concept:
8.1 Resolution and precision of DV based description and search

Due to its basal relevance the following fact is given first:

For a word which is more than grammatically different from other words we need an extra definition. But for all (different) DVs which belong to the same DS we need only one definition - the definition of the DS. This definition is usually also more precise than the definition of a word, and internationally valid. DSs can be created by web users according to their domains of interest. There can be much more different DVs (even in only one DS) than there are different words. Therefore DV based description and search has higher range and resolution than word based description and search.

As described in section 3, DV based description additionally provides information about similarity relations of resources.

8.2 Precise information exchange on the web

One important motivation for this approach is improvement of the availability (this means also searchability) of precise information on the web. For description of reality usually words of language are used, but they categorize the original quantitative features of reality. At this often interesting information gets lost. Even if someone wants to provide precise information and explicitly adds quantitative information, e.g. as numbers combined with text, up to now usually only the words of the text are searchable, not the numeric quantitative information. In many cases just this precise quantitative information is interesting for the readers. So it is reasonable to combine numbers with unique identifiers, that they are machine readable and searchable. This is done in a DV: It contains the DSI (HTTP URL of the DS), which together with the DI (Dimensions Identifier, see 7.2.1) uniquely identifies the numeric value of every given dimension.

The DS definition and its identifiers are also a guide for providers (writers) of numeric information. Often important numeric data are missing on the web, because the writer does not know well the expectations and interests of the reader. The DS definition shows the quantitative data, which are in a certain domain interesting for the readers. So it serves in this domain as standardized and expandable interface for exchange of precise numeric information between writer and reader. Later it is possible to provide to dimensions frequencies of usage.

8.3 Storage of DS definitions and DVs on the web

The implementation shows that realization of user defined DSs and DVs is also possible in a local online database with DV search engine. The Vector Location (VL, see 4.1) cannot be used in a local database, but the DVs can get a hyperlink to a location on the web, which they describe. So it would be technically feasible to realize DV search also in a local database.

This is better than nothing, especially if certain data cannot be published openly on the web, e.g. patient records with detailed quantitative data about medical findings, treatment and treatment outcome. Even if these data are not published, DV search and anonymous statistical output of the results (e.g. average values of dimensions) can be used for decision support.

But it requires considerable effort that a local (proprietary) database is used internationally. There is relevant probability that several competing databases arise, so that DV search is no more internationally complete but restricted to one of many proprietary databases (information silos). Therefore we recommend introduction of a
web standard for worldwide valid DSs, so that their elements (DV$s$) can be placed as open data on the web. Besides making quantitative data searchable, DV$s$ (identified numbers, see 7.4) can be also used for interoperable exchange of quantitative data in machine readable form.

8.4 Reliability of DS definitions
The numeric data in a DV are only meaningful together with their definition in the DS. Therefore every dimension definition of a DS must be stable. When in the course of time a dimension turns out to be no more recommendable for new data, it can be marked as "deprecated", and an explanation can be added. For this purpose we recommended usage of a keycomment pair (see section 7) with a fixed and a changeable part. A checksum can be calculated from every fixed part of a dimension definition, which can be also integrated in DV$s$. So change of a definition would be detectable. But recovery of a dimension definition is only possible from a backup copy. To guarantee reliability, DS definitions can be stored in reliable (official) web sites, which are open for read and which allow expansions and changes of DS definitions only in non-fixed parts. This signals to providers of numeric data (DV$s$) that the DS definitions are stable. Additionally DV search engines can create backup copies of (frequently used) DS definitions on the web and mark changes of DS definitions.

8.5 Definition of DSs by web users
To cover the range of topics on the web, those who create the web must be also able to routinely define DSs, so that they can make useful definitions about all topics which are of common interest. Appropriate Software can considerably facilitate generation of DS definitions and DV$s$, and interpret these. Web users can define DSs according to their expertise and domain of interest. They can define the numeric content and also the meaning of the hyperlinks (Fig. 7) of keycomments of DV$s$. The owner of a DS can for example define that the first hyperlink of the keycomment (see 6.1) of every contained DV points to a specific dataset (e.g. a picture, song, data generated by software of the DS owner) and the numeric content is a searchable specific feature extraction of this dataset. If the feature extraction is appropriate, the DV makes the dataset available to DS specific similarity search by all DV search engines.

