How to make quantitative data on the web searchable and interoperable part of the common vocabulary

Wolfgang Orthuber, University Clinic Schleswig-Holstein, UKSH
Department of Orthodontics,
Arnold Heller Str. 3, House 26, 24105 Kiel / Germany
orthuber@kfo-zmk.uni-kiel.de

Keywords:

- Quantitative Data
- Searchable Data
- Interoperable Data
- Common Vocabulary

Common Vocabulary

Definition: The common vocabulary is quickly known
by participants of a conversation (usually within 1 sec).

On the web:
Hyperlinks have been very successful, because their content is quickly viewable.

```
<a href="http://example.com">clickable text</a>
```

Similarly quantitative or numeric data can be made quickly viewable as elements (vectors) of metric spaces. At this every element contains

- the URL of its space's definition
- plus a sequence of values (vector).

Syntax example:

```
<v http://numericsearch.com/bw.xml; 2014-01-30; 83.914>clickable</v>
```

After click e.g.:
(possible in
system language)

	BodyWeight
2014-01-30	Date | yyyy-mm-dd
83.914	Weight-Morning | kg

Recall of the basics

Well defined information means selection from a well defined set.

The set should be the same for all
for interoperable information, for comparison of information
(equal, unequal; if sorted: smaller, greater)

In this approach every set is defined at one place (online).
The URL of the definition is also identifier.

Syntax example (without quotes):

```
<v http://numericsearch.com/bw.xml; 2014-01-30; 83.914>clickable</v>
```

After click e.g.:
(possible in
system language)

	BodyWeight
2014-01-30	Date | yyyy-mm-dd
83.914	Weight-Morning | kg

Recall of the basics

Well defined information means selection from a well defined set.

We can define the set according our needs on the web:
A quantitative space allows efficient handling and comparison,
a metric space also allows similarity search!

Quantitative data are relevant and fundamental

but (2015):

Quantitative Data are not searchable on the web!

Quantification and searchable quantitative data have great potential.
The current focus on text search is a far reaching restriction.

Current approaches

RDF or FHIR are well known approaches for machine readable data on the web. First we give an example for RDF of the above data (bodyweight).

```
<?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/Weight measured">
   <cd:date>2014-01-30</cd:date>
   <cd:weight>83.914</cd:weight>
</rdf:Description>
</rdf:RDF>
```

Modification from http://www.w3schools.com/webservices/ws_rdf_example.asp to code the observation of body weight.

There is no Link (URL) to a complete standardized definition of the data(set).

Current approaches

FHIR is a next generation standards framework created by HL7. At this interoperable data are coded in resources. The above data (bodyweight) are e.g.:

```
<Observation xmlns="http://hl7.org/fhir">
  <text>
    <status value="generated"/>
    <div xmlns="http://www.w3.org/1999/xhtml">
     Jan 30 2014: Body Weight = 185 lbs</div>
  </text>
  <name>
    <coding>
      <system value="http://loinc.org"/>
      <code value="3141-9"/>
      <display value="Weight Measured"/>
    </coding>
  </name>
  <valueQuantity>
    <value value="185"/>
    <units value="lbs"/>
    <system value="http://unitsofmeasure.org"/>
    <code value="[lb_av]"/>
  </valueQuantity>
</Observation>
```

Excerpt of http://www.hl7.org/fhir/observation-examples.html which codes the observation of body weight. There is no Link (URL) to a complete standardized definition of the data(set)

The proposed approach

In this approach a link to a complete standardized definition of the data ("Space") is essential

Data are elements of **"Domain Spaces" (DSs)**
which are online (globally) defined metric spaces.

Every data element is called **"Domain Vector" (DVs)**
and contains:

- the URL of the (complete) DS definition plus
- a sequence of values (vector), e.g.

```
<v http://numericsearch.com/bw.xml; 2014-01-30; 83.914>Bodyweight</v>
```

A DV can be clicked like a hyperlink using a browser which combines the definition with the values. These select from the dataset (information).

The proposed approach

Example of the **DS Definition** with URL http://numericsearch.com/bw.xml

```
<DS>
  <kw>BodyWeight</kw>
  <dim>
    <kw>Date</kw>
    <unit>yyyy-mm-dd</unit>
    <format>yyyy-mm-dd</format>
  </dim>
  <dim>
    <kw>Weight-Morning</kw>
    <co>Weight at morning directly after stand up</co>
    <unit>kg</unit>
    <format>float</format>
  </dim>
</DS>
```

After click on

```
<v http://numericsearch.com/bw.xml; 2014-01-30; 83.914>bodyweight</v>
```

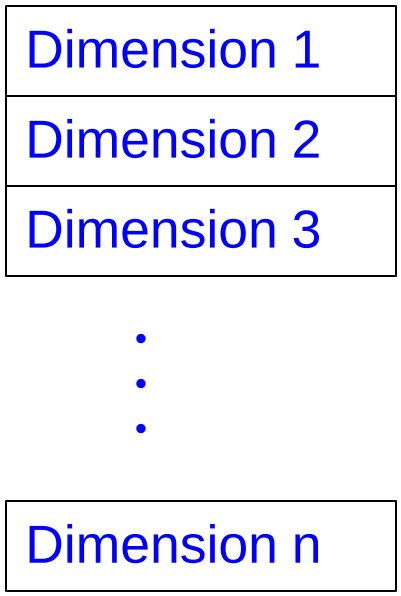
e.g.:

BodyWeight	
2014-01-30	Date | yyyy-mm-dd
83.914	Weight-Morning | kg

If wished: embedding in HTML conform syntax is possible:
http://html5doctor.com/html5-custom-data-attributes/

