gimp.org

ICDAR-OST 2019

CHRISTIAN CLAUSNER

PRIMA RESEARCH LAB, UNIVERSITY OF SALFORD, UK

University of Salford MANCHESTER

ABOUT ME

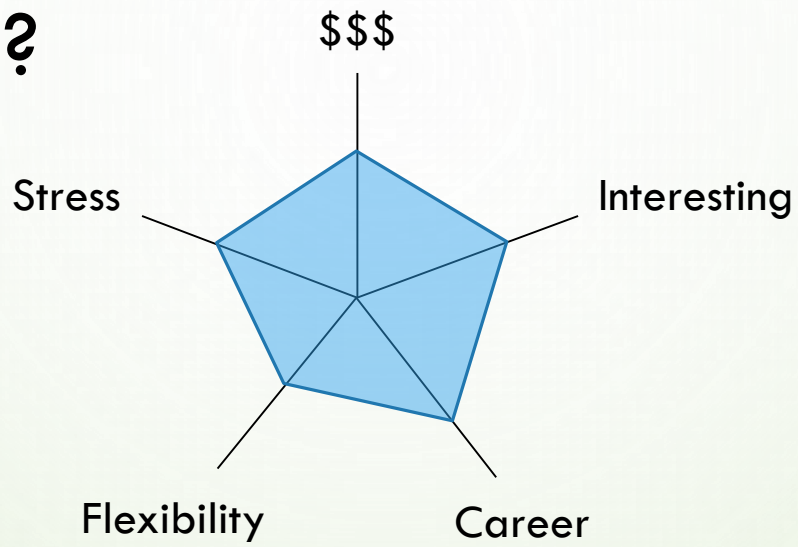
- 2000: Chemnitz University of Technology
  - Master's in Computer Science
- 2006: Industry (Software Dev)
- 2009: University of Salford (Research Fellow)
  - 2019: PhD ☑ (finally)

28/09/2009

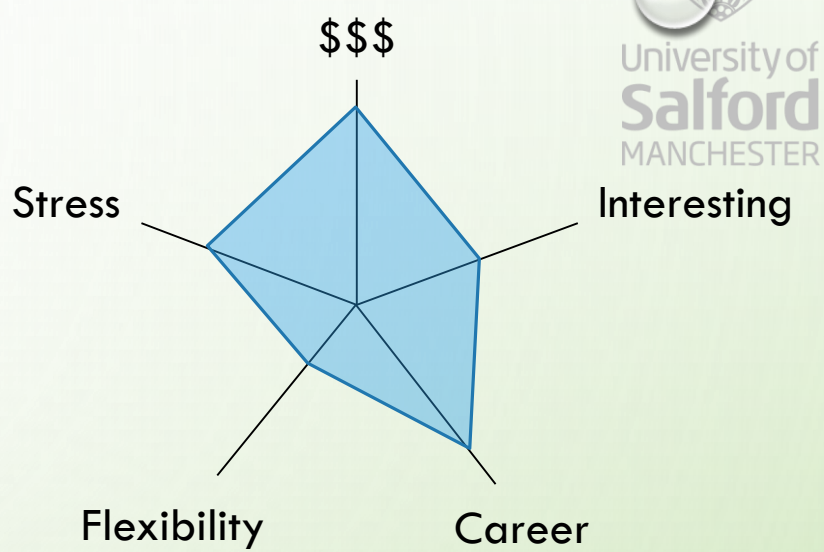
Pascal
Basic
ActionScript
C
C++
Java
PHP
SQL
Python
PowerShell
C#
JavaScript

WHY UNIVERSITY?

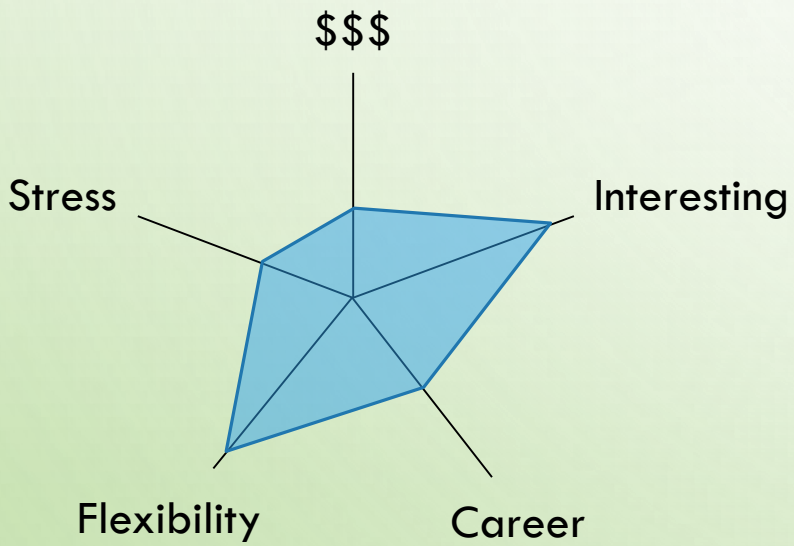
Lecturer



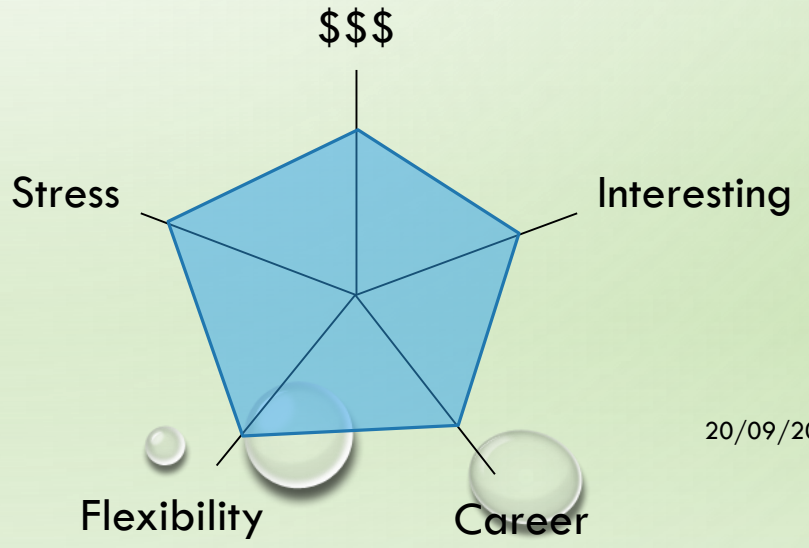
Industry



Researcher at university



Self-employed



Other factors:

- Job security
- Team
- …

Disclaimer: No scientific underpinning

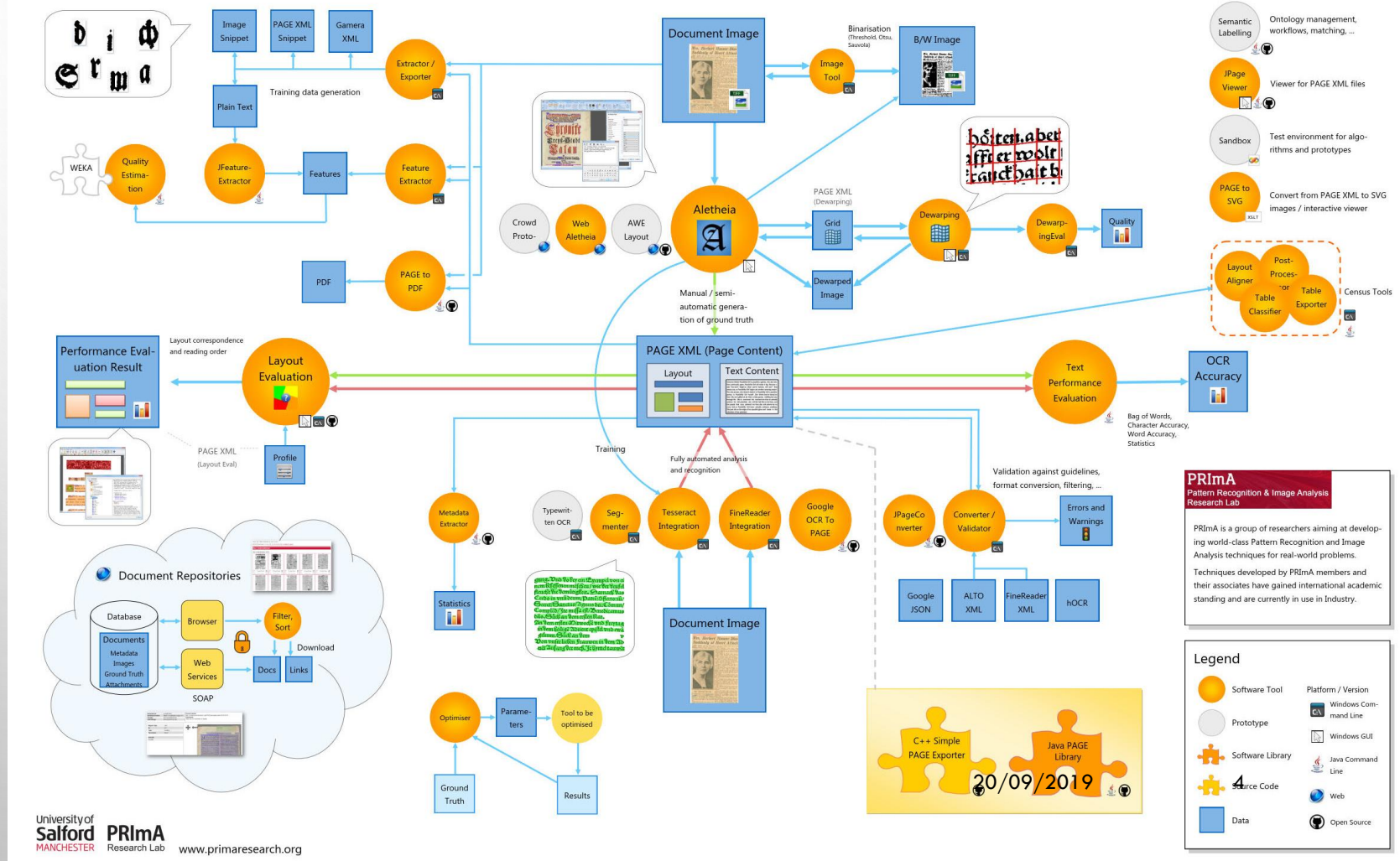
github.com/PRImA-Research-Lab/root/tree/master/docs

