

METADATEN

LUKAS BURGER, SHAYAN MOHAJERANI, NINA DREHER

Agenda

- Allgemeiner Überblick
 - Übungen 1-4
- Strukturelle Daten
 - Übung 5
- Semantische Daten
 - Übungen 6-10
- Quizz

METADATEN

“Metadaten sind strukturierte Informationen, die beschreiben, erklären, lokalisieren, oder es sonstwie einfacher machen, eine Informationsquelle abzurufen, zu verwenden, oder zu verwalten. Metadaten werden oft Daten zu bestimmten Daten oder Informationen zu bestimmten Informationen genannt.”

National Information Standards Organization



METADATEN

Liefern Informationen über:

- Daten (Dokumente, Bilder, Datensätze)
- Konzepte (Klassifikationen)
- Reale Begebenheiten (Personen, Organisationen, Standorte, Bilder)

AUFGABE 1: METADATENBEISPIEL AUS DER ANALOGEN WELT

Nicht nur digitale Erzeugnisse enthalten Metadaten. Welche Metadaten könnte es bei einem Buch geben?



LÖSUNG AUFGABE 1: METADATENBEISPIEL AUS DER ANALOGEN WELT

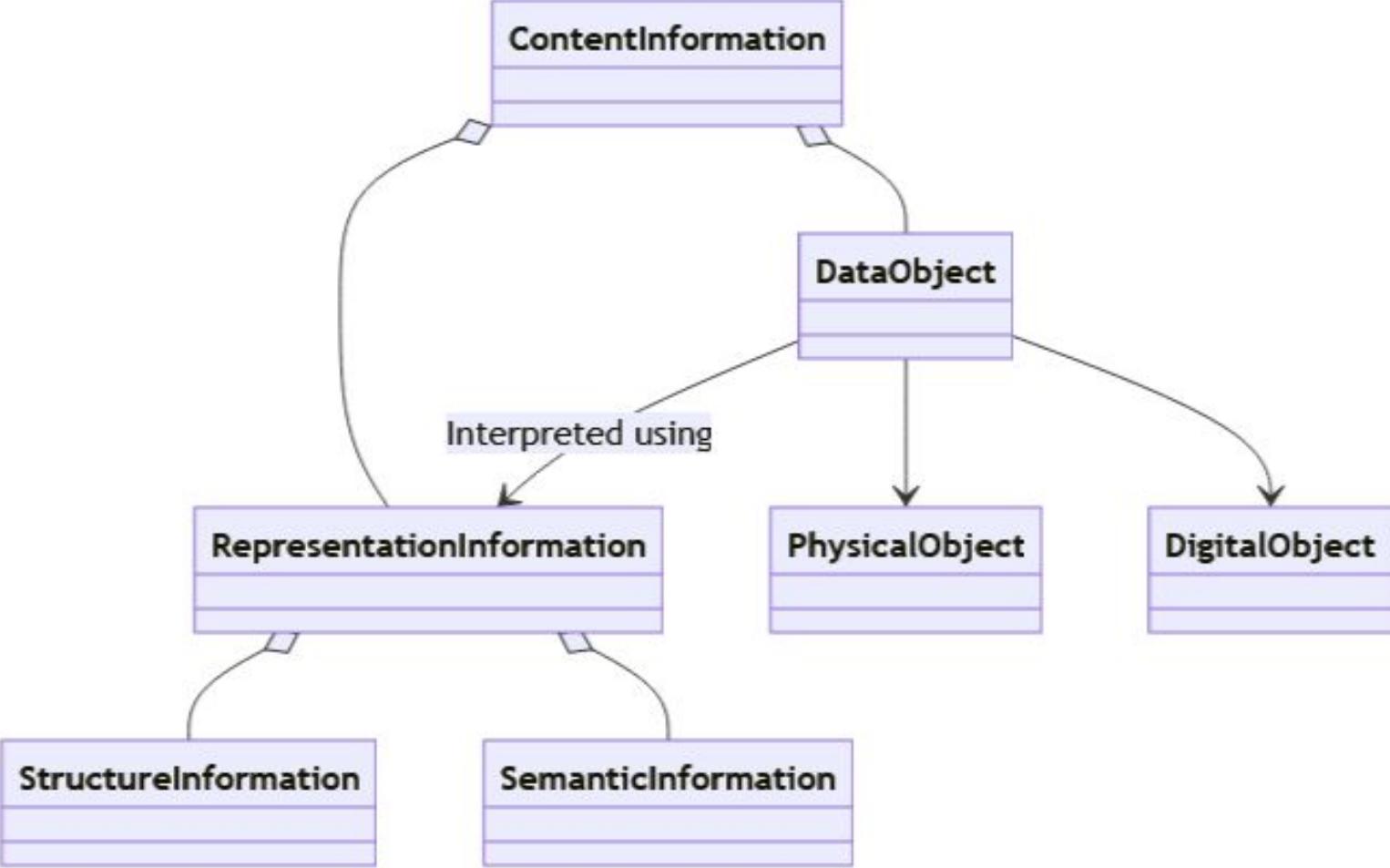
Welche Metadaten könnte es bei einem Buch geben?

- Name des Autors
- Auflage
- Erscheinungsjahr
- Verlag
- ISBN-Nummer

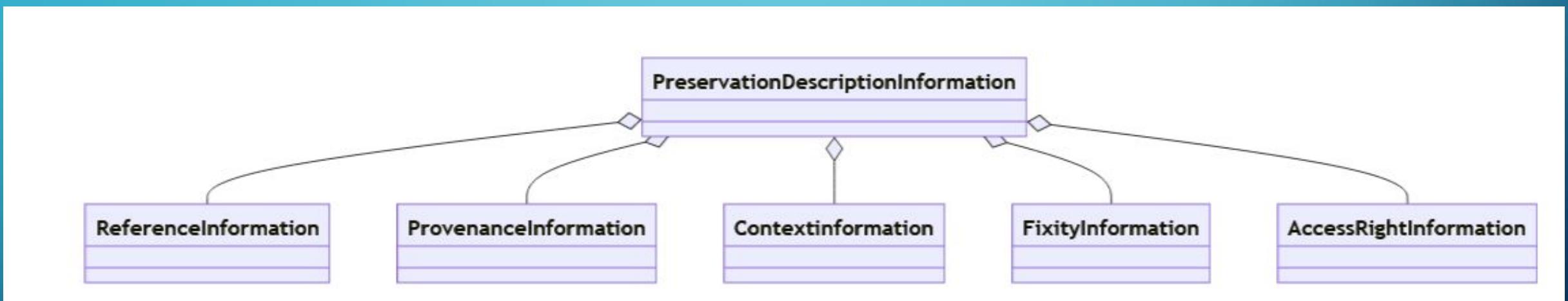
ARTEN VON METADATEN

- Beschreibende Metadaten
 - beschreiben eine Ressource für die Zwecke der Entdeckung und der Identifizierung
- Strukturelle Metadaten
 - z.B. Datenmodelle und Referenzdaten
- Administrative Metadaten
 - bieten Informationen zur Verwaltung einer Ressource