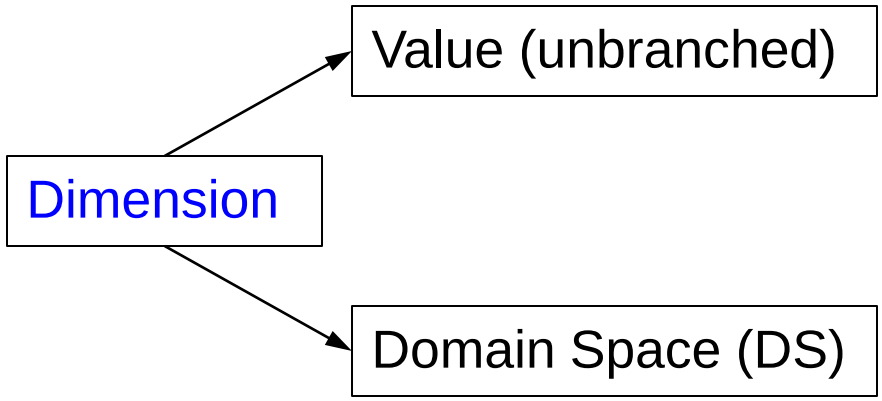
Domain Space Structure

Domain Space (DS)



The DS and every of its dimensions have a unique name (URL).

Dimension of a DS



Every dimension of a DS can represent an unbranched value or again a DS.
So external DSs can be integrated and nested.

Hierarchical Structure of Domain Spaces

Location=DS

DateTime

Temperature

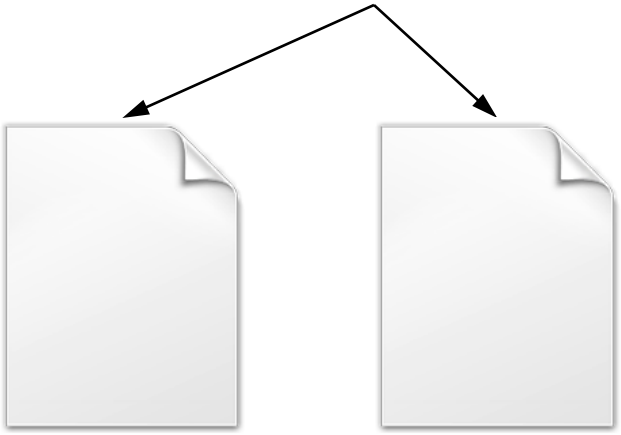
The structure of a DS definition is similar to that of an ordered directory tree. A DS combines dimensions which represent values or again DSs (like a directory contains files or again directories).

Hierarchical Structure of Domain Spaces

Location=DS

DateTime

Temperature



Latitude

Longitude

In a DS definition the order of dimension definitions is important because it determines the order in which values can be given in a DV without dimension identifier.

Definition of Domain Spaces (DSs)

To cover the range of topics on the web, all who create the web should be also able to 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 DVs, and interpret these. Web users can define DSs according to their expertise and domain of interest.

Handling of redundant definitions

So every user can generate DSs and searchable spaces with quantitative data. (Purposeful) Redundant definitions of dimensions are to be expected.

All redundant definitions can be connected. For this the (e.g. in http://www.w3.org/TR/owl-ref/#sameAs-def described) **sameAs** directive can be extended to the form:

this Dimension is sameAs (algebraic) expression of other DS Dimensions

Usually definers of DSs are interested to connect their definitions with other definitions, so that searches there can also include the own space.

Reliability of definitions

After "draft" state every dimension definition of a DS must be stable.

To guarantee reliability, certain DS definitions can be stored in *official web sites*.

Additionally Numeric Search engines can create dated backup copies of DS definitions on the web.

Implementation: http://numericsearch.com/

Dimension

keycomment of dimension owner

Keyword: Link:

Weight-Morning A

Unit: Link:

kg A

Comment:

Min: Max: Weight: 1 SetOneDivSd

representation: list tux

integer money floating point: medium length floating point: max. length

date in: yyyy-mm-dd hh:mm:ss yyyy-mm-dd hh:mm yyyy-mm-dd hh

yyyy-mm-dd yyyy-mm yyyy hh:mm:ss hh:mm

Implementation: http://numericsearch.com/

DS (Domain Space)

Definition of DS 1029 (BodyWeight) owner

|< << < > >> >| 0..0

0 BodyWeight
0 Date | yyyy-mm-dd
1 Weight-Morning | kg

Keyword: Link:

BodyWeight A

Comment:

This is: draft ok deprecated

Nested metric: Manhattan Euclidean Maximum

Implementation: http://numericsearch.com/

NumericSearch

NumericSearch allows high resolution search in user defined numeric Domain Spaces (DSs). This version is designed for demonstration purposes, the user generated DSs are stored in a local database.

If you have interesting numeric data which should be published freely in searchable form, please contact me

Click in the column below i7 on a index to search in a DS with count of resources r>0, e.g. click on 1005 or 1006

Motivation: For description of reality usually words of language are used, but they categorize the original quantitative features of reality. Depending on the domain of interest, specific features can be selected and represented more precisely by vectors ("Domain Vectors" = DVs) in "Domain Spaces" (DSs). DSs represent domain specific metric spaces. Every DS is unambiguously identified by a URI which is called "Domain Space Identifier" (DSI) and which can be the URL of the DS definition in a web standard. The DVs of a DS are accessible to similarity search with optional range restriction.

NumericSearch consists of 2 steps:

1. Selection of the DS directly or by text search of its DSI (first keyword kw0)
2. Similarity and/or regional search of DVs within this DS.

