# ANON

## a flexible tool for achieving optimal k-anonymous and l-diverse tables

Margareta Ciglic, Johann Eder, Christian Koncilia
Institut für Informatik-Systeme
Alpen-Adria-Universität Klagenfurt

# Problem and Goal Definition

- Problem: removing identifiers (ID, name, etc.) does not lead to individuals’ anonymity
  - Linking attack
  - Tracker
- Goal
  - Guarantee the required privacy
  - Maximize data utility (data quality)
  - Handle missing values

# k-Anonymity and l-Diversity

| Name | Zip | Age | Sex | Condition |
|---|---|---|---|---|
| Alice | 13053 | 28 | F | Hepatitis |
| Bob | 13068 | 29 | M | Hepatitis |
| Charlie | 13068 | 25 | M | Flu |
| Dolly | 14850 | 43 | F | Flu |
| Emma | 14853 | 50 | F | Cancer |
| Frank | 14853 | 48 | M | Hepatitis |
| George | 14850 | 79 | M | Flu |
| Harry | 27627 | 26 | M | Flu |

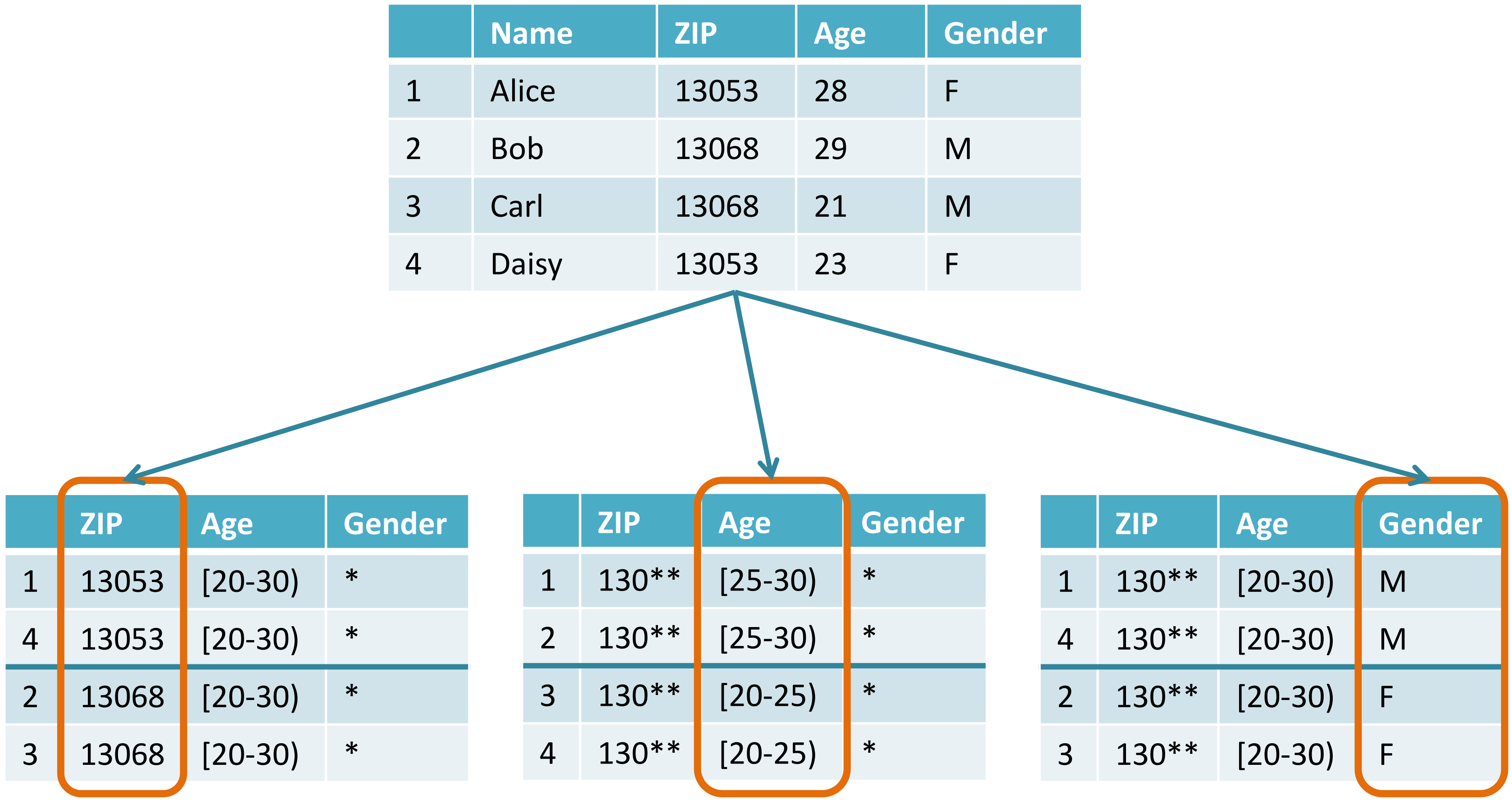
| | Zip | Age | Sex | Condition |
|---|---|---|---|---|
| A | 130** | 21-30 | * | Hepatitis |
| B | 130** | 21-30 | * | Hepatitis |
| C | 130** | 21-30 | * | Flu |
| D | 148** | 41-50 | * | Flu |
| E | 148** | 41-50 | * | Cancer |
| F | 148** | 41-50 | * | Hepatitis |
| G | | | | |
| H | | | | |

