

ICDAR-OST 2019

CHRISTIAN CLAUSNER

PRIMA RESEARCH LAB, UNIVERSITY OF SALFORD, UK





ABOUT ME

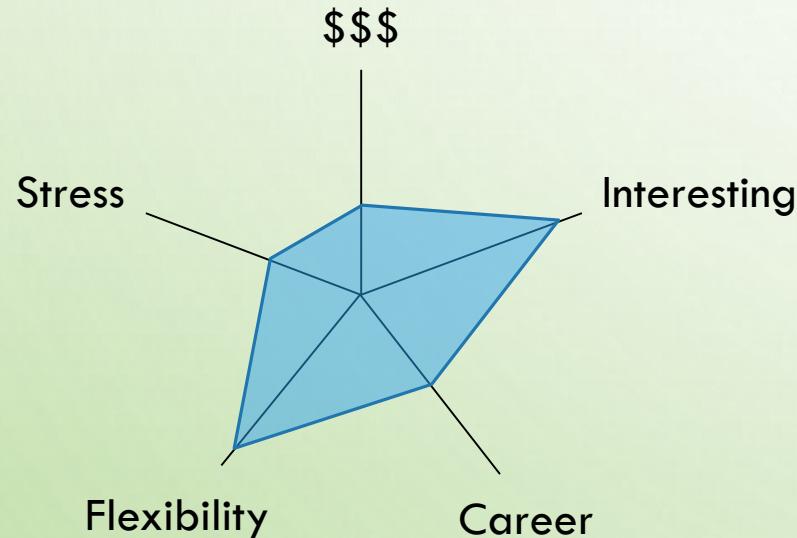
- 2000: Chemnitz University of Technology
 - Master's in Computer Science
- 2006: Industry (Software Dev)
- 2009: University of Salford (Research Fellow)
 - 2019: PhD (finally)



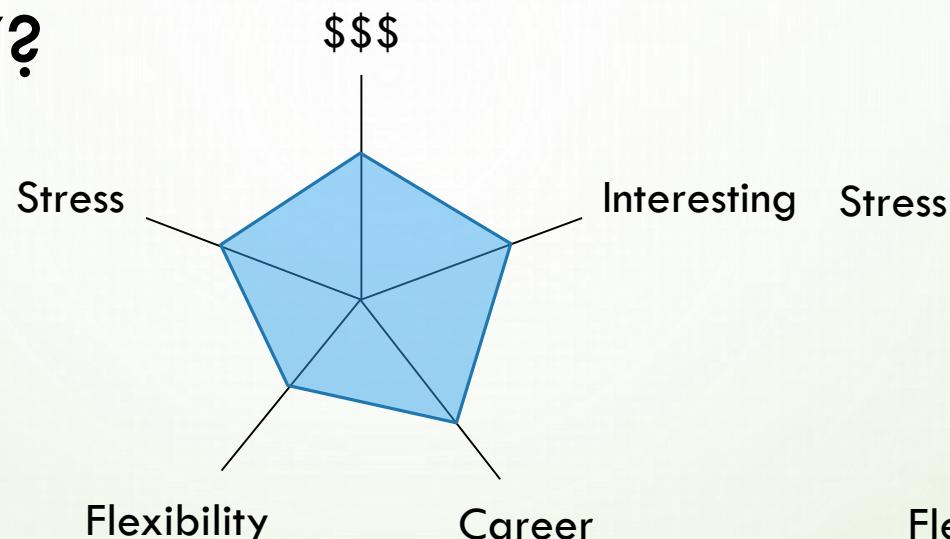
Pascal
Basic
ActionScript
C
Java
PHP
SQL
Python
JavaScript
c#
PowerShell

WHY UNIVERSITY?