OASI REFERENZMODELL



OASI REFERENZMODELL



METADATENSTANDARDS

- Interoperable Daten
 - Daten aus unterschiedlichen Quellen
 - ermöglichen gemeinsames Arbeiten
- Standards für interoperable Daten
 - Sicherstellen der Nutzbarkeit

BEISPIEL FÜR EINEN METADATENSTANDARD: DOI

- DOI = digital object identifiers
- Standard für bibliographische Beschreibung von Forschungsdaten
- Verfügbarkeit in XML-Format

BEISPIEL FÜR EINEN METADATENSTANDARD: DOI

- Vorgaben:

- Verpflichtende Informationen zu einem Datensatz
 - Autor
 - Titel
- Empfohlene Angaben
 - Fachbereich
 - Beschreibung
- Optionale Angaben
 - Finanzierung
 - Nutzungsrechte

METADATENSCHEMA

- Metadatenschemata sind Zusammenstellungen von Elementen zur Beschreibung von Daten.
- Da Informationen für das Auffinden, das Verständnis und die Nutzung von Daten essentiell sind, sollten standardisierte Metadatenschemata eine möglichst einheitliche und nachvollziehbare Beschreibung sicherstellen.
- Metadatenschemata legen fest, welche Informationen geliefert werden sollen.
- Disziplinunabhängige Schema
 - Dublin Core
 - MARC21
 - RADAR
 - MODS

METADATENSCHEMA DOKUMENTATION

- geht über die Beschreibung der Daten durch Metadaten hinaus
- tiefere Erschließung der Daten
- Betrachtung und ausführliche Beschreibung von:
 - Entscheidungskontexten
 - Variablen
 - Instrumente
 - Methoden

METADATENSCHEMA DOKUMENTATION

- Beschreibung der Daten unerlässlich
 - Überprüfbarkeit
 - Nachvollziehbarkeit
 - Nachnutzbarkeit

Metadaten

| Sichtungen von Sciurus vulgaris, Bearbeiter: Frank Forscher, Rachel Research | | | | |
|--|---------------------|------------|-------|-------|
| ID | Datum | Ort | Farbe | Größe |
| 10034 | 2017-04-29T17:03:07 | DE-NW-521 | r | 22 |
| 10035 | 2017-04-29T17:21:58 | DE-NW-505 | r | 24 |
| 10036 | 2017-04-29T17:44:23 | DE-RP-372 | b | 28 |
| 10037 | 2017-04-29T18:06:36 | GB-SCT-037 | b | 29 |
| 10038 | 2017-04-29T18:35:15 | GB-SCT-029 | r | 21 |
| 10039 | 2017-04-29T19:26:37 | DE-RP-312 | r | 22 |
| 10040 | 2017-04-29T19:47:09 | GB-WLS-317 | r | 25 |
| 10041 | 2017-04-30T06:42:26 | GB-SCT-014 | b | 26 |
| 10042 | 2017-04-30T06:58:11 | GB-SCT-117 | r | 25 |
| 10043 | 2017-04-30T07:39:34 | DE-SL-465 | b | 29 |
| 10044 | 2017-04-30T07:41:75 | DE-SL-497 | b | 27 |
| 10045 | 2017-05-01T10:46:02 | DE-NW-512 | b | 29 |

Daten

| Feld | Inhalt | Format | Referenz |
|-------|-------------------------------|-------------------------|--------------------------------|
| ID | laufende Nummer | integer | |
| Datum | Tag/Uhrzeit der Beobachtung | YYYY-MM-DD "T" hh:mm:ss | Notation nach ISO 8601 |
| Ort | ID der Beobachtungsstelle | string | siehe Blatt "Data-Messstellen" |
| Farbe | Farbe des Eichhörnchens | b=black, r=redbrown | |
| Größe | Größe des Eichhörnchens in cm | cm | |

Metadaten-schema

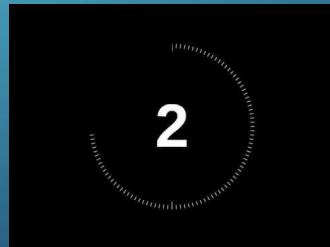
AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

Schaut euch den nachfolgenden Datensatz an:

<https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>

Versuche nun zunächst die folgenden allgemeinen Fragen zu beantworten:

- Wer besitzt die Daten
- Woher kommen die Daten?
- Zugriffrechte der Daten / Welche Lizenz?
- Wo werden die Daten genutzt/geteilt?
- Benutzung / Verwendung der Daten?
- Wann wurden die Daten erstellt/geändert?
- Datum der einzelnen Messungen vorhanden?



AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

- Wer besitzt die Daten
- Woher kommen die Daten?
- Zugriffrechte der Daten / Welche Lizenz?

AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

- Wer besitzt die Daten
 - Svetlana Ulianova (Owner)
- Woher kommen die Daten?
- Zugriffrechte der Daten / Welche Lizenz?

AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

- Wer besitzt die Daten
 - Svetlana Ulianova (Owner)
- Woher kommen die Daten?
 - Alle Datensatzwerte wurden zum Zeitpunkt der ärztlichen Untersuchung erhoben
- Zugriffrechte der Daten / Welche Lizenz?

AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

- Wer besitzt die Daten
 - Svetlana Ulianova (Owner)
- Woher kommen die Daten?
 - Alle Datensatzwerte wurden zum Zeitpunkt der ärztlichen Untersuchung erhoben
- Zugriffrechte der Daten / Welche Lizenz?
 - Public / Unknown

AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

- Wo werden die Daten genutzt/geteilt?
- Benutzung / Verwendung der Daten?
- Wann wurden die Daten erstellt/geändert?
- Datum der einzelnen Messungen vorhanden?

AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

- Wo werden die Daten genutzt/geteilt?
 - Kaggle
- Benutzung / Verwendung der Daten?
- Wann wurden die Daten erstellt/geändert?
- Datum der einzelnen Messungen vorhanden?

AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

- Wo werden die Daten genutzt/geteilt?
 - Kaggle
- Benutzung / Verwendung der Daten?
 - Klassifizierung von Herzkrankheiten
- Wann wurden die Daten erstellt/geändert?
 - Datum der einzelnen Messungen vorhanden?

AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

- Wo werden die Daten genutzt/geteilt?
 - Kaggle
- Benutzung / Verwendung der Daten?
 - Klassifizierung von Herzkrankheiten
- Wann wurden die Daten erstellt/geändert?
 - 2019
- Datum der einzelnen Messungen vorhanden?