GENERALIZATION

SUPPRESSION

Initial microdata and 3-anonymous, 2-diverse table
with 2 blocks (equivalence classes, partitions)

# (Some) Transformation Possibilities



| | Name | ZIP | Age | Gender |
|---|---|---|---|---|
| 1 | Alice | 13053 | 28 | F |
| 2 | Bob | 13068 | 29 | M |
| 3 | Carl | 13068 | 21 | M |
| 4 | Daisy | 13053 | 23 | F |

| | ZIP | Age | Gender |
|---|---|---|---|
| 1 | 13053 | [20-30) | * |
| 4 | 13053 | [20-30) | * |
| 2 | 13068 | [20-30) | * |
| 3 | 13068 | [20-30) | * |

| | ZIP | Age | Gender |
|---|---|---|---|
| 1 | 130** | [25-30) | * |
| 2 | 130** | [25-30) | * |
| 3 | 130** | [20-25) | * |
| 4 | 130** | [20-25) | * |

| | ZIP | Age | Gender |
|---|---|---|---|
| 1 | 130** | [20-30) | M |
| 4 | 130** | [20-30) | M |
| 2 | 130** | [20-30) | F |
| 3 | 130** | [20-30) | F |

# Utility Maximization

Utility maximization is reached with the calculation of the weighted information loss.

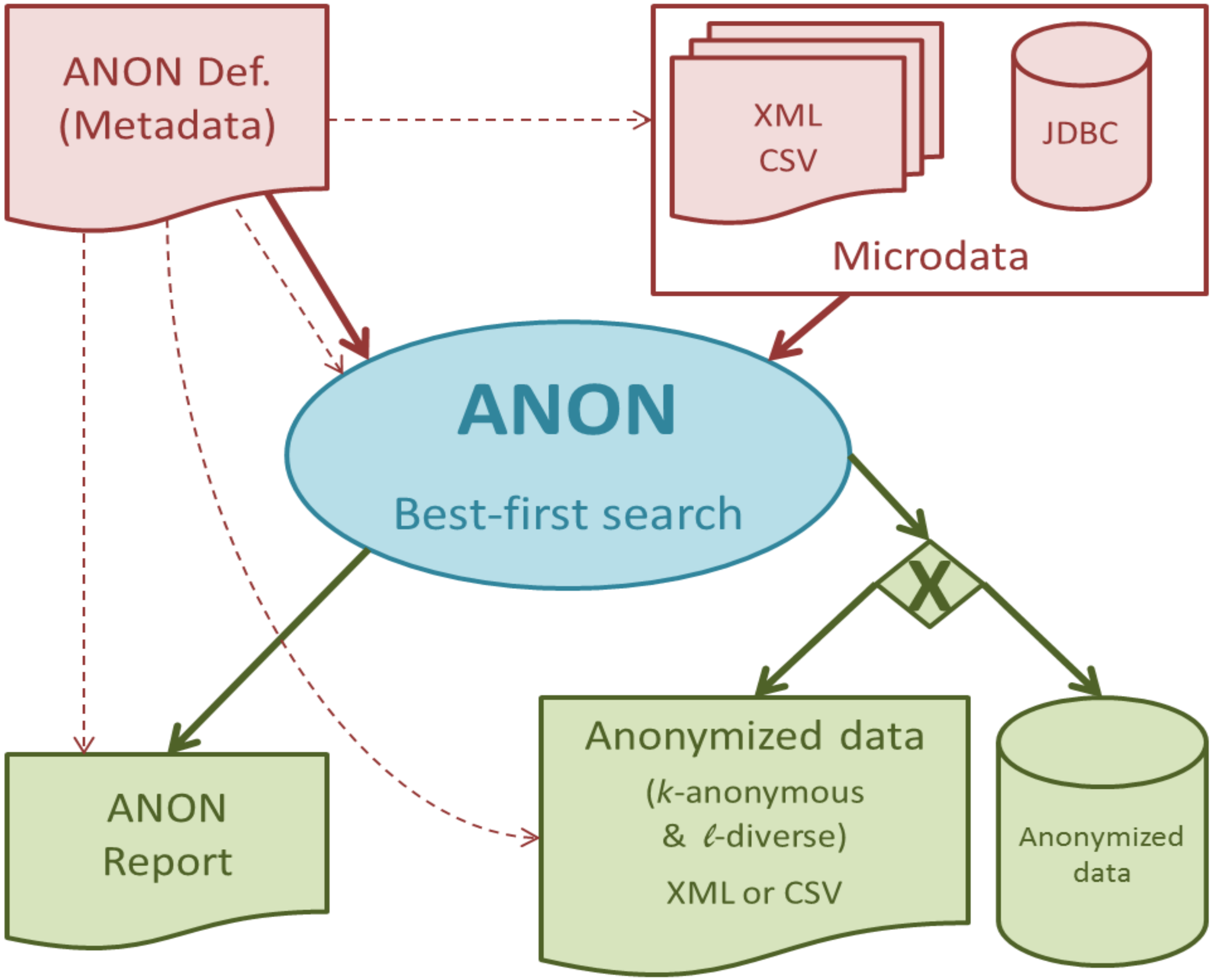
## Weighted Information Loss:

$$\sum_{i=1}^{n} P_{\alpha_i} * V_{\alpha_i}^{L}$$

$P_{\alpha_i}$ ………… Priority of the attribute $\alpha_i$

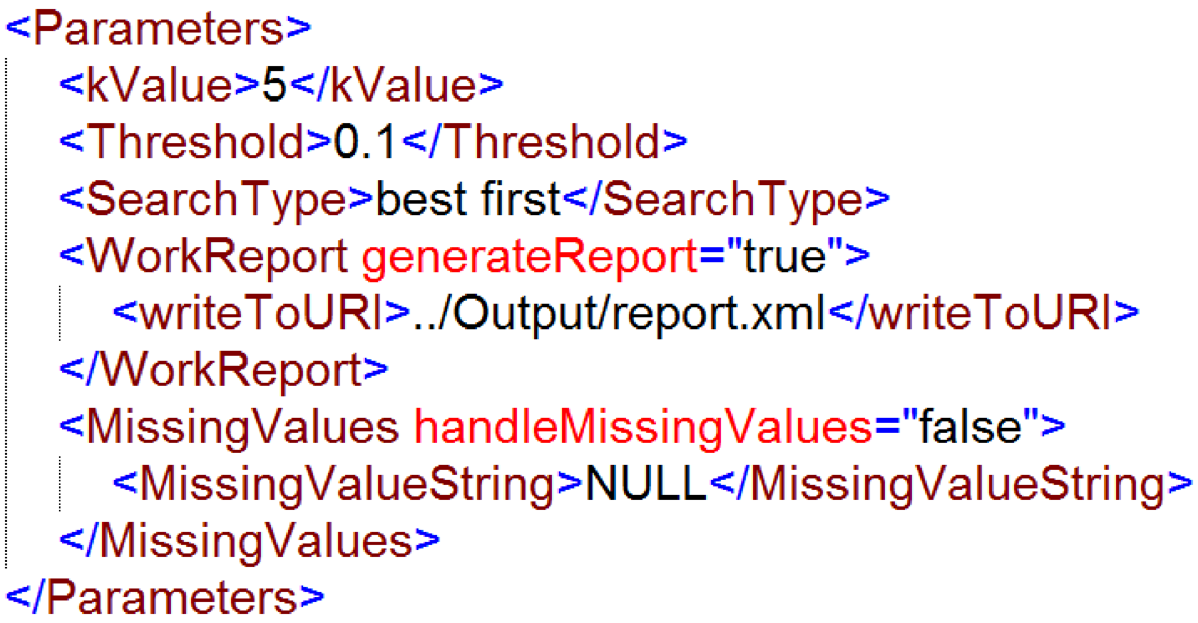
$V_{\alpha_i}^{L}$ ………… user-defined information loss of the attribute $\alpha_i$
if its values get transformed to the level $L$

# ANON - Principle and Implementation



# ANON – Anonymization Parameters

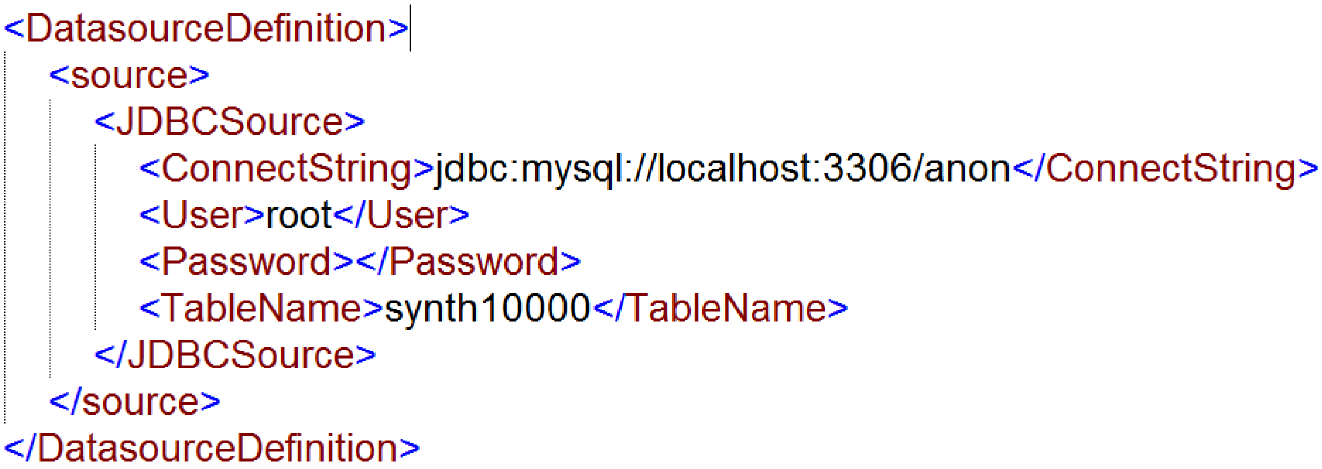
- kValue
- Threshold
- SearchType
- WorkReport
- MissingValues



```
<Parameters>
  <kValue>5</kValue>
  <Threshold>0.1</Threshold>
  <SearchType>best first</SearchType>
  <WorkReport generateReport="true">
    <writeToURI>../Output/report.xml</writeToURI>
  </WorkReport>
  <MissingValues handleMissingValues="false">
    <MissingValueString>NULL</MissingValueString>
  </MissingValues>
</Parameters>
```