TOOLS AND FORMATS

- Overview poster
- PAGE XML is central

www.primaresearch.org

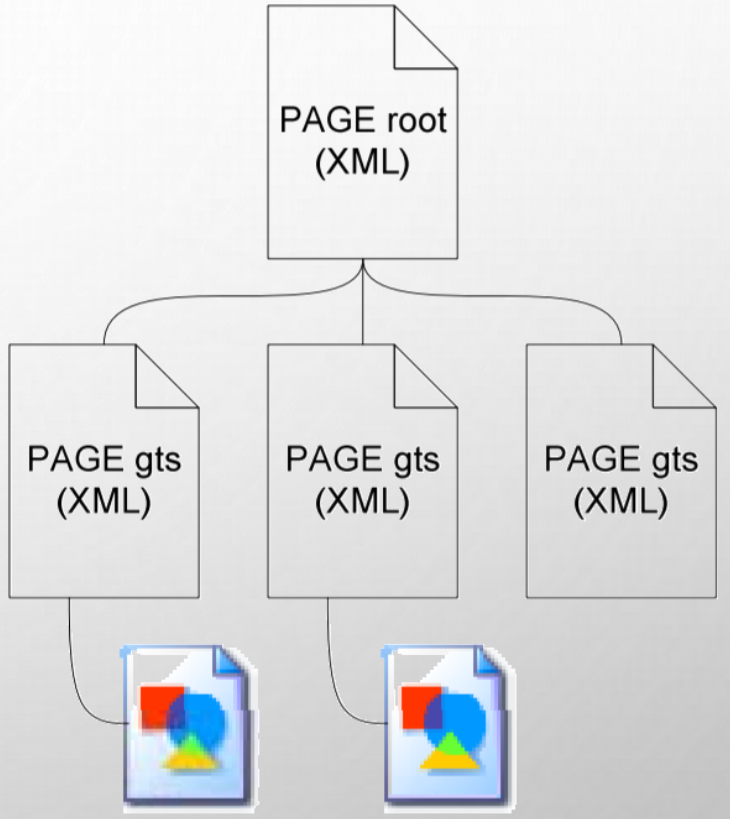
github.com/PRImA-Research-Lab



github.com/PRImA-Research-Lab/PAGE-XML

PAGE XML

- Page Analysis and Ground Truth Elements
- Long history, but main effort by Stefan Pletschacher (2009/2010)
- Clarification:
  - PAGE is a collection for formats
  - PAGE/gts/pagecontent is the main one
- Always open, but recently moved to GitHub



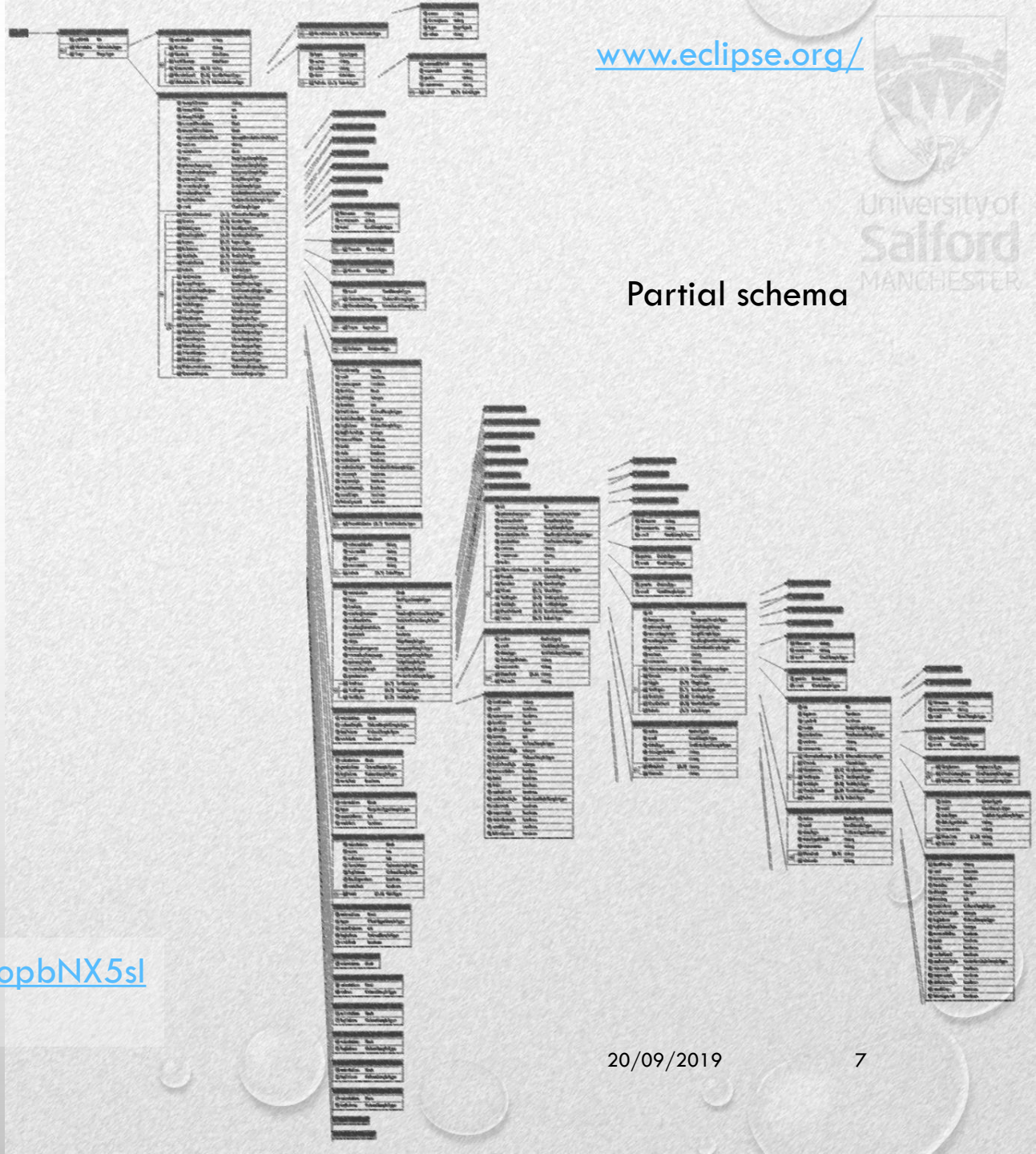
OPEN SOURCED, AND NOW?

- Just throw it out there or provide maintenance?
  - I try to:
    - Answer questions
    - Fix bugs
  - But within reason
- There can be disagreements with the community

PAGE (PAGECONTENT)

- More than just polygons
- Designed for detailed ground truthing of document content
- Schema useful for:
  - Validation
  - Automated model / GUI creation
- ALTO and PAGE are getting closer
  - ALTO to PAGE via converters
  - PAGE to ALTO not yet, but go ahead 😀
    - Comparison table:

https://docs.google.com/document/d/11GdAMJD5yDopbNX5sICgArBpcLroSITqK0EaIARi954/edit?usp=sharing



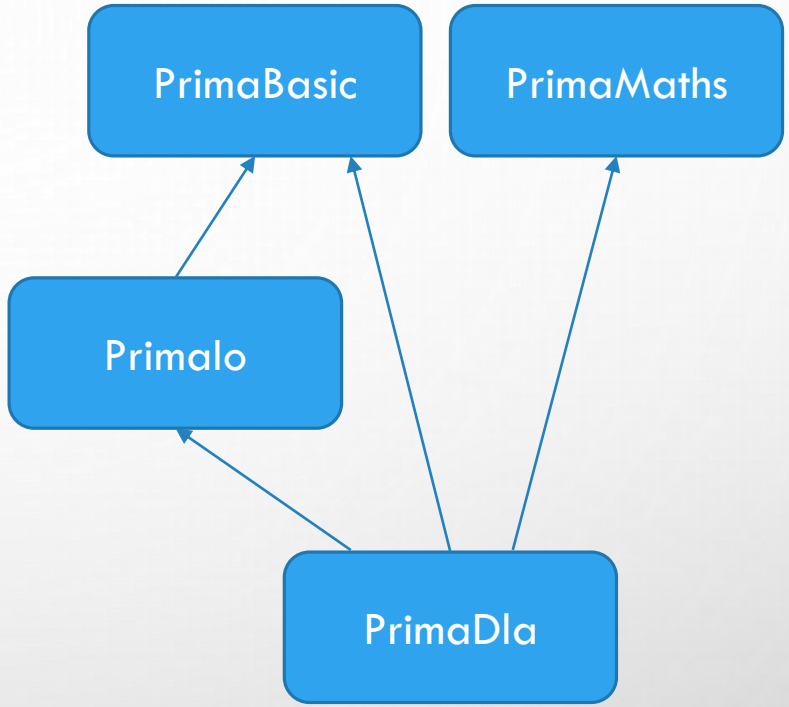
www.eclipse.org/

Partial schema

github.com/PRImA-Research-Lab/prima-core-libs

CORE LIBS

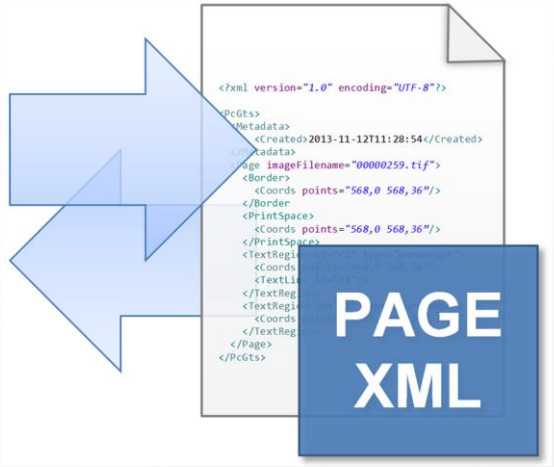
- Java
- Data model + Readers/Writers
  - Backwards-compatible to older PAGE versions
- Used by most other Java-based tools



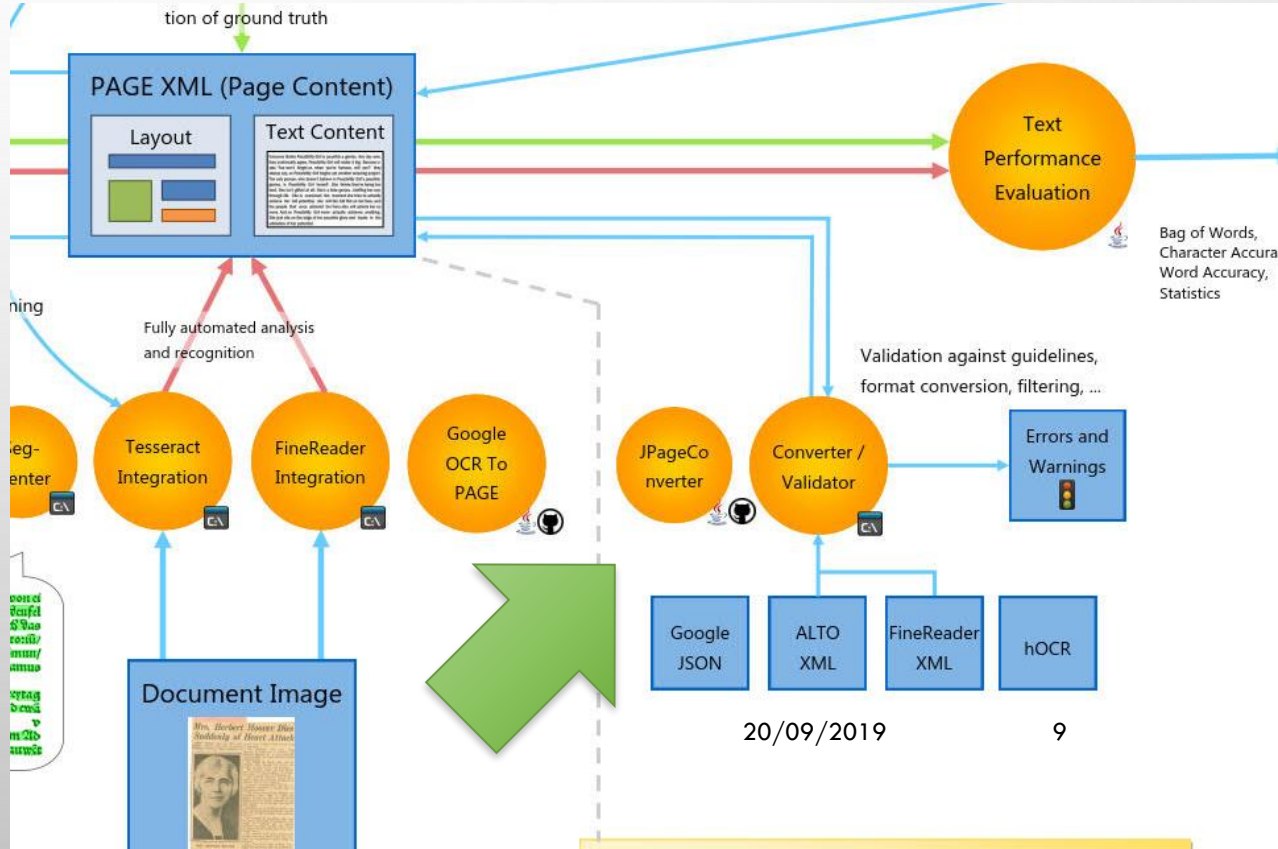
github.com/PRImA-Research-Lab/prima-page-converter