Different formats are possible as concrete syntax for DS definitions and DV$s$ [schema.org 2014], [WHATWG 2014], [Wikipedia: RDF 2014], [Wikipedia: Web Ontology Language 2014]. The main point is that the numbers in the DV$s$ are identified (see 7.4) so that they can be associated to the dimension of a DS. This can be done in varying environment (see e.g. 8.9).

8.6 Redundant DS definitions
Redundant DS definitions are not desirable. The main disadvantage of such definitions is the distribution of DV$s$ (their quantitative data) over many DSs. So DV search over only one DS could become incomplete.

Recommendations for creators of DS definitions: To avoid redundancy, existing DS definitions can be checked e.g. by text search within (selectable parts of) DS definitions, before defining a new DS. If there is a DS with satisfying definition (and cooperative owner), it should be used and not redefined. If there is an existing DS (let's call it "DS-owned") and few dimensions are missing, this can be told to the owner. The owner should regard this and add missing dimension definitions to keep
the DS attractive. But if the suggestions are ignored and a DS with satisfying
definition is finally not available, the definition of a new DS can be appropriate.
Instead of defining a DS completely new, it is preferable to include (see 5.2) suitable
and frequently used DS definitions (also from dimensions, see 7.1.1 - it is possible to
include the dimensions of "DS-owned"). The advantage is that later similarity search
over these included (established) DSs (or DS-dimensions) can cover the new DS plus
the other DSs which use these definitions. A DS which includes frequently used
(groups of) dimensions therefore can get higher search frequency than a DS with
(new) isolated dimension definitions.

**Recommendations for providers of DVs:** To find the most relevant suitable DS for
quantitative data, search engines can be asked using specific text search within
(keywords or comments of) DS definitions to get a list of DSs which touch a certain
Domain or topic. The list can be ordered e.g. by the size of the DSs (the count of
contained DVs) or search frequency. This can be used to find the most relevant DS
definition. A check of their definitions can help to find the best fitting DS.

### 8.7 Nested DS definitions

According to section 5.2 every DS definition combines dimensions which can
represent not only numeric values but also again a DS definition. This possibility
allows complex expansions and generates additional similarity relations. So e.g. an
included DS definition with URL http://example.org/DS1 defines additional similarity
relations to DVs of the DS http://example.org/DS1 and DVs of all other DSs which
include the DS http://example.org/DS1 . Because this included DS definition again
can be nested, ontological structures of DS definitions with high dimensionality are
possible. Such structured definitions of complex DSs are meaningful due to several
reasons. Besides the definition of additional connections (similarity relations) and
reusage of existing definitions they can provide a subdivided and structured
representation of the domain. So Domain Spaces can also represent user defined
conceptual spaces which were proposed by Gärdensfors [Gaerdenfors 2000; 2004].

A special case needs attention:

#### 8.7.1 Infinite nesting level

There is the possibility of a circular definition in a DS: A DS definition (e.g.
http://example.org/person) can include as dimension again (DS definitions which
include) the same DS definition (http://example.org/person as friend). Obviously a DV
of such a DS can provide content (numerical values) only up to a limited count of
circular expansions. There is, however, the possibility (see 7.2.1) to include into a DV
as dimension instance the HTTP URL of another DV instead of including directly the
numerical values. Expansion of the DV is in this case possible until the chain of
included (HTTP URLs of) DVs stops or (circularly) goes back into itself. For search
such deep expansions are usually not meaningful. A search engine can at index
calculation limit the count of circular expansions, or generally limit the nesting level
which is traced.

### 8.8 High dimensional DSs

Extensions and nesting of DSs definitions can soon lead to high dimensional DSs.
High dimensional similarity search in these spaces tends to become inefficient
[Aggarwal et al. 2001], but the dimensions can serve as large set of possibilities for
low dimensional search, and as basis for exchanging quantitative information. High
dimensional DS definitions can serve as container of many one dimensional DS definitions. These are accessible using the HTTP URLs of the dimensions, see 7.1.1. So it is possible to reuse and combine dimensions in new nested DSs.

