Faculty of Computer Science

universität wien

Quality Criteria for Controlled Web Vocabularies

Christian Mader
University of Vienna, Faculty of Computer Science
Research Group Multimedia Information Systems

Bernhard Haslhofer
Cornell University, Information Science

Motivation

- **Create new** and **find existing** vocabularies to adopt or align to
- Must fit to the needs of the developer/adopter
- Efficient automated method needed to support developers in
  - creating vocabularies to reach intended goals
  - Finding suitable vocabularies
- Focus on SKOS vocabularies in a Web of Data setting

Approach

Standards, best practices and metrics for vocabulary development exist

- ISO/DIS 25964-1
- ANSI/NISO Z39.19-2005
- Soergel, Kless2010, Stvilia2007,...

Approach

Standards, best practices and metrics for vocabulary development exist

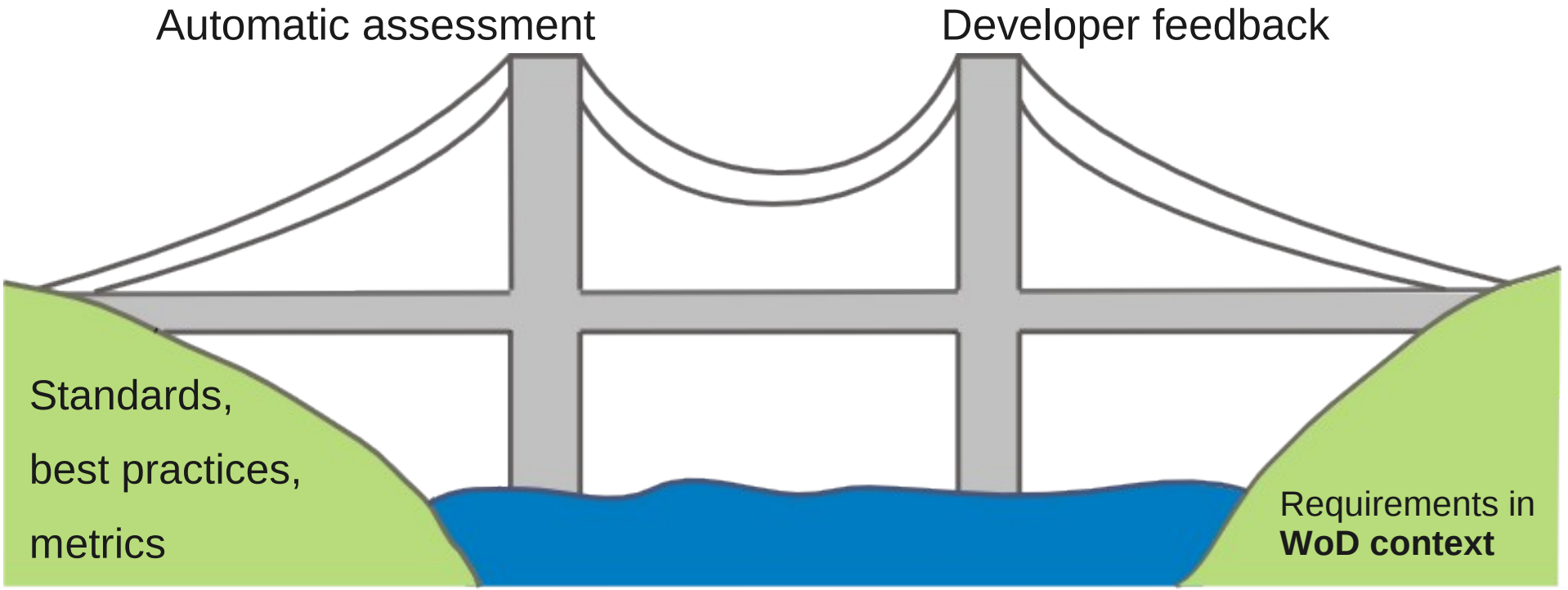
- ISO/DIS 25964-1
- ANSI/NISO Z39.19-2005
- Soergel, Kless2010, Stvilia2007,...

Requirements in **WoD context**

- Application-driven, e.g.,
  - Search & retrieval
  - Query expansion
- Maintained
- “Well-known”
- Documented
- Multilingual
- ...