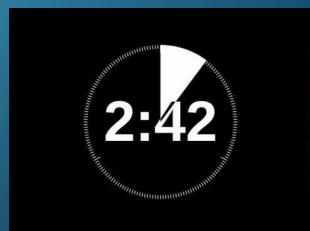
AUFGABE 2: ERSTE EXPLORATIVE ANALYSE DES DATENSATZES

- Wo werden die Daten genutzt/geteilt?
 - Kaggle
- Benutzung / Verwendung der Daten?
 - Klassifizierung von Herzkrankheiten
- Wann wurden die Daten erstellt/geändert?
 - 2019
- Datum der einzelnen Messungen vorhanden?
 - Nicht vorhanden

AUFGABE 3: METADATENSCHEMA AUSFÜLLEN

Analysiere nun den Datensatz genauer, welche Informationen kannst du finden? Erstelle hierzu eine Tabelle welche alle potentiellen Metadaten und deren Informationen enthält.

| Feld/Name | Inhalt | Format | Referenz | Zusatzinformationen |
|-----------|-------------------------|--------|----------|-----------------------|
| age | Age | int | - | Age specified in days |
| ap_hi | Systolic blood pressure | int | - | - |



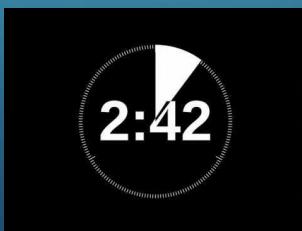
LÖSUNG AUFGABE 3: METADATENSCHHEMA AUSFÜLLEN

| Feld/Name | Inhalt | Format | Referenz | Zusatzinformationen |
|-------------|---|--------------------------|----------|--|
| age | Age | int | - | Age specified in days |
| ap_hi | Systolic blood pressure | int | - | - |
| height | Height | int | - | Specification in cm |
| weight | Weight | float | - | Specification in kg |
| gender | Gender | int (categorical code) | - | 1: women, 2: men |
| ap_lo | Diastolic blood pressure | int | - | - |
| cholesterol | Cholesterol | int (categorical code) | - | 1: normal, 2: above normal, 3: well above normal |
| gluc | Glucose | int (categorical code) | - | 1: normal, 2: above normal, 3: well above normal |
| smoke | Smoking | int (binary code) | - | - |
| alco | Alcohol intake | int (binary code) | - | - |
| active | Physical activity | int (binary code) | - | - |
| cardio | Presence or absence of cardiovascular disease | int (binary target code) | - | - |
| id | ID Number | int | - | - |

AUFGABE 4: WEITERE “VERSTECKTE” METADATEN

Es gibt noch weitere Daten, die nicht direkt ersichtlich sind, sich aber aus dem Datensatz generieren lassen.

1. Welche Daten könnte man aus den Spalten “Körpergröße” und “Gewicht” ermitteln?
2. Öffne das [Notebook](#) und füge zu dem Datensatz eine weitere Spalte mit den neuen Daten hinzu.



LÖSUNG - AUFGABE 4: WEITERE “VERSTECKTE” METADATEN

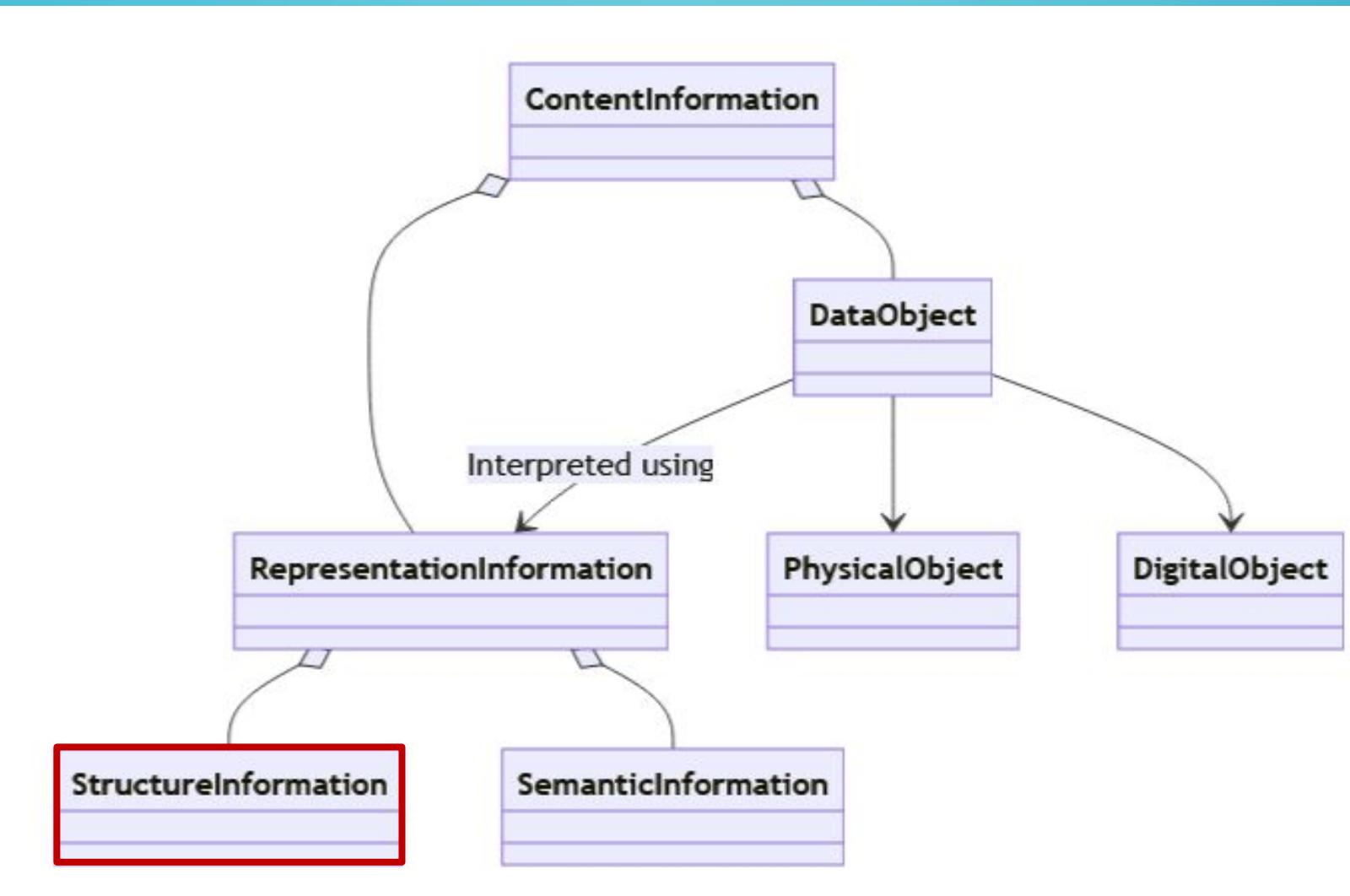
1. Welche Daten könnte man aus den Spalten “Körpergröße” und “Gewicht” ermitteln?
 - BMI (Body-Mass-Index)
 - $\text{BMI} = \text{Körpergewicht [kg]} / (\text{Körpergröße [m]})^2$

LÖSUNG - AUFGABE 4: WEITERE “VERSTECKTE” METADATEN

2. Öffne das Notebook und füge zu dem Datensatz eine weitere Spalte mit den neuen Daten hinzu.

```
# Füge hier deine Lösung ein  
  
# BMI = Körpergewicht [kg] / (Körpergröße [m])2  
df['bmi'] = df['weight'] / pow(df['height']/100,2)  
df.head(10)
```

| | age | height | weight | ap_hi | ap_lo | smoke | cardio | bmi |
|----|-----|--------|--------|-------|-------|-------|--------|-----------|
| id | | | | | | | | |
| 0 | 50 | 168 | 62.0 | 110 | 80 | 2 | 5 | 21.967120 |
| 1 | 55 | 156 | 85.0 | 140 | 90 | 0 | 5 | 34.927679 |
| 2 | 51 | 165 | 64.0 | 130 | 70 | 1 | 6 | 23.507805 |
| 3 | 48 | 169 | 82.0 | 150 | 100 | 1 | 0 | 28.710479 |
| 4 | 47 | 156 | 56.0 | 100 | 60 | 0 | 9 | 23.011177 |
| 8 | 59 | 151 | 67.0 | 120 | 80 | 2 | 3 | 29.384676 |
| 9 | 60 | 157 | 93.0 | 130 | 80 | 2 | 0 | 37.729725 |
| 12 | 61 | 178 | 95.0 | 130 | 90 | 0 | 5 | 29.983588 |
| 13 | 48 | 158 | 71.0 | 110 | 70 | 2 | 5 | 28.440955 |
| 14 | 54 | 164 | 68.0 | 110 | 60 | 2 | 4 | 25.282570 |



STRUCTURE INFORMATION, DEFINITION BY OPEN ARCHIVAL INFORMATION SYSTEM:

It does this by **describing the format**, or **data structure concepts**, which are to be applied to the bit sequences and that in turn result in more meaningful values such as characters, numbers, pixels, arrays, tables, etc.

