How to test and evaluate the results of the prototyping phase

Antonella Fresa
Promoter Srl
PREFORMA Technical Coordinator

EC Concertation Meeting for PCP projects
Brussels, 10 March 2016

Introduction

Project Identity Card

- PREFORMA is a **Pre-Commercial Procurement** project co-funded by the European Commission under FP7-ICT Programme.
- **Start date**: 1 January 2014
- **Duration**: 48 month (end date: 31 December 2017)
- **Total budget for the procurement**: 2.805.000 EUR
- **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 Partners

- RIKSARKIVET, Sweden **Project Coordinator and memory institution**
- PROMOTER SRL, Italy **Technical and Communication Coordinator**

- **Technical partners**
  - PACKED EXPERTISECENTRUM DIGITAAL ERFGOED VZW, Belgium
  - FRAUNHOFER-GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V, Germany
  - HOGSKOLAN I SKOVDE (University of Skovde), Sweden
  - UNIVERSITA DEGLI STUDI DI PADOVA, Italy

- **Memory institutions**
  - STICHTING NEDERLANDS INSTITUUT VOOR BEELD EN GELUID, Netherlands
  - Koninklijk Instituut voor het Kunstpatrimonium, Belgium
  - GREEK FILM CENTRE AE, Greece
  - LOCAL GOVERNMENT MANAGEMENT AGENCY-AN GHNIOMHAIREACHT BAINISTIOCHTA RIALTAIS AITIUIL, Ireland
  - STIFTUNG PREUSSISCHER KULTURBESITZ, Germany
  - AYUNTAMIENTO DE GIRONA, Spain
  - Eesti Vabariigi Kultuuriministeerium, Estonia
  - KUNGLIGA BIBLIOTEKET, Sweden

Overall R&D Objective
(The PREFORMA Challenge)

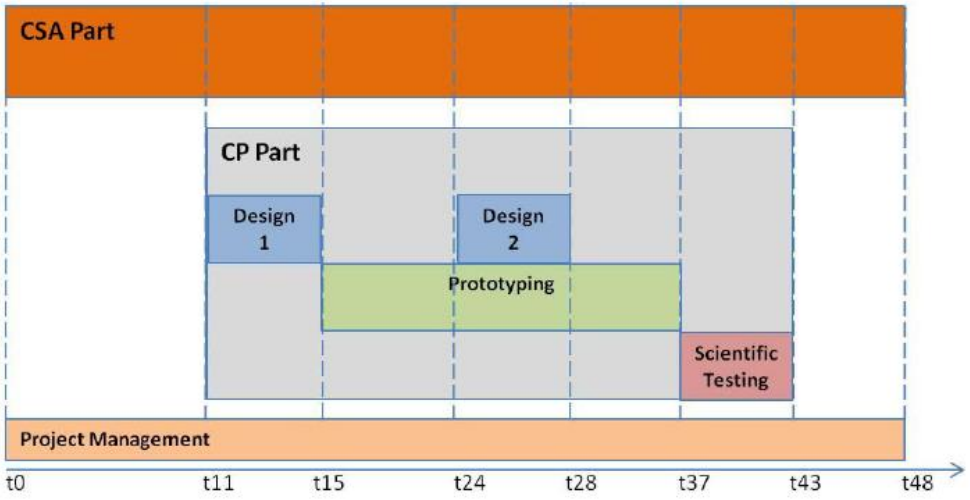
- Develop an **open source conformance checker** that:
  - checks if a file complies with standard specifications
  - checks if a file complies with the acceptance criteria of the memory institution
  - reports back to human and software agents
  - perform simple fixes

- Use cases:
  - Conformance Checking at **Creation Time**
  - Conformance Checking at **Transfer time**
  - Conformance Checking at **Digitization time**
  - Conformance Checking at **Migration time**

Standard specifications to be checked

Content type	Standard specifications
ELECTRONIC DOCUMENT	**PDF 1.7** (ISO 32000-1:2008) **PDF/A-1** (ISO 19005-1:2005) **PDF/A-2** (ISO 19005-2:2011) **PDF/A-3** (ISO 19005-3:2012)
IMAGE	**TIFF/EP** (ISO 12234-2:2001) **TIFF/IT** (ISO 12369:2004)
AUDIOVISUAL	**MKV** (http://www.matroska.org/technical/index.html) **Lossless FFV1** http://www.ffmpeg.org/~michael/ffv1.html **Linear PCM** (IEC 60958-1 ed3.1 Consol. with am1: 2014)

