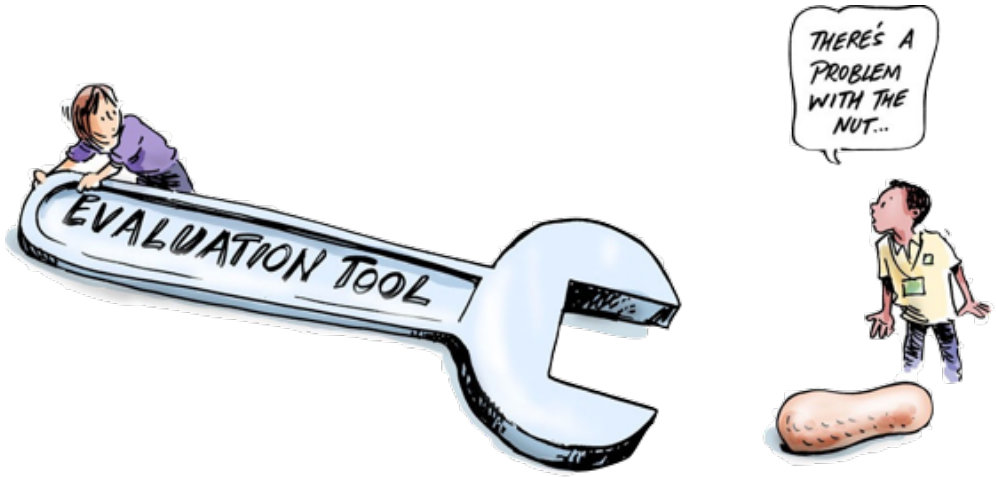
Proposal for an Evaluation Framework for Compliance Checkers for Long-term Digital Preservation

Nicola Ferro
University of Padua, Italy



12th Italian Research Conference on Digital Libraries - IRCDL 2016
Florence, Italy, 4-5 February 2016

UNIVERSITÀ DEGLI STUDI DI PADOVA

SEVENTH FRAMEWORK PROGRAMME

Project Identity Card

- PREFORMA is a **Pre-Commercial Procurement** project co-funded by the European Commission under its FP7-ICT Programme.
- **Start date**: 1 January 2014
- **Duration**: 48 month (end date: 31 December 2017)
- **Website**: www.preforma-project.eu
- **Contacts**
  - Project Coordinator: Borje Justrell, Riksarkivet, borje.justrell@riksarkivet.se
  - Technical Coordinator: Antonella Fresa, Promoter Srl, fresa@promoter.it
  - Communication Coordinator: Claudio Prandoni, Promoter Srl, prandoni@promoter.it

Project Aim and Objectives

- **The aim:** to address the challenge of implementing various good quality standardised file formats for preserving data content in the long term.

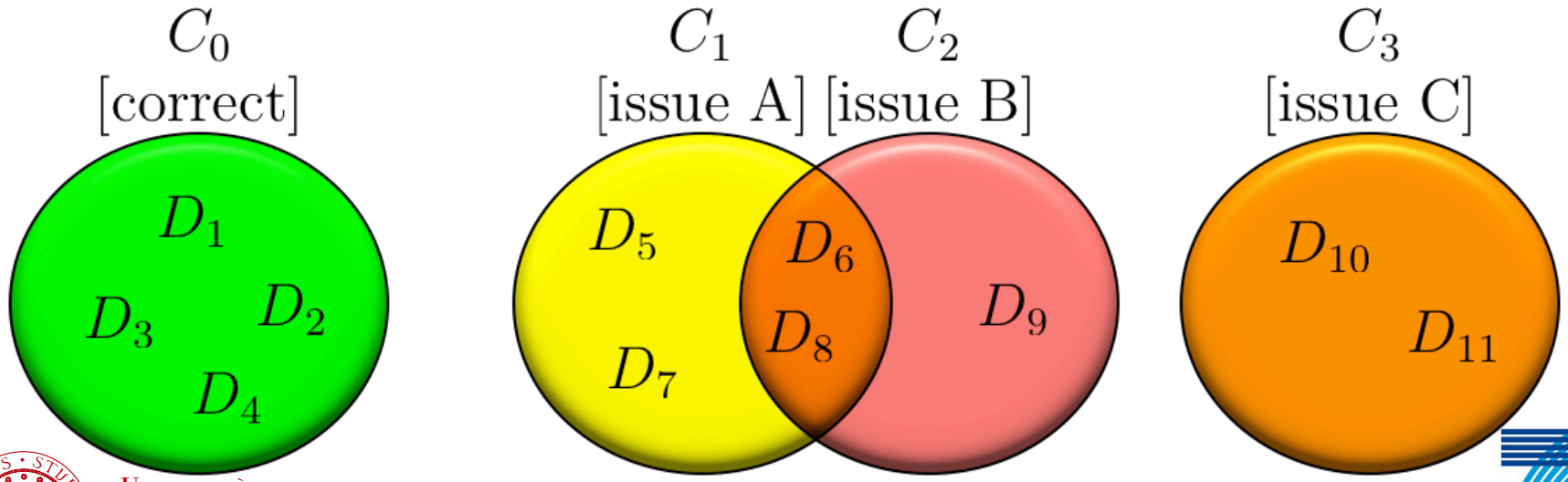
- **The main objective:** to give memory institutions full control of the process of conformity tests of files to be ingested into archives.