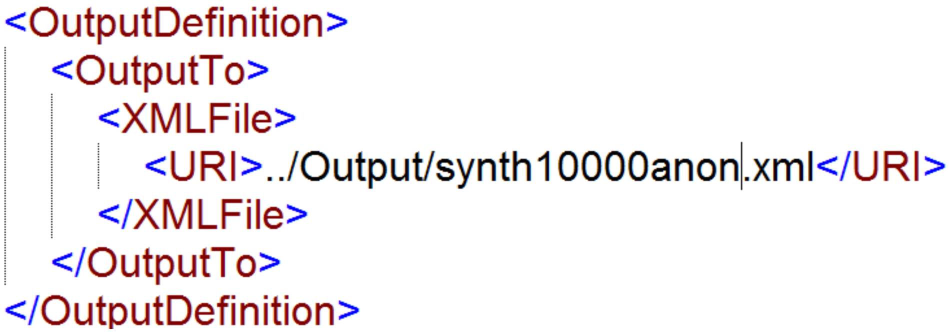
# ANON – Input/Output Parameters

- Input (microdata)
  CSV, XML and/or JDBC



```xml
<DatasourceDefinition>
    <source>
        <JDBCSource>
            <ConnectString>jdbc:mysql://localhost:3306/anon</ConnectString>
            <User>root</User>
            <Password></Password>
            <TableName>synth10000</TableName>
        </JDBCSource>
    </source>
</DatasourceDefinition>
```

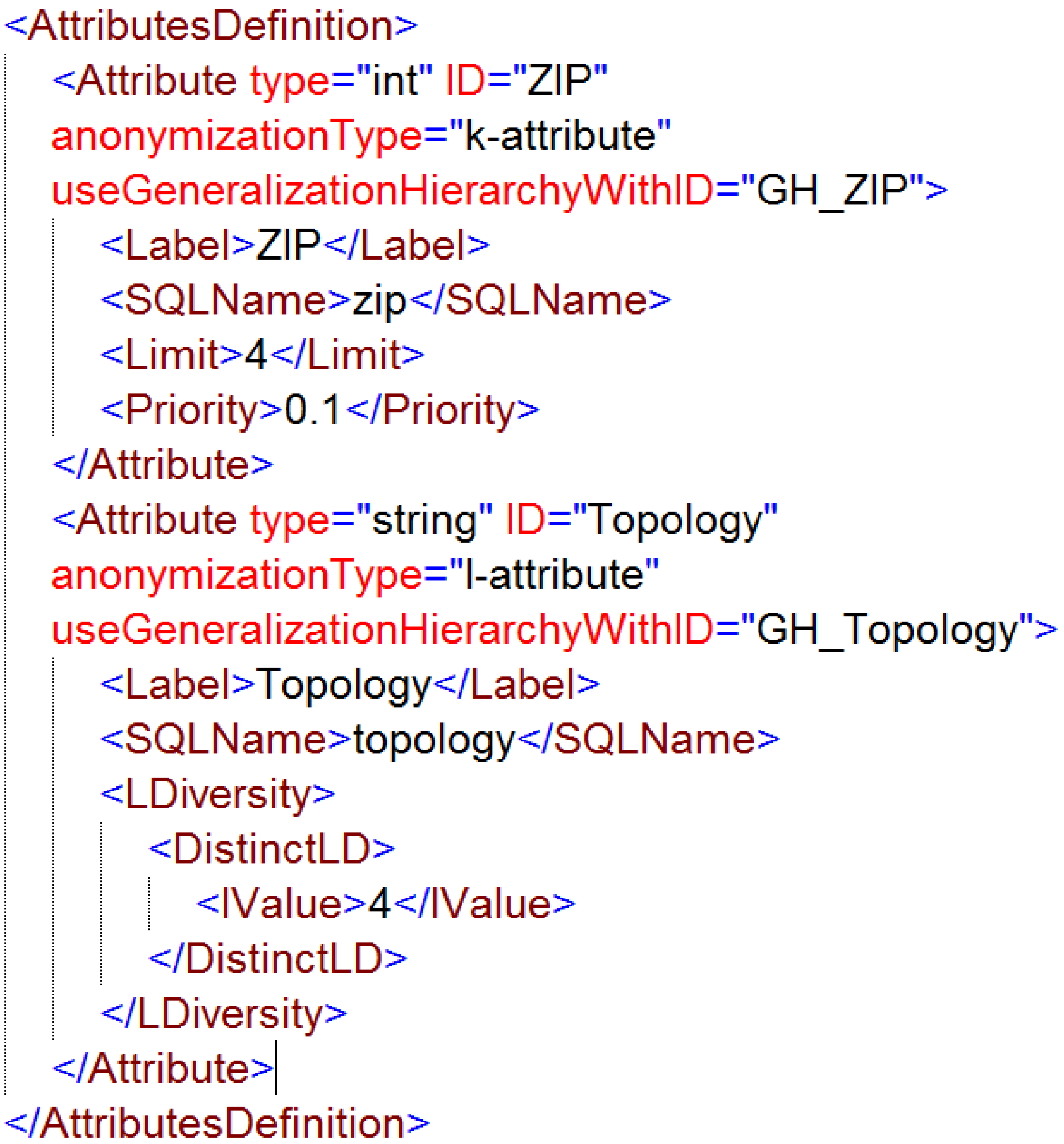
- Output (anonymized data)
  CSV, XML or JDBC



```xml
<OutputDefinition>
    <OutputTo>
        <XMLFile>
            <URI>../Output/synth10000anon.xml</URI>
        </XMLFile>
    </OutputTo>
</OutputDefinition>
```