STRUCTURE INFORMATION, DEFINITION BY OPEN ARCHIVAL INFORMATION SYSTEM:

These common computer data types, aggregations of these data types, and mapping rules which map from the underlying data types to the higher level concepts needed to understand the Digital Object are referred to as the Structure Information of the Representation Information object.

STRUCTURE INFORMATION, DEFINITION BY OPEN ARCHIVAL INFORMATION SYSTEM:

- Beispiel
 - Ein Verweis auf den ASCII-Standard (ISO 9660), um Bits in Characters zu interpretieren.
 - Ein Verweis auf ISO/TS 22028-4 (Digital images encoded using eciRGB) um Bits in Bilder zu interpretieren.

DATA EXPLORATION

Original Daten im CSV können, wie angesprochen unter

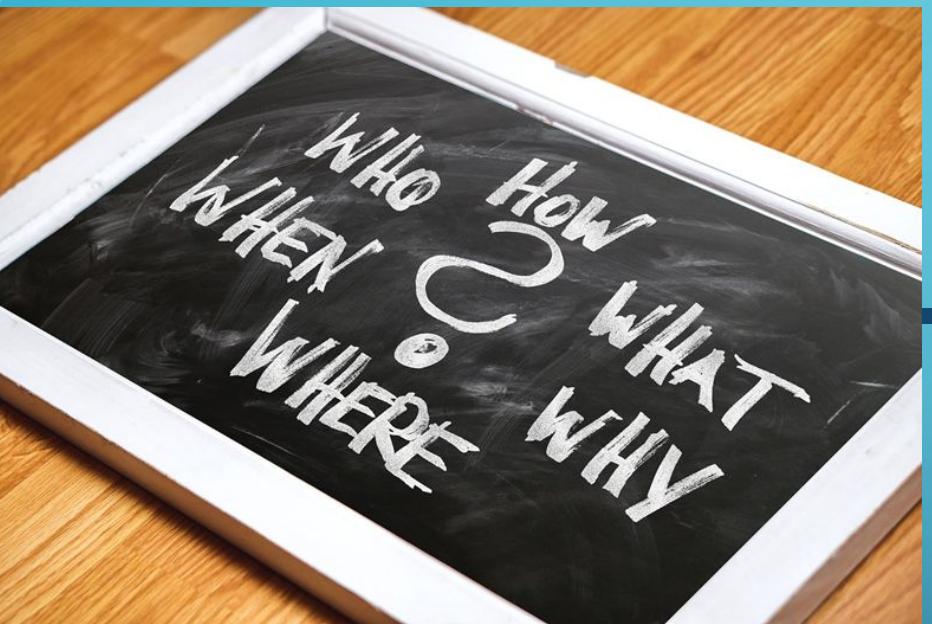
<https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset> gefunden werden.

- Features:

- Age | Objective Feature | age | int (days)
- Height | Objective Feature | height | int (cm)
- Weight | Objective Feature | weight | float (kg)
- Systolic blood pressure | Examination Feature | ap_hi | int
- Diastolic blood pressure | Examination Feature | ap_lo | int
- Smoking | Subjective Feature | smoke | Binär
- Presence or absence of cardiovascular | Binär

WAS BEDEUTEN DIESE DATEN?





Schema →



METADATENSCHEMA

Generische: Dublin Core, MODS (Metadata Object Description Schema) sind in der Regel einfach zu verwenden und weit verbreitet, müssen jedoch häufig erweitert werden, um spezifischere Informationen abzudecken.

METADATENSCHEMA

Domänenspezifische: Haben ein viel reichhaltiges Vokabular und eine viel umfangreichere Struktur, sind jedoch in der Regel hochspezialisiert und nur für Forscher auf diesem Gebiet verständlich. [Beispiele hier](#)

DUBLIN CORE

- Dublin Core geht auf die sogenannte Dublin Core Metadata Initiative(DCMI)
- 1995 in Dublin einheitliche Standards zur Auszeichnung von Metaangaben definiert.
- Ziel: Suchmaschinen das Durchsuchen von Dokumenten zu erleichtern, indem wichtige Inhalte bereits in den Metadaten hinterlegt sind.
- Heute werden die Standards von einer Gruppe aus Freiwilligen weiter bearbeitet.

EINTEILUNG IN 15 CORE ELEMENTS

- contributor
- coverage
- creator
- date
- description
- format
- identifier
- language
- publisher
- relation
- rights
- source

NAME SPACES

- abstract, accessRights, accrualMethod
- accrualPeriodicity, accrualPolicy
- alternative, audience, available
- bibliographicCitation, conformsTo, contributor
- coverage, created, creator,
- date, dateAccepted, dateCopyright
- dateSubmitted, description, educationLevel
- extent, format, hasFormat
- hasPart, hasVersion, identifier
-

CORE ELEMENTS

- **contributor (Beitragende) WHO?:**

Nennen der Person(en) oder Organisation(en), die bei der Erstellung der Ressource (Content) mitgewirkt haben.

- **coverage (Ort und Zeit) WHERE/WHEN?:**

An dieser Stelle werden Informationen zum Ort und zeitlichen Gültigkeitsbereich abgelegt. Hierbei sollen für Orte die gültigen Namen und für die temporäre Dauer Zeitintervalle (z.B. 07.07 - 12.07) verwendet werden.

CORE ELEMENTS

- **creator (Ersteller)** WHO?:

Nennen des ursprünglichen Autors einer Ressource. Autoren können Person(en) und Organisation(en) sein.

- **date (Datum)** WHEN?:

Hinterlegen von Informationen bezüglich Erstellungsdatum, Änderungsdatum, Sperrfrist und Löschdatum.

CORE ELEMENTS

- **description (Beschreibung)** WHY/WHAT?:

Zusätzliche Informationen, die die Ressource noch näher beschreiben. Hierzu zählen z.B. eine Kurzfassung oder ein Inhaltsverzeichnis.

- **format (Format)** WHAT/HOW?:

Angaben zum MIME-Typ der Ressource wie Pixelgröße, Dateiformat, Bearbeitungsdauer, usw.