primaresearch.org/tools/PAGEConverterValidator

PAGE CONVERTER AND VALIDATOR



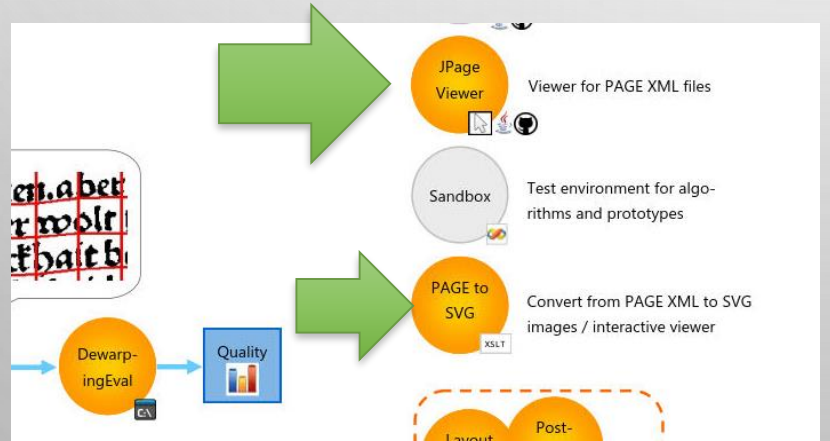
- Java and C++ version
- Slightly different features
- Conversion to PAGE
- Validation
  - Against schema
  - Against ground truthing rules
- Modification
  - E.g. removing certain content



github.com/PRImA-Research-Lab/prima-page-viewer

PAGE VIEWER

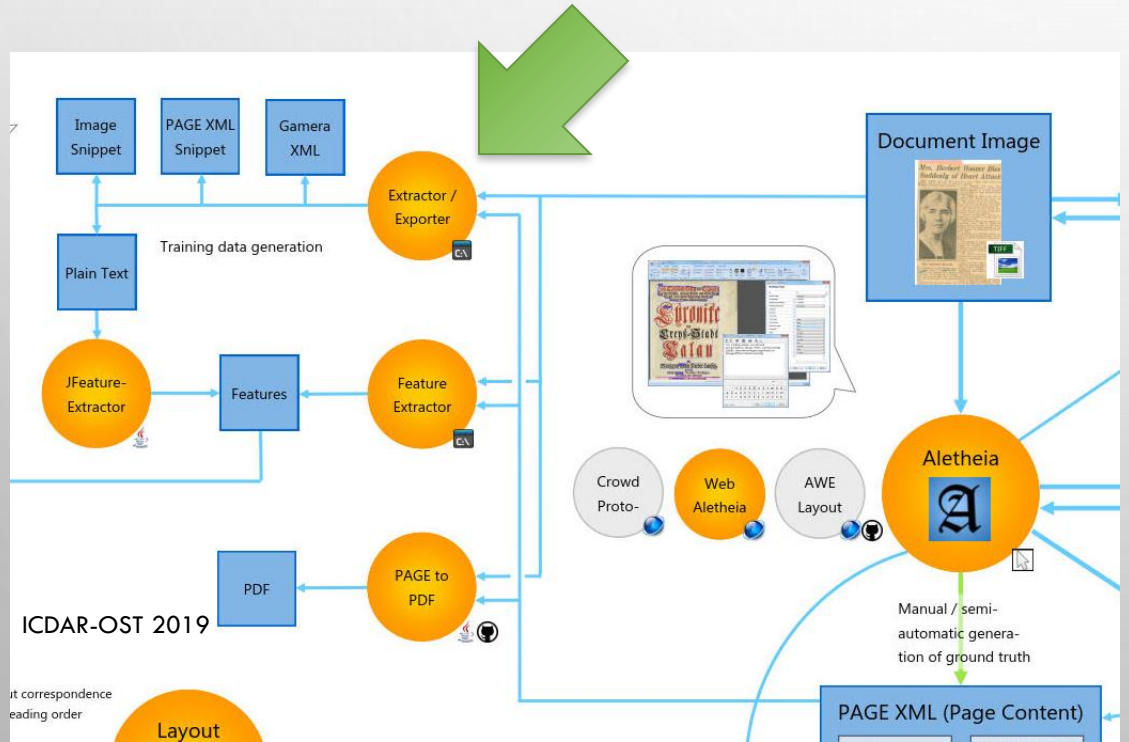
- Java
  - SWT dependency
- Most popular PRImA project on GitHub
- Was intended as editor (JAletheia)
  - Performance was an issue
- Quick and easy way to view PAGE, ALTO, …

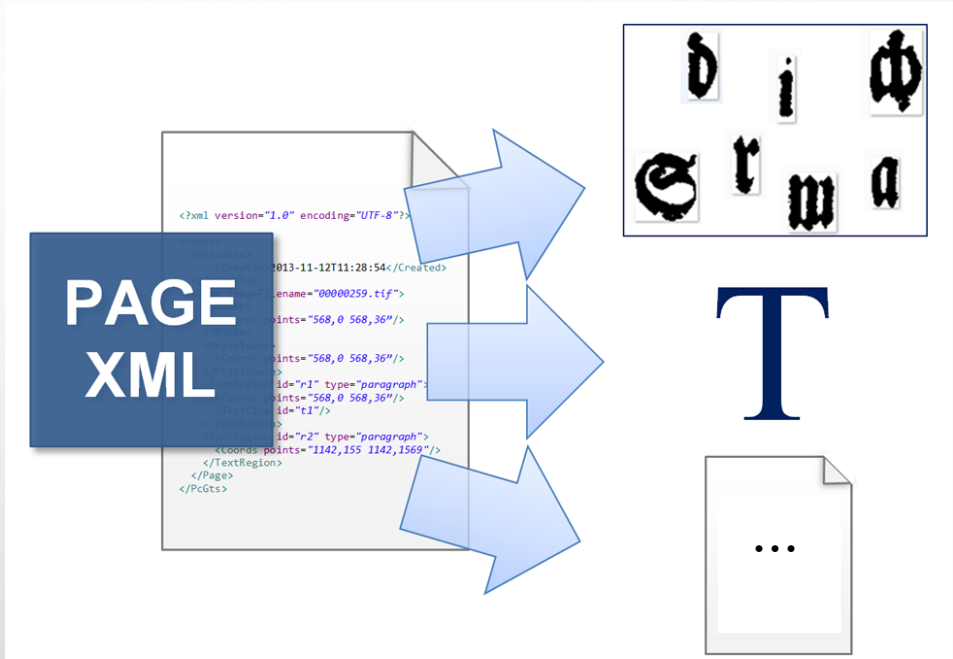


primaresearch.org/tools/ExtractorExporter

EXTRACTOR / EXPORTER

- A

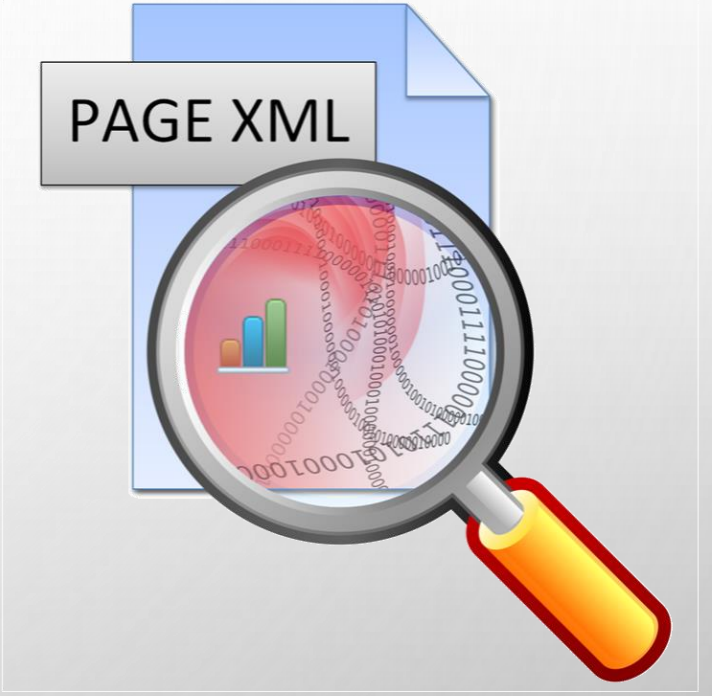


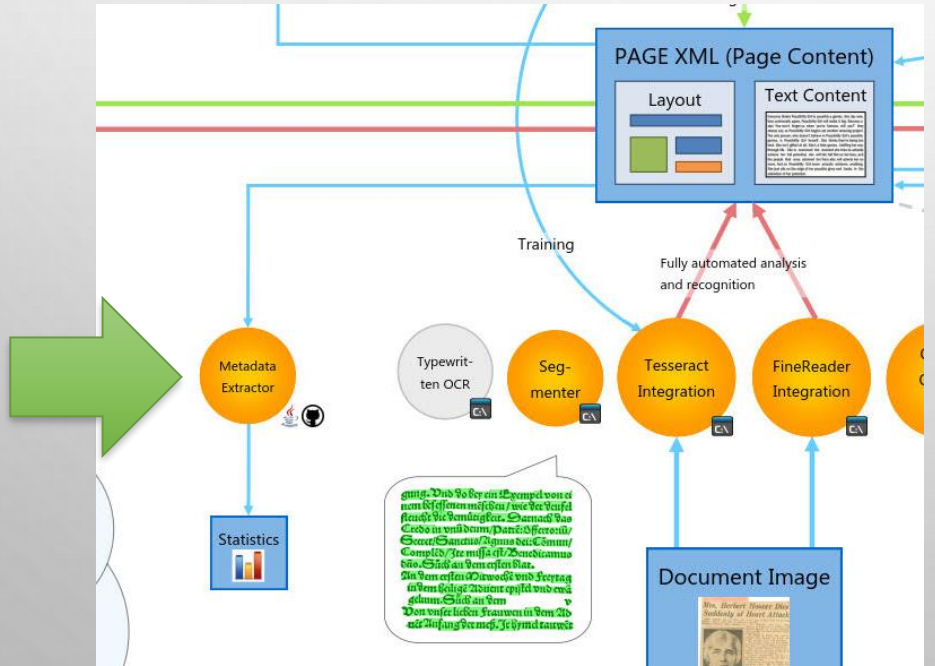


github.com/PRImA-Research-Lab/prima-page-metadata-scanner

PAGE METADATA EXTRACTOR

- Extracts various pieces of information from a PAGE file and outputs them in CSV format
- E.g. for use with machine learning