# ANON - Attributes

- anonymizationType
- HierarchyID
- SQLName
- Limit
- Priority
- Ldiversity
    - lValue



```
<AttributesDefinition>
  <Attribute type="int" ID="ZIP"
  anonymizationType="k-attribute"
  useGeneralizationHierarchyWithID="GH_ZIP">
    <Label>ZIP</Label>
    <SQLName>zip</SQLName>
    <Limit>4</Limit>
    <Priority>0.1</Priority>
  </Attribute>
  <Attribute type="string" ID="Topology"
  anonymizationType="l-attribute"
  useGeneralizationHierarchyWithID="GH_Topology">
    <Label>Topology</Label>
    <SQLName>topology</SQLName>
    <LDiversity>
      <DistinctLD>
        <lValue>4</lValue>
      </DistinctLD>
    </LDiversity>
  </Attribute>
</AttributesDefinition>
```

# Anonymization Types (Example)

| Attribute Name | Anonymization Type |
|---|---|
| Patient No. | ignore |
| Name | ignore |
| Address | ignore |
| ZIP | k-attribute |
| Age | k-attribute |
| Sex | k-attribute |
| Education | k-attribute |
| Pseudonym | dontcare |
| Topology | l-attribute |
| Staging | l-attribute |
| Grading | l-attribute |
| R | l-attribute |
| V | l-attribute |