Quality Criteria

- No absolute measure for vocabulary quality
- Each criterion covering specific requirement

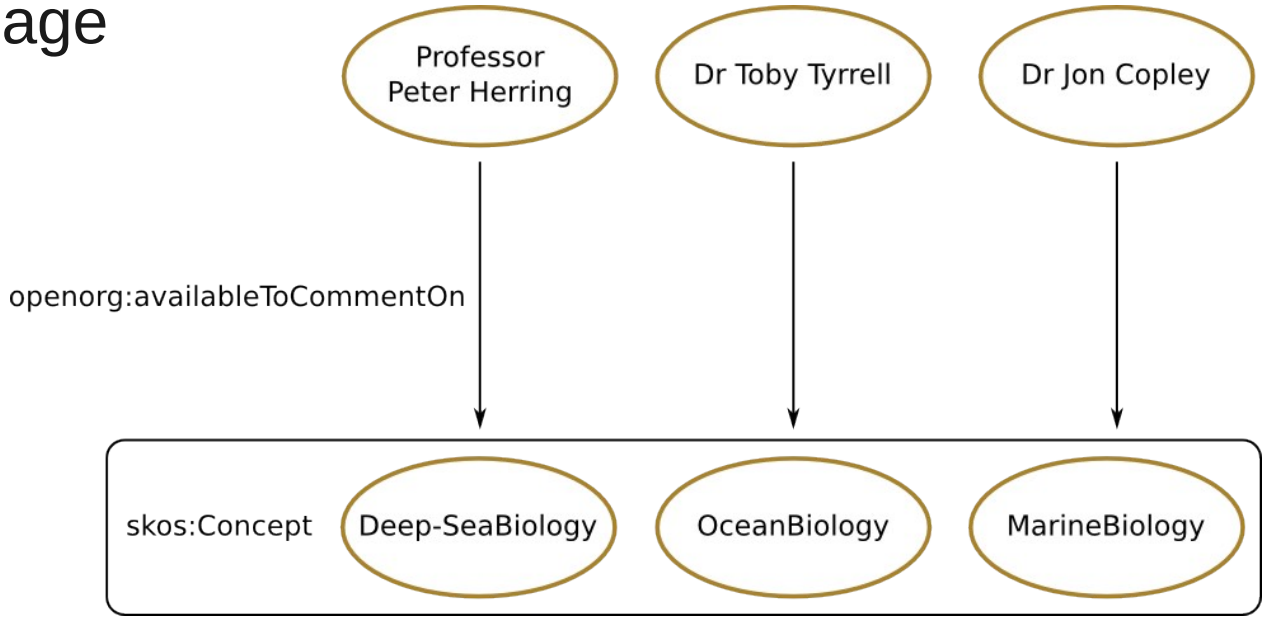


Methodology

- Identified criteria to support real-world use cases
- Reviewed occurrence in existing vocabularies
  - Press contacts information dataset (PCI), University of Southampton
  - STW Thesaurus for Economics, Leibniz Information Centre for Economics
  - New York Times People Vocabulary (NYTP)
  - LVAk Thesaurus, Austrian Armed Forces

Criterion – Loose Concepts

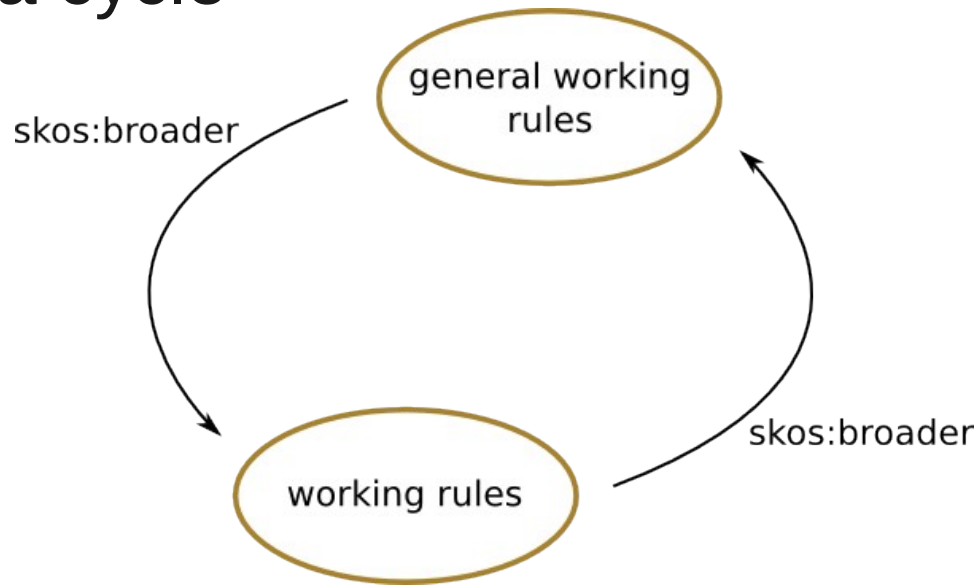
- Relative number of loose concepts
  - PCI
    - All concepts (1125) are loose concepts
    - Only for some concepts wikipedia references are defined, using foaf:page