github.com/PRImA-Research-Lab/prima-page-to-pdf

PAGE TO PDF

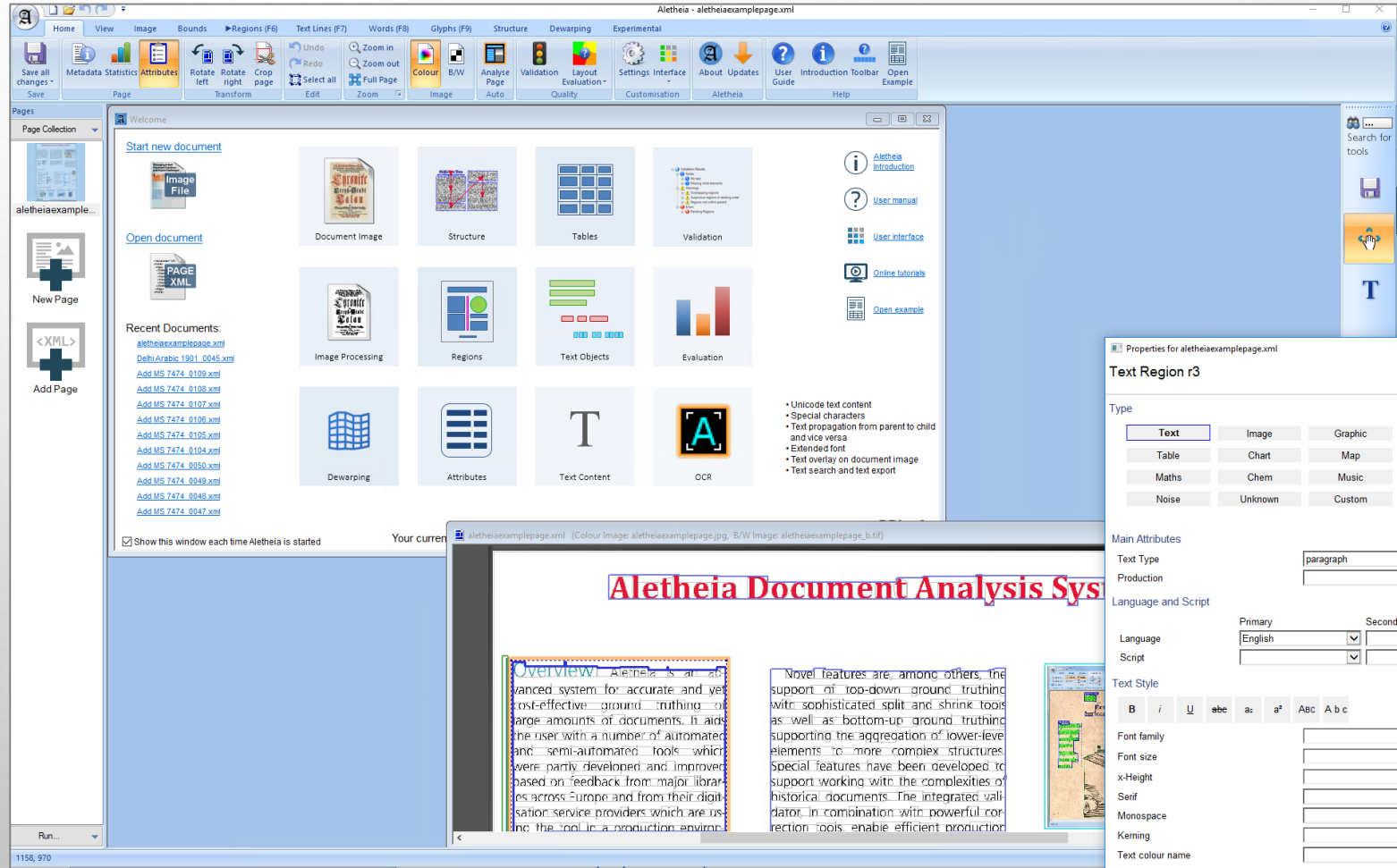
- Creates PDF with
  - Document image
  - Hidden text layer
  - Optional block outlines
- For visual inspection of documents, for example

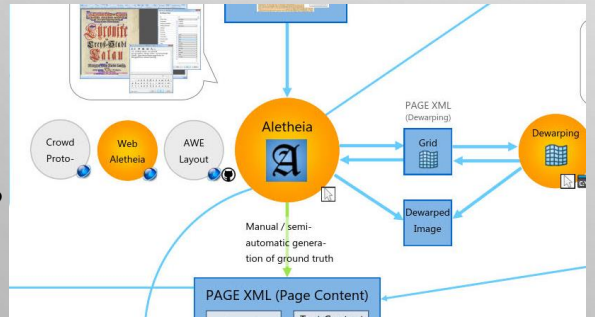
https://www.primaresearch.org/tools/Aletheia

See also “PRImA Research Lab” YouTube channel

ALETHEIA

- Aletheia = Truth (Greek)
- Platform that integrates many of the PRImA tools
- Windows
- Based on MFC
- Using OpenCV for image I/O and ops (opencv.org/)





ICDAR-OST 2019

www.wikipedia.org

MFC

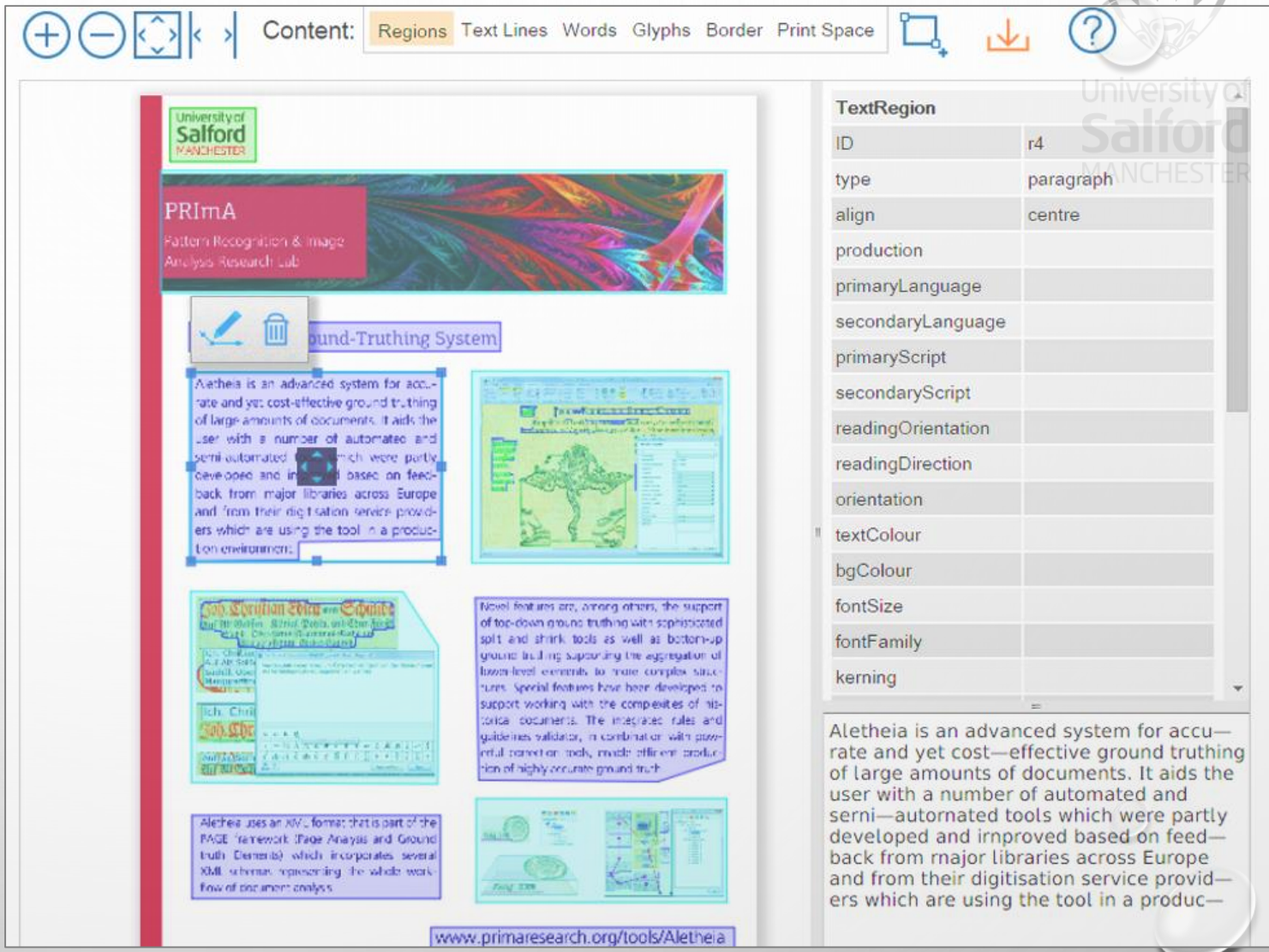
- Microsoft Foundation Class Library (C++)
  - Everything is called AFX…
  - Original name: "Application Framework Extensions"
- Initial release 1992!!!
- Visual Studio
- Outdated, but:
- Aletheia worked in Windows XP and still works in Windows 10
- Difficult to switch to something more modern

- Has a GUI builder
- Supports the ribbon toolbar
- Some newer features like HTML dialog

WEB ALETHEIA

- Based on GWT (pronounced [gwid])
  - Google Web Toolkit, but now community-run
  - Cross-compiles from Java to HTTP+JavaScript
  - Less and less activity
    - Dead end?!
- Runs on Tomcat
- Basic viewing and editing

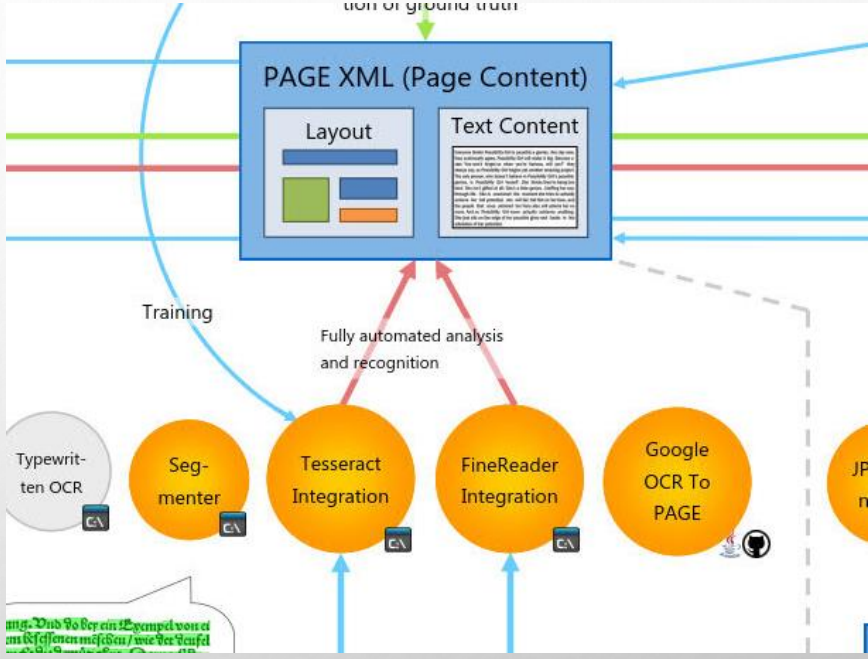
github.com/PRImA-Research-Lab/prima-aletheia-web
primaresearch.org/tools/WebAletheia