Project implementation schedule



- **Design phase** (4 months): November 2014 – February 2015
- **Prototyping phase** (22 months): March 2015 – December 2016
  - First prototypes: March 2015 – October 2015
  - Re-design: November 2015 – February 2016
  - Second prototype: March 2016 – December 2016
- **Testing phase** (6 months): January 2017 – June 2017

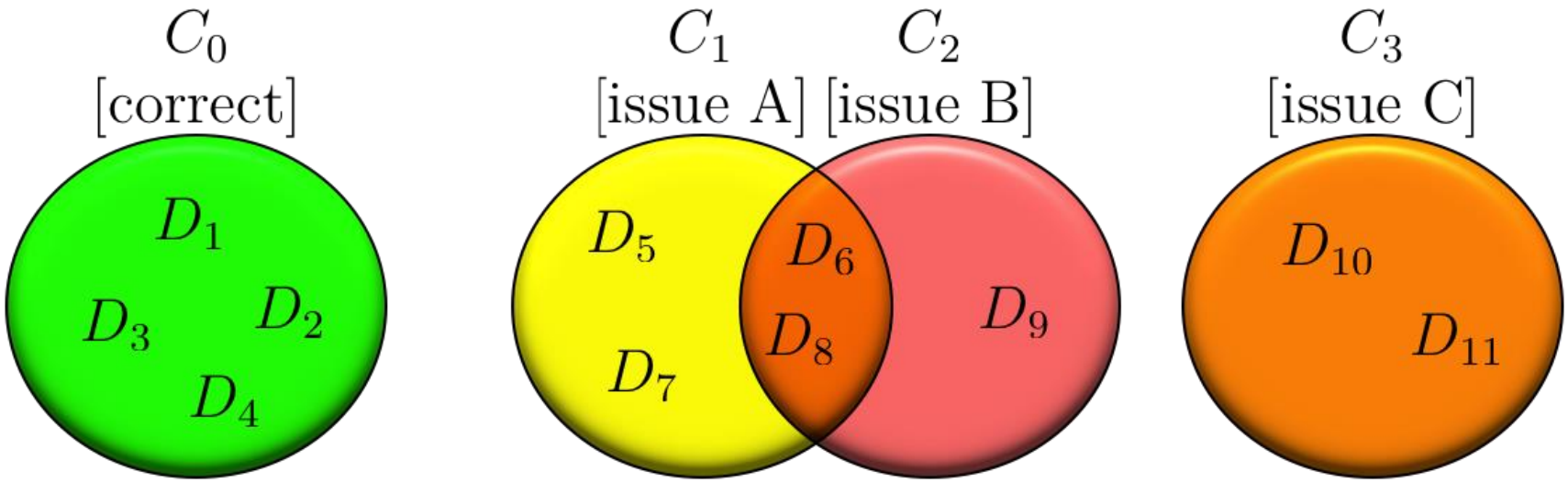
PREFORMA Suppliers in the prototyping phase

1. veraPDF Consortium (led by Open Preservation Foundation and PDF Association)
   - The PDF/A conformance checker accepted industry-wide (PDF/A)
2. EasyInnova
   - Digital Preservation Formats Manager (TIFF)
3. MediaArea
   - PREFORMA MediaConch - CONformance CHecking for audiovisual files (MKV|FFV1|LPCM)

Preparation of Phase 3

A classification task

- The goal of PREFORMA is to **validate file**(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