✹ i4 ✹ iu search kw0 I logout[10001] up own us

✹ i7=1029', o | 2013-02-09 BodyWeight

Select i7 (index of Domain Space)

kw0 add X copy nx

|< <<< << < > >> >>> >| 1000..1024..4011

i7	s	r	
1000	91	68	space-of-spaces || v386743
1001	30	9	o ride
1002	29	4	o my-location
1003	28	1	o real-estate
1004	7	0	o car
1005	518	10001	o test-space || try search 0..10: subv1, subv2 filled with pseudo random numbers 0..10
1006	48	23	o Cupboard | Schrank
1007	108	11	o Diode || (for rectification)
1008	183	1500001	o 260dim-demo || try search 0..10: subv1, subv2 filled with pseudo random numbers 0..1
1009	202	57	o text-as-dimension-example || dimensions (not used for similarity comparison) can also
1011	3	1	o cardiovascular-disease
1012	20	2	o Tamiflu-Test || Is Tamiflu indicated? For answer of this question we could fill this space
1013	40	85	o NOx-Pollution-in-1000tons || exemplary data from Australia, Austria, Belgium, Germany
1014	2	8	o Screw | Schraube
1015	135	24	o datacube-example-as-TS || data like "The RDF Data Cube vocabulary" example chapter
1016	0	0	o opinion-about-xx
1017	0	0	o climate-fluctuations
1018	4	0	o Meeting | Treffen
1019	44	11	o Kugellager-Edelstahl
1020	0	0	o Help || Search help (kind of help, time, location, duration etc.)
1021	0	0	o SleepDay || Documentation of one day sleep with result, result optionally after daycount
1022	0	0	o MRT-usage-year || yearly usage data about one magnetic resonance tomograph
1023	428	100001	o test-150dim || try search 0..10 in subv1(Euclidean metric) and subv2(Manhattan metric
1024	8	1	o traffic-accident || DVs can become increasingly part of legislative vocabulary, existing ju

w0 1 2 3 4 5 6

one page sketch

Technical background and details

Help

Introduction

Search demo

imprint, contact

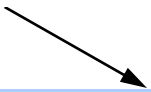
Implementation: http://numericsearch.com/

NumericSearch in DS 1029 (BodyWeight)

|< << < > >> >| 0..0 search spar

	sim	min	max	g
0				BodyWeight
0				☐ Date | yyyy-mm-dd
1	80			☐ Weight-Morning | kg

d	a	
0	17	o Bodyweight-1 | 2014-02-05, 80,
3.914	55	o Bodyweight | 2014-01-30, 83.914,
5	26	o Bodyweight-2 | 2014-03-10, 75,
10	19	o Bodyweight-3 | 2014-04-20, 70,



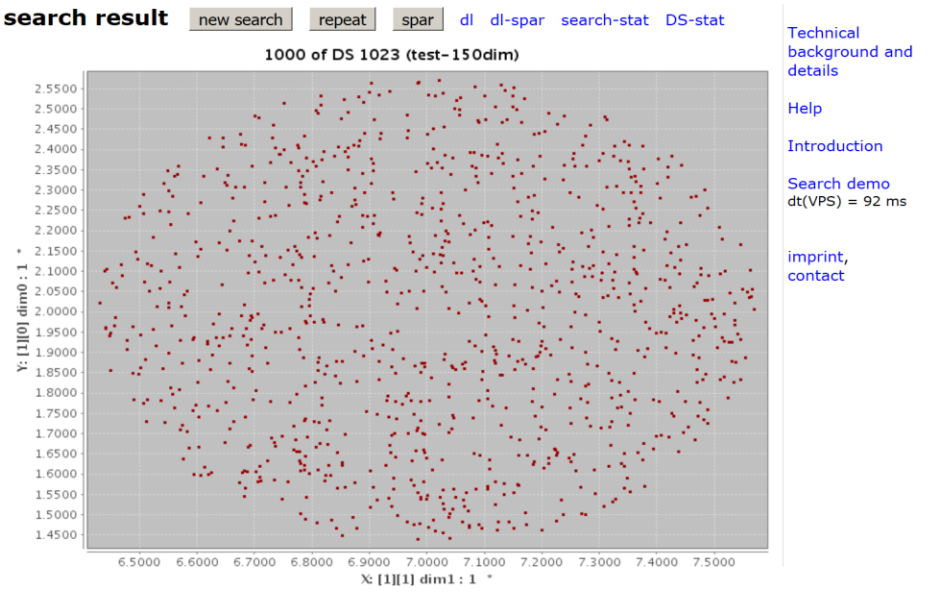
BodyWeight	
2014-04-20	Date | yyyy-mm-dd
70	Weight-Morning | kg