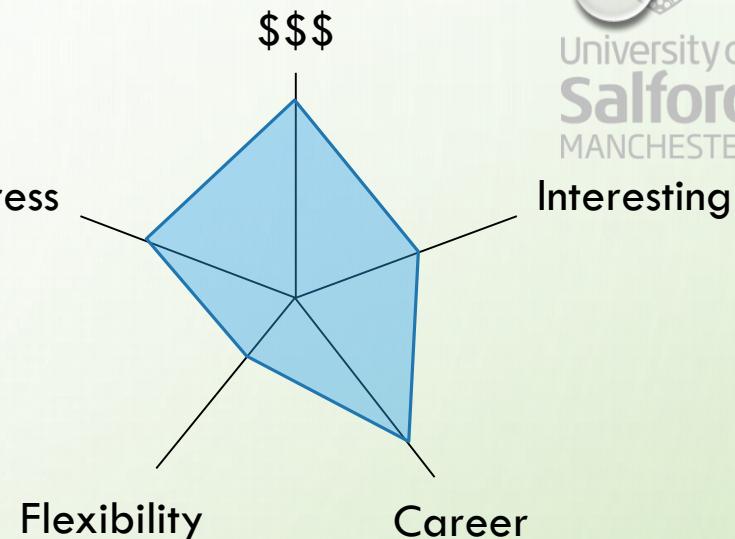
Researcher at university



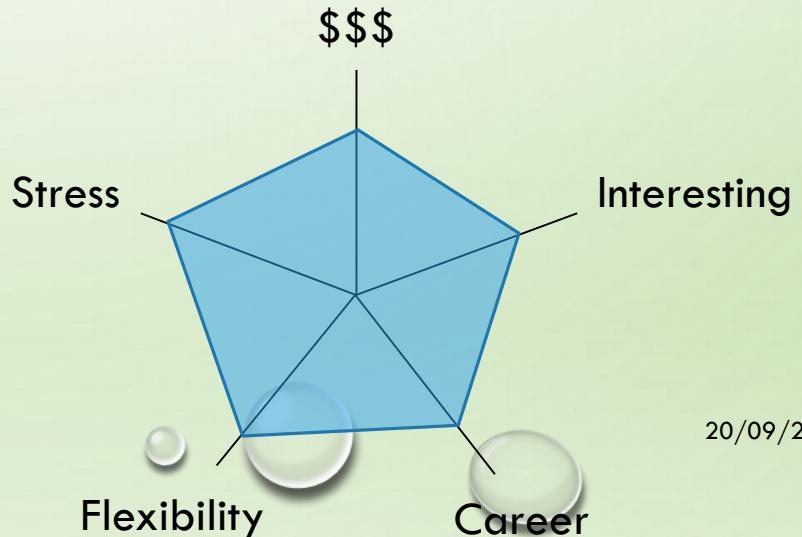
Lecturer



Industry



Self-employed



Other factors:

- Job security
- Team
- ...



