Modeling Quality Problems for a Generic Data Quality Management Process for Research Data

Markus Matoni – CIDOC 2020

KONDA

Philipps Universität Marburg

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

SUB

Bundesministerium für Bildung und Forschung

Contents

- KONDA – What is that?
- Motivation & Aim
- Methods: How did we get to the Profiles?
- Examples of the Profiles for Quality Problems
- Wrap up & Future Work

KONDA - what is that?

Continuous quality management process of dynamic research data on objects of material culture using the LIDO standard

- 3 year project funded by BMBF
- 2019 - 2022
- 5 researchers

Bundesministerium
für Bildung
und Forschung

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

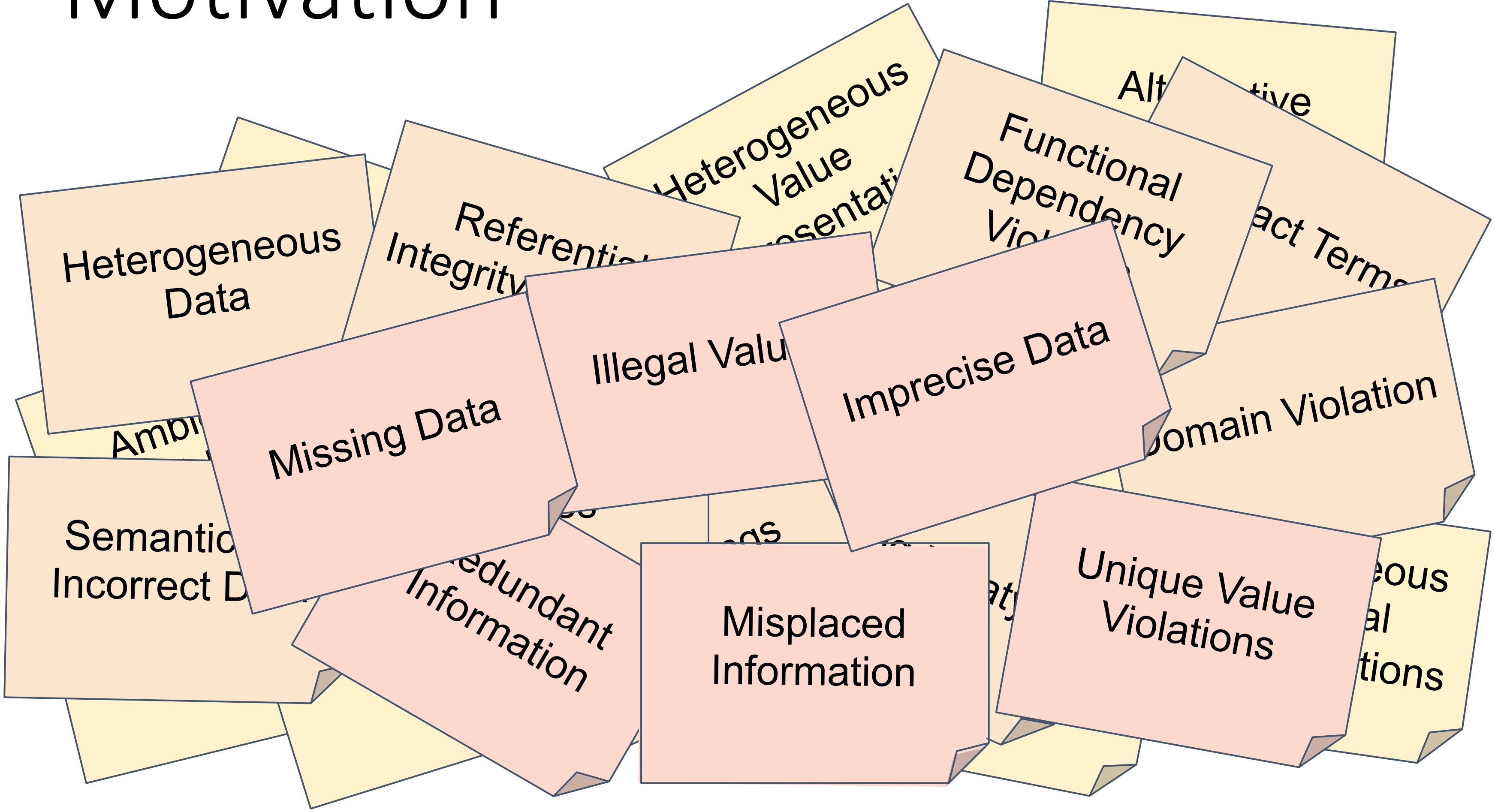
NIEDERSÄCHSISCHE STAATS- UND
UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN | SUB

Philipps Universität Marburg

KONDA - what is that?