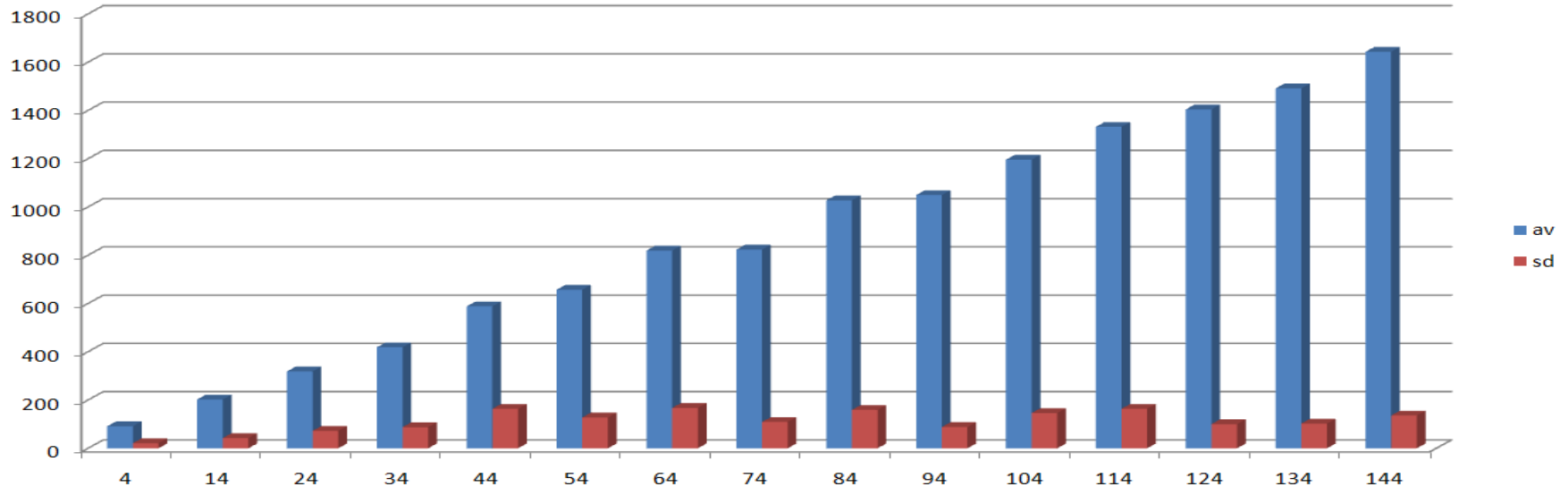
Performance of synchronized index

NumericSearch in DS 1023 (test-150dim)

|< << < > >> >| 0..2..2 search spar

	sim |	min	max |	g
0				subv0
0				☐ dim0
1				☐ dim1
1				subv1
0	2			☑ dim0
1	7			☑ dim1





The search time within 100001 DVs in ms (vertical) in dependence of searched dimensionality (x 64 bit).

Resolution of quantitative data

Quantitative data have important technical advantages: Only one definition is necessary for a complete quantitative space with all its elements. The definition of the proposed "Domain Spaces" (DSs) is online so that everywhere the same definition is accessible.

The simultaneous definition of all elements of a DS is the technical reason for the *high resolution* of DV based description. This resolution is not nearly reachable by non quantitative language.
Quantitative (numeric) data are used already today on the web. But without standard and not as elements of globally defined spaces. Therefore:

Today (2015) quantitative data are not interoperable and not searchable on the web.

2015: Quantitative Data are not searchable!

There are many important quantitative data.

Quantitative data can become interoperable and searchable on the web as elements of DSs (DVs)

Standardization of (DVs and) DS definitions is necessary.

------- Applications -------

DSs are also a

Guide for data providers

Search engines can provide selective search within DSs definitions. They can show interesting DSs and size of DSs.

Definition of a DS shows data which are relevant in a certain domain. This increases motivation for writers to provide these data, to make the web more informative. Also from BIG DATA we can only extract existing data.

Application: Guide for data providers, e.g.

Real Estate

|< << < > >> >| 0..5

0		gps-coordinates
0		latitude | degree
1		longitude | degree
1		financial
0		price | euro || (if for sale)
1		price-per-square-meter-living-area | euro / square-meter
2		monthly-rent | euro || (if renting)
3		monthly-rent-per-square-meter-living-area | euro / square-meter
4		maintenance-costs-per-month-average | euro
5		this-per-square-meter-living-area | euro / square-meter
2		energy-efficiency
0		energy-costs-per-year | euro
1		this-per-square-meter-living-area | euro / square-meter
3		age
0		build | year
1		last-renovation | year
4		size
0		count-of-living-rooms
1		living-area | square-meter
2		area-of-corridors | percent-of-living-area

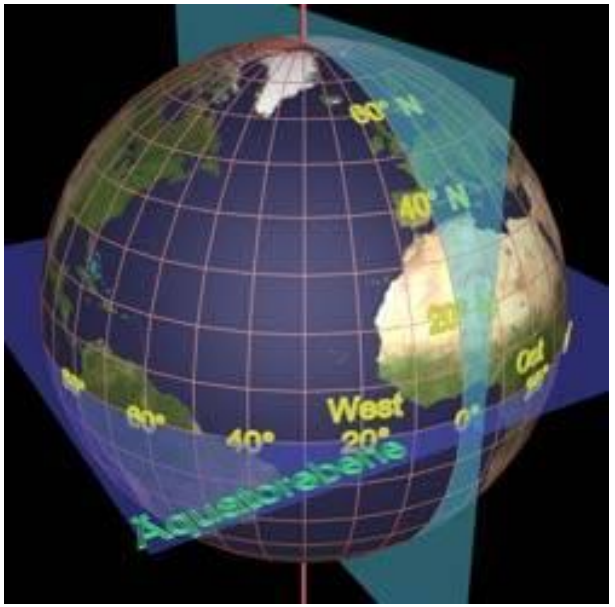
DV Examples

GPS-Coordinates

Feature Vector:

a_1 = Latitude

a_2 = Longitude



DV Examples

Industrial products, e.g. electric motors

Feature Vector:

a_1 = power (in Watt)

a_2 = rpm (revolutions per minute)

a_3 = energy efficiency (in percent)

a_4 = axial dimeter in mm

a_5 = length in mm

a_6 = height in mm

a_7 = weight in kg

DV Examples

Customized clothes

Feature Vector:

a_1 = collar size (in cm)

a_2 = abdominal girth (in cm)

a_3 = chest measurement (in cm)

...

This DV be also used for ordering clothes.