primaresearch.org/tools/TesseractOCRToPAGE

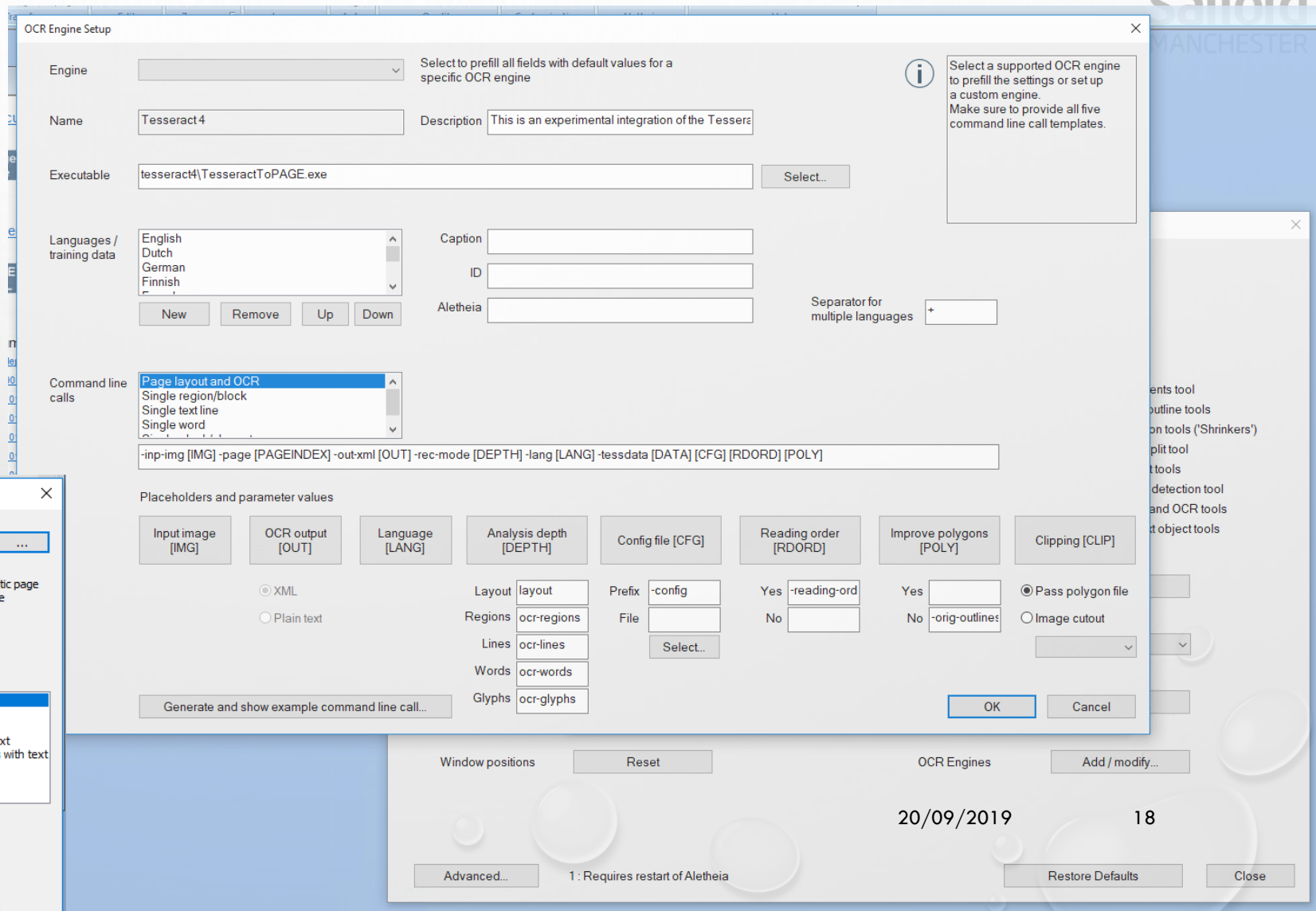
OCR INTEGRATION

- Run open source / commercial OCR engine and export PAGE XML
  - Not just conversion from the native output formats
  - Rather API calls and direct export to PAGE
    - Less loss of information
- Tesseract, Google Cloud Vision ABBYY FineReader,
- Command line tools



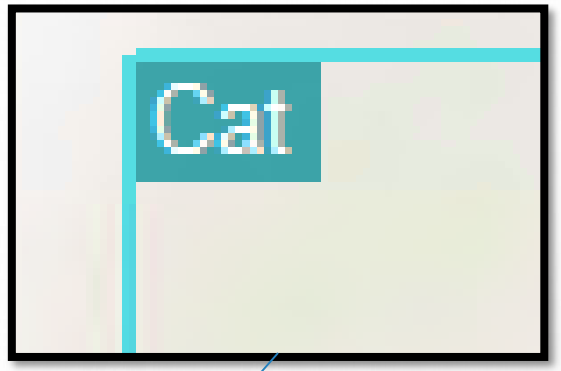
OCR IN ALETHEIA

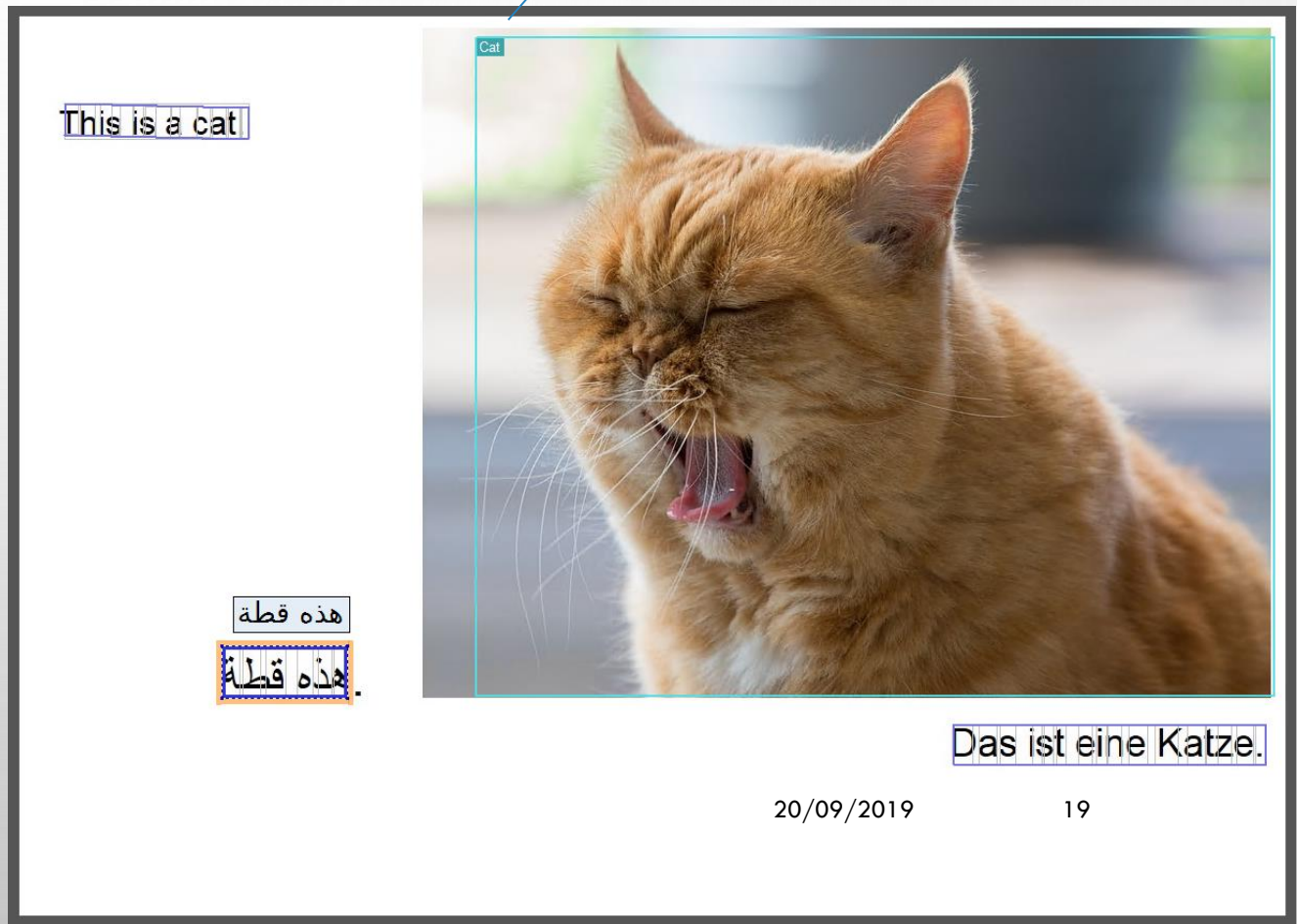
- Add any OCR engine via CLI
- Templates for
  - TesseractToPAGE
  - GoogleCloudOctToPage
  - ABBYY Cloud Client
  - PRImA FRE Integration



github.com/PRImA-Research-Lab/cloud-vision-ocr-to-page

GOOGLE CLOUD VISION





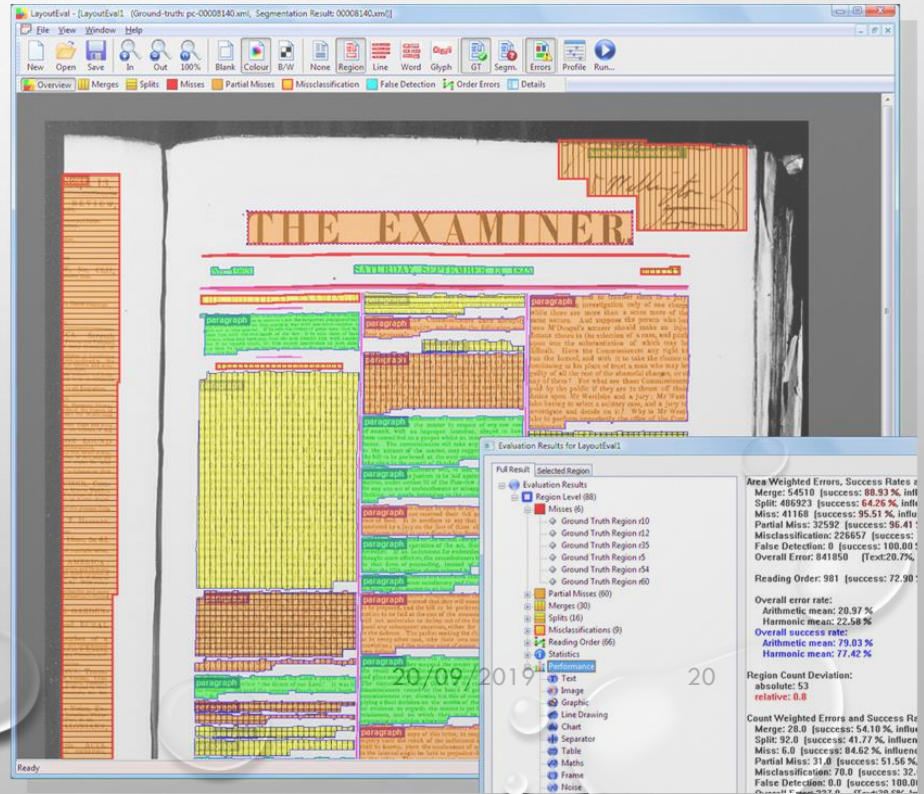
PERFORMANCE EVALUATION

- Page Layout Evaluation (Segmentation, Classification)
- Text-based OCR result evaluation
  - Standard metrics + a couple of new ones