CORE ELEMENTS

- **identifier (Identifizierer) WHAT?:**

Dieses Element enthält einen eindeutigen Bezeichner für die Ressource z.B. eine URL([DOI](#)), Artikelnummer oder UID.

- **language (Sprache) WHAT/HOW?:**

Hinterlegen eines Sprachecodes. Hierfür sollen Sprachcodes nach [ISO 639](#) oder RFC 3066 verwendet werden.

CORE ELEMENTS

- **publisher (Verlag/Herausgeber) WHO?:**
Enthält Informationen über den Verleger. Der Verleger können Person(en) oder Organisation(en) sein.
- **relation (Beziehungen) WHAT?:**
Hier werden Informationen über Beziehungen zu anderen Ressourcen festgehalten.

CORE ELEMENTS

- **rights (Rechte)** WHO/WHERE?:

An dieser Stelle werden Informationen bezüglich den Rechten an Ressourcen hinterlegt. Zum Beispiel über den Urheber oder die Lizenzart (GPL, LGPL, ZPL usw.)

- **source (Quelle)** WHAT?:

Eine verwandte Ressource, von der die beschriebene Ressource abgeleitet ist. Die beschriebene Ressource kann ganz oder teilweise von der verwandten Ressource abgeleitet sein.

CORE ELEMENTS

- **subject (Stichwörter) WHAT?:**

Hier können Stichwörter oder ganze identifizierende Phrasen zu einer Ressource hinterlegt werden.

- **title (Titel) WHAT?:**

Hinterlegen des Ressourcentitels (z.B. Dokumenttitel).

- **type (Typ) WHAT/HOW?:**

Über den Typ wird einer Ressource eine Medienkategorie wie Bild, Artikel, Ordner usw. zugeordnet.

AUFGABE 5

Fassen Sie die noch fehlenden "Core Elemente" unter der Verwendung der verlinkten Codierung-Standards für den vorgestellten Datensatz zusammen.
Zusätzliche Informationen zum Datensatz können [hier](#) entnommen werden.

Beispielhafte Codierung-Standards:

- [Thesaurus of Geographic Names \(TGN\)](#)
- [Date and Time Formats](#)
- [Media types](#)
- [Codes for the Representation of Names of Languages \(ISO 639-2\)](#)
- [List of popular Licenses](#)

Geschichte

Der Chefarzt Dr. Müller aus unsere Geschichte ist für die Erstellung des zusammengeführten Datensatzes und für seiner Klinik Mannheim verantwortlich. Das Forschungsprojekt läuft unter der Organisation "Daten ohne Grenzen", welcher die Rechte über die Daten hält und diese unter der MIT-License veröffentlichen will. Zusätzlich hat die Organisation folgenden Regularien für die Daten aufgestellt:

- Daten müssen in Englisch codiert/beschrieben werden
- Daten müssen nach 10 Jahren ab Erstellungsdatum gelöscht werden
- Zentrale Speicherung der Daten erfolgt in Mannheim

Lösung AUFGABE 5

Fassen Sie die noch fehlenden "Core Elemente" unter der Verwendung der verlinkten Codierung-Standards für den vorgestellten Datensatz zusammen.
Zusätzliche Informationen zum Datensatz können [hier](#) entnommen werden.

Beispielhafte Codierung-Standards:

- Thesaurus of Geographic Names (TGN) ⇒ ID: 7005179 (<http://vocab.getty.edu/page/tgn/7005179>)
- Date and Time Formats ⇒ YYYY-MM-DD ⇒ “2021-05-18” ⇒ “2021-05-28”
- Media types ⇒ `text` ⇒ `csv`
- Codes for the Representation of Names of Languages (ISO 639-2) ⇒ `deu`
- List of popular Licenses ⇒ “copyright owner” ⇒ MIT license

XML-SCHEMA

- Guidlines für XML- & RDF- Format
- Unterschied **Simple** und **Qualified** Dublin Core

SIMPLE DUBLIN CORE

- Besteht aus einer oder mehreren Eigenschaften und den zugehörigen Werten.
- Jede Eigenschaft ist ein Attribut der beschriebenen Ressource.
- Jede Eigenschaft muss eines der 15 DCMES [DCMES]-Elemente sein.
- Eigenschaften können wiederholt werden.
- Jeder Wert ist ein String.
- Jeder String-Wert kann eine zugeordnete Sprache haben (z.B. en-GB).

```
<?xml version="1.0"?>

<metadata
  xmlns="http://example.org/myapp/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://example.org/myapp/ http://example.org/myapp/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <dc:title>
    UKOLN
  </dc:title>
  <dc:description>
    UKOLN is a national focus of expertise in digital information management. It provides policy, research and awareness services to the UK library, information and cultural heritage communities. UKOLN is based at the University of Bath.
  </dc:description>
  <dc:publisher>
    UKOLN, University of Bath
  </dc:publisher>
  <dc:identifier>
    http://www.ukoln.ac.uk/
  </dc:identifier>

</metadata>
```

QUALIFIED DUBLIN CORE

- Besteht aus einer oder mehreren Eigenschaften und den zugehörigen Werten. ✓
- Jede Eigenschaft ist ein Attribut der beschriebenen Ressource. ✓
- Jede Eigenschaft muss entweder:
 - eines der 15 DC-Elemente, ✓
 - eines der anderen vom DCMI empfohlenen Elemente (z. B. Publikum) [DCTERMS],
 - eine der Elementverfeinerungen, die in der Empfehlung der DCMI-Metadatenbedingungen [DCTERMS] aufgeführt sind.

QUALIFIED DUBLIN CORE

- Eigenschaften können wiederholt werden. ✓
- Jeder Wert ist eine String. ✓
- Jeder Wert kann ein zugeordnetes Codierungsschema haben.
- Jedes Kodierungsschema hat einen Namen.
- Jeder String-Wert kann eine zugeordnete Sprache haben (z. B. en-GB). ✓

