A simple tool to "validate" metadata content

DC Checker

Dr. Heinrich Stamerjohanns
Institute for Science Networking
at the
University of Oldenburg

Current OAI-PMH

- many implementations exist to support protocol
- people are trying to share their existing data through OAI-PMH protocol
- with the help of the Repository Explorer by Hussein Sulemann most implementations create formally correct XML metadata records
- yet problems still exist for Service Providers

OAI-PMH is not just replication

- now able to read data, but we want to have information
- obeying the protocol is not enough
  - correct xml encoding (and cleaning) of existing content
  - metadata format conversion
  - character format encoding
    - especially (limited) conversion of existing metadata (markup and word format) to utf8

Service Provider has to look at content...

- Dublin Core metadata format does not have strict content regulation
- SP cannot “trust” the incoming metadata
  - normalization necessary
  - metadata is formally correct, but lack of shared semantics
  - simple examples
    - DC.language "deutsch" → "ger"
    - DC.date "1.02.1999" → "1999-02-01"

DC Checker

- to check for correctness of protocol use Repository Explorer
- DC Checker does not aim to check formal correctness of metadata records (although implicitly it does)
- looks at content of metadata records, checks whether some simple rules are followed
- creates statistics of repositories

Dublin Core

- 15 basic elements
- No regulations, but (finally) recommendations
- DC Checker checks whether these recommendations are followed
- additional checks
  - detailed analysis of UTF8 problems
  - markup

Dublin Core metadata records

- dc:title
  - markup?
- dc:creator
  - Lastname, Firstname
  - one author in each record
- dc:subject,
  dc:publisher,
  dc:description,
  dc:contributor
  - markup?

Dublin Core metadata records

- dc:date
  - profile of ISO8601, **W3CDTF**
    YYYY (eg 1997)
    YYYY-MM (e.g. 1997-07)
    YYYY-MM-DD (e.g. 1997-07-16)
    YYYY-MM-DDThh:mmTZD
    (e.g. 1997-07-16T19:20+01:00)
    YYYY-MM-DDThh:mm:ssTZD
    (e.g. 1997-07-16T19:20:30+01:00)
    YYYY-MM-DDThh:mm:ss.sTZD
    (e.g. 1997-07-16T19:20:30.45+01:00)
  - markup

Dublin Core metadata records

- dc:type
  - controlled vocabulary really used?
  - no mimetypes, use dc:format
- dc:format
  - mimetypes
- dc:identifier
  dc:source
  - currently none

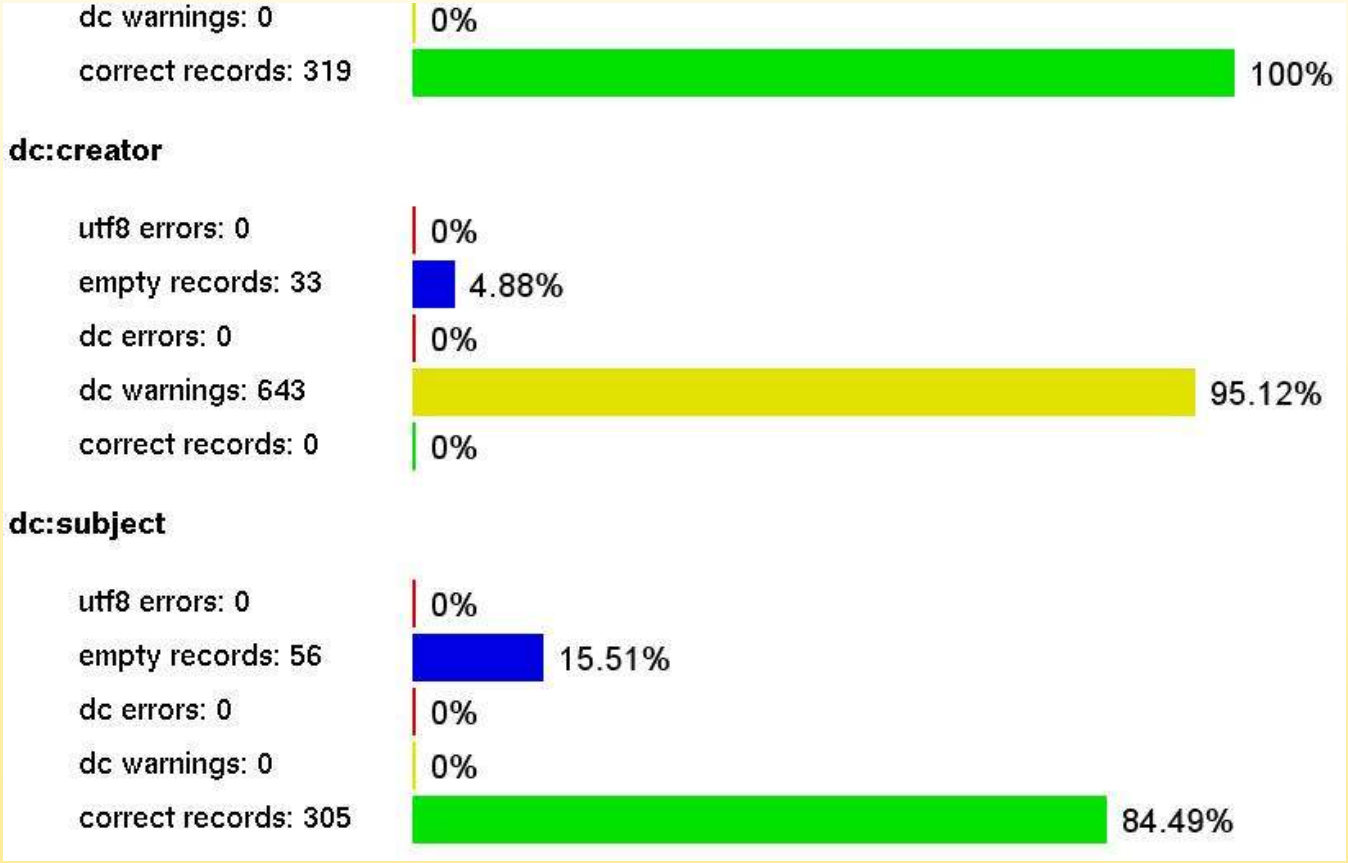
Collection
Dataset
Event
Image
InteractiveResource
PhysicalObject
Service
Software
Sound
Text