University of
Salford
MANCHESTER

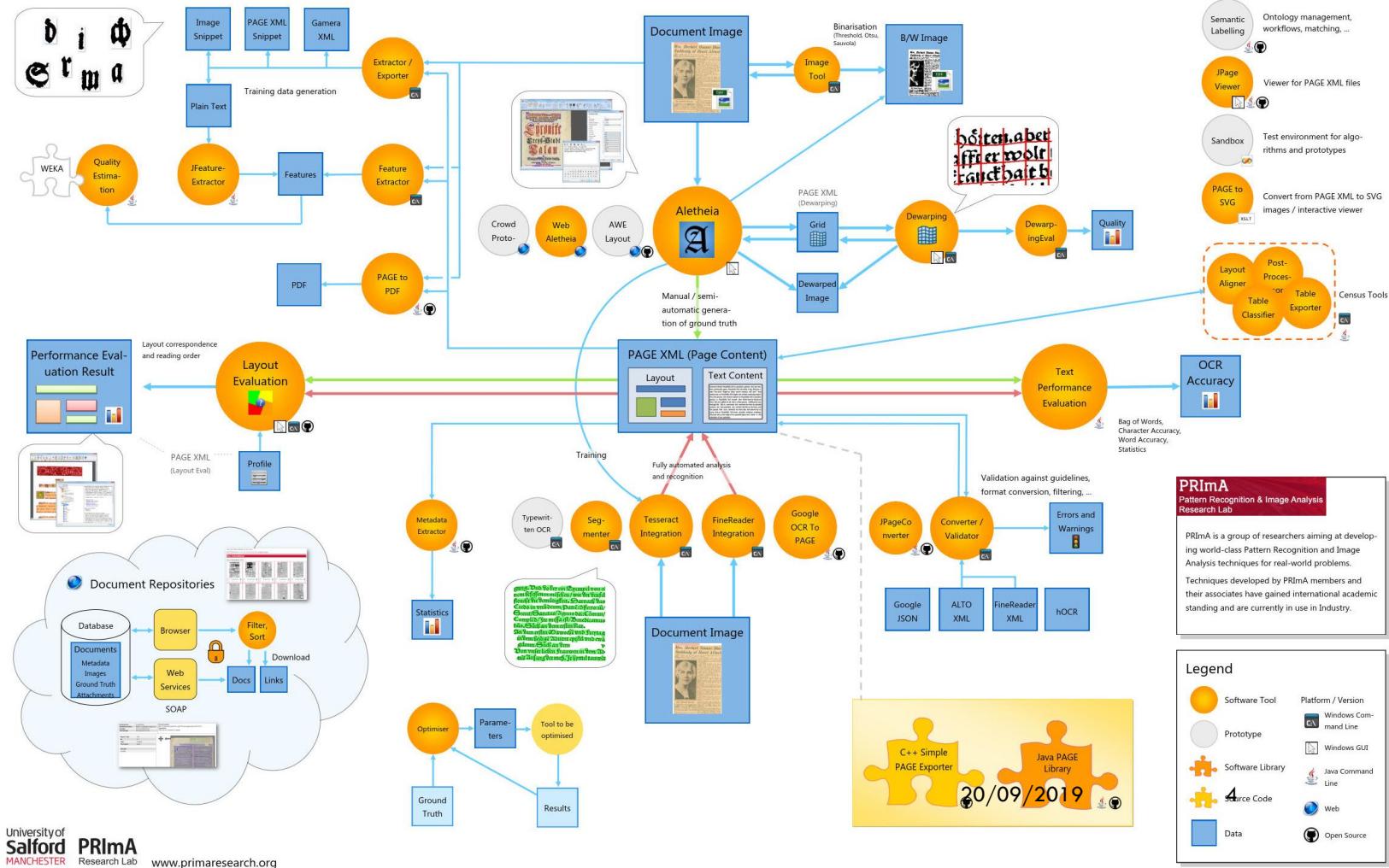
TOOLS AND FORMATS

- Overview poster
- PAGE XML is central

www.primaresearch.org

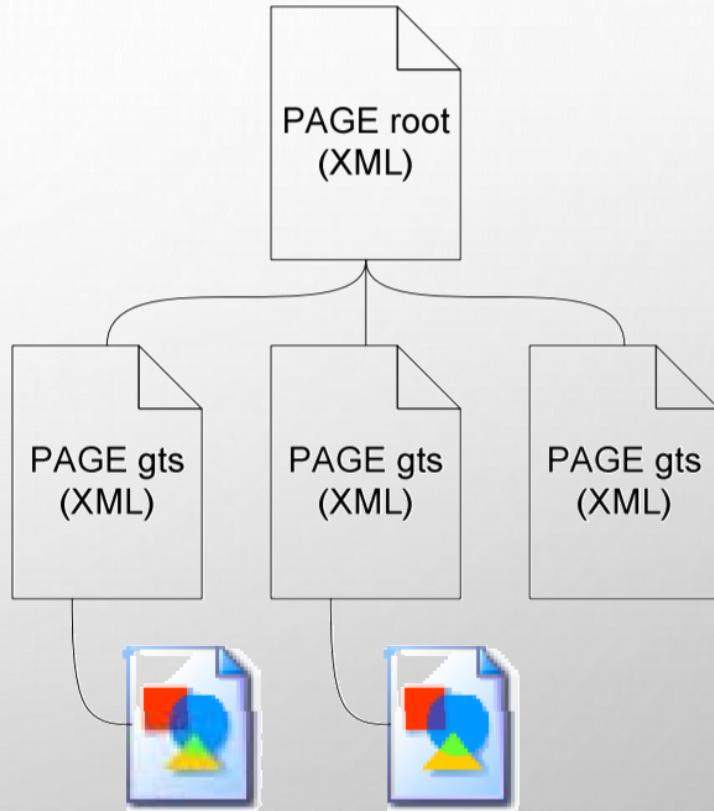
github.com/PRImA-Research-Lab

ICDAR-OST 2019



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





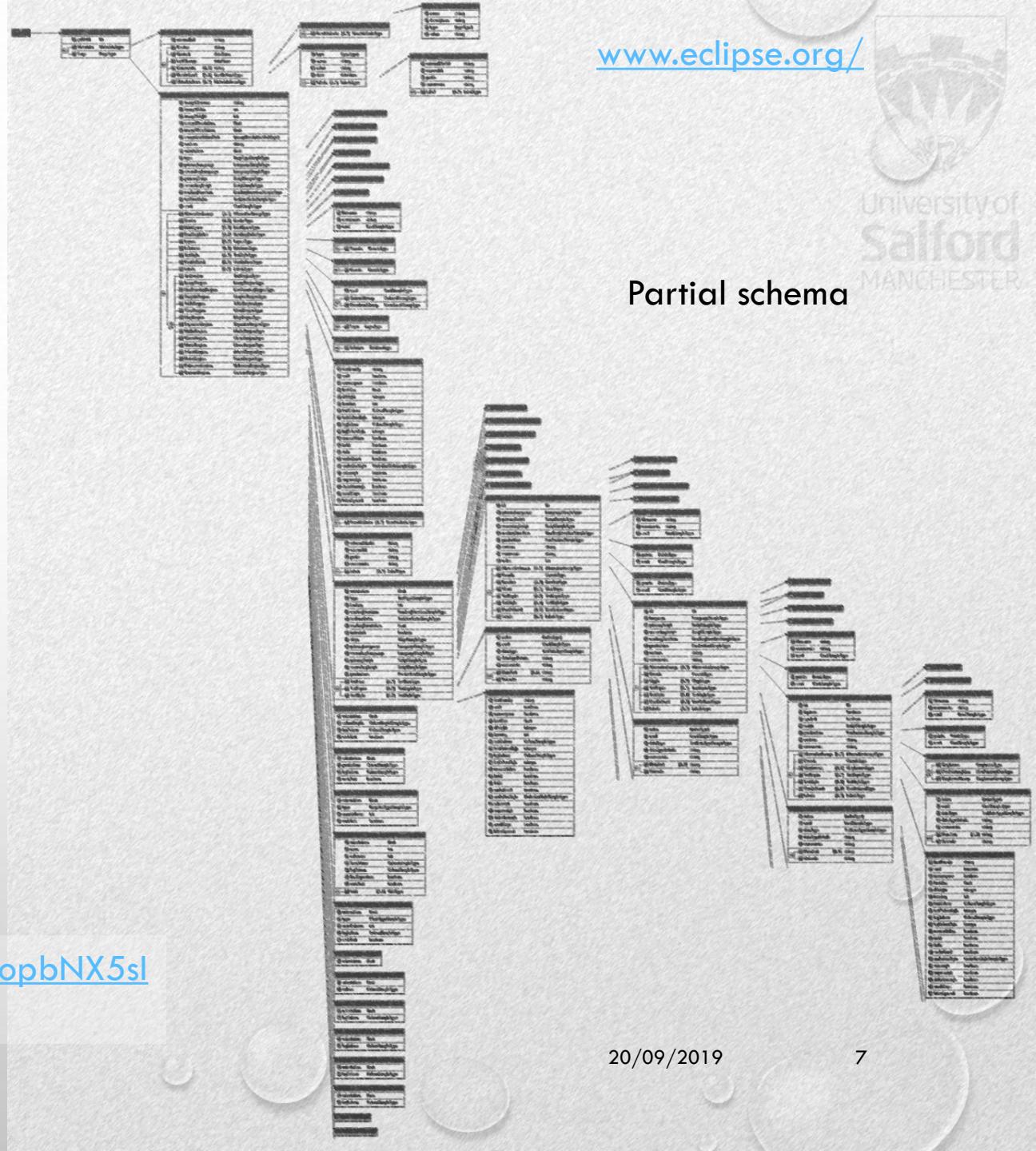
OPEN SOURCED, AND NOW?