primaresearch.org/tools/PerformanceEvaluation

github.com/PRImA-Research-Lab/prima-layout-eval

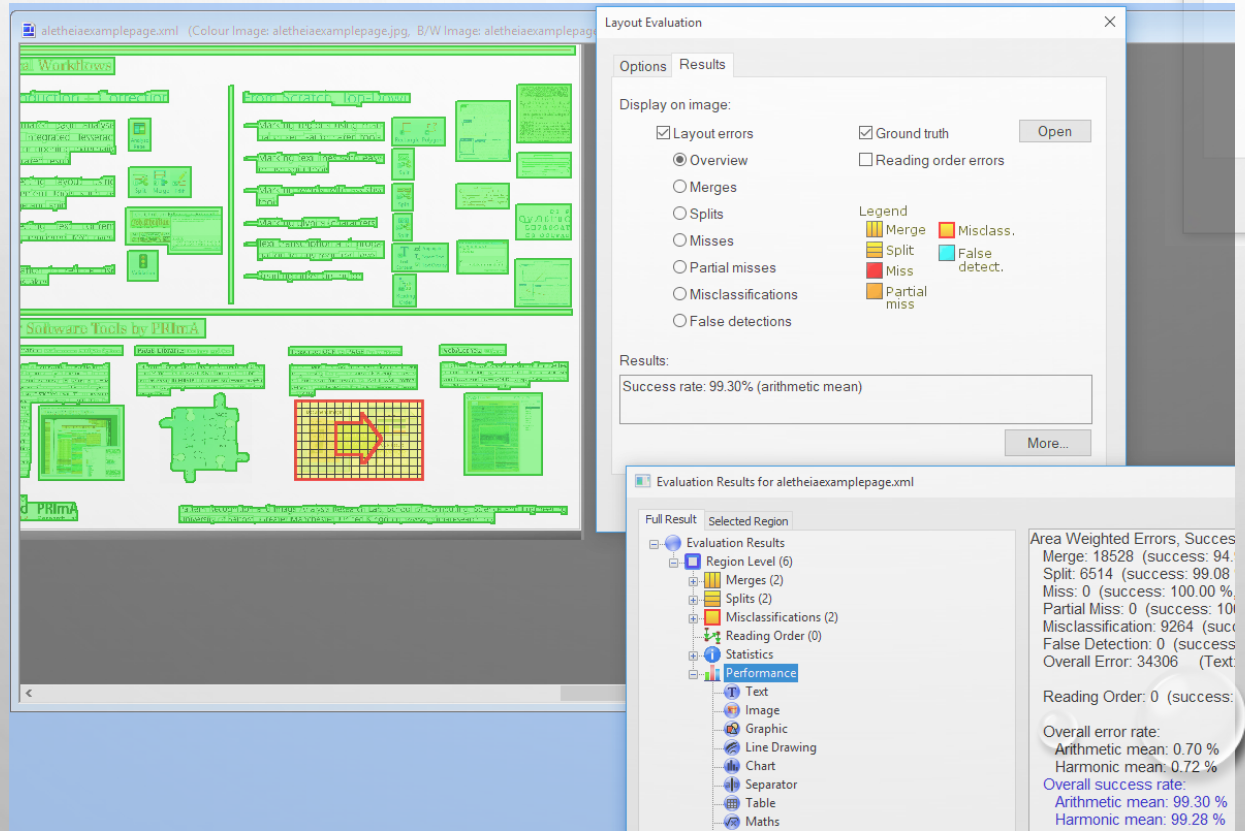
github.com/PRImA-Research-Lab/prima-text

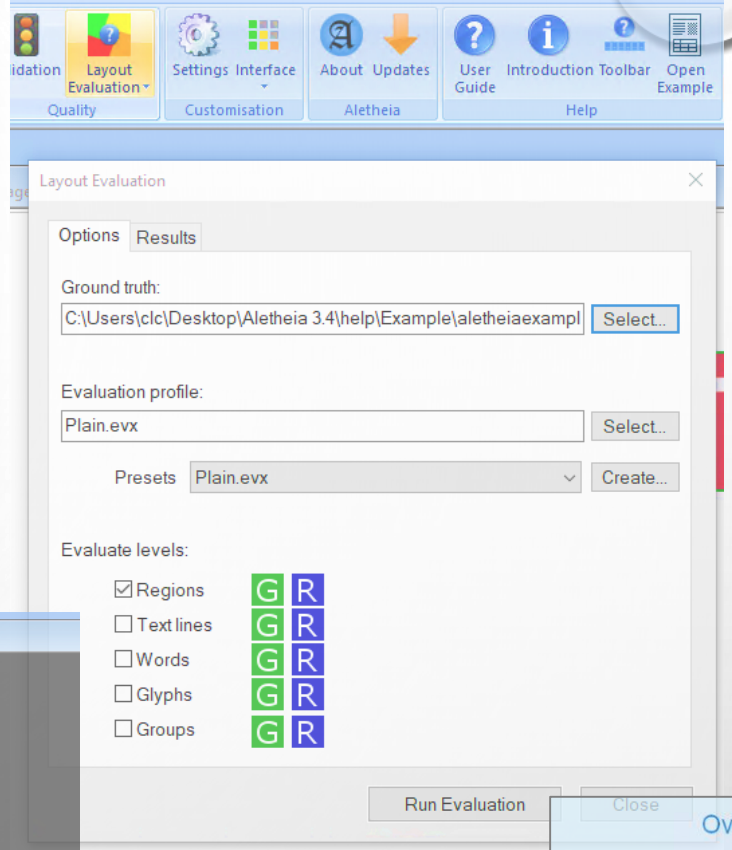


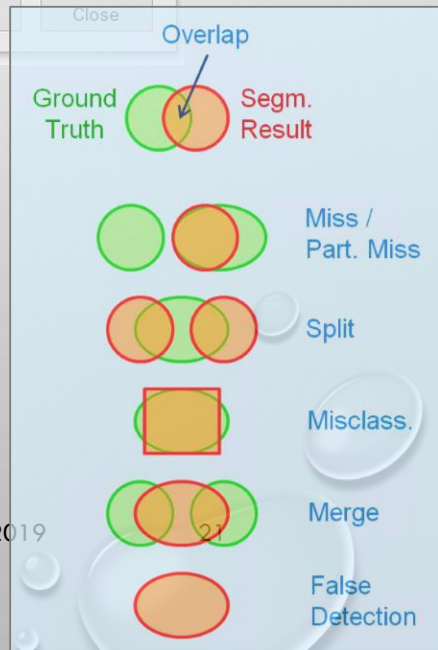
github.com/PRImA-Research-Lab/prima-layout-eval

LAYOUT EVALUATION

- Dedicated tool, but most features also in Aletheia
- Detailed results







OCR-D

ocr-d.de/eng

ocr-d.github.io/

ocr-d.github.io/projects

github.com/OCR-D/

github.com/topics/ocr-d

- Effort to reimplement the layout evaluation in python github.com/PRImA-Research-Lab/prima-layout-eval
- OCR-D is role model project for openness
- Talk to Clemens Neudecker or colleauges

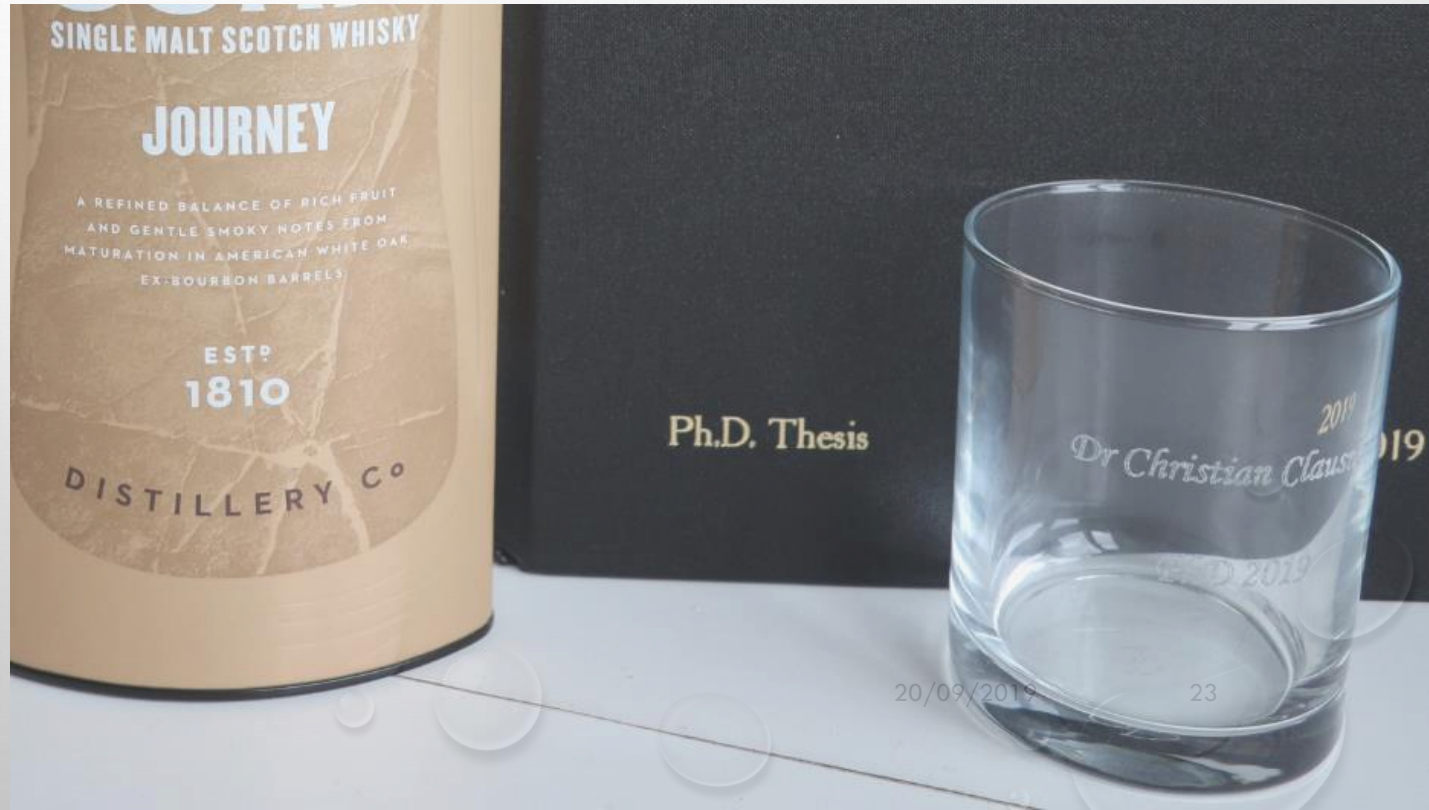
github.com/PRImA-Research-Lab/semantic-labelling

http://usir.salford.ac.uk/id/eprint/51450

University of Salford MANCHESTER

SEMANTIC LABELLING

- PhD project
- Ontology for DIA
- Algorithms and tools
- Prototype!
- OCR-D adopted part of the ontology