DV Examples

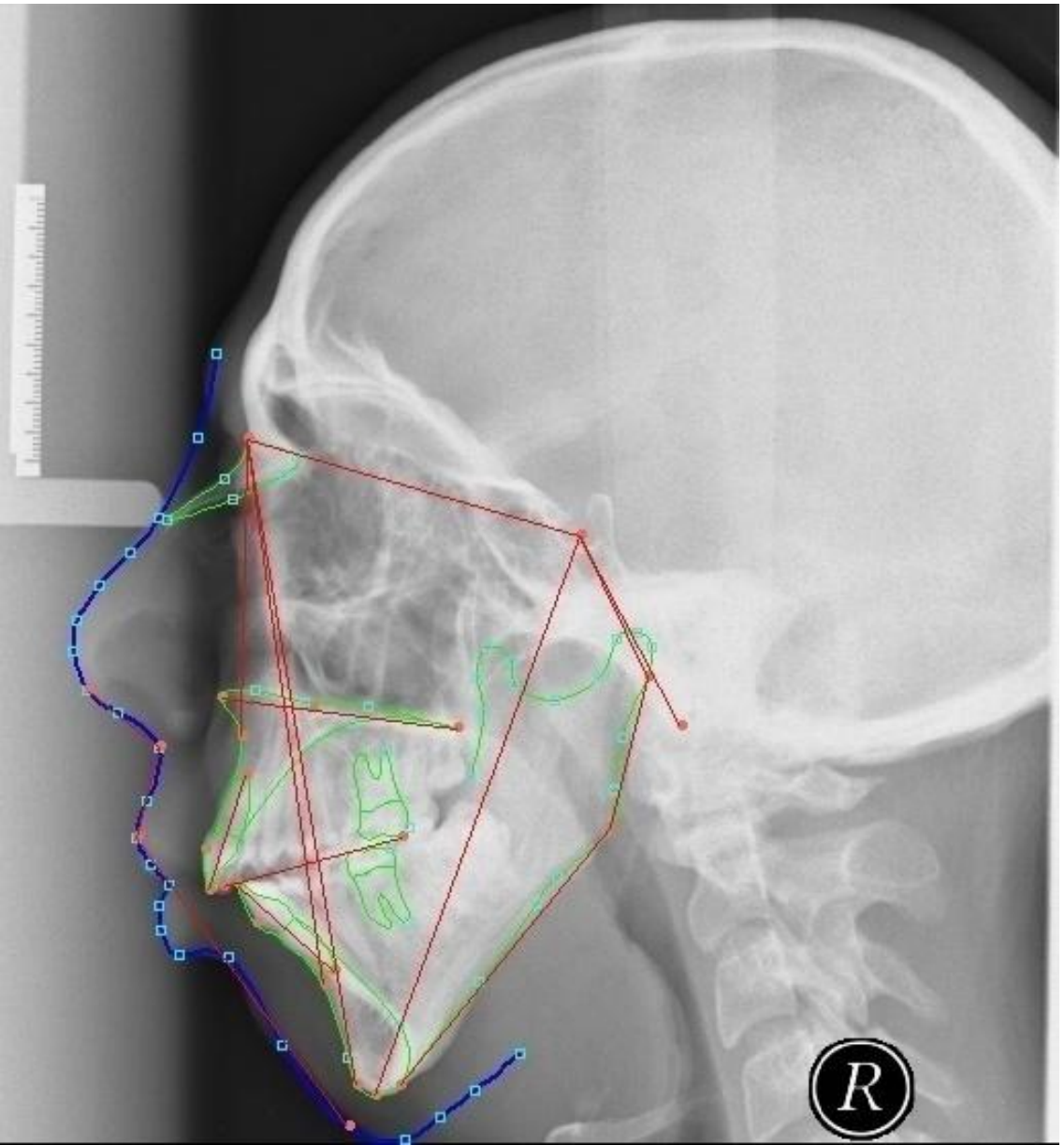
Searchable original scientific data

Scientific original data are usually detailed quantitative data.

As DVs on the web these would be searchable and interoperable.

Quantitative data could be defined that automatic combination is possible.

DV Examples (Medicine)



History of idea:
Medical applications

e.g. Cephalometry:

- A scientific study of the measurements of the head with relation to specific reference points
- utilizing a fixed, reproducible position for lateral radiographic exposure of skull
- used for orthodontic treatment planning, for evaluation of facial growth and development, including soft tissue profile.

SureSmile 5.6 -

File Edit Go Case-Mgmt Tools Help

TS

Upper Occlus

Reference Model: Diagnostic Model 1 (07/16/07 19:21:16)
Reference Arch: Occlusal Plane: Upper, Arch Form: Lower

Bolton Ratio (6..6): 86.3 (3..3): 65.3
Maxilla Sum (6..6): 92.1 mm (3..3): 44.7 mm
Mandible Sum (6..6): 79.5 mm (3..3): 29.2 mm
Surplus (6..6): Maxilla 5.1 mm
Surplus (3..3): Maxilla 5.9 mm
Measured Overjet: -0.5 mm
Arch Length Discrepancy:
UR: 0.0
UL: 0.0
LR: 0.1
LL: 0.0

8.3
8.1
7.0
6.9
7.5
6.9
7.7
6.8
6.4
6.8
10.4
10.2
10.2
9.6

Camera Navigation

[20.7.2007] Intra-oral Right Buccal

[20.7.2007] X-ray Lateral Ceph

Global Registration | References | Occlusal Plane | Arch Form | Upper Disp. | Lower Disp. | Fixed Teeth | Notes

U

Midline: r. 0.2 / r. 0.2

Arch Width

	(27.0)	(27.4)
Molar	27.0	27.4
R / L		
Canine	16.9	17.6
	(16.9)	(17.6)

AP Position: R none | L none

L

Midline: r. 2.2 / r. 2.2

Arch Width

	(11.5)	(12.2)
Canine	11.5	12.2
R / L		
Molar	24.3	25.0
	(24.3)	(25.0)

AP Position: Front none; R / L Molar none none

Extract / Space (+) / IPR (-)

