**XML Projekt**

**für Datenbanksysteme ILV 2020**

**an der FH Wiener Neustadt**

Gruppe 4 (Brandstätter, Forsthuber, Jernej)

**Aufgabenstellung:**

Each group chooses a dataset from Open Data Austria - https://www.data.gv.at/ - and defines a use case (data based question)

• **Task 1:** Deploy the data from the file in an XML enabled database (MSSQL, Oracle, PostGre, …) **– Jernej**

• **Task 2:** Define an XML schema for the data (or a subset of interest) **– Jernej**

• **Task 3**: Produce via the SQL/XML support of your choosen RDBMS an XML version of it adhering to your schema **– Forsthuber**

• **Task 4:** Register your XML schema and afterwards store (validate) the related data in the native XML database exist - <http://exist-db.org/> **- Brandstätter**

• **Task 5**: answer your data based question using XQuery via API usage in a supported programming language **– Forsthuber**

Upload your documentation (PDF between 10 and 15 pages) including

• detailed documentation of each task so somebody else could re-do it

• issues and solutions on the way

**and source code all together in one zip file onto moodle!**

**Suche ein geeignetes Dataset über den Datenprovider Open Data Austria unter -** [**https://www.data.gv.at**](https://www.data.gv.at/)**.**

**Formuliere und definiere auf den Daten dementsprechende Use Cases, welches du mit Queries beantworten sollst.**

Unser Team hat sich einen Datenset über das Impfverhalten der Stadt Linz ausgesucht. Dieses Datenset bietet eine Aufzeichnung von 20 Jahren von über 28 möglichen Impfungen. In Zuge der Analyse der Datenstruktur sowie des Dateninhalts haben wir folgende Use Cases definiert, die es zu beantworten gilt:

1. Wie hat sich das Impfverhalten über alle Impfmöglichkeiten generell entwickelt?
2. Wo liegt der absolute Höchstwert von Impfungen & wann begann eine mögliche Trendwende nach unten oder umgekehrt?

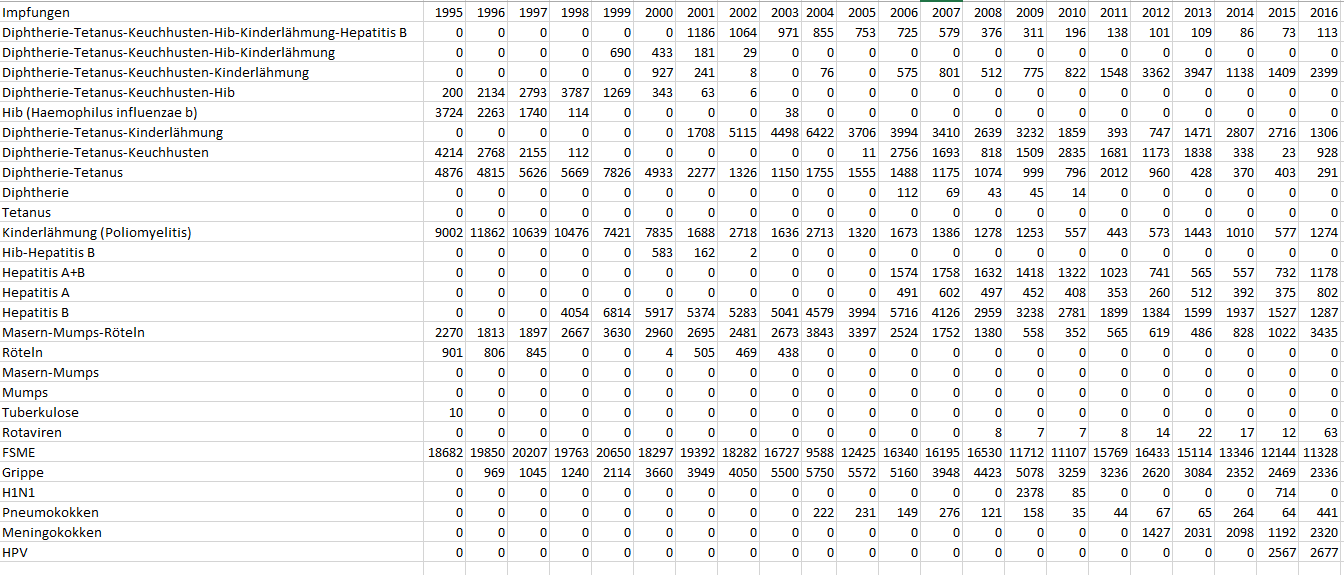
Github-Repo zum Projekt:

<https://github.com/ThePeziBear/XML_Project_Open_Data_Austria>

**Task 1: Erstellung einer XML fähigen Datenbank – Auswahl Postgres**

Um die von opendata heruntergeladenen Daten sinnvoll in eine Datenbank zu übertragen, wurden diese zuerst von einem wide auf ein long-Format umgewandelt.

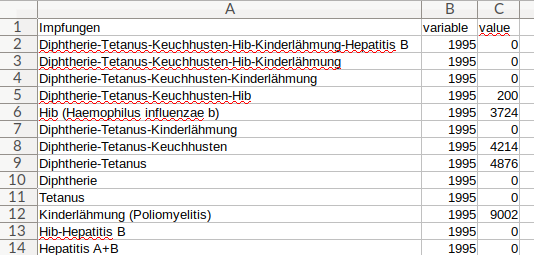
Auszug aus dem Datenset:



Transformieren mittels R-Code des Datenset für das Einlesen in Postgres:

|  |  |
| --- | --- |
|  |  |
|  | # File: impf\_recode.R |
|  | # Date: 3.03.2020 |
|  | # Copyright (C) 2020 |
|  | # Description: R-Script zum Einlesen und transformieren der Impfdaten |
|  |  |
|  | library(reshape2) |
|  |  |
|  | # Einlesen |
|  | impfdat <- read.csv2("./Impf.csv", fileEncoding = "Windows-1252") |
|  | # von wide auf long |
|  | impfdat\_long <- melt(impfdat) |
|  |  |
|  | # entfernen von X im Spaltennamen |
|  | impfdat\_long$variable <- as.character(impfdat\_long$variable) |
|  | impfdat\_long$variable <- gsub("X", "", impfdat\_long$variable) |
|  |  |
|  | # Schreiben ohne Index |
|  | write.csv(impfdat\_long, "./Impfung\_Recode.csv", row.names = FALSE) |

Output nach Transformation des Datensets in eine neue csv-Datei Impfung\_Recode.csv:



Die Datei wurde in utf-8 kodiert. Wie man sieht gibt es Umlaute, die später noch Probleme verursachen werden. Wir haben in zwei unterschiedlichen Betriebssystemen gearbeitet, Ubuntu und Windows. Unter Windows gibt es die Enkodierung Win-1252 unter Linux ist UTF-8 häufig Standard.

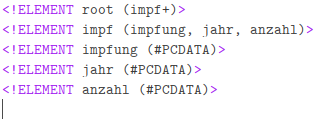
Nach erfolgreichem Setup der Installationsdatei von Postgres(pgAdmin4) wurde über die Konsole eine entsprechende Database und ein Table erstellt:

* Formularbeginn
* Formularende

|  |  |
| --- | --- |
|  | #Einloggen mit Benutzer 'postgres' |
|  | psql -U postgres  (Für Zugriff mit der Console muss die Systemsumgebungsvariable Path für PSQL erweitert werden. (C:/Programfiles/Postgres/bin) |
|  |  |
|  | #Passwort für Benutzer |
|  | Test |
|  |  |
|  | #Datenbank anlegen |
|  | CREATE DATABASE XML WITH ENCODING 'UTF8' LC\_COLLATE='German\_Germany' LC\_CTYPE='German\_Germany';  (Bei Fehlermeldung LC\_COLLATE='German-Austria & LC\_CTYPE='German\_Austria) |
|  |  |
|  | #scope auf erstellte db legen |
|  | \c xml |
|  |  |
|  | #Tabelle impf anlegen |
|  | CREATE TABLE impf (id serial PRIMARY KEY, impfung character varying(100), jahr smallint, anzahl Integer); (Für Zugriff von Postgres auf Datei muss der Zugriff unter Dateieigenschaften/Sicherheit ein neuer Benutzer "jeder" angelegt werden mit Schreib & Leserechte) |
|  |  |
|  | #csv file in die tabelle impf laden |
|  | COPY impf(impfung, jahr, anzahl) FROM 'C:/Users/Dell/Github/XML\_Projekt/Impfung\_Recode.csv' DELIMITER ',' CSV Header; |
|  |  |
|  | #alle daten aus impf anzeigen |
|  | SELECT \* FROM impf  Darstellung der Daten in pgAdmin4 unter Query Editor: |