- 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:

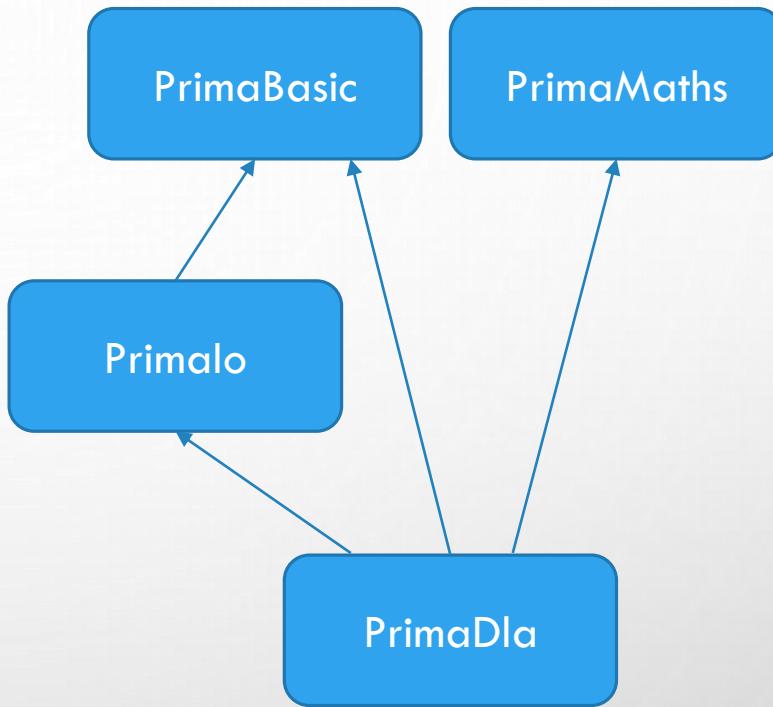
<https://docs.google.com/document/d/11GdAMJD5yDopbNX5lCgArBpcLroSITqKOElARI954/edit?usp=sharing>