- **The main objective of the PCP launched by PREFORMA:** to develop and deploy an open source software licensed reference implementation for various file format standards, aimed for any memory institution (or other organisation with a preservation task) that wish to check conformance with a specific standard.

PREFORMA: A classification task

- The goal of PREFORMA is to **validate documents** (files) against their respective standards
  - this turns into determining for each document (file) whether it is correct, it has issue A, issue B, and so on

- We can frame this as a **classification** task where you label documents according to their characteristics
  - each label (correct, issue A, issue B, …) is a **class**
  - in general **classes** may **intersect** but the **correct** class must be **separate**



Critical Issues in Evaluation

- It must be scientifically valid
  - valid metrics, methodology, and statistics
  - large-enough scale to be statistically valid
  - must be "repeatable" if possible
- It must be realistic
- It must be understandable to your audience/client

How Does Experimental Evaluation Work

- Cranfield Paradigm
  - Dates back to mid 1960s
- Makes use of **experimental collections**
  - documents (corpora)
  - information needs
  - ground-truth
- Ensures **comparability** and **repeatability** of the experiments

Evaluation@PREFORMA: Information Needs/**Classes**

- For each media type, we need **domain experts** who determine the list of classes for that media type
  - known validation issues, potential validation issues, preservation issues, …
  - asking for classes to our suppliers may introduce a bias

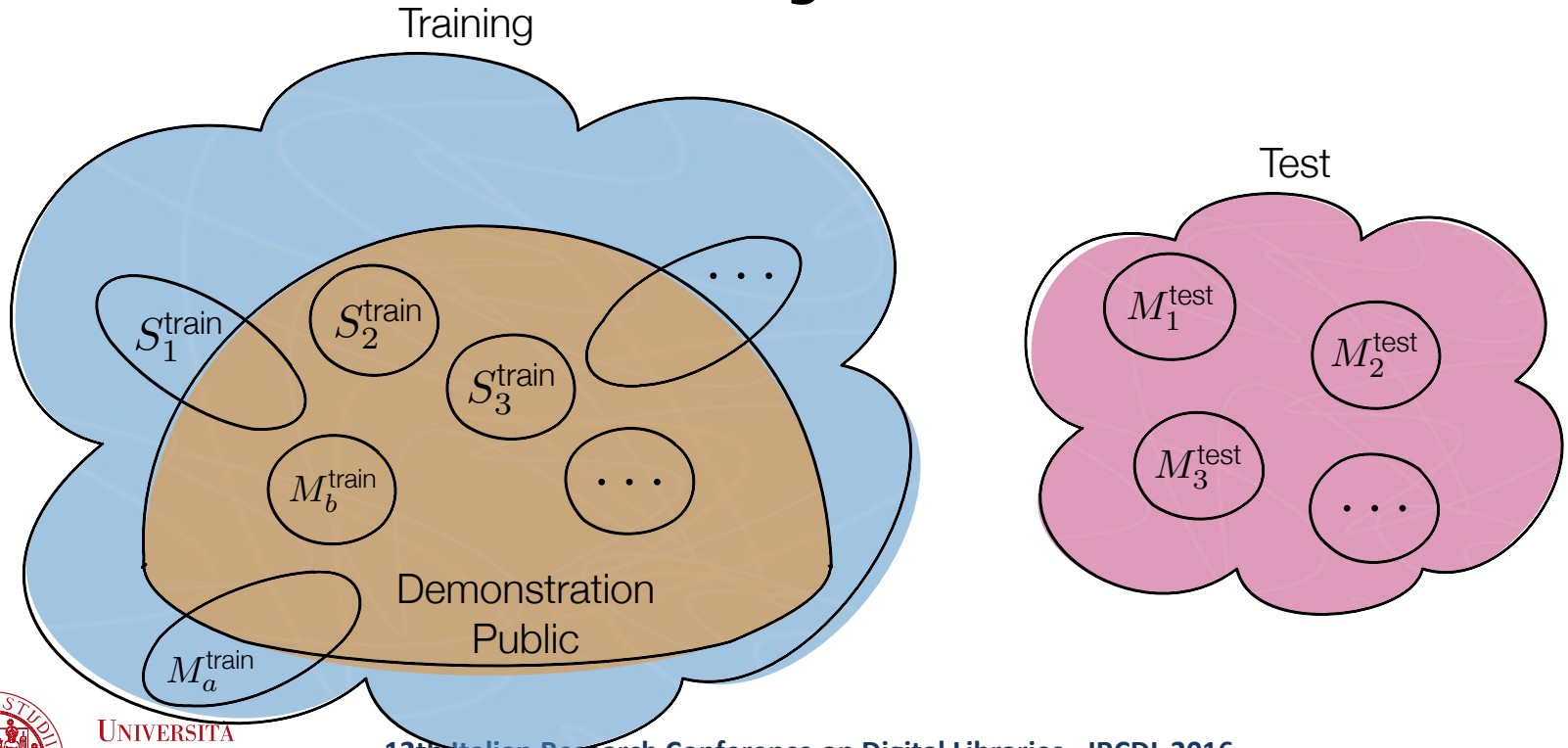
- We may also attach a **severity** to each class
  - some issues are errors, some others are warnings, some others are mis-conformances to policies and best practices