Abgesehen vom erfolgreichen Import fallen die seltsamen Sonderzeichen auf → Enkodierungsproblem.

**Task 2: Definition eines XML-Schema für die Daten**



Das Schema wurde als Validierung\_Datenschema.dtd gespeichert. Nach einigen missglückten Versuchen der Validierung, wurde es etwas vereinfacht (Header entfernt).

**Task 3: Erstelle über den SQL/XML-Support der RDBMS eine XML-Version der Daten die dem XML-Schema entspricht.**

Query über Query Editor pgAdmin4:

select xmlelement(name impf, xmlelement(name impfung, impfung), xmlelement(name jahr, jahr), xmlelement(name anzahl, anzahl)) FROM impf;

Ausgabe der XML-Datei:

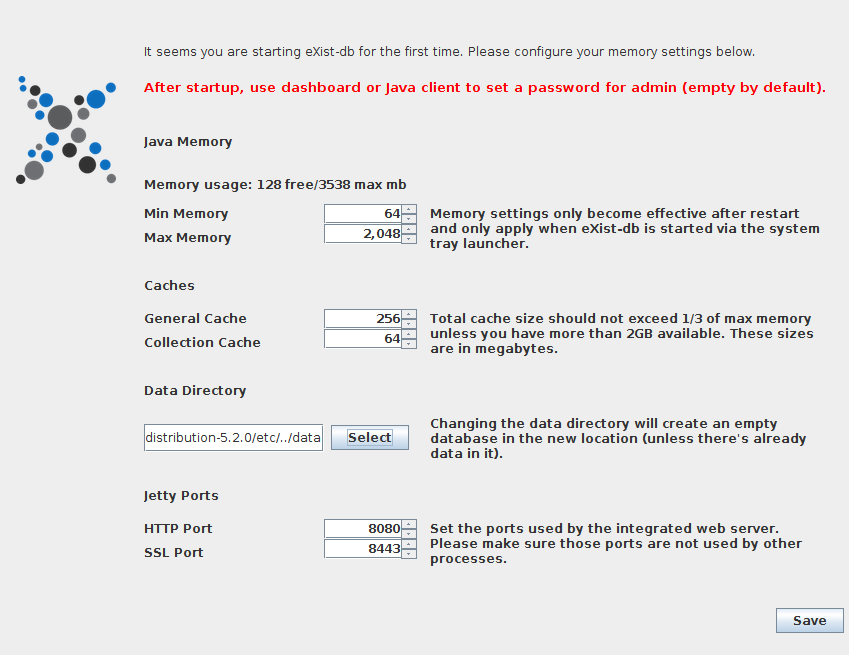
\copy (select table\_to\_xml('impfung', true, false, '')) TO '/home/christian/wd/fhwn/db\_systeme/XML\_Project\_Open\_Data\_Austria/result\_neu3.xml';