QUALIFIED DUBLIN CORE

```
<?xml version="1.0"?>

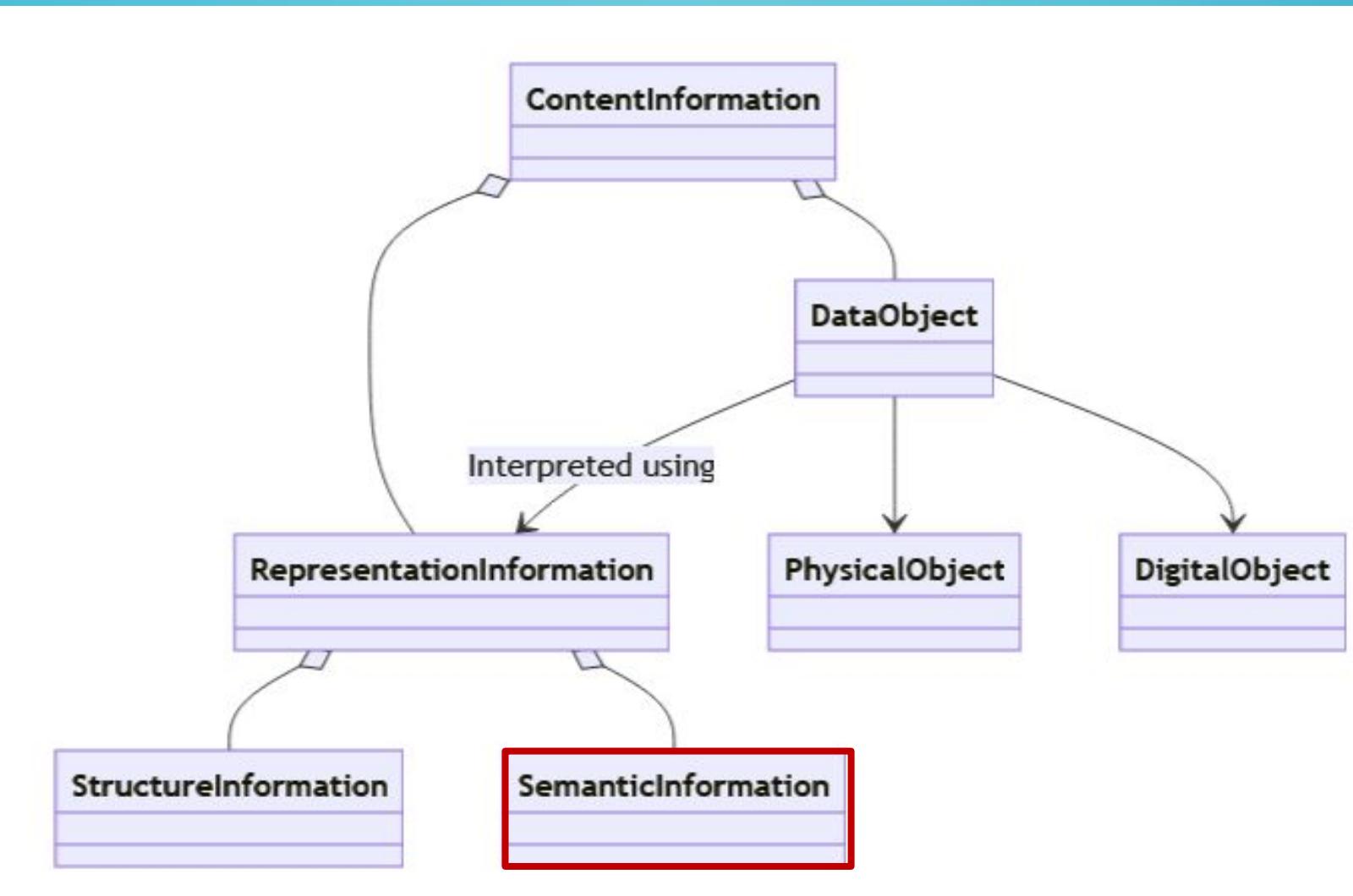
<metadata
    xmlns="http://example.org/myapp/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://example.org/myapp/ http://example.org/myapp/schema.xsd"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/">

    <dc:title>
        UKOLN
    </dc:title>
    <dcterms:alternative>
        UK Office for Library and Information Networking
    </dcterms:alternative>
    <dc:subject>
        national centre, network information support, library
        community, awareness, research, information services,public
        library networking, bibliographic management, distributed
        library systems, metadata, resource discovery,
        conferences,lectures, workshops
    </dc:subject>
    <dc:subject xsi:type="dcterms:DDC">
        062
    </dc:subject>
    <dc:subject xsi:type="dcterms:UDC">
        061(410)
    </dc:subject>
```

```
    <dc:description>
        UKOLN is a national focus of expertise in digital information
        management. It provides policy, research and awareness services
        to the UK library, information and cultural heritage communities.
        UKOLN is based at the University of Bath.
    </dc:description>
    <dc:description xml:lang="fr">
        UKOLN est un centre national d'expertise dans la gestion de l'inf
        ormation
        digitale.
    </dc:description>
    <dc:publisher>
        UKOLN, University of Bath
    </dc:publisher>
    <dcterms:isPartOf xsi:type="dcterms:URI">
        http://www.bath.ac.uk/
    </dcterms:isPartOf>
    <dc:identifier xsi:type="dcterms:URI">
        http://www.ukoln.ac.uk/
    </dc:identifier>
    <dcterms:modified xsi:type="dcterms:W3CDTF">
        2001-07-18
    </dcterms:modified>
    <dc:format xsi:type="dcterms:IMT">
        text/html
    </dc:format>
    <dcterms:extent>
        14 Kbytes
    </dcterms:extent>

```

```
</metadata>
```



SEMANTISCHE DATEN

Daten werden für einen menschlichen Benutzer erst durch Interpretation zu Information, und diese Interpretation erfordert sowohl Wissen über die konkrete Anwendung als auch Weltwissen.

~ *Informationsintegration : Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen / Ulf Leser, Felix Naumann*

SEMANTISCHE DATEN

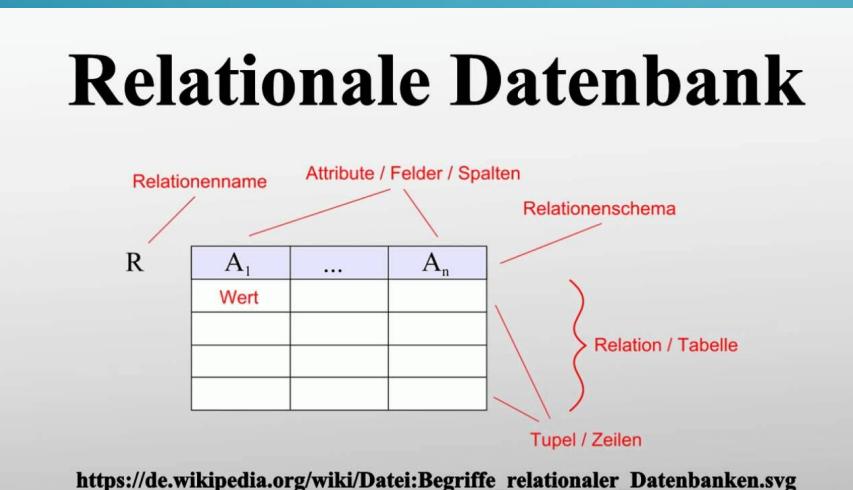
Wir fassen diese Informationen unter dem Begriff Kontext zusammen. Daten erhalten ihre Bedeutung erst durch Berücksichtigung des Kontextes. Dabei ist zu beachten, dass manche Teile des Kontextes unmittelbar in computerlesbarer Form vorliegen, wie das Schema, während andere Teile nicht explizit modelliert wurden und für Programme nicht erreichbar sind, wie das externe Wissen über die Anwendung

~ *Informationsintegration : Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen / Ulf Leser, Felix Naumann*

SEMANTISCHE DATEN

Da eine Menge strukturierter Daten durch die Namen der Strukturelemente beschrieben wird, sind Schemata Metadaten für Instanzen dieses Schemas; Metadaten zu einem Schema wiederum sind das Datenmodell, in dem das Schema erstellt ist