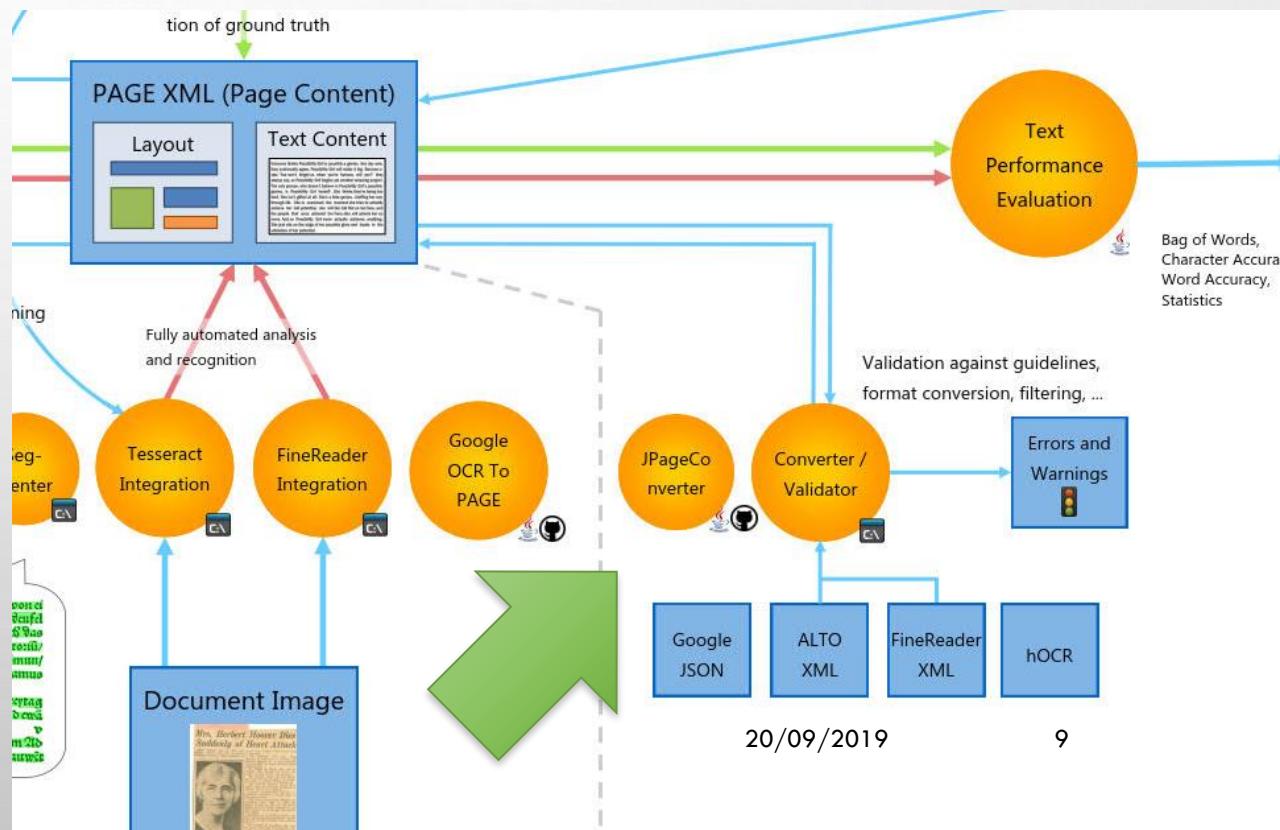
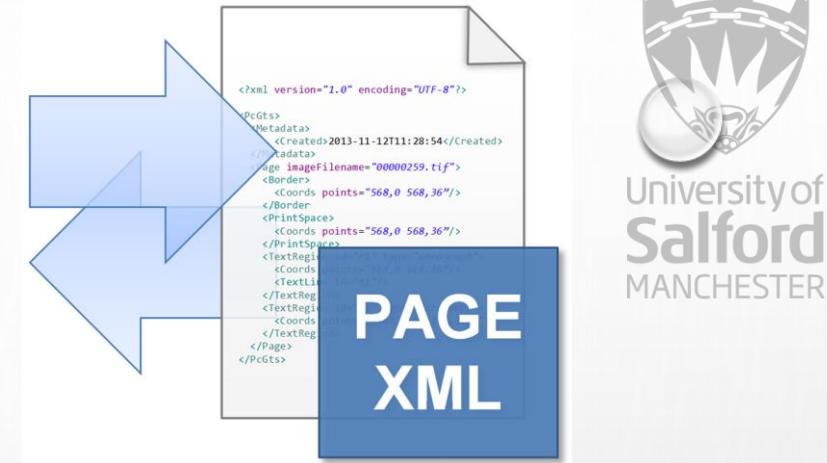
CORE LIBS

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



PAGE CONVERTER AND VALIDATOR

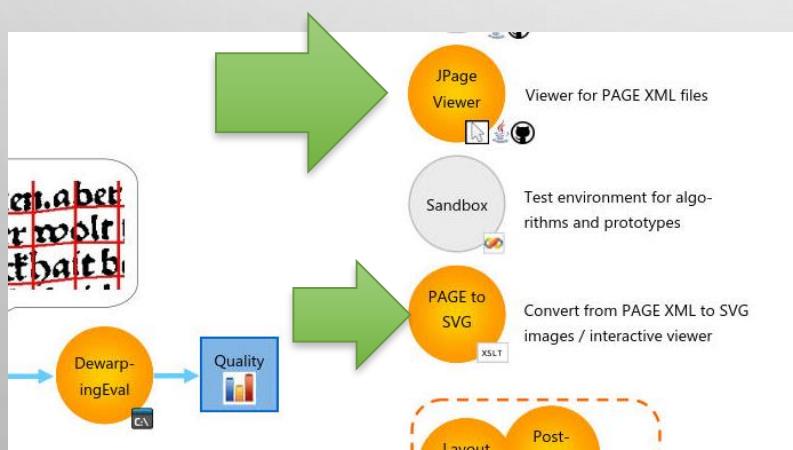
- Java and C++ version
- Slightly different features
- Conversion to PAGE
- Validation
 - Against schema
 - Against ground truthing rules
- Modification
 - E.g. removing certain content