- **Systematic quality assurance**
- **Continuously** over the entire data life cycle
- **Support development of LIDO** in terms of quality assurance
- **Improving the quality of research data** on objects of material culture
- **Interdisciplinary collaboration** (Community workshops, engagement in working groups (Deutscher Museumsbund, ICOM, …)

Motivation



Quality Problems: listen, discuss, collect and analyze

Aim

- Identify Quality Problems (catalog of profiles)
- 1st step to quality management process
  - Requirements for
    - Research data quality management process
    - Quality assurance techniques
  - Support the Development of LIDO

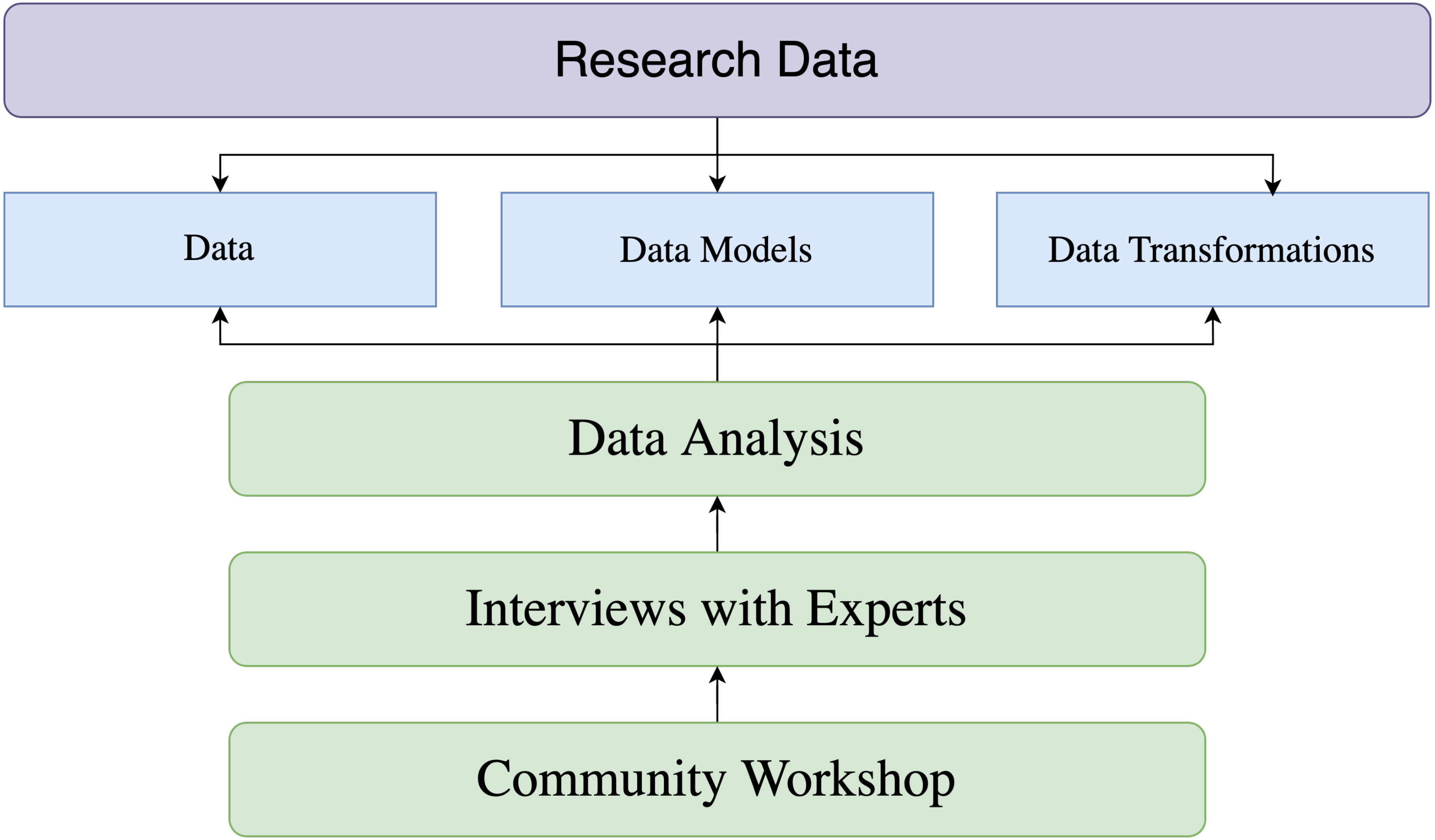
LIDO v1.1

- Downwards compatible to LIDO v1.0
- New Elements/Attributes
- Schematron rules (quality assurance)
- Public GitLab Repository & CI Workflow for Documentation

Catalog of Quality Problems

… how did we get there?

