Opening Schrödingers Library: Semi-automatic QA Reduces Uncertainty in Object Transformation

Lars Clausen
State and University Library
Århus, Denmark

ECDL 2007, Budapest, Hungary

Preservation of digital objects

- Recognized as a long-term problem
- Planets, DPE, ...
- Planning, characterization, actions
- Registries for formats and tools

Transformation as preservation

- One of several possible strategies
- Many known issues:
  - Format differences
  - Format specification issues
  - File quality issues
  - Transformation tool issues
  - System differences
  - Choice of format

Current status of QA

- Mostly manual inspection
- Most objects not inspected
- Quality is uncertain unless inspected
- In 100 years, are the objects “digital dust”?
- Focus of this paper
  - Not file format choice
  - Not planning or characterization

Aspects

- Suggested by Anders Johansen
- Generalized notion of “essential properties”
- Overlapping, abstract “views” of objects
- Independent of file formats

Semi-automatic QA

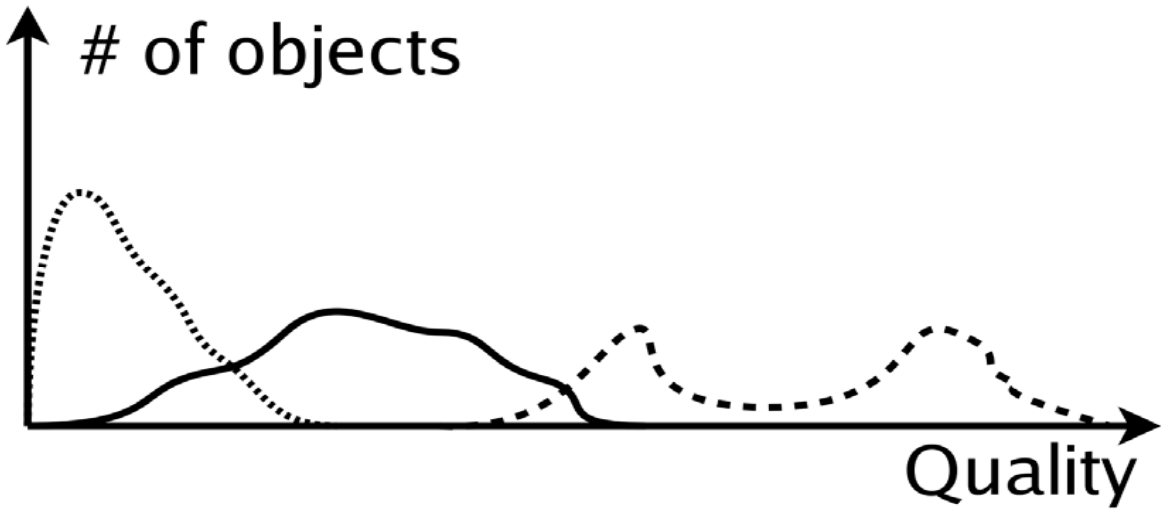
- Quality assessment system
- Uses many small comparisons (*measures*)
- Minor errors average out
- Gives a rough idea of quality
- Identifies good and bad examples
- Particularly useful in tool building/evaluation

Use of intermediate results

- Multiple transformations bad in preservation
- In QA, multiple transformations are useful
- Each transformation extracts partial data
- Simplifies comparison and understanding
- Allows use of simple third-party tools