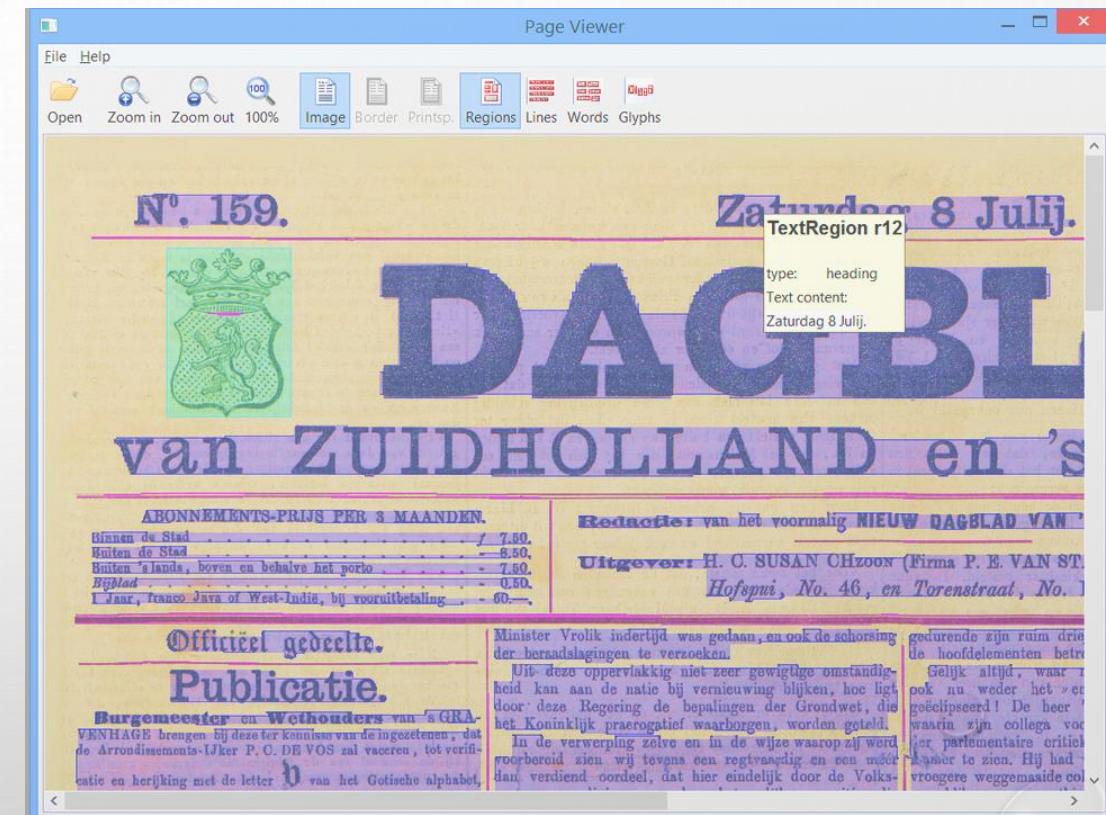


PAGE VIEWER

- 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, ...