8.9 Application in Linked Open Data (LOD)

A simple way to realize the concept in the LOD cloud [Bizer et al. 2009] is to define appropriate DSs for the interesting numeric data, and to identify these (e.g. using the form "DSI#DI=number", see 7.4), and to group the data of the same DV together. Then additionally to the connections via hyperlinks these numeric data (of a DV) define bidirectional similarity relations ("numeric links") to the (numeric data of) other DVs of the same DS (Fig. 4) and (in case of a nested DS definition) to the DVs of included DSs. Identifying numbers (as components of DVs) provides these connections generally between numeric data on the web.

8.10 Constructing DSs from frequently used dimensions (with medical example)

Building high dimensional DSs can be reasonable e.g. for providing patterns of commonly used dimensions within a large domain, for later construction of derived DSs and for special conversation within this large domain. An example: Medicine deals with a lot of quantitative data which can be derived from diagnostic measurements on the patient, treatment data, result data, derived data etc.. These quantitative data can be combined as dimensions of large expandable DSs which serve as standardized initial container. Then statistical data can be obtained by observing, which dimensions (quantitative data) of these containers are used by physicians (users) in case of which situation, e.g. ICD diagnosis [WHO 2014]. Then we get for different situations (diagnoses) frequencies of used dimensions (quantitative data). This information can be used again by physicians (users). If the information about frequently used quantitative data in a certain situation is taken into account, there is less probability that important data (measurements) are neglected. Additionally the statistics shows natural connections between different situations: Situations (diagnoses) with the same most frequently used quantitative data (dimensions) can be grouped together. Later DSs can be defined from these dimensions and associated to these groups of situations (diagnoses). Because these DSs are derived from natural frequencies of used quantitative data (dimensions), they are less dependent on the initially used nomenclature for certain situations (human created names of diagnoses). They depend on the original natural situation, and they can serve as interface for exchange of searchable objective quantitative data in this situation.

8.11 Decision support (with medical example)

Decision support (concerning measurable reality) is primary motivation for information processing. Due to the basal importance of this topic we provide a medical example. For decision support the dimensions of a Domain Space can be subdivided into 3 parts:

1. **Preconditions** (In Medicine: Findings)
2. **Decision** (In Medicine: Treatment)
3. **Result**

In daily practice much valuable information arises in medicine, especially information about the results after this or that treatment (decision). Up to now most of this information gets lost after some time and it is no more available for the
community. To make it available we need to store a description of 3 states or procedures: {1} The precondition (e.g. medical finding), {2} the decision (e.g. selection of treatment) and {3} the situation (result) enough time afterwards. (The dimensionality of the result {3} can be extended retrospectively, e.g. if additional consequences of a treatment become known.) The description of {1}{2}{3} should be reproducible, precise and searchable in sufficient resolution. DVs fulfill the requirements. If they contain respectively the sequences {1}{2}{3}, users can get decision support by searching descriptions of {1} and/or varying {2} and looking for the result {3}. Because this deals with quantitative data, immediate statistics "near" the searched description are possible, to find the variant {2} which leads to the best result {3}.

Medical example:

DS: Influenza

{1} Preconditions (Findings)
{2} Decision (Treatment)
{3} Result

DS: Findings (part of definition)
- Systolic blood pressure
- Diastolic blood pressure
- Age
- Body weight

DS: Treatment (part of definition)
- Dose of a certain medication
- Dose / Body weight

DS: Result (part of definition)
- Relative change of systolic blood pressure
- Relative change of diastolic blood pressure

Fig. 20. Exemplary DS definition to evaluate the relative change of blood pressure{3} in dependence of medication {2} in case of a certain individual situation {1}. Only a part of relevant dimensions is shown.

So it is e.g. possible to check in case of individual preconditions {1} (Fig. 20) immediate statistics of result dimensions {3} (e.g. change of blood pressure in Fig. 20) in dependence of selected ranges of influencing therapeutic dimensions (e.g. decisions {2} like dose of medication in Fig. 20).

