Component Metadata Infrastructure Best Practices for CLARIN

CMDI and Metadata Curation Task Forces

Thomas Eckart, Twan Goosen, Susanne Haaf,
Hanna Hedeland, Oddrun Ohren, Dieter Van Uytvanck
and Menzo Windhouwer

CLARIN Annual Conference
18-20 September 2017
Budapest, Hungary

History

- CMDI 1.2 technical specification
  - CE-2016-0880 (via clarin.eu/cmdi1.2)
  - What is valid CMD? (baseline)
  - Generic, i.e. leaves a lot of freedom for specific implementations/uses
- CMDI Best Practice Guide
  - How does CLARIN deal with CMD in its (CMD) Infrastructure?
  - How can you optimize your CMD records for CLARIN?
  - Loosely based on some previous attempts
  - Gather scattered/implicit knowledge
  - Group of editors, but input from all Task Force members

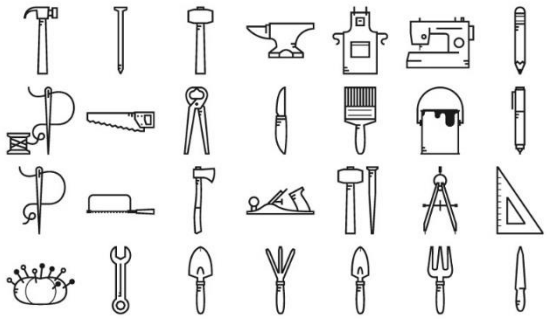
Draft available: CE-2017-1076

clarin.eu/content/cmdi-best-practice-guide

  - **Work in progress!**
  - **But your feedback is very welcome!**

Scope

- Audience
  1. CMD modellers, who select or create components and profiles to describe certain language resources
  2. Developers, who create metadata convertors, forms or editors to create records that comply with these profiles
  3. CMD creators, who create records by hand



Basic Structure of a Best Practice

E2: The MdSelfLink should be a Persistent Identifier (PID) [priority: medium] [CLARIN B Centre requirement: 6.7] [*TODO: check: CMDI Instance Validator*]

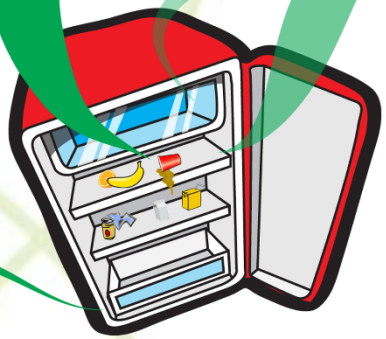
By giving a CMD record a PID and including it in the MdSelfLink it becomes possible for tools that present the records to the user to provide them with a persistent link suitable as a bookmark that will stand the test of time. This is another requirement for CLARIN B Centres

- ID and Title
- Priority
  - **High**: important ground principle of CMDI or the Web
  - **Medium** (default): only break if you have good (and documented) reasons
  - **Low**: good for CLARIN, but no bad consequences if you break it
- Relation to a CLARIN B Centre Requirement
  - **Important** if you are a CLARIN B Centre
- Is validation technically (partially) possible?
- Description
  - Rationale
    - good data modelling
    - infrastructure needs
    - community needs
  - Hints, tips & tricks

Workflow and Smells

- What is a good workflow?
- Which tools can help you?

- What are problem indicators (“smells)?
  - Results of design or implementation decisions that need special attention or be re-evaluated



Outline of the Guide

- Introduction
- *Workflow*
- Modelling Component Metadata
  - Components
  - Profiles
  - Workflow
  - Smells
- Authoring Component Metadata Records
  - General XML
  - The Envelope
  - The Component Section
  - Workflow
  - Smells
- *Common approaches/problems*
- *(Recommendations)*

A Sample of Best Practices for Modellers

- M1: Make your components, profiles and concepts “as generic as possible and as specific as needed”
  - **High**, but fuzzy
- C1: Provide detailed documentation
  - **High**, but non-conformance due to CMDI 1.1 legacy
- C2: Component, element and attribute names should be in English
  - **High**, but can’t be checked automatically
- C8: Prefer elements over attributes
  - **Low**, can be checked automatically
- C14: Add concept links to all elements, attributes, and vocabulary items
  - **High**, but hard to meet
- P1: Reuse Components
  - **High**, can be checked automatically

Workflow and Smells for Modellers

1. Specify the requirements
2. Checkout existing profiles and components
   - CLARIN Curation Module
   - SMC Browser (clarin.oeaw.ac.at/smc-browser)
   - cmdi@clarin.eu
3. If necessary, edit/create components and profile
4. Assess the private profile
   - CLARIN Curation Module (clarin.oeaw.ac.at/curate/)
   - VLO Mapping inspector (vlo.clarin.eu/mapping)
5. Publish the profile with development status
6. Before promoting the profile to production status: provide records and do a VLO test harvest
   - vlo@clarin.eu
7. Corrections or updates? Deprecate the profile and create a correct/up-to-date successor