Evaluation@PREFORMA: Documents (1/2)

- **Huge sample** (ten thousands) for each media type (text, image, audio)
  - memory institutions, suppliers, community
  - each document must be uniquely identified
- Documents can be **real** or **synthetic**
  - see, e.g. Becker and Duretec JCDL 2013 templating approach
- Documents must be **representative** of the different classes we experiment
  - we cannot have empty class
  - the cardinality of each class should make sense
- The whole process must be driven by **domain experts**

Evaluation@PREFORMA: Documents (2/2)

- Critical split: **training** vs **test** set
  - to avoid bias, supplier should not provide documents for testing



Evaluation@PREFORMA: Ground Truth

- **Manual assessment**, i.e. determining for each document to which classes it belongs to, is typically **not avoidable**
- **Domain experts** are crucial
- **Automatic assessment** is often hoped for but it risks to introduce **bias** towards existing tools and suppliers tools

Evaluation@PREFORMA
What to Measure?

- Evaluating suppliers tools is not just going through an expected feature list and check it

Evaluation@PREFORMA Confusion Matrix

Class C_i		Ground-Truth	
		Positive	Negative
Conformance Checker	Positive	True Positive (TP_i)	False Positive (FP_i)
	Negative	False Negative (FN_i)	True Negative (TN_i)

Evaluation@PREFORMA Typical Measures

- The confusion matrix allows us to compute several measures, e.g.
  - **Accuracy**: overall effectiveness of a supplier tool

$$\text{Accuracy}_i = \frac{|TP_i| + |TN_i|}{|TP_i| + |TN_i| + |FP_i| + |FN_i|}$$

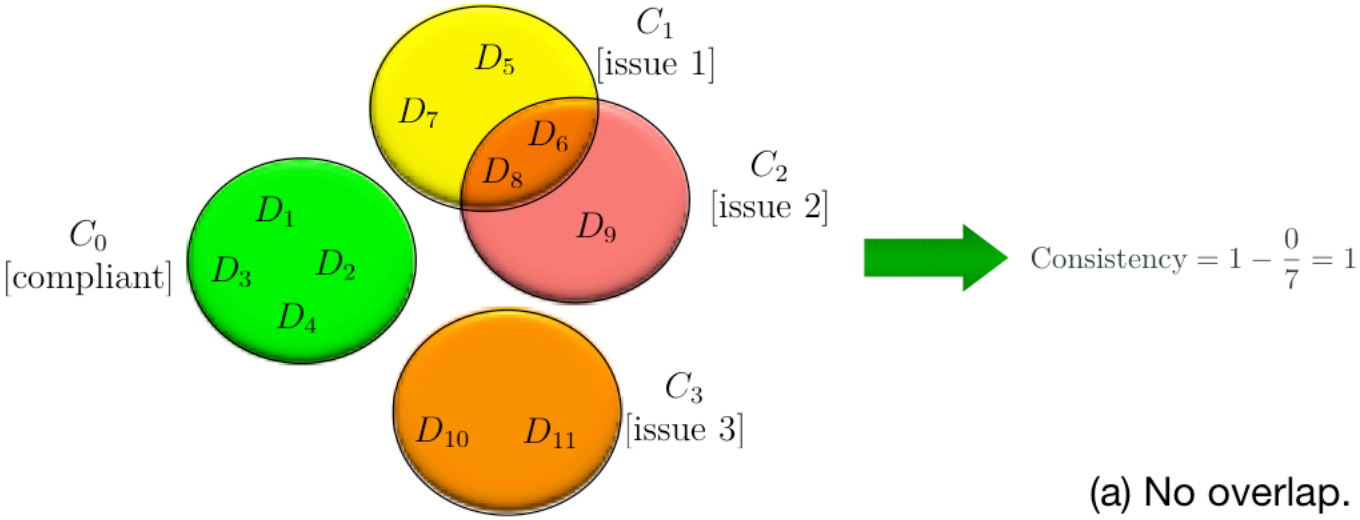
  - **Area Under the Curve (AUC)**: supplier tool's ability to avoid false classification