Output der XML-Datei (mit etwas manueller Nachbesserung):

|  |
| --- |
| Da es immer noch encoding-Probleme gab, wurden kurzerhand die Umlaute in ASCI-taugliche Formate ersetzt. Da die Validierung in der existdb auf „auto“ gesetzt wurde, musste das Validierungs-Schema mitgegeben werden. Außerdem wurde ein für den später notwendigen ExistDB-Import ein root-Tag hinzugefügt. <?xml version="1.0" encoding="UTF-8"?>  <!DOCTYPE root SYSTEM "Validierung\_Datenschema.dtd">  <root> |
| <impf><impfung>Diphtherie-Tetanus-Keuchhusten-Hib-Kinderlaehmung-Hepatitis B </impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Diphtherie-Tetanus-Keuchhusten-Hib-Kinderlaehmung</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Diphtherie-Tetanus-Keuchhusten-Kinderlaehmung</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Diphtherie-Tetanus-Keuchhusten-Hib</impfung><jahr>1995</jahr><anzahl>200</anzahl></impf> |
| <impf><impfung>Hib (Haemophilus influenzae b)</impfung><jahr>1995</jahr><anzahl>3724</anzahl></impf> |
| <impf><impfung>Diphtherie-Tetanus-Kinderlaehmung </impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Diphtherie-Tetanus-Keuchhusten</impfung><jahr>1995</jahr><anzahl>4214</anzahl></impf> |
| <impf><impfung>Diphtherie-Tetanus</impfung><jahr>1995</jahr><anzahl>4876</anzahl></impf> |
| <impf><impfung>Diphtherie</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Tetanus</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Kinderlaehmung (Poliomyelitis)</impfung><jahr>1995</jahr><anzahl>9002</anzahl></impf> |
| <impf><impfung>Hib-Hepatitis B </impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Hepatitis A+B</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Hepatitis A</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Hepatitis B</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Masern-Mumps Roeteln</impfung><jahr>1995</jahr><anzahl>2270</anzahl></impf> |
| <impf><impfung>Roeteln</impfung><jahr>1995</jahr><anzahl>901</anzahl></impf> |
| <impf><impfung>Masern-Mumps</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Mumps</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>Tuberkulose</impfung><jahr>1995</jahr><anzahl>10</anzahl></impf> |
| <impf><impfung>Rotaviren</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf> |
| <impf><impfung>FSME</impfung><jahr>1995</jahr><anzahl>18682</anzahl></impf> |
| <impf><impfung>Grippe</impfung><jahr>1995</jahr><anzahl>0</anzahl></impf>  ……………………………………………………………………………………………………………………………………………………………………………… |
|  |
|  |

**Task 4: Register your XML schema and afterwards store (validate) the related data in the native XML database exist -** [**http://exist-db.org/**](http://exist-db.org/)

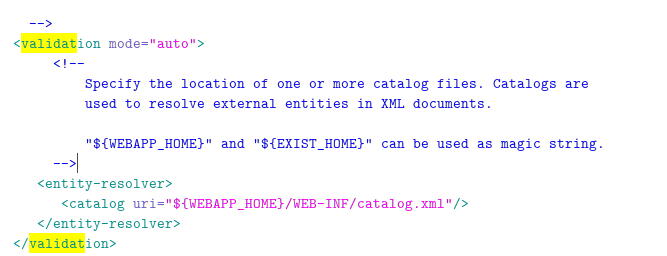
Installation Datenbank (Unix/Windows)

Alle Einstellungen wurden auf Standard gesetzt.

Um das bearbeitete XML-File hochladen und (implizit) validieren zu können, war es notwendig:

1. ein dtd-Schema zu erstellen (Siehe Task 2)