ICDAR-OST 2019



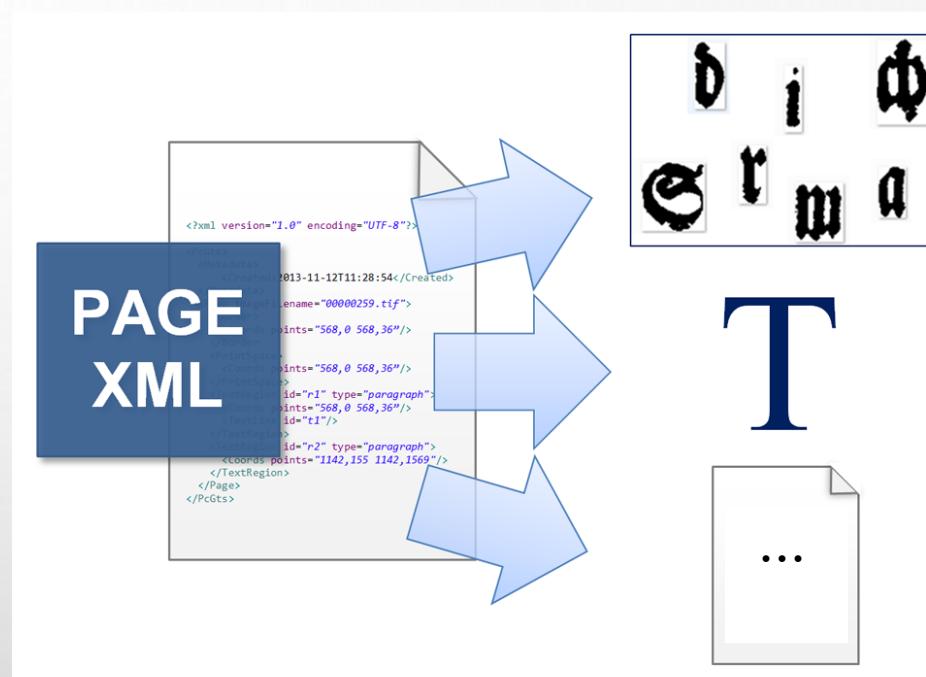
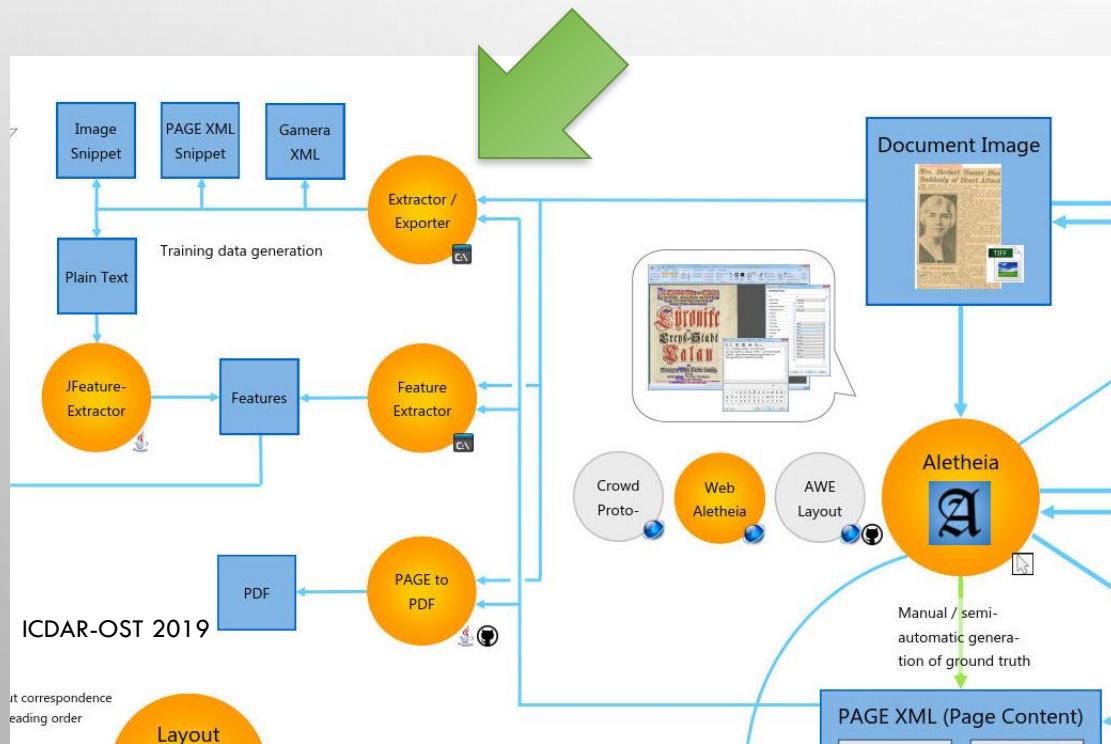
20/09/2019