primaresearch.org/competitions

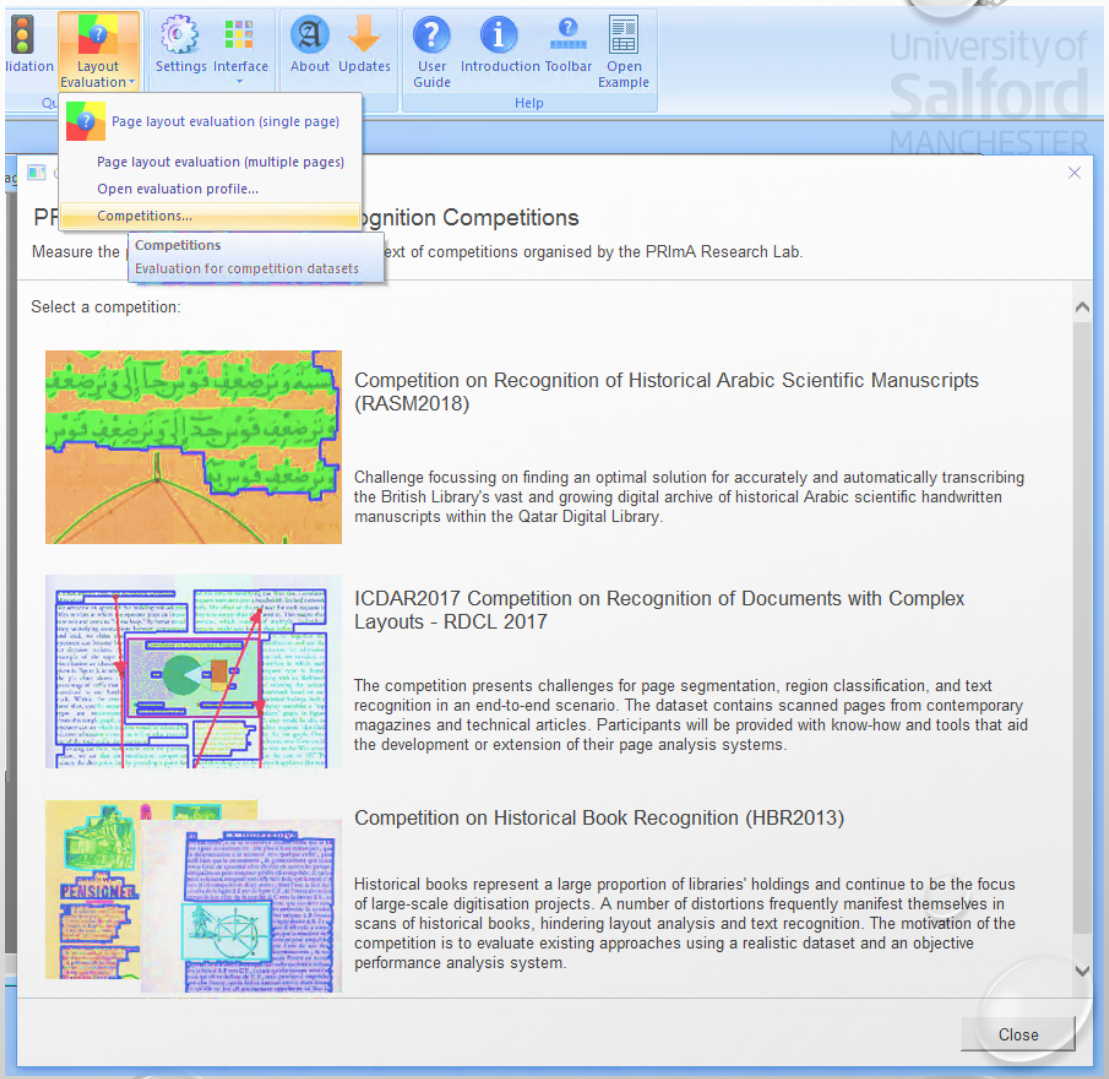
COMPETITIONS

- Long-standing series of DIA competitions (since 2001!)
- Evaluation of segmentation, classification and OCR
- Historical and contemporary documents
- Latest instalments:
  - RDCL 2019
  - REID 2019
  - RASM 2018, 2019

Presented at ICDAR

blogs.bl.uk/digital-scholarship/2019/09/rasm2019-results.html

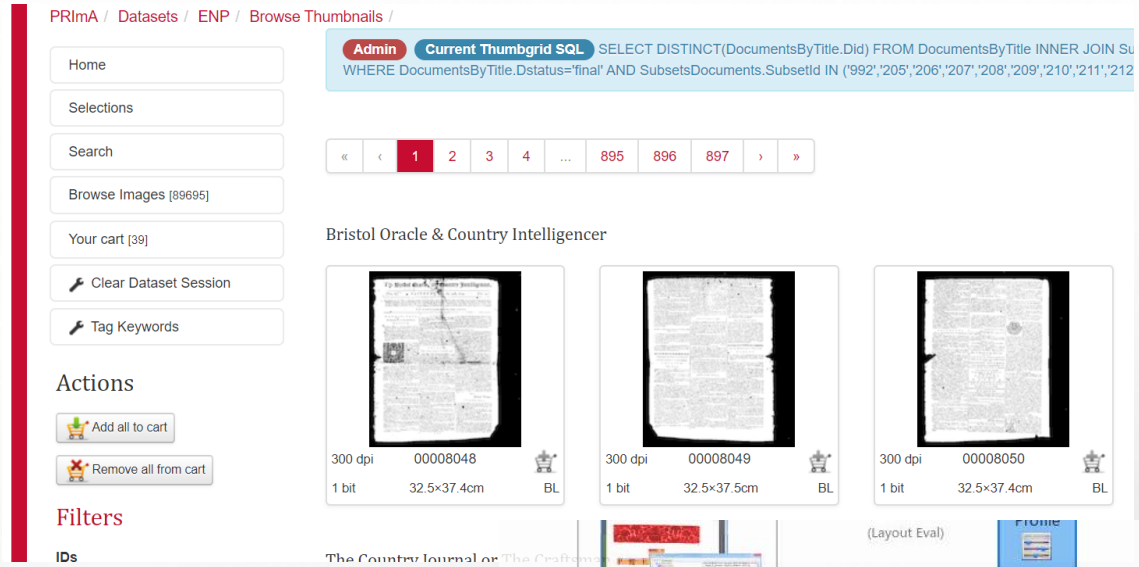
Evaluate results in Aletheia

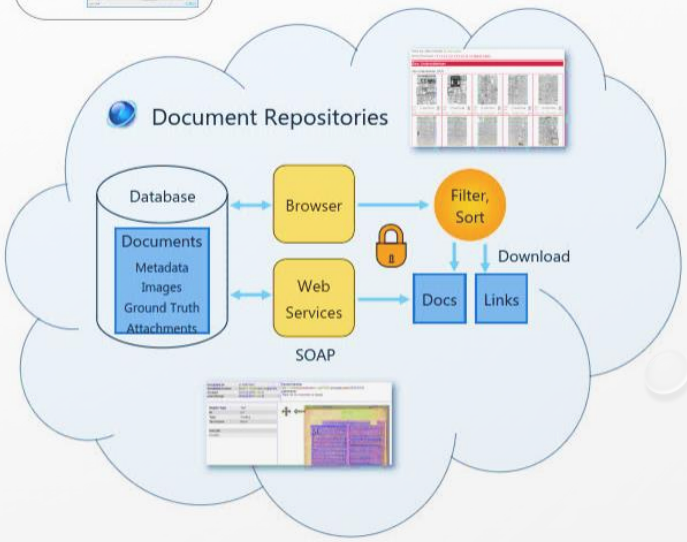


primaresearch.org/datasets

DATASETS

- Coded in PHP, stored in MySQL database
  - Christos Papadopoulos



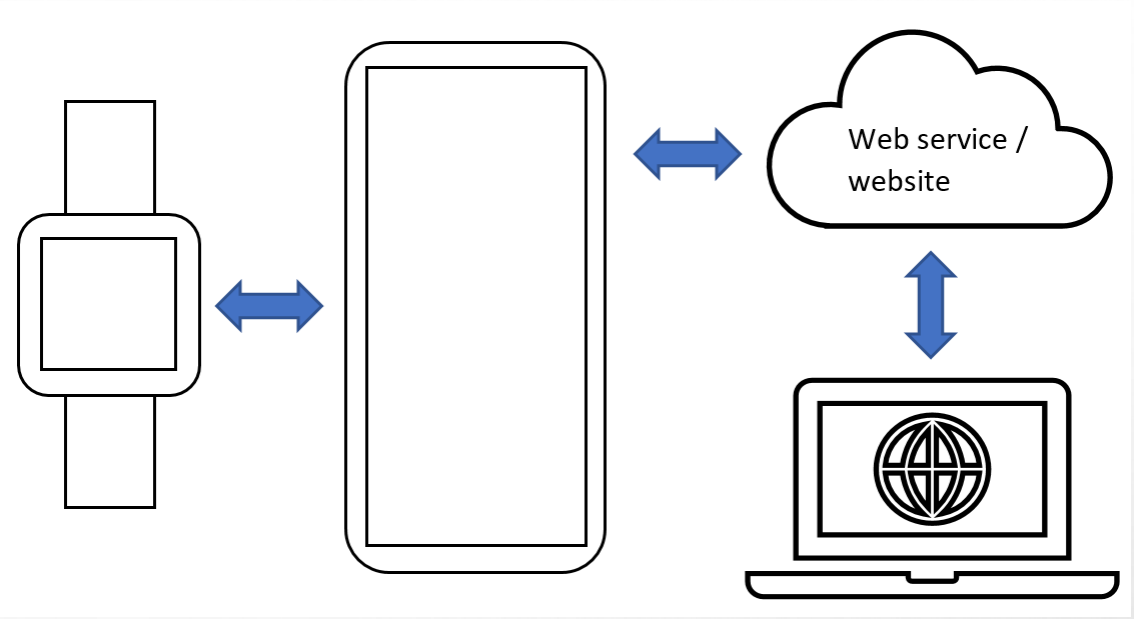


www.ref.ac.uk/

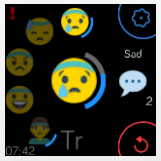
UNIVERSITY ENVIRONMENT

- People coming and going
  - Loss of know-how
    - E.g. website dev stalling
- University changes
  - Politics
  - Structure
- REF – Research Excellence Framework
  - Need impact case studies
    - Need to show impact (hence we ask for user information)