Dublin Core metadata records

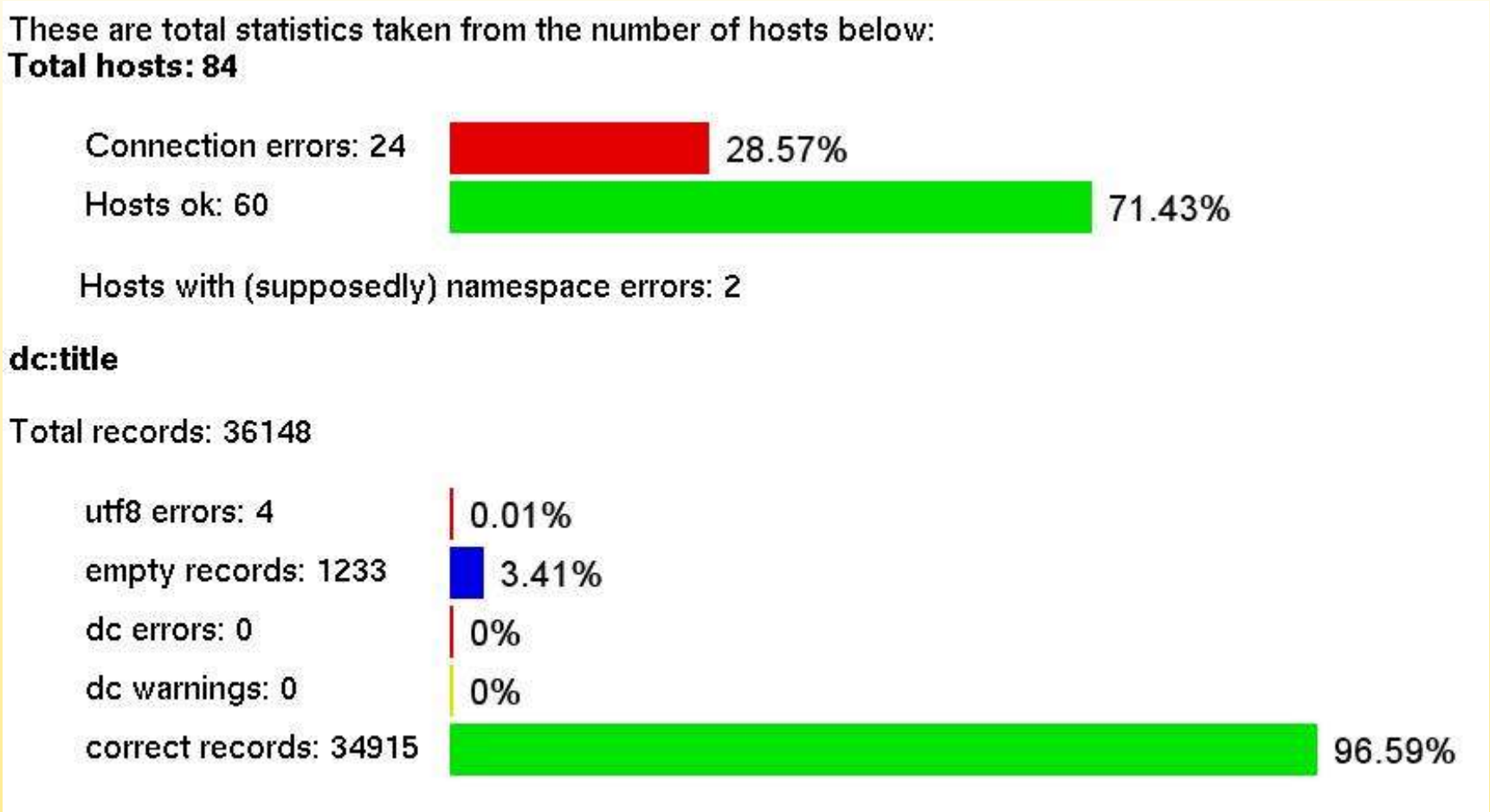
- dc:language
  - 2-letter tag from ISO639-1
    - e.g. de, en
  - 3-letter tag from ISO639-2
    - e.g. deu, eng
- dc:relation
  dc:coverage
  - markup?
- dc:rights
  - creative common license
    - no one uses it...

- DC Checker

http://harvest.physik.uni-oldenburg.de/dc/dcchecker.php



- Statistics
  http://harvest.physik.uni-oldenburg.de/dc/statistics.php
  - genererates total statistics of hosts



Thank you

- DC Checker at the Institute for Science Networking, Oldenburg:
  http://harvest.physik.uni-oldenburg.de/dc/

- stamer@uni-oldenburg.de