~ Informationsintegration : Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen / Ulf Leser, Felix Naumann



| Sichtungen von Sciurus vulgaris, Bearbeiter: Frank Forscher, Rachel Research | | | | |
|--|---------------------|------------|-------|-------|
| ID | Datum | Ort | Farbe | Größe |
| 10034 | 2017-04-29T17:03:07 | DE-NW-521 | r | 22 |
| 10035 | 2017-04-29T17:21:58 | DE-NW-505 | r | 24 |
| 10036 | 2017-04-29T17:44:23 | DE-RP-372 | b | 28 |
| 10037 | 2017-04-29T18:06:36 | GB-SCT-037 | b | 29 |
| 10038 | 2017-04-29T18:35:15 | GB-SCT-029 | r | 21 |
| 10039 | 2017-04-29T19:26:37 | DE-RP-312 | r | 22 |
| 10040 | 2017-04-29T19:47:09 | GB-WLS-317 | r | 25 |
| 10041 | 2017-04-30T06:42:26 | GB-SCT-014 | b | 26 |
| 10042 | 2017-04-30T06:58:11 | GB-SCT-117 | r | 25 |
| 10043 | 2017-04-30T07:39:34 | DE-SL-465 | b | 29 |
| 10044 | 2017-04-30T07:41:75 | DE-SL-497 | b | 27 |
| 10045 | 2017-05-01T10:46:02 | DE-NW-512 | b | 29 |

Metadaten

| Feld | Inhalt | Format | Referenz |
|-------|-------------------------------|-------------------------|--------------------------------|
| ID | laufende Nummer | integer | |
| Datum | Tag/Uhrzeit der Beobachtung | YYYY-MM-DD "T" hh:mm:ss | Notation nach ISO 8601 |
| Ort | ID der Beobachtungsstelle | string | siehe Blatt "Data-Messstellen" |
| Farbe | Farbe des Eichhörnchens | b=black, r=redbrown | |
| Größe | Größe des Eichhörnchens in cm | cm | |

Daten

Metadaten-schema

AUFGABE 6

1. WAS VERSTEHT IHR UNTER DEM BEGRIFF SEMANTIK?
2. WIE STEHEN SEMANTIK UND METADATEN IN BEZUG ZUEINANDER?

HINWEIS: SEMANTIK = INTERPRETATION VON DATEN

LÖSUNG AUFGABE 6

1. WAS VERSTEHT IHR UNTER DEM BEGRIFF SEMANTIK?

- im Prinzip ein “Verständnis” der Daten → der Sinn/Kontext und Inhalt der Daten

2. WIE STEHEN SEMANTIK UND METADATEN IN BEZUG ZUEINANDER?

- Semantische Metadaten verleihen den “Daten” eine Bedeutung sodass diese nicht nur durch einen Menschen, sondern auch einer Maschine Interpretiert werden können. Es wird somit ein Kontext geschaffen der kein menschliches Vorwissen benötigt bzw. voraussetzt.

AUFGABE 7

1. Welche Datenfelder weisen keinen semantischen Hintergrund auf?
 - <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>

LÖSUNG AUFGABE 7

1. Welche Datenfelder weisen keinen semantischen Hintergrund auf?
 - <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>
 - Smoking → Wer raucht? Wie viel wird geraucht? Was wird geraucht?
 - Alcohol intake → Wer trinkt? Wie viel wird getrunken? Was wird getrunken?
 - Physical activity → Wer? Was ? Wie viel?

KODIERUNGSSYSTEME

- Ziel ist eine Standardisierung von medizinischen Daten und somit dem Datenaustausch
- Ermöglichen eine konsistente und einheitliche Darstellung
- Nomenklaturen, Ontologien und Klassifizierungen sind das Werkzeug
- Bekannte und wichtige Kodierungssysteme
 - Ontologien/Datenbanken/Klassifikationsliste:
 - SNOMED (<https://www.snomed.org/>) □ Ontologie
 - LOINC (<https://loinc.org/>) □ Datenbank
 - ICD10 (<https://www.icd-code.de/icd/code/ICD-10-GM.html>) □ Klassifikationsliste

LOINC

- Internationaler Standard zur Festhaltung von Untersuchungs- und Testergebnissen aus Labor und Klinik
- 1994 wurde die Datenbank Logical Observation Identifiers Names and Codes (LOINC) aufgebaut
- Abkürzungen, Bezugswörter, Synonyme und Kommentare für alle Untersuchungen
- Viele zusätzliche Informationen wie z.B Maßeinheiten
- Zum Beispiel die Palpation (Abtastung) des Pulses : <https://loinc.org/8865-8/>

AUFGABE 8

→ <https://loinc.org/55284-4/>

1. Wie lautet die Empfehlung von LOINC bzgl. des Blutdrucks?
2. Welche LOINC Codes finden sich jeweils für den systolischen und diastolischen Blutdruck?
3. In welchen Maßeinheiten kann man den Systolischen Blutdruck festhalten

AUFGABE 8

1. Wie lautet die Empfehlung von LOINC bzgl. des Blutdrucks?
 - Discouraged Nicht den systolischen und diastolischen Blutdruck zusammenfassen als ein Schemaelement
2. Welche LOINC Codes finden sich jeweils für den systolischen und diastolischen Blutdruck?
 - 8480-6 Systolic blood pressure
 - 8462-4 Diastolic blood pressure
3. In welchen Maßeinheiten kann man den Systolischen Blutdruck festhalten
 - mm[Hg] Example UCUM Units
 - Mm Hg REGENSTRIEF

ICD-10

- International Classification of Diseases 10th Revision (ICD-10)
- Wurde ab 1983 entworfen und 1990 von der WHO gebilligt
- Verwendung ab 1994
- Weltweit anerkanntes System zur Klassifizierung von Diagnosen im medizinischen Umfeld.
- Ursache, Manifestation, Ort, Schwere und Art der Verletzung oder Erkrankung werden festgehalten
- Hierarchisch aufgebaut
- In USA wird ICD-10-CM (eigene erweiterte Version genutzt)
 - Code W55.2 □ Contact with cow
- In Deutschland ICD-10-GM
 - T63.3 □ Spinnengift

AUFGABE 9

Unter folgendem Link findet Ihr eine Klassifizierung der Krankheiten:

<https://www.icd-code.de/icd/code/ICD-10-GM.html>

1. Versucht einen ICD-10 Code Wertebereich zu finden der allgemein Herzkrankheiten beschreibt. Welche Wertebereich findet sich?
2. Der Datensatz beschreibt somit nur allgemein eine Herzkrankheit aber nicht genau welche. Welcher ICD-10 Code würde eine Primäre Rechtsherzinsuffizienz beschreiben?

Hinweis = Herzkreislaufkrankheiten (Cardiovascular Disease) zählen zu Krankheiten des Kreislaufsystems.

AUFGABE 9

Unter folgendem Link findet Ihr eine Klassifizierung der Krankheiten:

<https://www.icd-code.de/icd/code/ICD-10-GM.html>

1. Versucht einen ICD-10 Code Wertebereich zu finden der allgemein Herzkrankheiten beschreibt. Welche Wertebereich findet sich?
 - I00-I99
 2. Der Datensatz beschreibt somit nur allgemein eine Herzkrankheit aber nicht genau welche. Welcher ICD-10 Code würde eine Primäre Rechtsherzinsuffizienz beschreiben?
 - I50.00
- Hinweis = Herzkreislaufkrankheiten (Cardiovascular Disease) zählen zu Krankheiten des Kreislaufsystems.