Apply | Cancel

0.1

Molar Class: Right / Left — Maint., I, II, III

Overjet: cur. 3.3 — 0.0

Arch Selection: Currently: Natural — Natural; Symmetric / Asymmetric

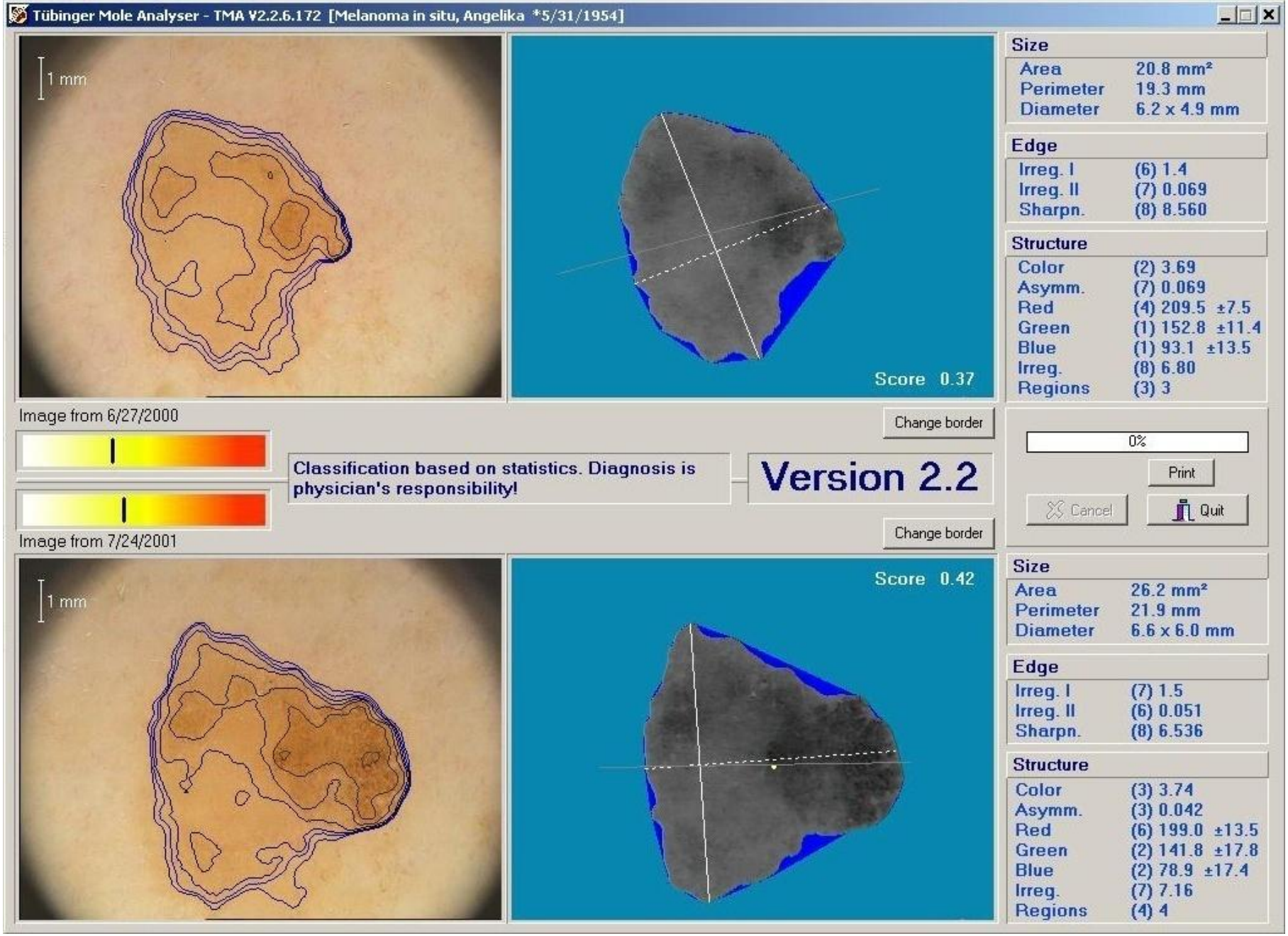
Setup: Arch: Upper, Lower; Vertical, Vert. Contact, Horizontal, Horiz. Contact; Go; Reset

Max. IPR [mm]: 0.0 — 3 - 3 / 4 - 4 — Apply

Align: Front — Align

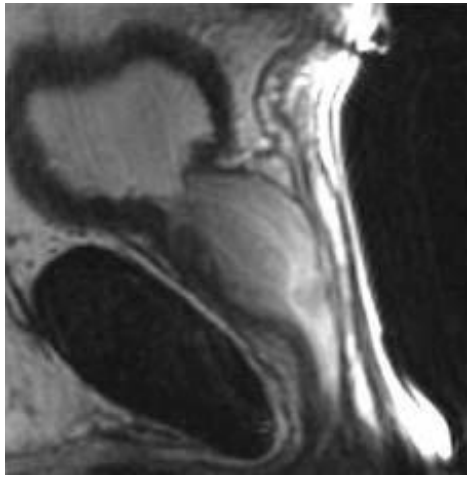
For Help, press F1

Elapsed Treatment Time: 7 Months, 3 Weeks | Admin | Active | Real



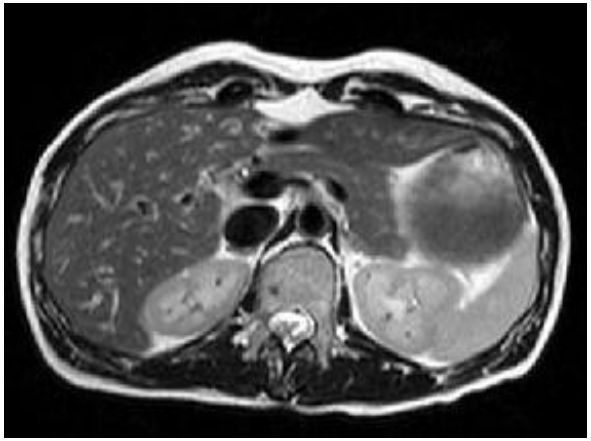
Source: Tübinger Moleanalyzer © FotoFinder Systems GmbH 2008