$$\text{AUC}_i = \frac{1}{2}\left(\frac{|TP_i|}{|TP_i| + |FN_i|} + \frac{|TN_i|}{|TN_i| + |FP_i|}\right)$$

Evaluation@PREFORMA Consistency

$$\text{Consistency} = 1 - \frac{\sum_{i=1}^{N} |(TP_0 \cup FP_0) \cap (TP_i \cup FP_i)|}{\sum_{i=1}^{N} |(TP_i \cup FP_i)|}$$

$$= 1 - \frac{\sum_{i=1}^{N} |C_0 \cap C_i|}{\sum_{i=1}^{N} |C_i|}$$

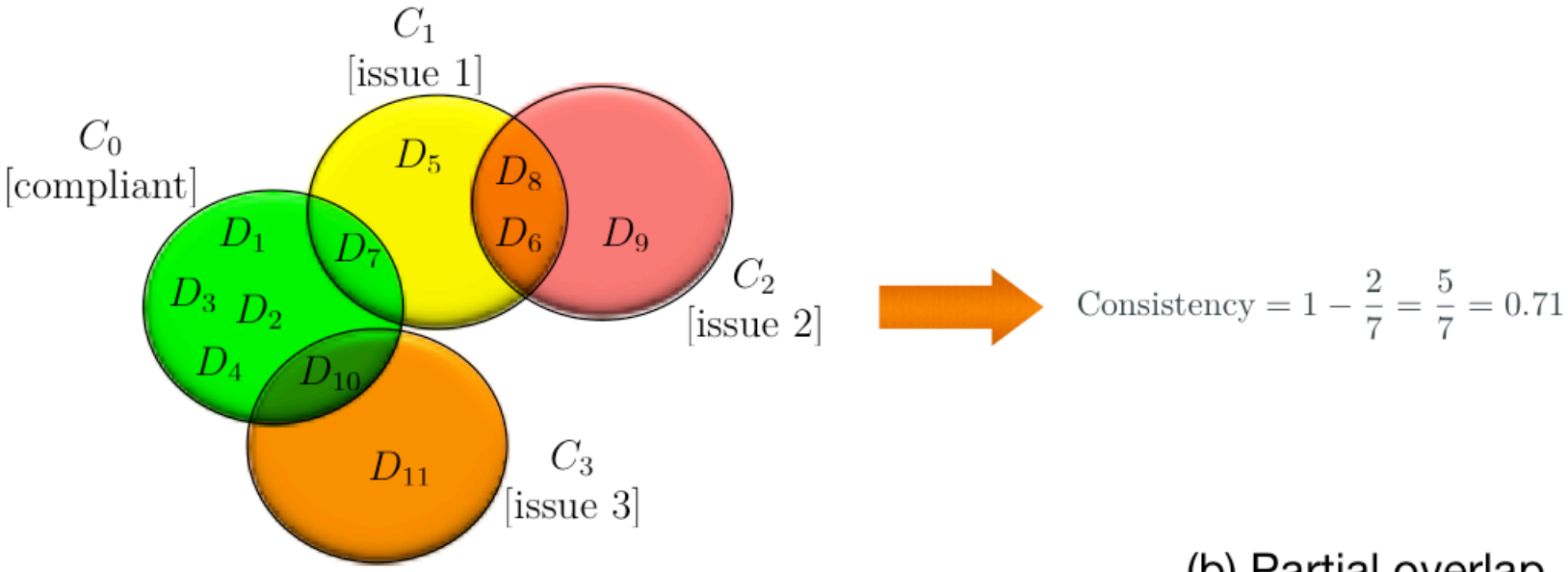


(a) No overlap.