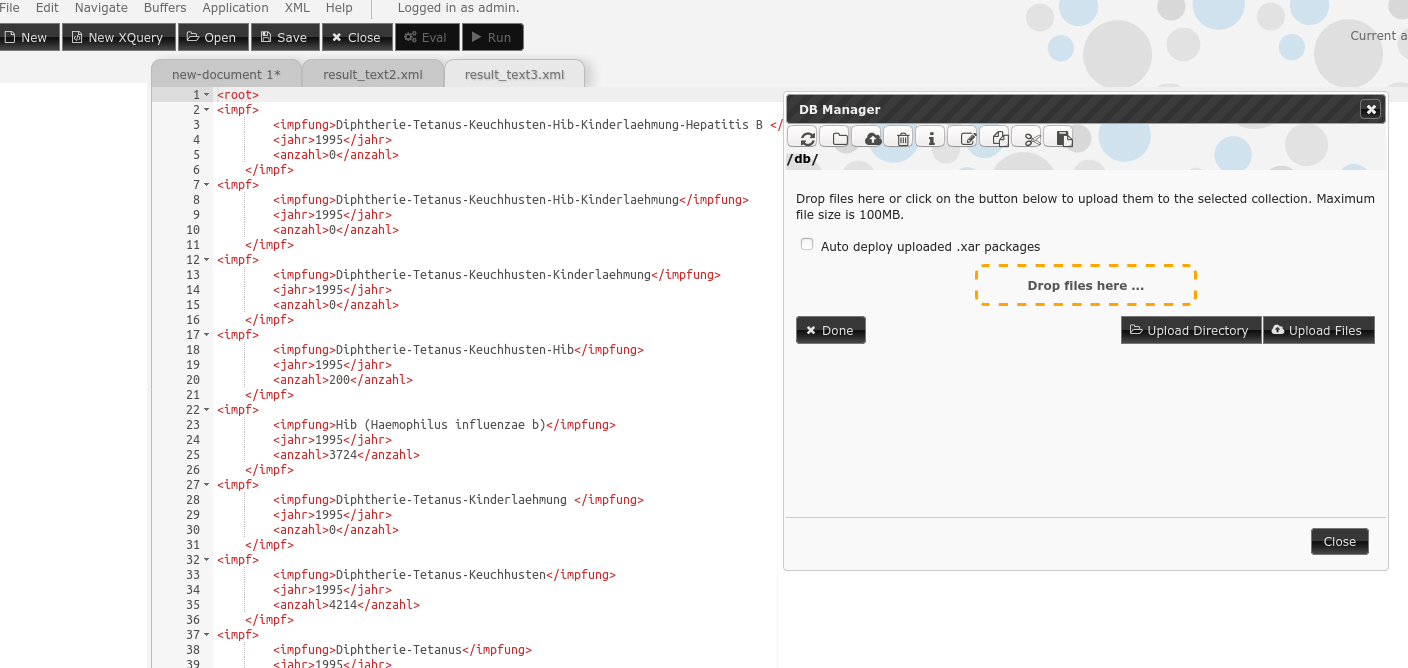
2. die Datei conf.xml im Verzeichnis /etc zu ändern (validation mode = „auto“)

3. Die hochzuladende xml-Datei noch einmal zu modifizieren:

<!DOCTYPE root SYSTEM "Validierung\_Datenschema.dtd">

4. die Datei catalog.xml im Verzeichnis ./*etc/*webapp/WEB-INF zu editieren, damit das Schema gefunden wird:

5. den eigentlichen Upload auszuführen.

Das ganze Prozedere war etwas Trial und Error, die Log-Datei exist.loc im Verzeichnis ./logs war dabei sehr hilfreich. Das Encoding-Problem erschwerte die Sache etwas.

Wenn man die Validierung auf „Auto“ stellt muss man explizit das Schema und dessen Pfad mitgeben mittels <system systemID=““ uri =““>. Alternativ könnte man in der config-Datei „yes“ auswählen, aber wir hatten mit „auto“ erste Erfolge beim Import erzielt. Der Grund war, dass ohne Angabe des Schemas keine Validierung stattgefunden hat.

**Task 5: answer your data based question using XQuery via API usage in a supported programming language**

Als Programmiersprache der Wahl wurde Python gewählt. Als Entwicklungsumgebung wurde Anaconda3 mit Jupyther Notebook verwendet.

Die folgend aufgeführten Packages wurden importiert.

pyexistdb wird für die Schnittstelle zur exist Datebank verwendet.