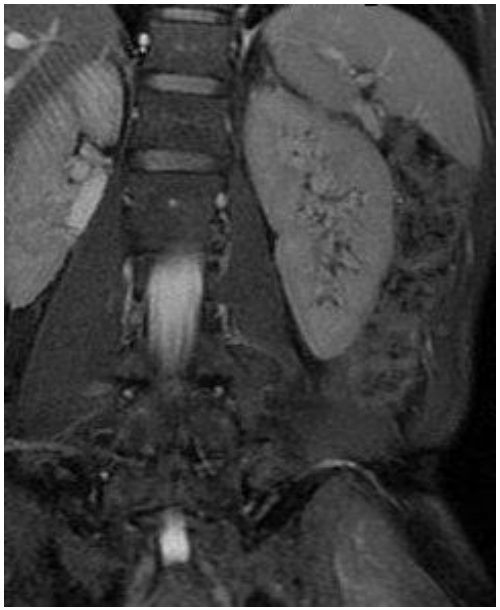
DV Examples (Medicine)

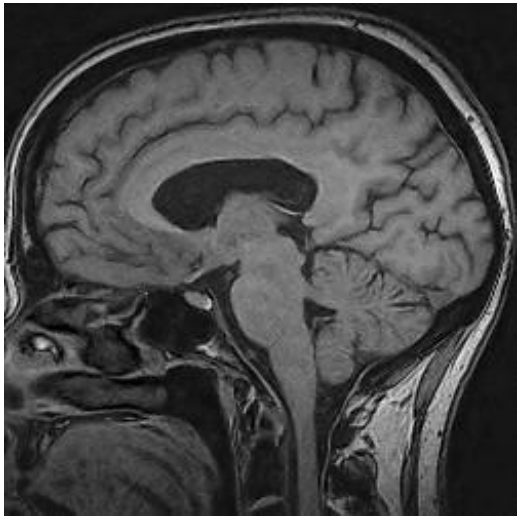


MRI prophylaxis

- Selection of frequent and serious diseases which are best detectable by MRI
- Description and quantification of decision relevant features (initially 2D, later 3D)
- Comparison with previous findings and cases



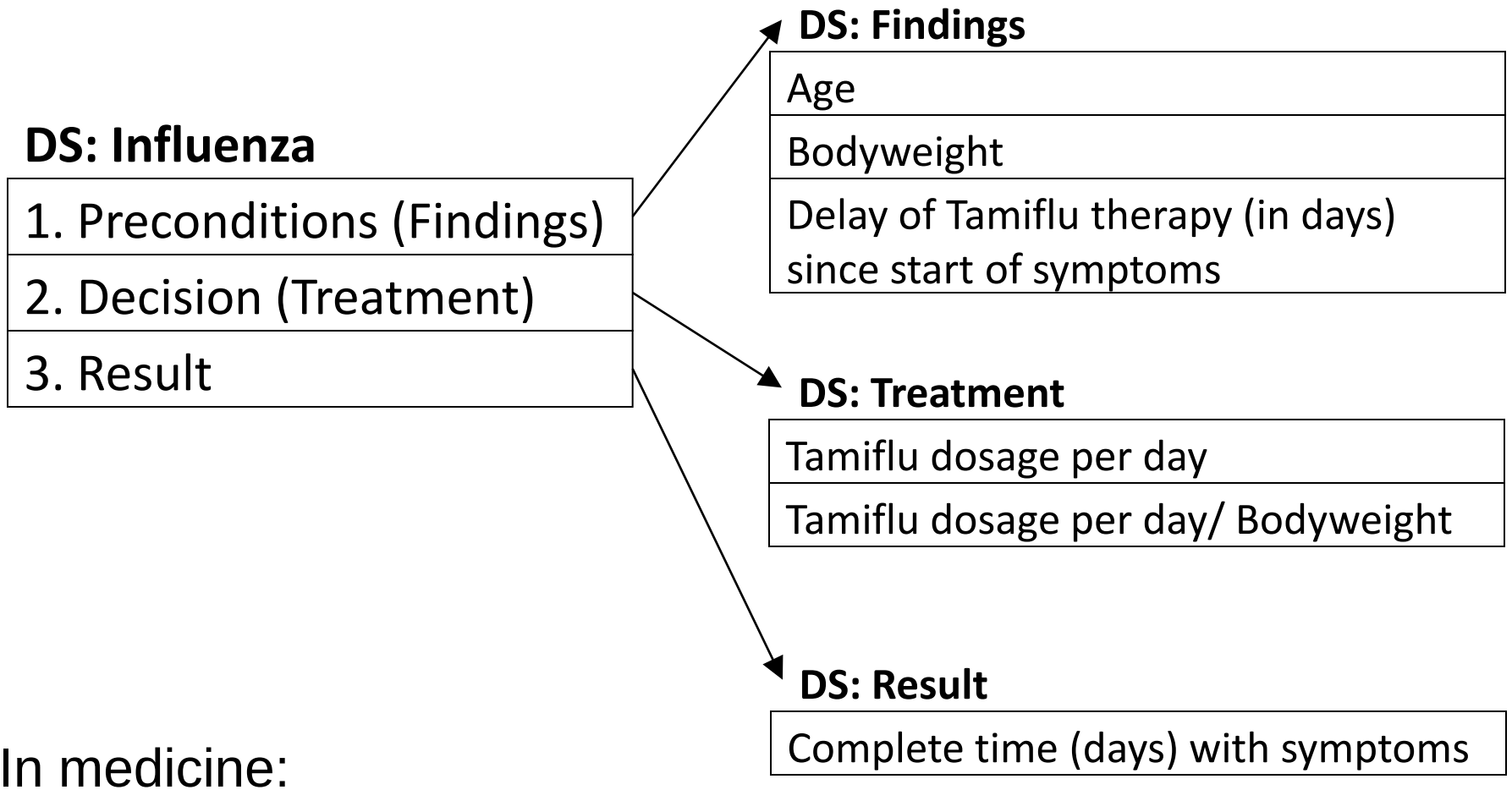




Application: Decision support

- A decision means a selection within a given domain (value or definition set).
- So precondition of well defined decision support is that all speak and think about the same domain.
- So a common standardized Domain Space definition (and with this the definition of the domain) on the web is natural also for decision support.