A Sample of Best Practices for Authors

- E5: There should be at least one resource proxy
  - **High**, can be checked automatically
- E9: Specify the MIME type of a resource
  - **High**, can be checked automatically
- CS3: In case of multilinguality, explicitly specify the language used
  - **Medium**, to be checked automatically

Workflow and Smells for Authors

- More dependable on the centres
- Always
  1. Validate the CMD records against their profile XSD
  2. Assess the records using the CLARIN Curation Module
  3. “Smell”
- Also
  1. Check consistency, esp. for manually created records
  2. Do automated checks, e.g., via a set of Schematron rules
- After the VLO test import assess
  - appropriateness of all normalised and mapped values for VLO facets (contact cmdi@clarin.eu if needed)
  - appropriateness of the display of resource hierarchies

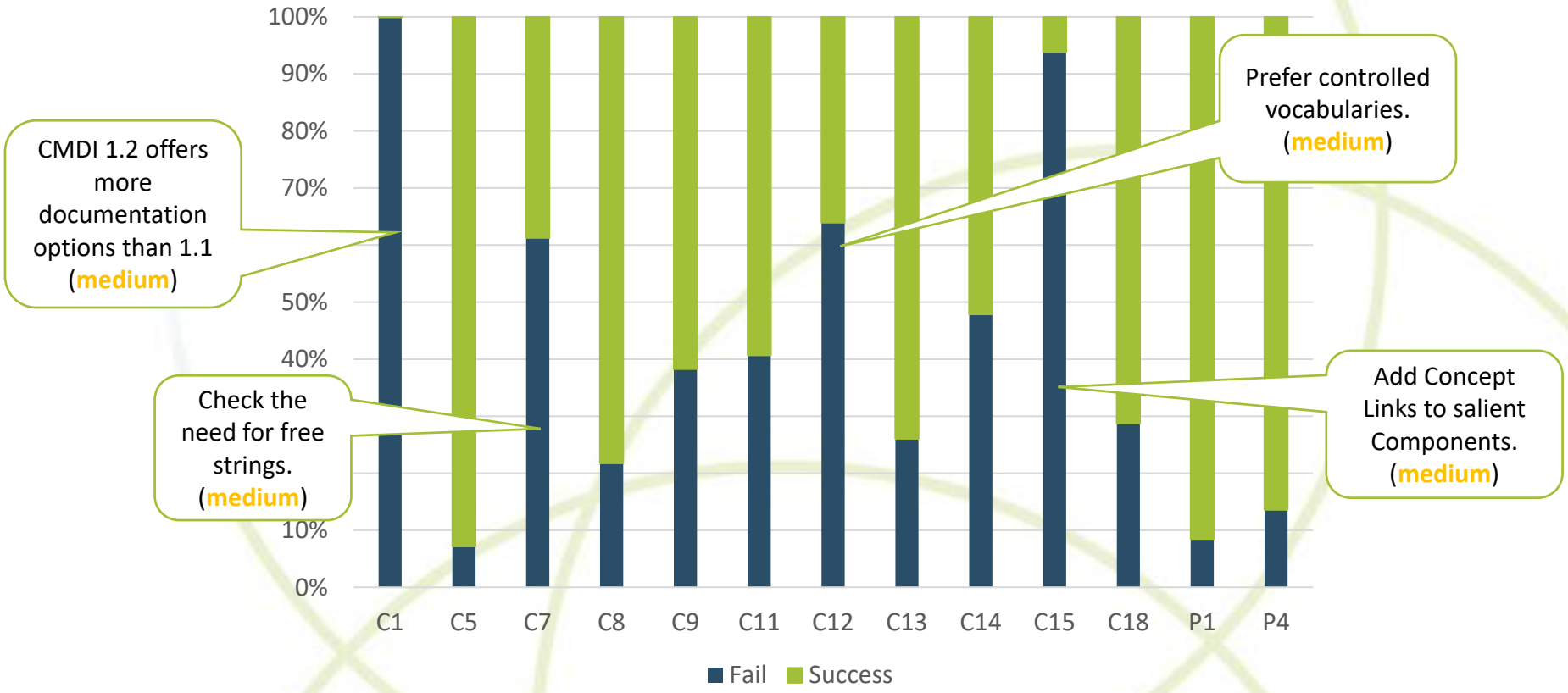
Smells, e.g.
1. very large records
2. no title or description
3. sparse records
4. empty elements or attributes
5. information entropy
6. structured strings
7. URLs in the payload

Checking Best Practices

- Many Best Practices need human interpretation!
  - but still many Best Practices can be (partially) technically supported
- Profiles and Components
  - CMDValidator, which is embedded in the Component Registry
  - Schematron rules based on the Best Practices:
    - To be part of the CMDI Toolkit (infra.clarin.eu/CMDI/1.2/sch/)
    - Currently under construction (`develop` at github.com/clarin-eric/cmdi-toolkit)
- Records
  - CMD Instance Validator (github.com/clarin-eric/cmdi-instance-validator)
  - Schematron rules based on the Best Practices
- In due time the CLARIN Curation Module will be the tool/service to assess CMD Profiles, Components and Records!