ONTOLOGIE

- Ontologien definieren ein Vokabular mit dem alle Konzepte der Anwendung beschrieben werden können
 - Konzept steht für ein bestimmtes Ding oder eine Klasse von Dingen
 - Z.b Jaguar bezeichnet eine Automarke und ein Raubtier
 - Somit auch individuell
- Durch Konzeptualisierung wird das Vokabular festgelegt,
- Dazu zählen auch Beziehungen der Elemente zueinander

SNOMED CT

- Systematisierte Ontologie der Medizin (SNOMED)
 - Anwendung von standardisierten medizinischen Begriffen
 - Ursprünglich eine Nomenklatur aber mittlerweile eine Ontologie
 - Erste Testversion im Jahre 1974
 - Veröffentlicht in 1979
 - Seit 1984 in Deutschland
 - Verweise zu ICD-10 und LOINC
- <https://bioportal.bioontology.org/ontologies/SNOMEDCT/?p=classes&conceptid=root>

AUFGABE 10

→ <https://bioportal.bioontology.org/ontologies/SNOMEDCT/?p=classes&conceptid=root>

1. Durchforstet die Ontologie nach "Tobacco use and exposure".
2. Versucht durch die Basisklasse "Tobacco use and exposure" die Subklasse "Smoker" zu finden und notiert bitte die ID der Klasse.
3. Würdet Ihr sagen das "Smoker" eine korrekte Klassifizierung darstellt oder könnte man noch das Element noch genauer definieren?
4. Wie müsste man in den Daten einen moderaten Raucher festhalten, der am Tag 10-19 Zigaretten raucht?

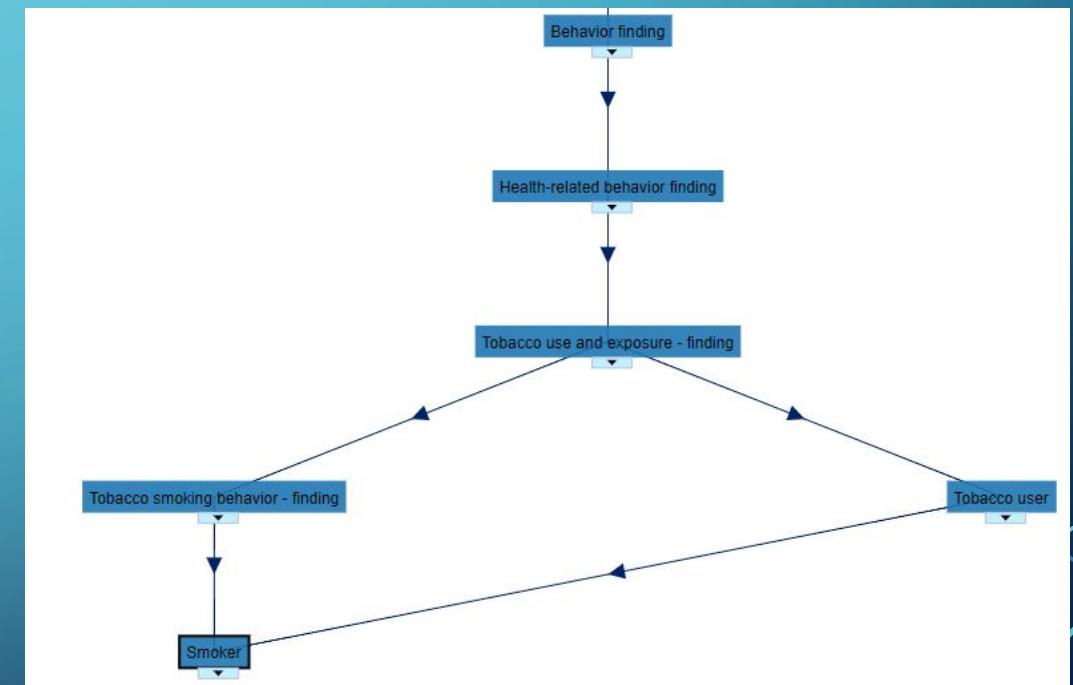
LÖSUNG AUFGABE 10

1. Durchforstet die Ontologie nach "Tobacco use and exposure".

- https://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes&conceptid=http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FSNOMEDCT%2F229819007&jump_to_nav=true

2. Versucht durch die Basisklasse "Tobacco use and exposure" die Subklasse "Smoker" zu finden und notiert bitte die ID der Klasse.

- <https://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes&conceptid=http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FSNOMEDCT%2F77176002#details>
- 77176002



LÖSUNG AUFGABE 10

3. Würdet Ihr sagen das "Smoker" eine korrekte Klassifizierung darstellt oder könnte man noch das Element noch genauer definieren?

- Es fehlt der Kontext

4. Wie müsste man in den Daten einen moderaten Raucher festhalten, der am Tag 10-19 Zigaretten raucht?

- Moderate cigarette smoker (10-19 cigs/day)
- <https://bioportal.bioontology.org/ontologies/SNOMEDCT/?p=classes&conceptid=http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FSNOMEDCT%2F160604004#details>

ZUSAMMENFASSUNG

- Heutzutage können Daten aus dem medizinischen Umfeld standardisiert und sehr präzise definiert werden
- Es ist wichtig den semantischen Kontext der Daten so genau wie möglich anzugeben,
 - Sonst besteht Freiraum für Interpretationen der Daten
 - Verlust des Kontexts
- Kodierungssysteme, Ontologien, Klassifizierungslisten und standardisierte Datenbank knüpfen an den Punkt an.
 - Ermöglichen es den Kontext der Daten klar zu definieren
 - Somit kann auch können maschinell beziehungsweise durch Anwendungen die Daten ausgelesen, interpretiert und ausgewertet werden können.

ZUSAMMENFASSUNG

- Loinc wird verwendet um klinische Ergebnisse standardisiert festzuhalten
 - Laborergebnisse
 - Klinische Observationen
- ICD-10 dient zur Klassifizierung des Krankheitsbildes, Symptomen, Verletzungen und Anzeichen
- SNOMED zur standardisierung von klinischen Bezeichnungen
 - Gibt den medizinischen Daten ein Kontext

QUIZTIME

The image shows a mobile phone on the left and a computer monitor on the right, both displaying a Kahoot! quiz interface.

Mobile Phone Screen (Left):

- Top bar: "1 von 1" and "Quiz".
- Middle text: "Wähle eine oder mehrere Antworten aus!"
- Four answer options:
 - Red square with white triangle (top-left)
 - Blue square with white diamond (top-right)
 - Yellow square with white circle (bottom-left)
 - Green square with white square (bottom-right)
- Bottom buttons: "Senden" (Send) and "0" (score).
- Bottom status: "Tester" and "0".

Computer Monitor Screen (Right):

- Top bar: "Ist dies eine gute Testfrage?" and a "Skip" button.
- Top right: A small clock icon showing approximately 10:10.
- Score: "16" in a purple circle on the left and "0 Answers" on the right.
- Question text: "Ist dies eine gute Testfrage?"
- Image: A large teal rectangle with the word "Kahoot!" in white.
- Answer options (each with a radio button):
 - Red bar: "▲ Ja, super!"
 - Blue bar: "◆ Definiere gut..."
 - Yellow bar: "● Mir egal :P"
 - Green bar: "■ Neee!"
- Bottom right: "81".
- Bottom status: "1/1".