Methods



Research Data

Data

Data Models

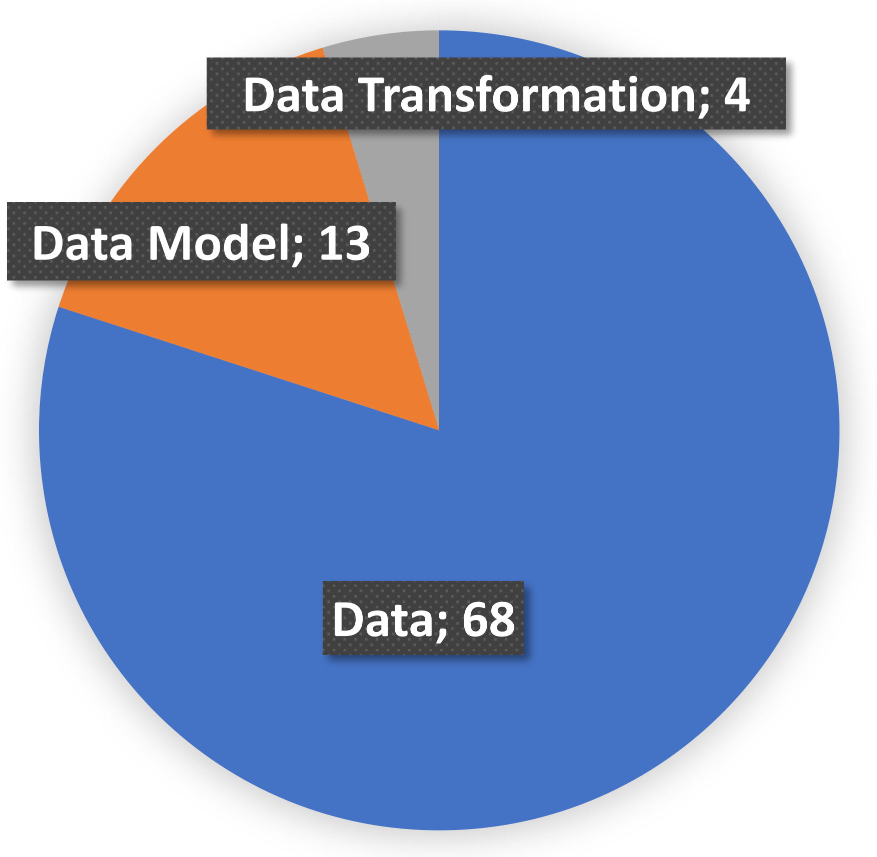
Data Transformations

Data Analysis

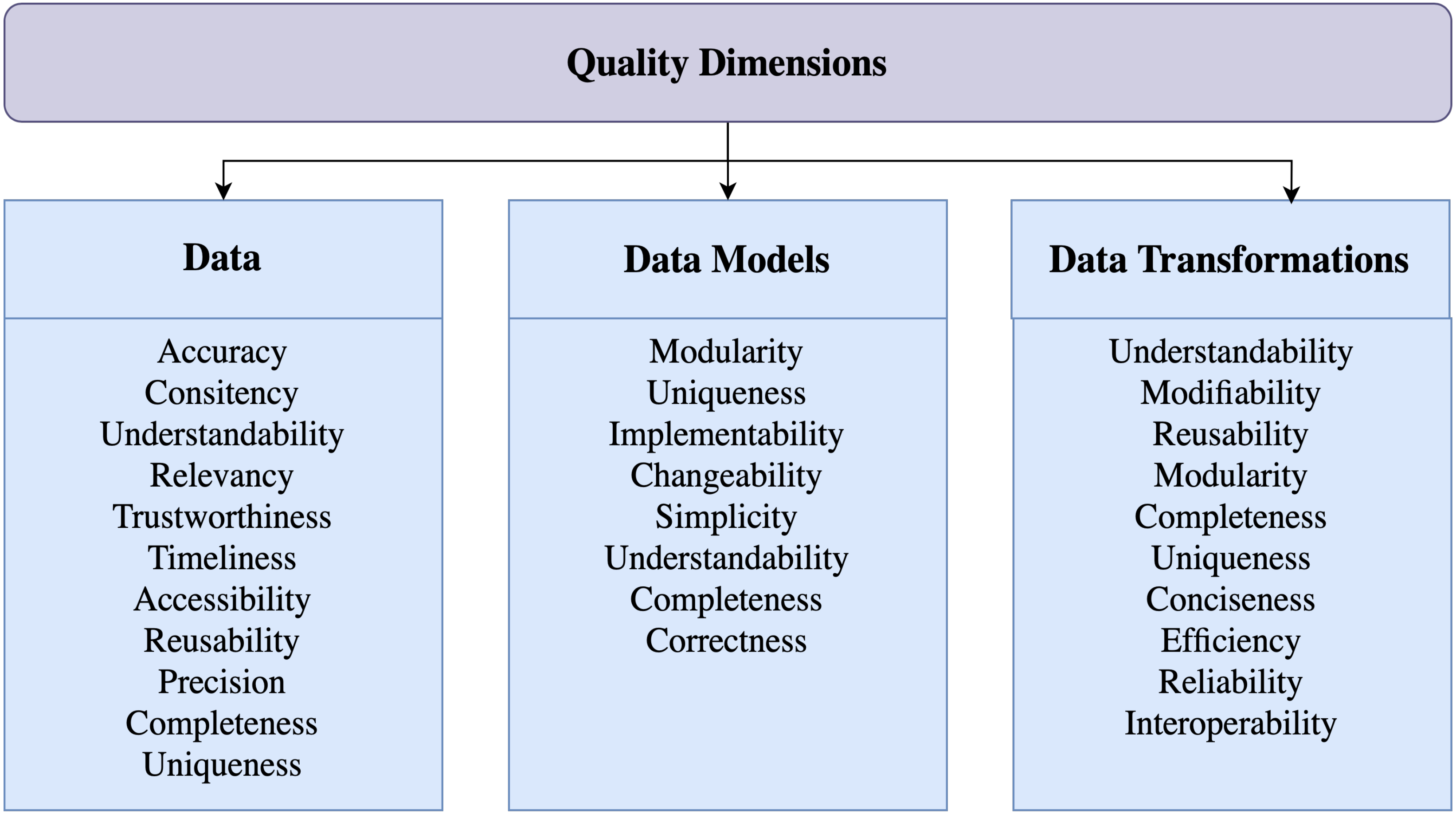
Interviews with Experts

Community Workshop

Quality Problems in



Data Quality Dimensions



Quality Dimensions

Data

Accuracy
Consitency
Understandability
Relevancy
Trustworthiness
Timeliness
Accessibility
Reusability
Precision
Completeness
Uniqueness