Evaluation@PREFORMA Consistency

$$\text{Consistency} = 1 - \frac{\sum_{i=1}^{N} |(TP_0 \cup FP_0) \cap (TP_i \cup FP_i)|}{\sum_{i=1}^{N} |(TP_i \cup FP_i)|}$$

$$= 1 - \frac{\sum_{i=1}^{N} |C_0 \cap C_i|}{\sum_{i=1}^{N} |C_i|}$$

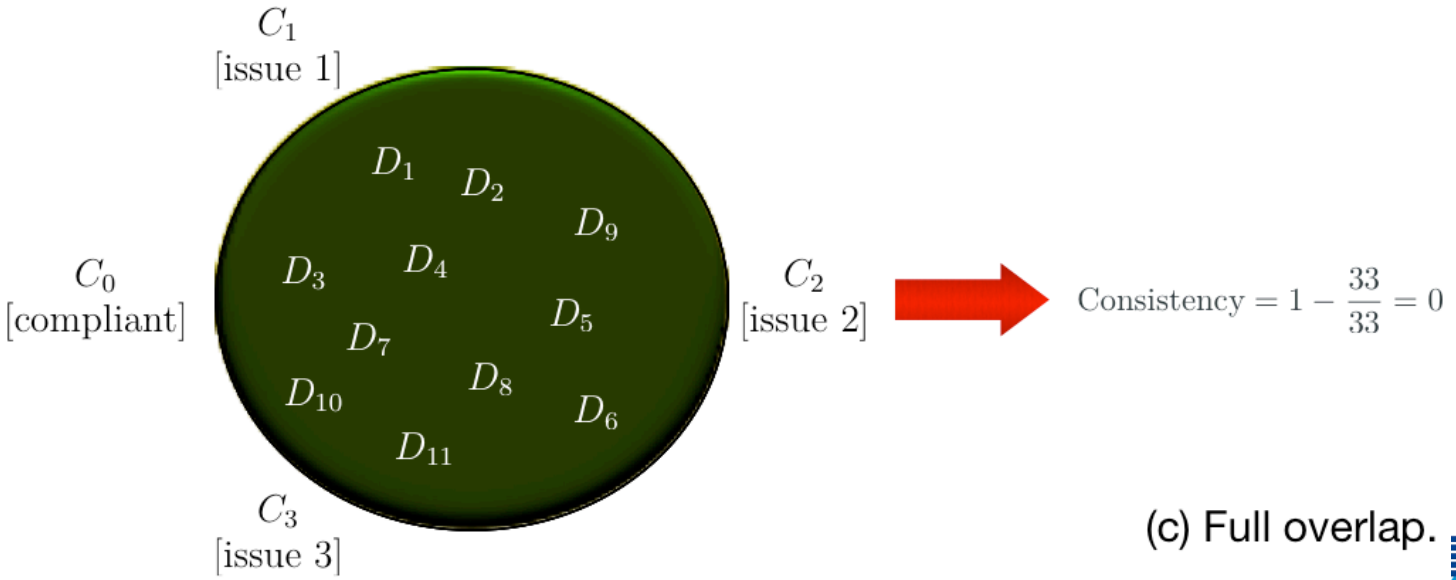


(b) Partial overlap.

Evaluation@PREFORMA Consistency

$$\text{Consistency} = 1 - \frac{\sum_{i=1}^{N} |(TP_0 \cup FP_0) \cap (TP_i \cup FP_i)|}{\sum_{i=1}^{N} |(TP_i \cup FP_i)|}$$

$$= 1 - \frac{\sum_{i=1}^{N} |C_0 \cap C_i|}{\sum_{i=1}^{N} |C_i|}$$



(c) Full overlap.

Conclusions

- We discussed how to model the process of conformance checking for long-term digital preservation and, consequently how to evaluate it
- We then discussed how to instantiate the Cranfield paradigm for the specific purpose of evaluating conformance checkers
- Feedback from the research community is much appreciated before we take the next step, which is to instantiate the proposed approach to evaluate PREFORMA suppliers

Follow us!

PREFORMA

PREFORMA Website

www.preforma-project.eu

PREFORMA Blog

www.digitalmeetsculture.net/projects/preforma/


ANY
QUESTIONS
?