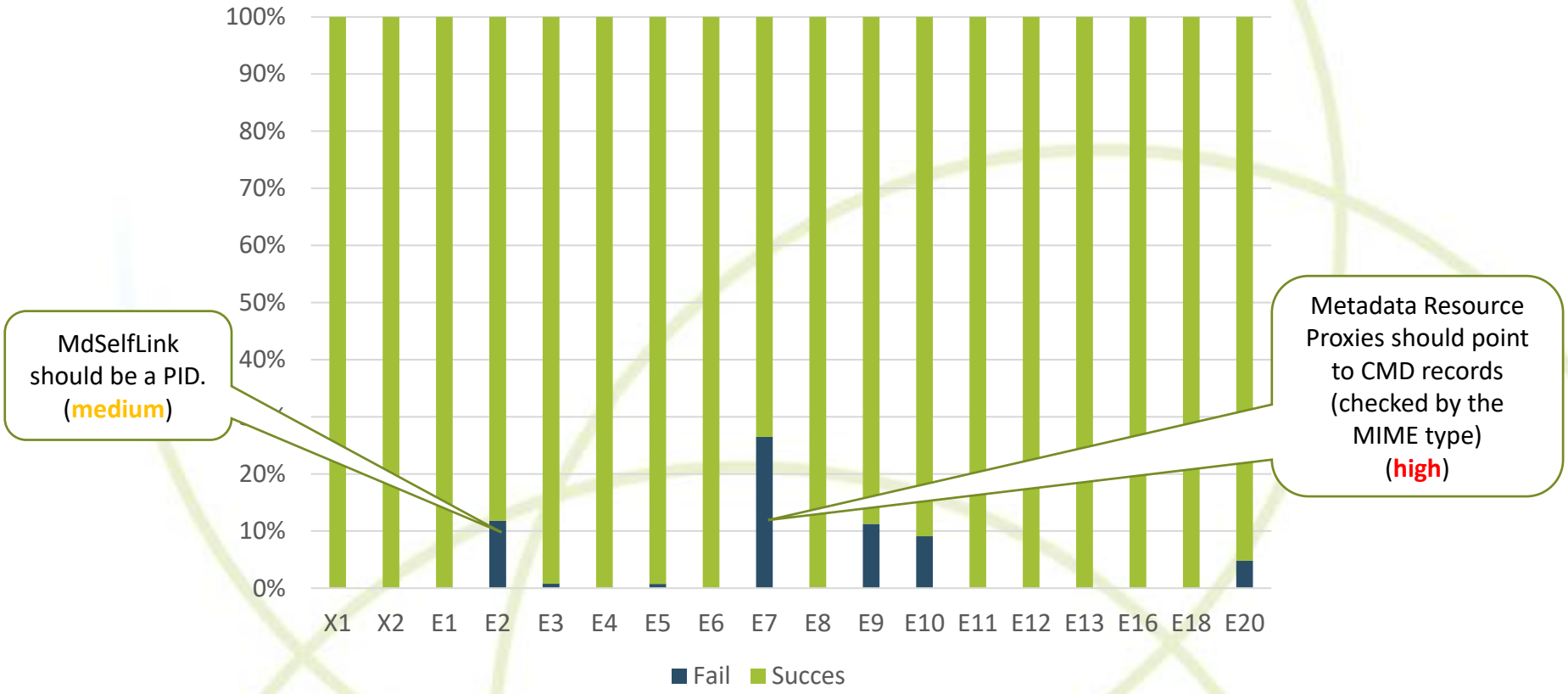
First Results for Technical Best Practice Checks

- Profiles and Components (1,448)
  - Public workspace of 2017, June 19



First Results for Technical Best Practice Checks

- CMD Records (661,409)
  - CLARIN OAI harvest of 2017, September 8
  - Baseline: 8,31% (54,933) records are **invalid**, i.e., don't adhere to their profile! 😟



Issues

- Numbering scheme?
  - Not persistent, better use short mnemonics?
- Utopian Best Practices
  - Encourage better practices! 👍
  - Or is it actually discouraging? 😟
    - You may never get a green slate ☑
- Best Practices are often hints
  - Please consider carefully all alternatives and make a conscious decision
    - Mark profiles and records if you've done so?
- CMDI 1.2 opened up new possibilities, but there is a lot of "legacy"
  - Documentation
  - Lifecycle management
    - Clean-up by admins (as far as possible)?
- VLO is not the ultimate goal, but a help
  - The goal is high quality metadata for the CLARIN community
  - Aim for minimal semantic coverage as defined by the community
    - expressed by identified (core) concepts
- Where do recommendations live?
  - Currently in the FAQ, which is too far away from everyday use
  - Preferably indicated in the Component Registry?

Feedback

- Thumbs up/down
- Missing Best Practices
- Conflicting Best Practices
- Understandability
- …

CE-2017-1076

clarin.eu/content/cmdi-best-practice-guide

- Talk to:
  - The editors: Thomas Eckart, Twan Goosen, Susanne Haaf, Hanna Hedeland, Oddrun Ohren, Dieter Van Uytvanck and Menzo Windhouwer
  - A CMDI or Metadata Curation Task Force member
- Or cmdi@clarin.eu