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

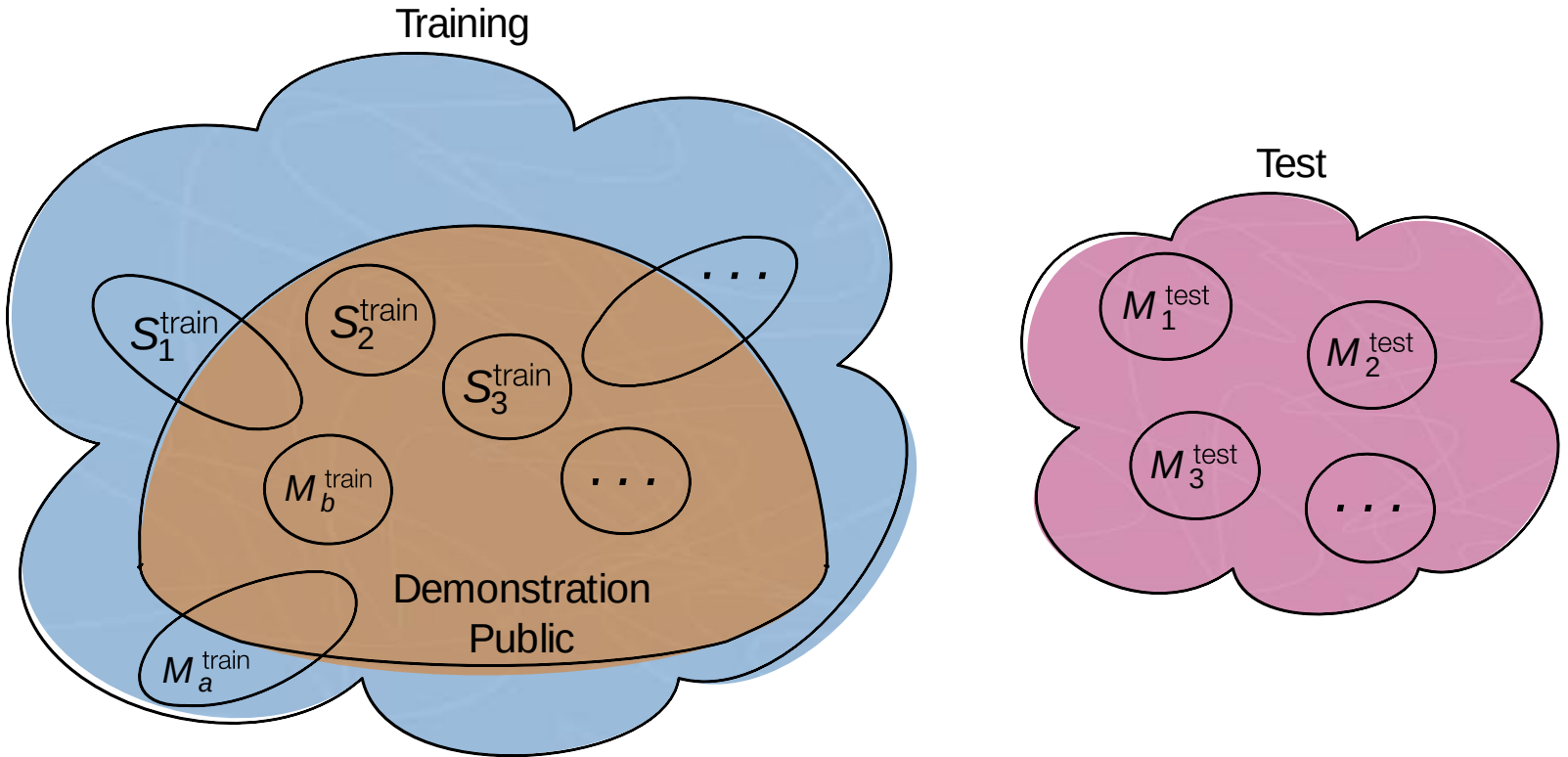
- We may also attach a **severity** to each class
  - some issues are errors, some others are warnings, some others are mis-conformances to policies and best practices

Datasets (1/2)

- **Huge sample** for each media type (electronic document, image, audio)
  - memory institutions, suppliers, community
  - each document must be uniquely identified
- Documents can be **real** or **synthetic**
- Documents must be **representative** of the different classes we experiment
  - we cannot have empty class
- The whole process must be driven by **domain experts**

Training, test and demonstration files

- **Training** vs **test**: critical split
  - to avoid bias, supplier should not provide documents for testing
- **Demonstration** files
  - should be public and made available together with the open source project
- The **community of public procurers** is invited to contribute training and demonstration files



Ground Truth

- **Manual assessment** is typically **not avoidable**
  - determining for each document to which classes it belongs to
  - **Domain experts** are crucial

- **Automatic assessment** is often hoped for but it risks to introduce **bias** towards existing tools and suppliers tools

Confusion Matrix

PREFORMA

Class C_i		Ground Truth	
		Positive	**Negative**
Supplier Tool	**Positive**	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**True Negative (TN)**