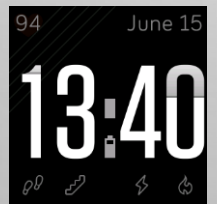
SELF-HARM PROJECT



- Diversification
- Project in early stage
- Interesting technologies (new for me)
  - Wearable app dev
  - NodeJS
  - Cloud-based app service and DB



github.com/chris1010010/Fitbit-Behind-the-Numbers

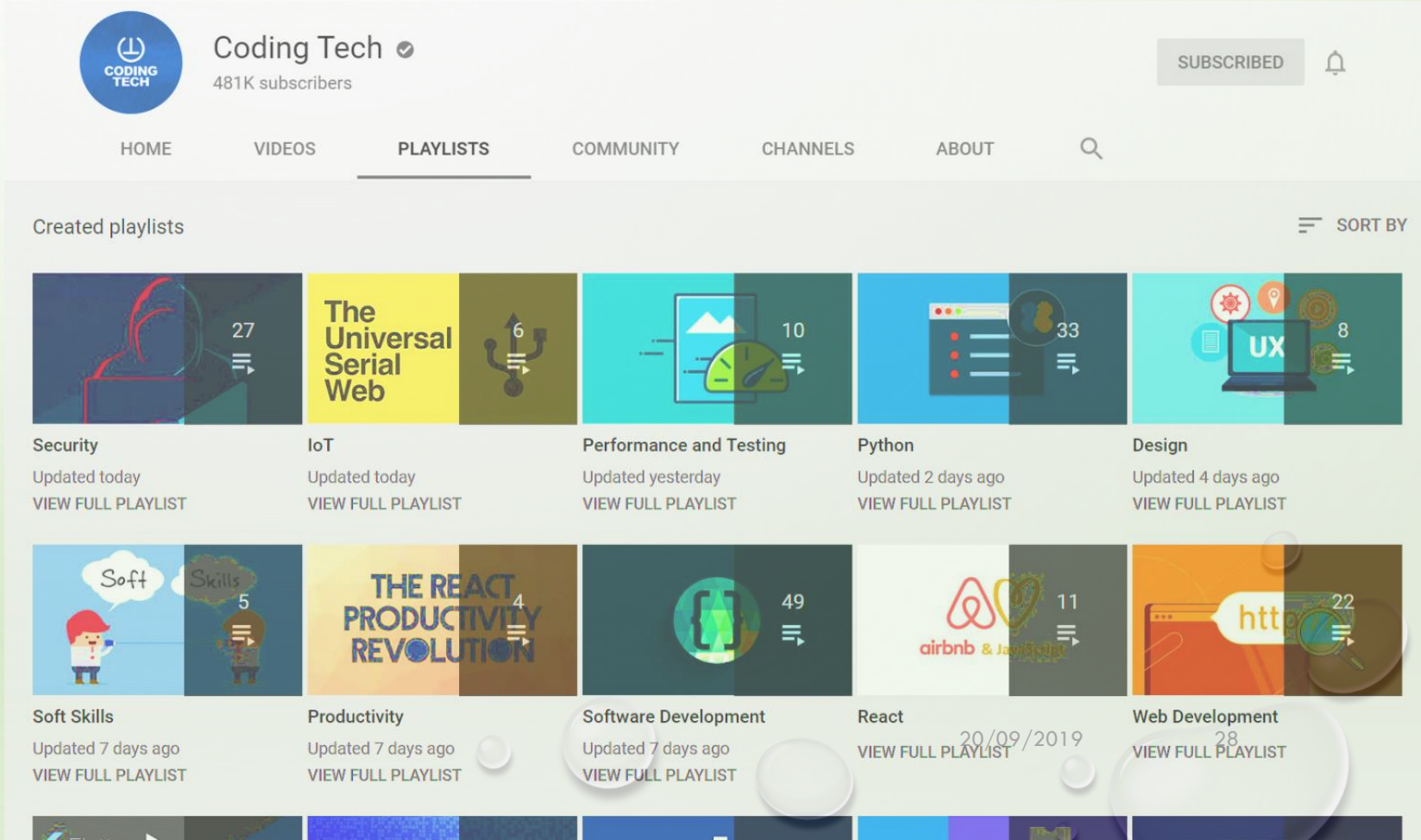


youtube.com/channel/UCtxCXg-UvSnTKPOzLH4wJaQ

Search for “Coding Tech” on YouTube

CODING TECH

- Great source of information
- Very diverse content, something for everyone
- Talks, tutorials, …



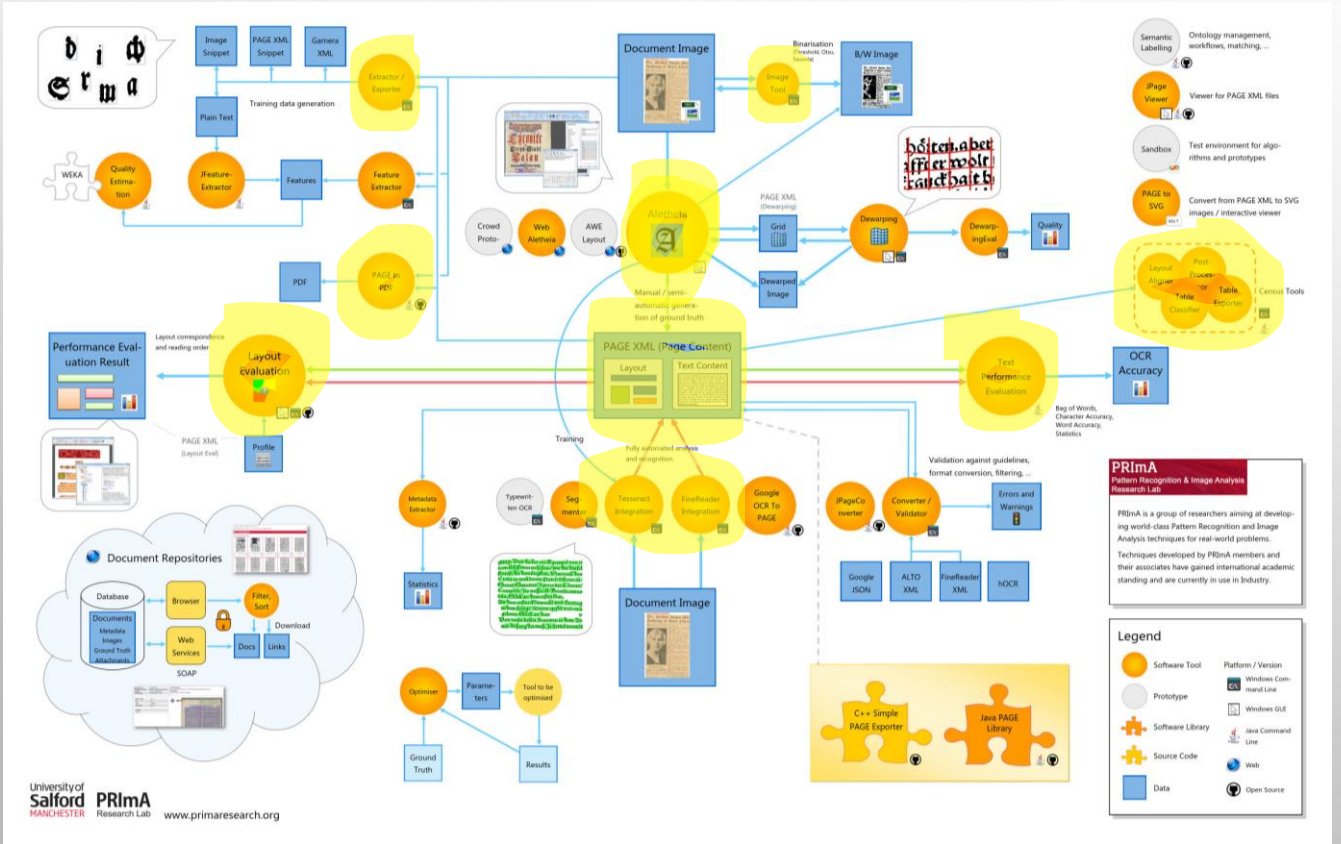
imagemagick.org

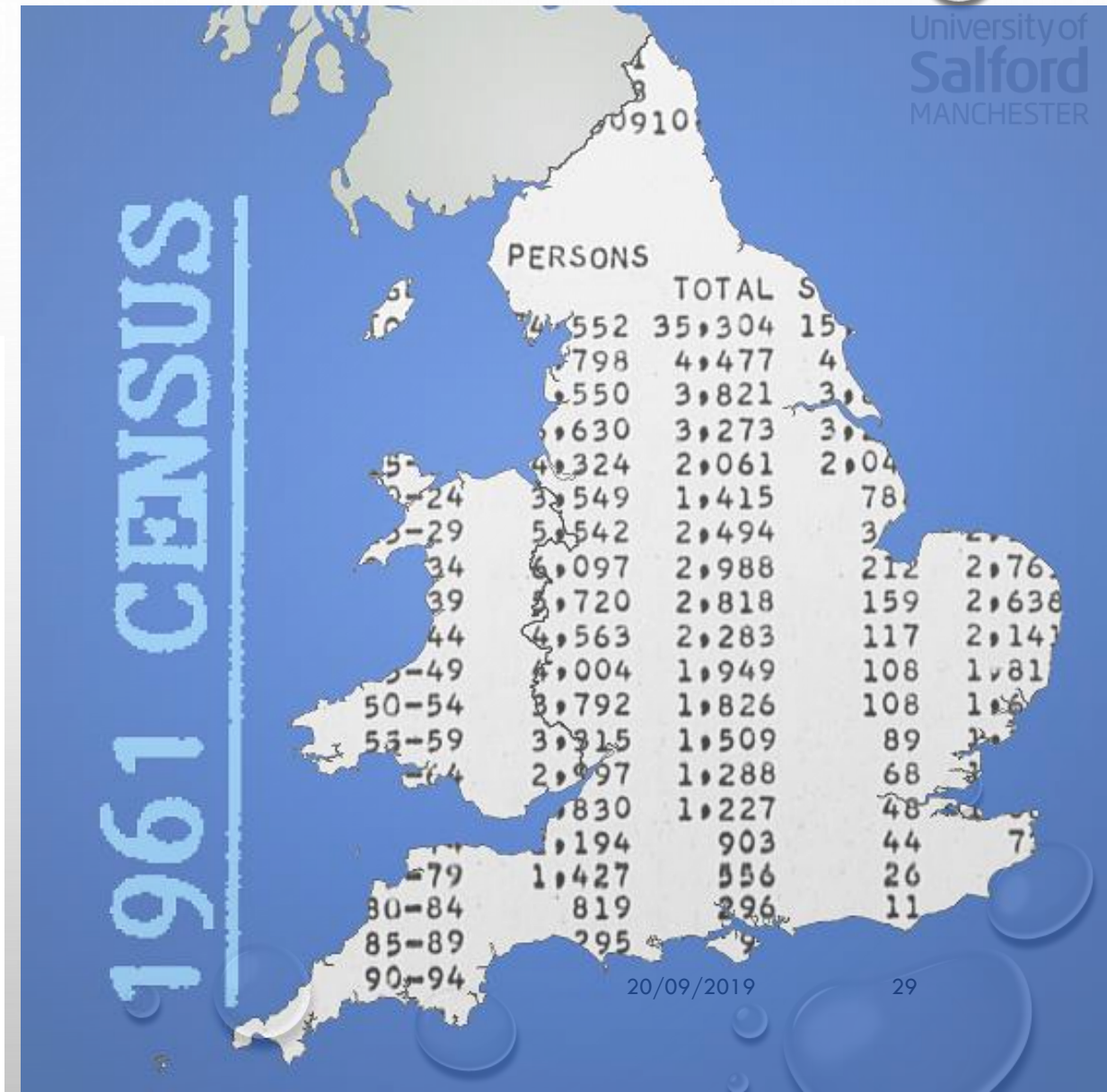
primaresearch.org/projects/Census1961

www.zooniverse.org/

CENSUS

- Used many of the shown tools + some new ones





OPENNESS

- There has been a rethinking of openness in the PRImA lab
- More open source or open binaries
- With constraints
  - University wants to retain intellectual property
    - Performance measure of uni
  - Evaluation of impact
    - Also leads to funding for the university
- There's room for IP
  - Dual approach to open parts and retain others

THE END


Document Image
Aletheia
PAGE XML (Dewarping)
Grid
Dewarping
Dewarped Image
Manual / semi-automatic genera-tion of ground truth
PAGE XML (Page Content)
Layout
Text Content
Training
Fully automated analysis and recognition
Tesseract Integration
FineReader Integration
Google OCR To PAGE
Document Image