Pandas wird zur Verwendung von DataFrames importiert.

lxml wird verwendet um die Daten aus der Datenbank verarbeiten zu können.

mathplotlib wird für das Auswertungsdiagramm gebraucht.

from pyexistdb import db

import pandas as pd

from lxml import etree,objectify

import matplotlib.pyplot as plt

Mit folgendem Befehl wir die Verbindung zur Datenbank hergestellt. So kann über die Variable „exdb“ auf die Datenbank zugegriffen werden. Es wird die Adresse der Datenbank, der Benutzername und das entsprechende Passwort übergeben.

exdb = db.ExistDB("http://127.0.0.1:8080/exist", username = "admin", password = "")

Die Methode „QueryToTree“ schickt einen xquery Ausdruck an die Datenbank und gibt den erhaltenen xml-Baum aus und zurück. Die Abfrage wird in die <result> Tags gepackt, um zu vermeiden, dass die Datenbank nur 10 Ergebnisse liefert. Der Parameter „pretty\_print“ wird True gesetzt, damit der xml-Baum durch den „tostring“ Befehl in eine lesbare Form umgewandelt wird.

def QueryToTree(query):

result = exdb.query("<result>{"+query+"}</result>")

print(etree.tostring(result.results[0],pretty\_print=True).decode())

return result.results[0]

Folgende Abfrage würde den Inhalt des Datenbankeintrags „result\_text3.xml“ ausgeben.

QueryToTree("for $x in doc(\"result\_text3.xml\") return $x")

Folgender Befehl beantwortet die oben in der Aufgabenstellung formulierten Fragen die wie folgt lauten:

**Wie verhält sich die Gesamtzahl der Impfungen über die Jahre?**

**Wo liegt der absolute Höchstwert von Impfungen & wann begann eine mögliche Trendwende nach unten oder umgekehrt?**

root = QueryToTree("for $x in doc(\"result\_text3.xml\")/root/impf let $jahr := $x/jahr group by $jahr order by sum($x/jahr) return <impf><impfung>{data($jahr)}</impfung><impfgesamtzahl>{sum($x/anzahl)}</impfgesamtzahl></impf>")

Anschließend wird der Antwort-Baum zur besseren Übersicht in ein DataFrame umgewandelt und die Daten in einem Diagramm dargestellt.

impfungen = [child.getchildren() for child in root.getchildren()]

jahre = [child[0].text for child in impfungen]

anzahl = [int(child[1].text) for child in impfungen]

df = pd.DataFrame(anzahl,jahre, columns=['Anzahl'])

df.plot(kind='bar')