Generally the practical benefit of decision support depends on the completeness of available data. In case of simple situations which depend on few dimensions, e.g. certain technical data, it is easier to provide enough data for decision support than in case of situations which depend on many dimensions. Because real life situations contain (of course) too many dimensions for a complete acquisition and description using available hardware, approximation is necessary using few dimensions within a restricted domain of interest. Finding the most relevant influencing dimensions for a certain domain of interest (a DS) can be a demanding process. The completeness of these dimensions can be estimated. It is the better, the smaller in case of controlled dimensions {1}{2} the standard deviations of the dependent result dimensions {3} are. Statistics can be used to detect dependencies. Statistical results are the more reliable, the larger the database is. This is one of many arguments for worldwide DSs.
8.12 Comparison to the Vector Space Model in information retrieval

Up to now the Vector Space Model in information retrieval is mainly used to compare text documents [Wikipedia: Vector space model 2014]. Vectors which represent text documents represent quantitative data of these texts (e.g. frequencies of keywords), so these can be seen a special case of Domain Vectors.

8.13 Motivation for owners of DSs

Motivation for the owners of DSs definitions is better communication within their domain of interest, and the possibility to expand the DSs subsequently by additional interesting dimensions. Moreover owners can modify the changeable part of DS definitions (see 7) and provide links e.g. to their web pages.

Patents on DS definitions, however, should not be possible. The reason for this is given in the following remark:

8.14 Language is not patentable

The proposed standard for worldwide valid DSs allows to include (reuse) DS definitions in new definitions and to extend definitions a posteriori. The approach is designed for free and efficient usage of data on the web. Patents on DS definitions would contradict this purpose. Moreover: DSs define precise quantitative descriptions worldwide. This can be seen as an extension of language. Patents on (parts of) language are not possible. Therefore patents on DSs and contained DVs should not be possible.

9. CONCLUSION

DSs can be defined by web users according to their domains of interest. Their elements, the DVs, identify quantitative data and so make these data machine readable and searchable. Nested DS definitions realize hierarchical ontologic structures, where common identifiers (HTTP URLs) of dimensions provide connections between different DSs and enable general similarity search of quantitative data on the web. DV based description and search can become an important addition to usual word based description and search on the web. Therefore the introduction of a web standard for worldwide valid DS definitions and DVs is recommendable.

REFERENCES

Aggarwal, C. Hinneburg, and A. Keim, D. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space. ICDT 2001, Vol. 1973, 420-434. DOI: http://dx.doi.org/10.1007/3-540-44503-x_27

Berners-Lee, T., Fielding, R., & Masinter, L. 1998. Uniform resource identifiers (URI): generic syntax. Retrieved April 1, http://www.hjp.at/doc/rfc/rfc3986.html

Bizer, C., Heath, T., & Berners-Lee, T. 2009. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), 1-22. http://dx.doi.org/10.4018/ijswis.2009081901

Black M. 1963, Reasoning with Loose Concepts. Dialogue, 2, 1-12. http://dx.doi.org/10.1017/S001221730004083X

Fontoura M., Lempel R., Qi R., Zien J. 2006, Inverted Index Support for Numeric Search, Internet Mathematics, 3(2), 153-185. http://projecteuclid.org/euclid.im/1204906137

Gärdenfors, P. 2000. Conceptual Spaces: The Geometry of Thought. MIT Press, Cambridge, MA.
Gärdenfors, P. 2004. How to make the semantic web more semantic. In A.C. Vieu and L. Varzi, editors, Formal Ontology in Information Systems, pages 19–36. IOS Press, 2004.
http://yaxu.org/tmp/Gardenfors04.pdf

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

Haas, P. 2005. Medizinische Informationssysteme und Elektronische Krankenakten. Springer, Berlin, Heidelberg.

Haveliwala T. H., Gionis A., Kleinx D., Indyk P. 2002. Evaluating Strategies for Similarity Search on the Web. In Proc. WWW 2002, 432-442, ACM Digital Library.