- For each class (issue) we create a **confusion matrix**
  - **Positive** = the problem is detected by the tool or exists in the file
  - **Negative** = the problem is not detected by the tool or does not exist in the file
- **False Negatives** are the worst error for the tool, since they are **not detected in a not conforming documents**
- False Positives are just false alarms

Performance Measures

- The confusion matrix allows us to compute several measures, e.g.
  - **Accuracy**: overall effectiveness of a supplier tool
  - **Area Under the Curve (AUC)**: supplier tool's ability to avoid false classification

Preliminary activities are very important

- Preparation of a **Data management plan** for training, testing and demonstration files
- Set up the **PREFORMA Vault**, the central logical point where all providers submit their files
- Launch a **call for participation** to involve content provider outside the PREFORMA Consortium to participate in the testing phase: so far **16** external memory institutions are uploading files in the Vault
- Prepare a specific **form** to collect information that needs to be attached to the files to be able to understand and analyse the results of the tests (descriptive metadata, technical properties, copyright restrictions, expected behavior, …)

Cooperation with other projects

- **BenchmarkDP**
  - Use shared methodologies and approaches to establish an objective frame of reference for the evaluation of the conformance checkers
- **Europeana Space**
  - Integrate the conformance checkers in the Technical Space, a web based application for the development of applications and services based on digital cultural content
- **E-ARK**
  - Use the PREFORMA tools in E-ARK pilot archival services
- **AppHub**
  - Additional channel to distribute the conformance checkers
  - Evaluate and incorporate the code quality and OSS risk management best practices developed by the AppHub community

Next appointments

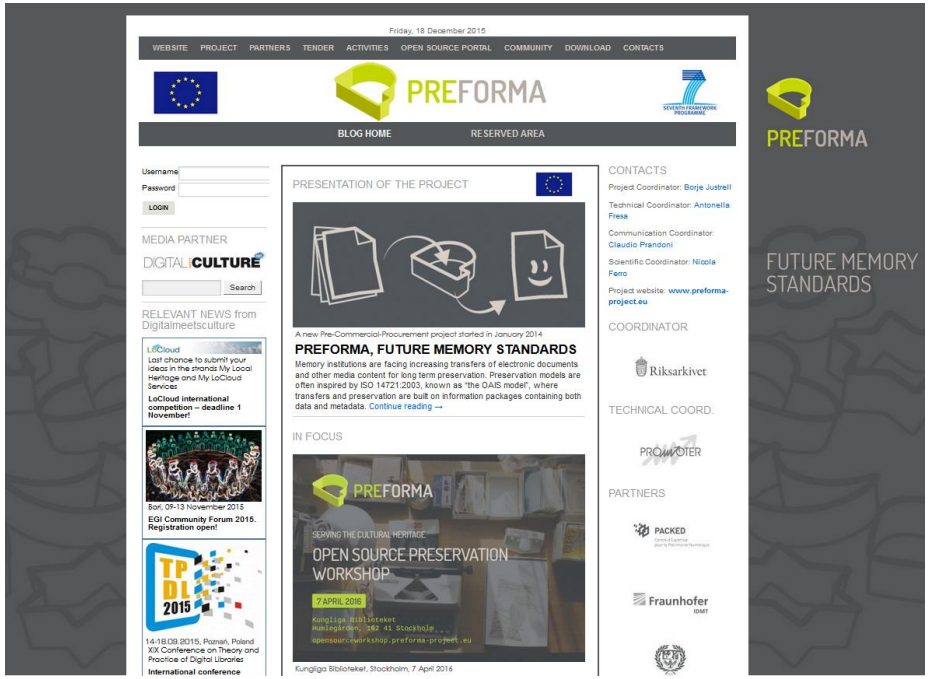
- **Open Source Workshop**, Stockholm, 7 April 2016
- **Experience Workshop for memory institutions**, Berlin, December 2016

Follow us!

PREFORMA

PREFORMA Website

www.preforma-project.eu



PREFORMA Blog

www.digitalmeetsculture.net/preforma

PROMOTER
Information technology, research and innovation

EC Concertation Meeting for PCP projects
Brussels, 10 March 2016

SEVENTH FRAMEWORK PROGRAMME

Thank you!

Antonella Fresa
PREFORMA Technical Coordinator

Promoter Srl

fresa@promoter.it

PREFORMA

PROMOTER
Information technology, research and innovation

EC Concertation Meeting for PCP projects
Brussels, 10 March 2016

SEVENTH FRAMEWORK PROGRAMME