Unique identifiers

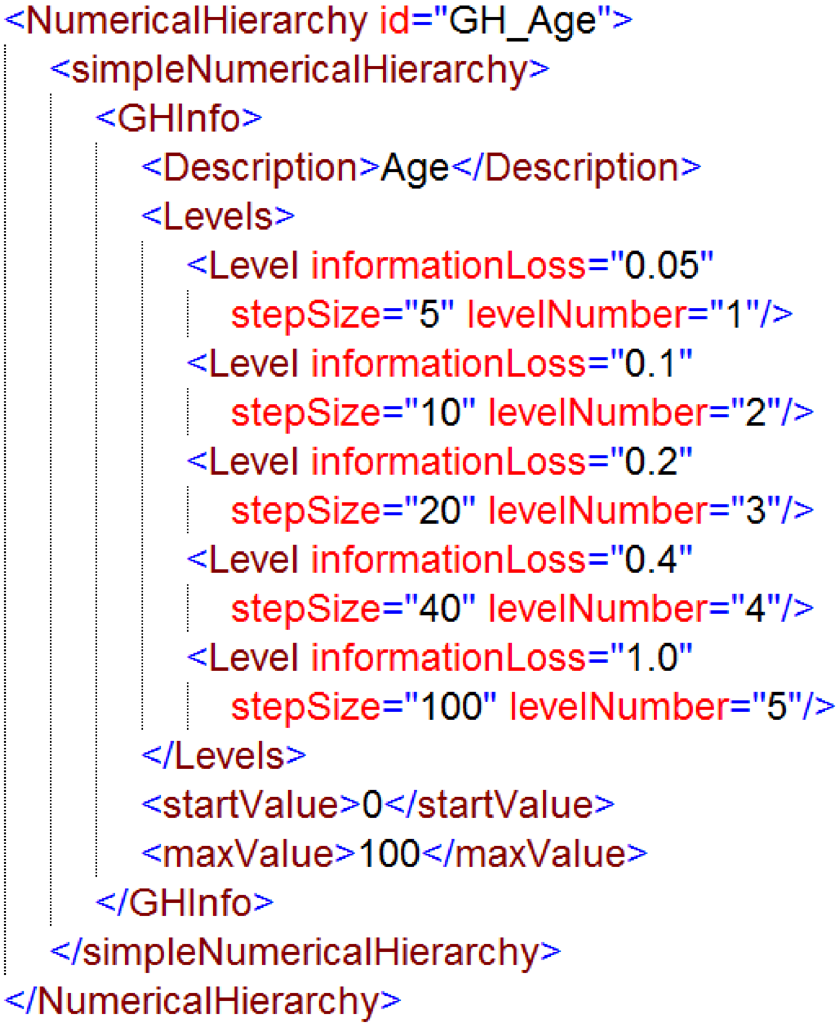
Quasi-identifiers

No identifier;
not sensitive

Sensitive attributes

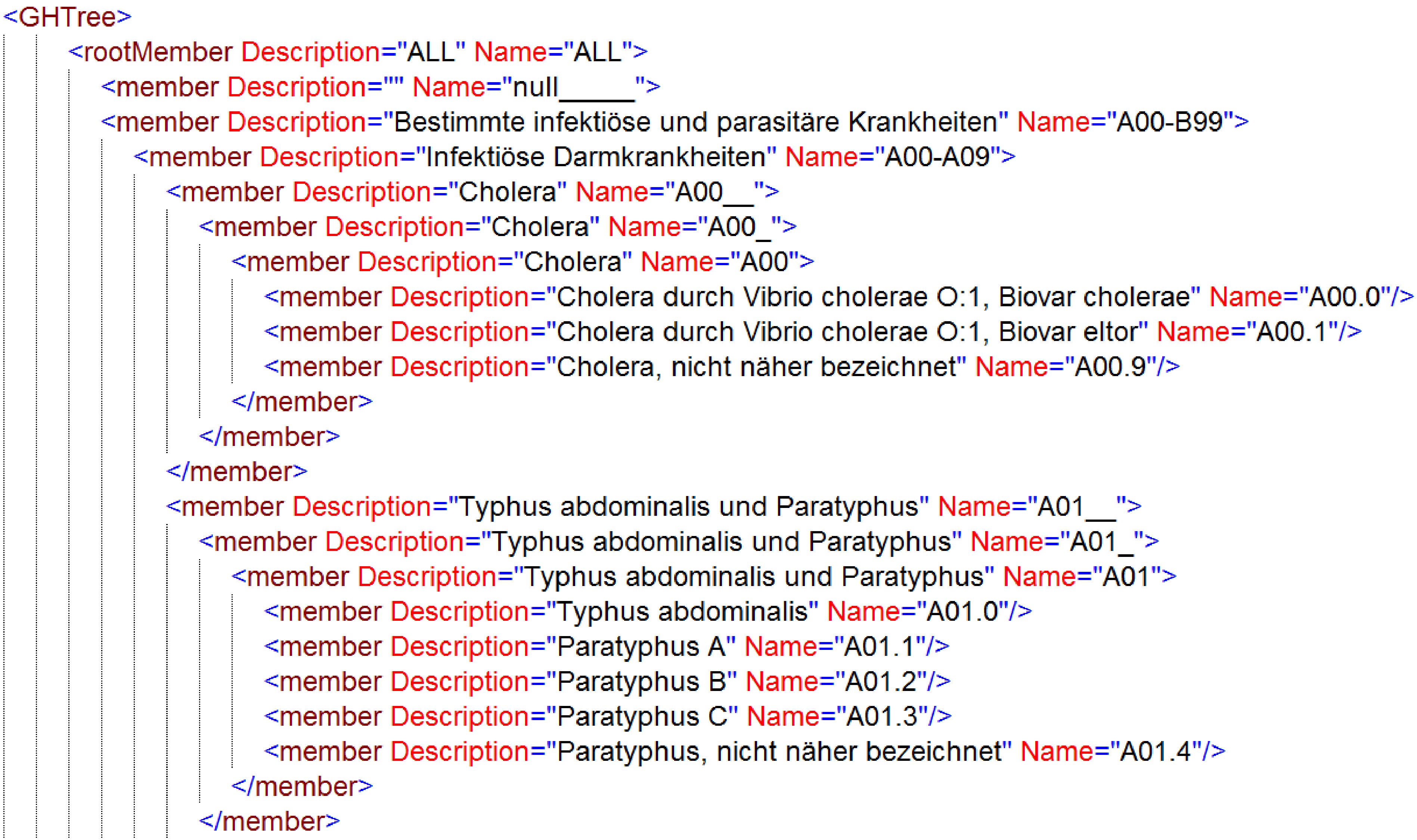
# ANON – Generalisation Hierarchies

- NumericalHierarchy
  - ○ Levels
    - levelNumber
    - informationLoss
    - Intervals (stepSize)
- CategoricalHierarchy
  - ○ Levels
    - levelNumber
    - informationLoss
  - ○ GHTree



```
<NumericalHierarchy id="GH_Age">
  <simpleNumericalHierarchy>
    <GHInfo>
      <Description>Age</Description>
      <Levels>
        <Level informationLoss="0.05"
          stepSize="5" levelNumber="1"/>
        <Level informationLoss="0.1"
          stepSize="10" levelNumber="2"/>
        <Level informationLoss="0.2"
          stepSize="20" levelNumber="3"/>
        <Level informationLoss="0.4"
          stepSize="40" levelNumber="4"/>
        <Level informationLoss="1.0"
          stepSize="100" levelNumber="5"/>
      </Levels>
      <startValue>0</startValue>
      <maxValue>100</maxValue>
    </GHInfo>
  </simpleNumericalHierarchy>
</NumericalHierarchy>
```