Statistics on measurements

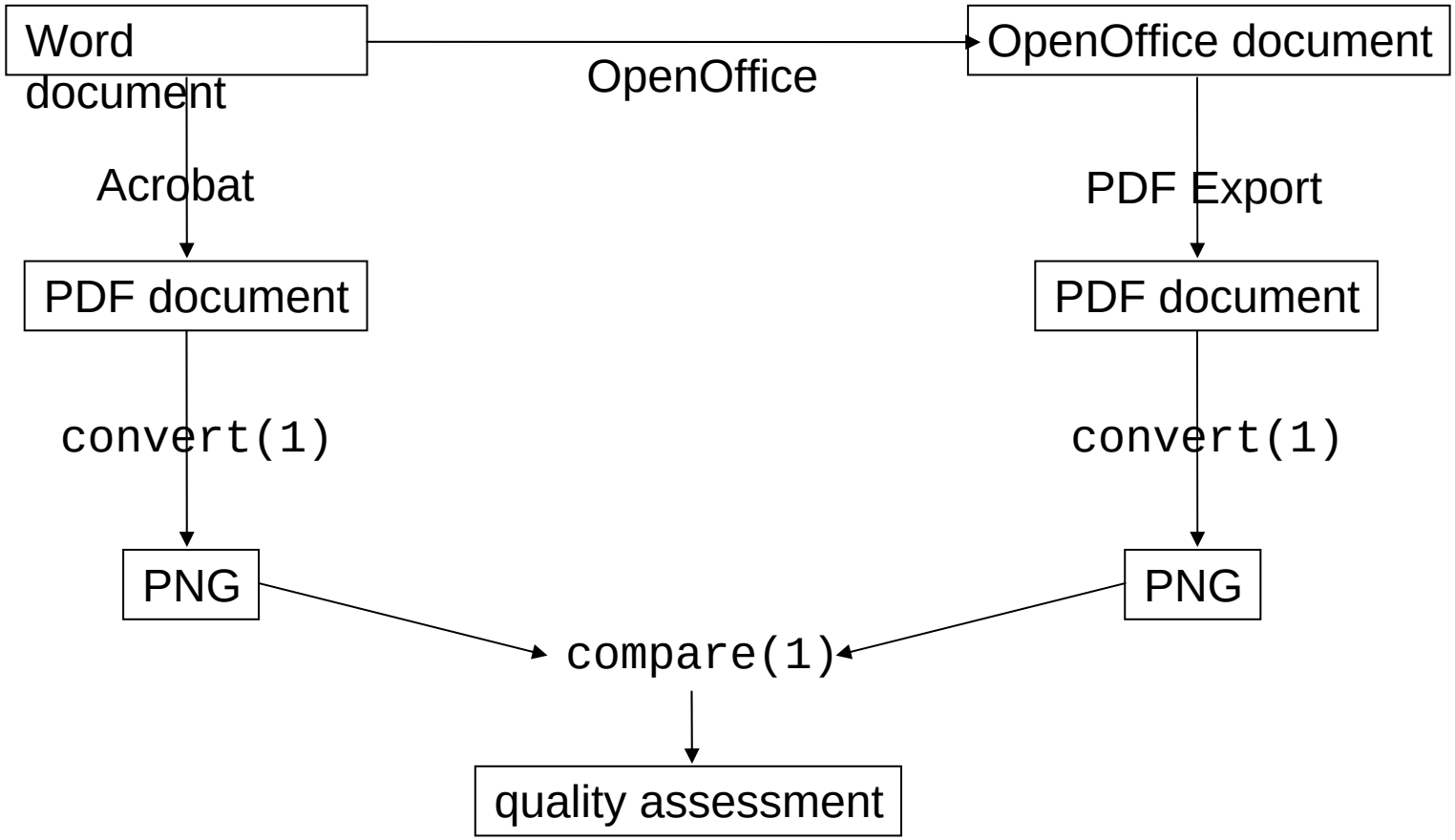
- Measurements are within arbitrary range
- End-points should be found manually
- Normalization destroys relative importance
- Potentially large differences in severity



Example: Word documents

- Approach tested with random Word files
- Test of quality of OpenOffice import
- Converted to PDF with Adobe Acrobat
- Converted to PDF with OpenOffice
- PDFs compared with multiple tools

Extraction chain



Not a Panacea

- There Is No Silver Bullet
- Manual inspection needed
  - Sanity check, check best & worst
- “Computer-assisted QA”
  - Present original and transformed to user
  - Side-by-side, flipping, diff, ???

Planets Developers QA Framework

- Framework for managing measure data
- Caches intermediate results
- Allows flexible data extraction
- First prototype works, only one measure
- More measures trivial to add
- Soon available under LGPL license?