Criterion – Cyclic Relations

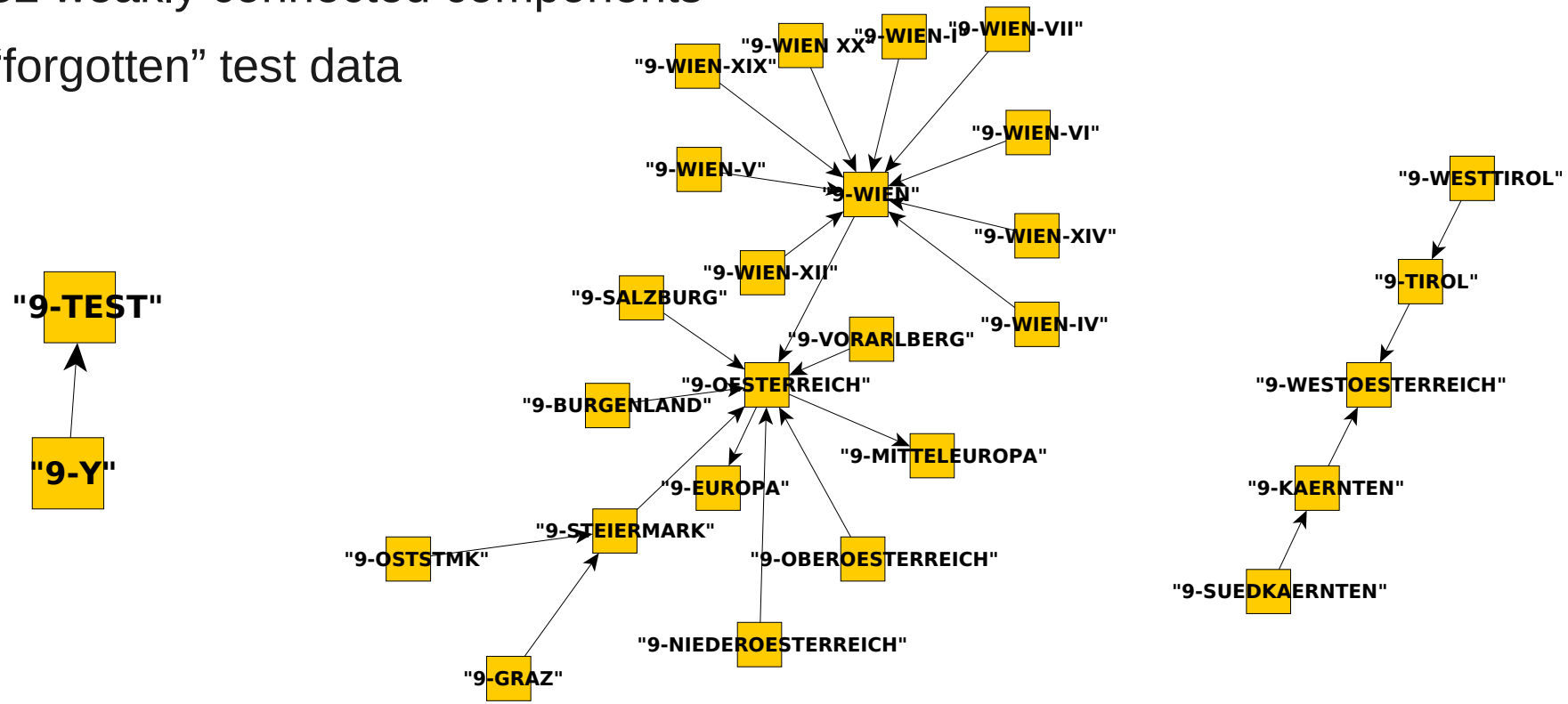
- LVAk
  - >13400 concepts
  - 6 cycles invoking hierarchical relations
  - 2-5 concepts involved in a cycle
  - Creator's feedback:

    “Relations need revision”



Criterion – Weakly Connected Components

- LVAk
  - 1 giant component (>13000 concepts)
  - 32 weakly connected components
  - “forgotten” test data



Criterion - Ambiguous labeling

- STW
  - Duplicates (7 concept pairs labeled identical)
  - Manifestation:

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Criterion - Multilinguality

- Increases potential vocabulary user base
  - PCI
    - none of the plain text literals have a language tag
  - STW
    - 3 languages available “en”, “de”, “x-other”
    - Majority of concepts (99%) labeled in 2 languages, rest (only 68) in 3 languages

Criterion – Linked Data Issues

- Degree of external links
  - PCI: 9 of 1125 concepts have foaf:page to wikipedia
  - STW and LVAk: none defined
  - NYTP: avg. 2.9 external links per concept
- Link Target Availability
  - PCI: 7 of 1669 total links not dereferenceable (99.6% dereferenceable)
  - NYTP: 90% dereferenceable
  - Indicator for vocabulary maintenance

Conclusions

- Found criteria have practical relevance
- Further criteria, documentation, in-depth coverage
  - qSKOS project

    https://github.com/cmader/qSKOS/wiki/Quality-Criteria-for-SKOS-Vocabularies
  - Work in progress
- Thank You!
  - christian.mader@univie.ac.at
  - Questions welcome!