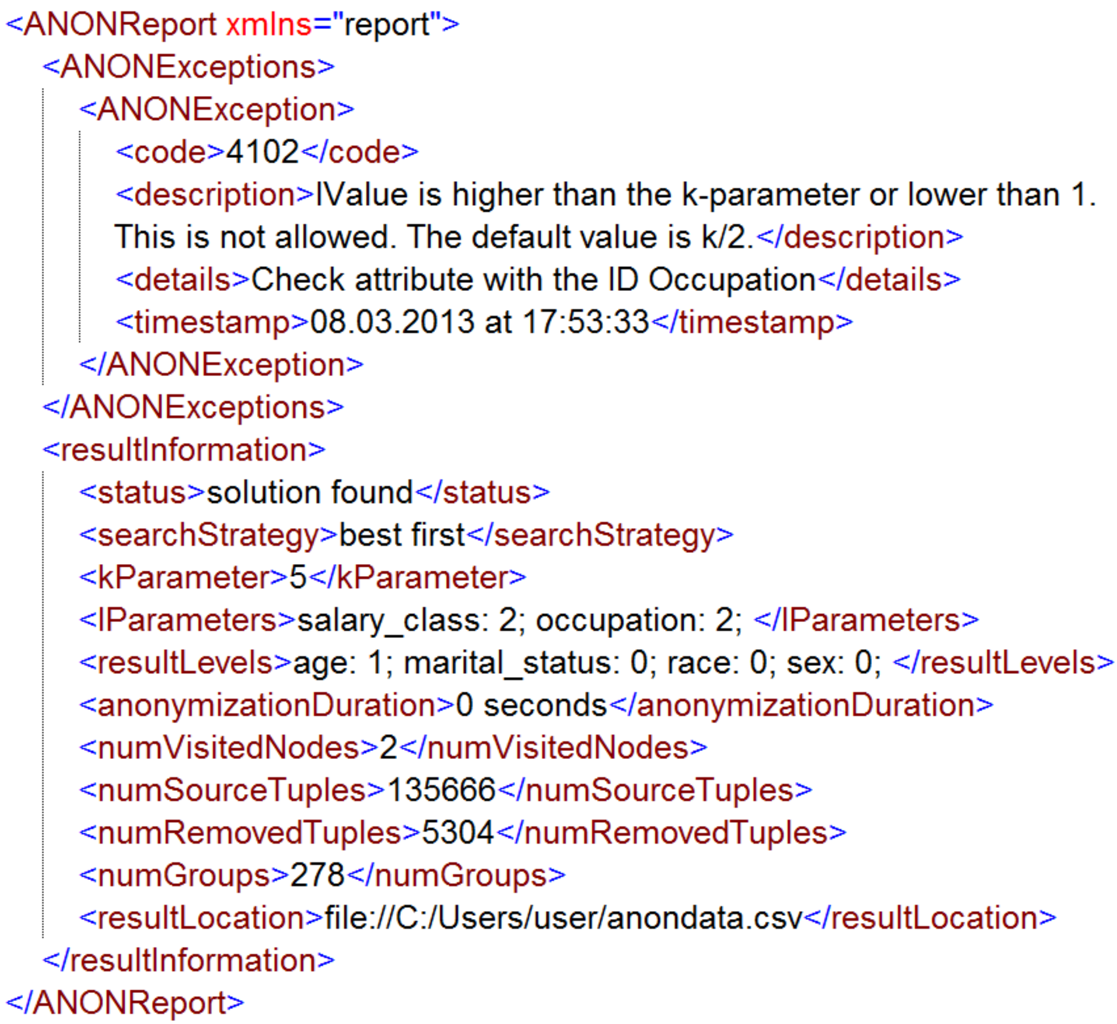
# Anonymization Tool ANON



```
<GHTree>
    <rootMember Description="ALL" Name="ALL">
        <member Description="" Name="null_____">
        <member Description="Bestimmte infektiöse und parasitäre Krankheiten" Name="A00-B99">
            <member Description="Infektiöse Darmkrankheiten" Name="A00-A09">
                <member Description="Cholera" Name="A00__">
                    <member Description="Cholera" Name="A00_">
                        <member Description="Cholera" Name="A00">
                            <member Description="Cholera durch Vibrio cholerae O:1, Biovar cholerae" Name="A00.0"/>
                            <member Description="Cholera durch Vibrio cholerae O:1, Biovar eltor" Name="A00.1"/>
                            <member Description="Cholera, nicht näher bezeichnet" Name="A00.9"/>
                        </member>
                    </member>
                </member>
                <member Description="Typhus abdominalis und Paratyphus" Name="A01__">
                    <member Description="Typhus abdominalis und Paratyphus" Name="A01_">
                        <member Description="Typhus abdominalis und Paratyphus" Name="A01">
                            <member Description="Typhus abdominalis" Name="A01.0"/>
                            <member Description="Paratyphus A" Name="A01.1"/>
                            <member Description="Paratyphus B" Name="A01.2"/>
                            <member Description="Paratyphus C" Name="A01.3"/>
                            <member Description="Paratyphus, nicht näher bezeichnet" Name="A01.4"/>
                        </member>
                    </member>
```