Heinz, S. and Zobel, J. 2003, Efficient single-pass index construction for text databases. J. Am. Soc. Inf. Sci. 54, 713–729. http://dx.doi.org/10.1002/asi.10268

Holsgrove, G. and Elzubeir, M. 1998. Imprecise terms in UK medical multiple-choice questions: what examiners think they mean. Medical Education 32, 343–350. http://dx.doi.org/10.1046/j.1365-2923.1998.00203.x

Kriegel, H.P., Kröger, P., Renz, M., Schubert, M. 2010. Metric spaces in data mining: applications to clustering. SIGSPATIAL Special Volume 2 Issue 2 (July 2010), 36-39. http://dl.acm.org/citation.cfm?doid=1862413.1862423

MATH41002: LINEAR ANALYSIS, viewed May 2014, Minkowski inequality http://www.maths.manchester.ac.uk/~nikita/31002/minkowski.pdf

Nakao M. A., Axelrod S. 2001. A fast and reliable SVM decision function. Journal of Machine Learning Research, 2, 242-281.

Orthuber, W. 2013. General approach to similarity search of resources with numeric features on the web. 2013.

Orthuber, W. 2010. Standardized Vectorial Representation of Medical Data in Patient Records, Medical and Care Compunetics 6, 153-166. http://science.imm.org/2010/06/20/standardized-vectorial-representation-of-medical-data-in-patient-records/

Orthuber, W. 2010. Towards Standardized Vectorial Resource Descriptors on the Web. 2010. In Lecture Notes in Informatics, INFORMATIK 2010. Service Science - Neue Perspektiven für die Informatik. Band 2 P-176, 453-458 http://nbs.emis.de/LNI/Proceedings/Proceedings2/176/469.pdf

Orthuber, W., Papavramidis E. 2010. Standardized Vectorial Representation of Medical Data in Patient Records, Special Volume 2 Issue 2 (July 2010), 36. IOS Press, 2004.

Orthuber, W. 2010. Movable Type Scripts, viewed May 2014, online implementation of vectorial description and search. 2012. (Online since July 2012) http://nummel.com/

Orthuber, W. 2010. General approach to similarity search of resources with numeric features on the web. 2013. Open Data on the Web, http://www.w3.org/2013/04/odw/papers#all

WHATWG community, HTML Living Standard, viewed May 2014. Microdata, in http://www.whatwg.org/specs/web-apps/current-work/multipage/microdata.html#microdata

Wikipedia, viewed May 2014. Domain model, in http://en.wikipedia.org/wiki/Domain_model

Wikipedia, viewed May 2014. Ontology, in http://en.wikipedia.org/wiki/Ontology_(information_science)

Wikipedia, viewed May 2014. Feature extraction, in http://en.wikipedia.org/wiki/Feature_extraction

Wikipedia, viewed May 2014. Linear subspace, in http://en.wikipedia.org/wiki/Linear_subspace

Wikipedia, viewed May 2014. Metric Space, in http://en.wikipedia.org/wiki/Metric_space

Wikipedia, viewed May 2014. Minkowski Distance, in http://en.wikipedia.org/wiki/Minkowski_distance

Wikipedia, viewed May 2014. Quantity, in http://en.wikipedia.org/wiki/Quantity

Wikipedia, viewed May 2014. Resource Description Framework (RDF), in http://de.wikipedia.org/wiki/Resource_Description_Framework

Wikipedia, viewed May 2014. Semantic Web, in http://en.wikipedia.org/wiki/Semantic_Web

Wikipedia, viewed May 2014. Vector Space Model, in http://en.wikipedia.org/wiki/Vector_space_model

Wikipedia, viewed May 2014. Web Ontology Language, in http://de.wikipedia.org/wiki/Web_Ontology_Language

World Health Organization, viewed May 2014. International Classification of Diseases (ICD), in http://www.who.int/classifications/icd/en/

Zezula, P., Amato, G., Dohnal, V., Batko, M. Similarity Search. The Metric Space Approach. 2005. Series: Advances in Database Systems, Vol. 32, Springer, Berlin, Heidelberg.