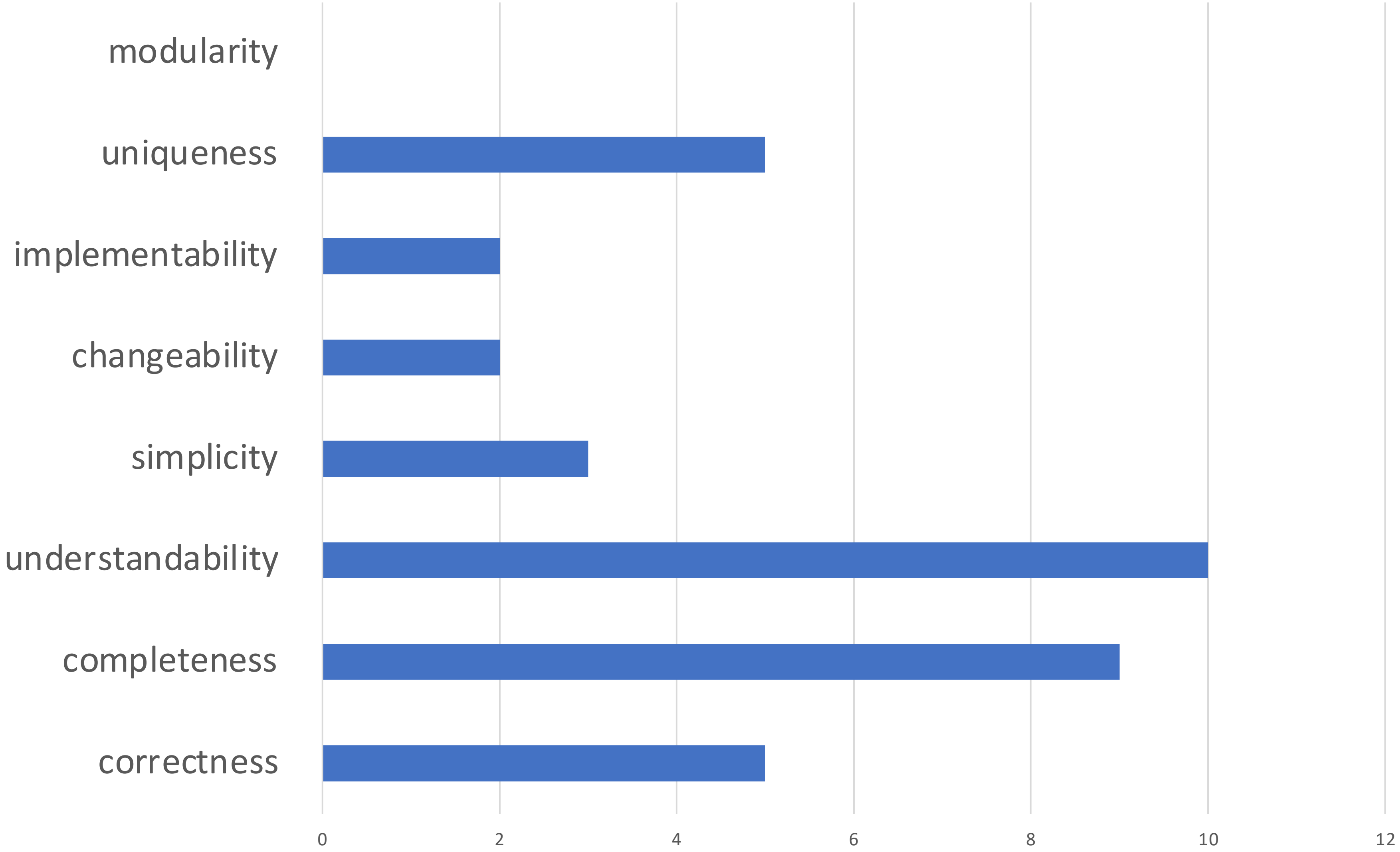
Data Models

Modularity
Uniqueness
Implementability
Changeability
Simplicity
Understandability
Completeness
Correctness

Data Transformations

Understandability
Modifiability
Reusability
Modularity
Completeness
Uniqueness
Conciseness
Efficiency
Reliability
Interoperability

Occurrences of data model quality problems per dimension



Examples

- DATA11 Violation of Formal Specifications
- MODEL03 Use of Free Text Fields Instead of LOD References

DATA11 Violation of Formal Specifications

Examples
- Exact Datings

12.03.2002 03/12/2002 12.3.02

12/03/2002 12.3.2002 12-03-2002

Examples
- Postcode, street names
- Numbers (binary, decimal, hexadecimal) (Letters, Signs)
- No long text in classifying terms (terms, classification)

DATA11 Violation of Formal Specifications

- **Description**: Some fields have concrete syntactical specifications, which have to apply to all fields of a kind. Invalid data in those fields does violate these specifications.
- Mainly affected quality **dimension**: Consistency
- Data life cycle: Collect
- **Impact** on data quality: The data are not explicitly readable, neither from employees nor automatically.
- Causes: Human error; No format check on data acquisition