Application: Decision support



In medicine:

- search patients with similar findings
- at this vary possible treatment decisions
- look for decision with best result

Medical data collection with search engine

Application:

1. The doctor makes a first principal diagnosis, e.g. ICD.
2. Using the code (later also additional finer quantitative findings) the search engine shows frequencies of fine diagnostics chosen by other doctors in such a case.
3. The doctor decides about finer diagnostics. This includes collection of relevant collateral data e.g. about daily food intake, sports etc.. The multidimensional results of finer diagnostics are provided to the software (if possible, more and more automatically).
4. Most important quantitative results are provided to the search engine. If there are enough patients with similar data in the database, anonymously frequencies of further diagnoses with treatment decisions and associated results can be shown in this group of patients, like in a scientific study ("individual study").
5. Decision about further diagnostics or treatment is done and provided to software which can prepare, if wished, the draft of a medical report.
6. If necessary, later (with new data) continuation at 2. or even 1.

Applications

DVs can be grouped together, so that one group describes the same resource. The index is not restricted to one DS.

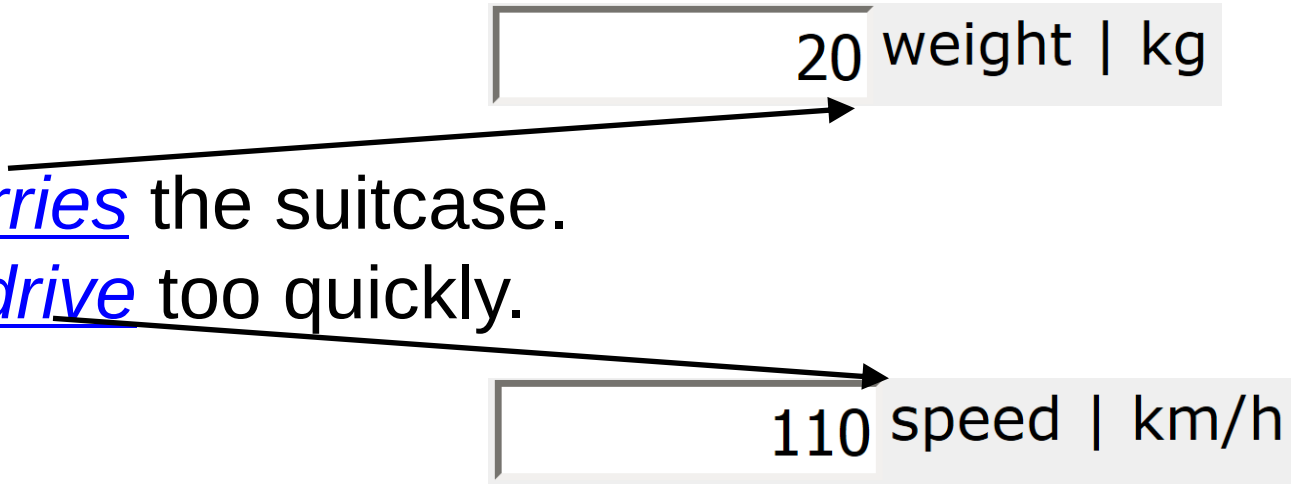
So data providers can select dimensions which they group together.

Later it is possible to combine dimensions of different DSs also for search.

Applications

Quantified text

Words of language can be made more precise by additional quantitative features shown after click, e.g.:



If the DS definition is done in multiple languages, automatically the definition in (by user selected) system language can be shown.

Quantified legislative text

DVs can make legislative text more precise. Similarly like for description of medical decisions DVs can be also used for description of judgments and (internationally) large searchable web collections of judgments can be built. So it would be possible for judges to compare existing cases to past cases in the collections more precisely and to check past judgments. This could help jurisdiction towards better reproducibility and precision.

Searchable Feature Extraction

There are uncountable many possibilities for Feature Extraction. Representation of important features of a resource as dimensions of a DS is an important application. It could be used to make complex resources identifiable and searchable.

Applications

Correctness is precondition of precision. From original DSs automatically evaluation DSs can be derived, with evaluation dimensions for every (unbranched) value of the original.

- correct value
- |value| / |correct value|
- subjective grading of precision (0..15)
- subjective grading of reliability (0..15)

etc.

Remark

Patents on DS definitions not recommended

The proposed standard for worldwide valid DSs allows to include (reuse) DS definitions in new definitions and to extend definitions subsequently. The approach is designed for free and efficient usage of data on the web. Patents on DS definitions would contradict this purpose and therefore are counterproductive.

DS definitions and DVs can be seen as part of language and patents on parts of language should not be possible.

2015: Quantitative Data isolated

Quantitative data can become interoperable and searchable

Important steps towards this are:

- Efficient **Standard** for DVs and definition of DSs.
- "**Official Domains**" for controlled definition of DSs.
- Integration into HTML editors and browsers (as clickable data)

- Support by search engines
- Expandable topic for research

Also interested in precise searchable information?

Interested to contribute?

Please contact me: orthuber@kfo-zmk.uni-kiel.de

Further Information: **http://numericsearch.com**

Thank you for your attention!