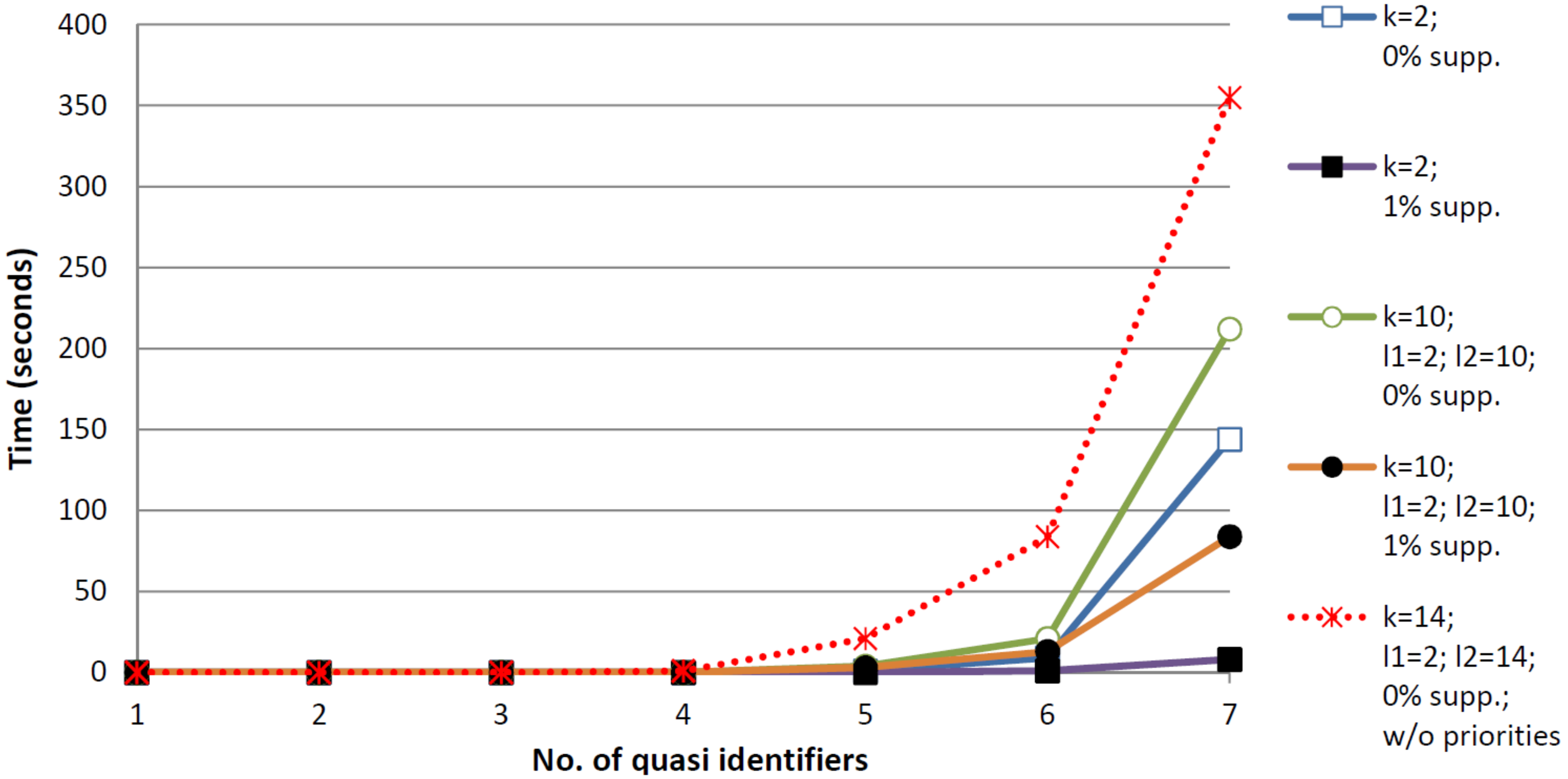
# ANON - Report

- ANONExceptions
- Status
- kParameter
- lParameter
- resultLevels
- numRemovedTuples
- resultLocation



```
<ANONReport xmlns="report">
  <ANONExceptions>
    <ANONException>
      <code>4102</code>
      <description>lValue is higher than the k-parameter or lower than 1.
      This is not allowed. The default value is k/2.</description>
      <details>Check attribute with the ID Occupation</details>
      <timestamp>08.03.2013 at 17:53:33</timestamp>
    </ANONException>
  </ANONExceptions>
  <resultInformation>
    <status>solution found</status>
    <searchStrategy>best first</searchStrategy>
    <kParameter>5</kParameter>
    <lParameters>salary_class: 2; occupation: 2; </lParameters>
    <resultLevels>age: 1; marital_status: 0; race: 0; sex: 0; </resultLevels>
    <anonymizationDuration>0 seconds</anonymizationDuration>
    <numVisitedNodes>2</numVisitedNodes>
    <numSourceTuples>135666</numSourceTuples>
    <numRemovedTuples>5304</numRemovedTuples>
    <numGroups>278</numGroups>
    <resultLocation>file://C:/Users/user/anondata.csv</resultLocation>
  </resultInformation>
</ANONReport>
```

# ANON - Performance



Adult Data Set, UCI, 45.222 records

# Conclusions

- Privacy is the major concern in microdata publishing
- BUT: data quality must be considered, too!

| | PRIVACY vs. | DATA QUALITY |
|---|---|---|
| Stakeholder | Individuals (Patients) | (Medical) researcher |
| Algorithm | Privacy test (goal test) of the search algorithm | Anonymization algorithm (Best-first search) |
| Method | k-anonymity (size of a group)<br>quasi identifiers<br>l-diversity (diversity of a group)<br>sensitive Attributes | Utility function (weighted information loss)<br>Missing value handling |

# References

1. Margareta Ciglic, Johann Eder, Christian Koncilia: ANON - a flexible tool for achieving optimal k-anonymous and l-diverse tables. Technical report, 2014. http://isys.uni-klu.ac.at/PDF/2014-ANON-Techreport.pdf

2. Margareta Ciglic, Johann Eder, and Christian Koncilia: k-anonymity of microdata with null values. In Proc. of the 25th International Conference on Database and Expert Systems Applications - DEXA 2014, 2014.

3. Margareta Ciglic, Johann Eder, and Christian Koncilia: ANON User Manual, 2013

4. Latanya Sweeney: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 5 (October 2002), 557-570.