plt.show

<result xmlns:exist="http://exist.sourceforge.net/NS/exist">

<impf>

<impfung>1995</impfung>

<impfgesamtzahl>43879</impfgesamtzahl>

</impf>

<impf>

<impfung>1996</impfung>

<impfgesamtzahl>47280</impfgesamtzahl>

</impf>

<impf>

<impfung>1997</impfung>

<impfgesamtzahl>46947</impfgesamtzahl>

</impf>

<impf>

<impfung>1998</impfung>

<impfgesamtzahl>47882</impfgesamtzahl>

</impf>

<impf>

<impfung>1999</impfung>

<impfgesamtzahl>50414</impfgesamtzahl>

</impf>

<impf>

<impfung>2000</impfung>

<impfgesamtzahl>45892</impfgesamtzahl>

</impf>

<impf>

<impfung>2001</impfung>

<impfgesamtzahl>39421</impfgesamtzahl>

</impf>

<impf>

<impfung>2002</impfung>

<impfgesamtzahl>40833</impfgesamtzahl>

</impf>

<impf>

<impfung>2003</impfung>

<impfgesamtzahl>38672</impfgesamtzahl>

</impf>

<impf>

<impfung>2004</impfung>

<impfgesamtzahl>35803</impfgesamtzahl>

</impf>

<impf>

<impfung>2005</impfung>

<impfgesamtzahl>32964</impfgesamtzahl>

</impf>

<impf>

<impfung>2006</impfung>

<impfgesamtzahl>43277</impfgesamtzahl>

</impf>

<impf>

<impfung>2007</impfung>

<impfgesamtzahl>37770</impfgesamtzahl>

</impf>

<impf>

<impfung>2008</impfung>

<impfgesamtzahl>34290</impfgesamtzahl>

</impf>

<impf>

<impfung>2009</impfung>

<impfgesamtzahl>33123</impfgesamtzahl>

</impf>

<impf>

<impfung>2010</impfung>

<impfgesamtzahl>26435</impfgesamtzahl>

</impf>

<impf>

<impfung>2011</impfung>

<impfgesamtzahl>29112</impfgesamtzahl>

</impf>

<impf>

<impfung>2012</impfung>

<impfgesamtzahl>30481</impfgesamtzahl>

</impf>

<impf>

<impfung>2013</impfung>

<impfgesamtzahl>32714</impfgesamtzahl>

</impf>

<impf>

<impfung>2014</impfung>

<impfgesamtzahl>27540</impfgesamtzahl>

</impf>

<impf>

<impfung>2015</impfung>

<impfgesamtzahl>28019</impfgesamtzahl>

</impf>

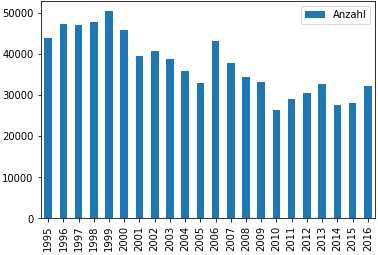
<impf>

<impfung>2016</impfung>

<impfgesamtzahl>32178</impfgesamtzahl>

</impf>

</result>



Dabei zeigt sich eine Abnahme der absoluten Anzahl der Impfungen in Linz. Es gab 1999 den Höchstwert. So eine Abnahme kann viele verschiedene Ursachen haben: es könnten etwa mehrere Impfungen zu einer zusammengefasst worden sein. Tatsächlich zeigt sich etwa, dass zum Beispiel ab etwa 2000 eine Kombinationsimpfung „Diphtherie-Tetanus-Keuchhusten-Hib-Kinderlaehmung-Hepatitis B“ vorhanden war, die es vorher nur ohne Hepatitis B gab, und vor 1999 auch noch ohne Kinderlaehmung.

Eine weitere Ursache für die Abnahme könnte die demographische Veränderung der Bevölkerungsstruktur sein, d.h. es wurden weniger Kinder geboren. Das müsste man mit demographischen Daten abgleichen.

Es könnte auch daran liegen, dass Familien mit Kindern vermehrt aufs Land ziehen, statt in der Stadt zu bleiben. Um das zu verifizieren, könnte man die Geburtenanzahl gemeinsam mit den Schulgängern/Kindergartenkindern (idealerweise korrigiert um das Alter) in OÖ betrachten. Eine mögliche Frage wäre: sind die Schulgänger/Kindergartenkinder in Linz zurückgegangen und die Geburten gleichgeblieben? Das könnte unter Umständen diese Hypothese belegen. Man müsste natürlich auch bedenken, dass möglicherweise die Hausgeburten zurückgegangen sind, und auch die Kinder vom Land vermehrt in der Stadt geboren wurden. Daher sollte man die Schulbesucher/Kindergartenkinder in ganz OÖ betrachten.

Eine weitere Hypothese könnte die Zunahme der Impfgegnerschaft untersuchen: sollten sich die demographischen und impftechnischen Faktoren als vernachlässigbar erweisen (oder auch nicht), müsste man die einzelnen Impfarten einzeln betrachten. Vielleicht zeigen sich Abnahmen bei Impfungen, die Impfskeptikern besonders gefährlich erscheinen.

Hier noch eine abschließende Betrachtung nach Impfungen:

Etwa bei Meningokokken hat es 2010 einen ziemlichen Rückgang gegeben, was 2016 wieder deutlich anstieg. Auch in 2016 gab es einen starken Anstieg bei der Masern-Impfung. FSME, die deutlich häufigste Anzahl, wirkte einigermaßen konstant. Die Kinderlähmungsimpfung war bis etwa 2000 separiert und dürfte dann in die Diphtherie-Tetanus-Kinderlähmung-Impfung aufgenommen worden sein.

So auf den ersten oberflächlichen Blick ergeben sich die größten Unterschiede aufgrund der Impftechnik, kleinere Unterschiede aufgrund von Saisonalen Trends (etwa 2016 die Masernepidemie). Insgesamt bleibt jedoch ein schwach abnehmender Trend.