DATA11 Violation of Formal Specifications

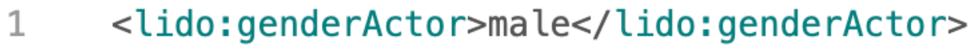
- **Identification:** MATCH-Pattern (regular expression)
- **Target state**: All fields do match their syntactical specifications.
- **Preventive** improvement
- **Retrospective** improvement

DATA11 Violation of Formal Specifications

lido:displayDate (earliestDate and latestDate)

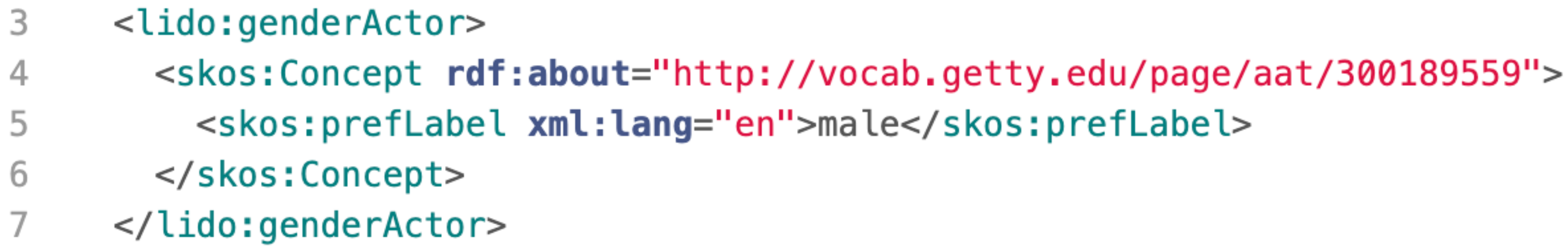
- Definition according to ISO 8601
- Schema allows xs:string
- LIDO 1.1 adds Schematron rule (id="sch_DateTime")
- XSD downward compatible + quality assurance

MODEL03 Use of Free Text Fields Instead of LOD References



```
<lido:genderActor>male</lido:genderActor>
```

vs



```
<lido:genderActor>
  <skos:Concept rdf:about="http://vocab.getty.edu/page/aat/300189559">
    <skos:prefLabel xml:lang="en">male</skos:prefLabel>
  </skos:Concept>
</lido:genderActor>
```

MODEL03 Use of Free Text Fields Instead of LOD References

- **Mainly affected quality dimension:** precision
- **Other affected quality dimensions:** completeness, correctness, simplicity, implementability
- **Impact on data quality:**
  - Heterogeneous, unstructured
  - reusing the data difficult
  - changes in information have to be added manually

MODEL03 Use of Free Text Fields Instead of LOD References

- **Causes:**
  - using LOD/authority files is not possible due to the data model (e.g. no URIs allowed)
  - using LOD/authority files is not implemented in the cataloguing software
  - lack of knowledge: LOD/authority files generally unknown; which vocabulary is suitable for which use case?
- **Root in the data life cycle:** plan
- Identification:
- Target state:
- Preventive improvement:
- Retrospective improvement:

MODEL03 Use of Free Text Fields Instead of LOD References

- lido:genderActor
  - Definition according to sex or gender specification
  - Schema allows both: free-text and skos:Concept
  - LIDO v1.1 adds Schematron rule (id="sch_MixedContentInfo")
  - XSD downward compatible + quality assurance

Wrap up

- Catalog of Quality Problems in Data, Data Models and Data Transformations
  - doi.org/10.5281/zenodo.3955500
  - zenodo.org/communities/konda-project
- LIDO v1.1
  - lido-schema.org
  - gitlab.gwdg.de/lido/lido-publication

Future Work

zenodo.org/communities/konda-project

Contribution

- ✓ Define Quality Dimensions
- ✓ Catalog of Quality Problems
- ✓ First Ideas on preventive & retrospective improvement
- ✓ Detecting Quality Problems in Research Data: A Model-Driven Approach

Future Work

- Identification of further Methods for Analysis and Improvement
- Tool support (semi & full) and workflow specifications
- Define a Generic Data Quality Management Process for Research Data
- Define Requirements for LIDO v2.0 in terms of quality