10



EXTRACTOR / EXPORTER

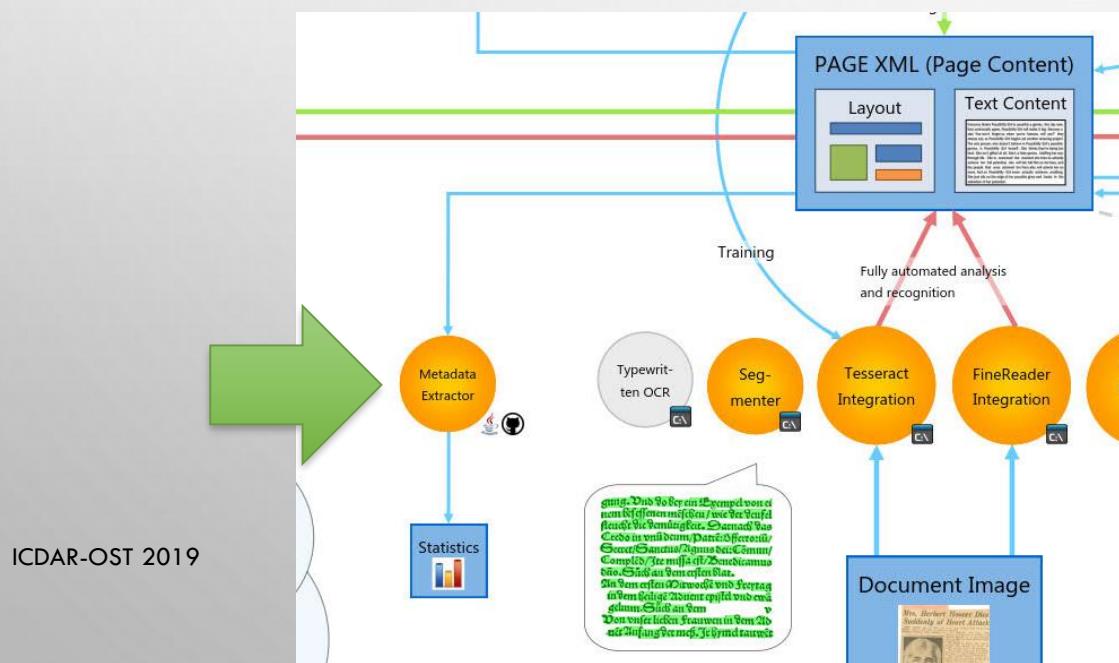
- A





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





PAGE TO PDF

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

ALETHEIA

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

See also “PRIMa Research Lab” YouTube channel

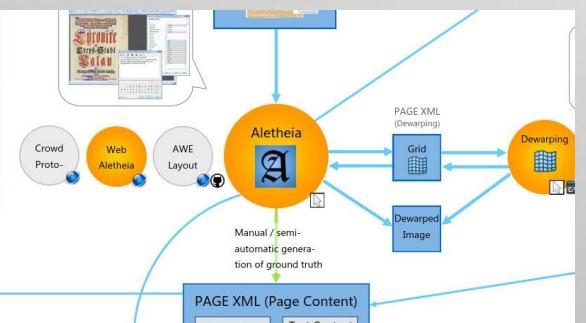


University of
Salford
MANCHESTER

- 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/)

The screenshot shows the Aletheia software interface. The top menu bar includes Home, View, Image, Bounds, Regions (F6), Text Lines (F7), Words (F8), Glyphs (F9), Structure, Dewarping, Experimental, Save all changes, Save, Metadata Statistics, Attributes, Undo, Redo, Rotate left, Rotate right, Crop page, Select all, Full Page, Colour, B/W, Analyse Page Auto, Validation, Layout Evaluation, Settings Interface, About, Updates, User Guide, Introduction Toolbar, Open Example, Help, and Aletheia. The main window has a "Welcome" tab showing "Start new document" options like Image File, Document Image, Structure, Tables, Validation, and links to Aletheia Introduction, User manual, User interface, Online tutorials, and Open example. It also shows "Open document" options like PAGE XML and "Add Page" options like <XML> and <X>. A "Recent Documents" list includes aletheiaexamplepage.xml, DehiArabic_1901_0045.xml, Add MS 7474_0109.xml, Add MS 7474_0050.xml, Add MS 7474_0049.xml, Add MS 7474_0049.xml, and Add MS 7474_0047.xml. The bottom right panel displays the "Properties for aletheiaexamplepage.xml" for a "Text Region r3", showing attributes like Type (Text), Main Attributes (Text Type: Production, Language and Script: Primary: English, Secondary:), and detailed settings for Text Style, Font family, Font size, x-Height, Serif, Monospace, Kerning, and Text colour name. The bottom center panel shows the "Aletheia Document Analysis System" interface with an "Overview" section containing text about Aletheia's features and its support for ground truthing and validation.

ICDAR-OST 2019



- 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

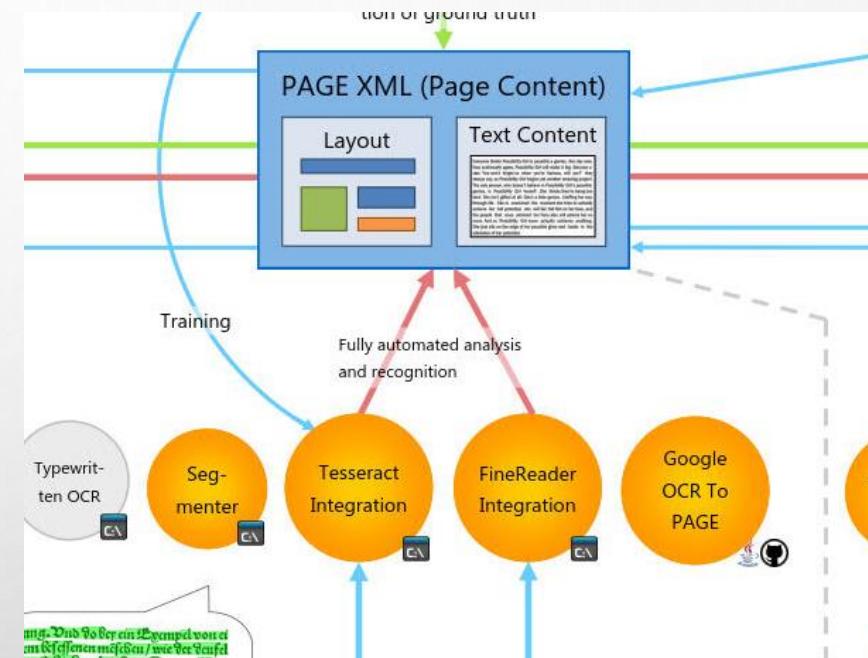
- Based on GWT (pronounced [gwid])
 - Google Web Toolkit, but now community-run
 - Cross-compiles from Java to HTTP+JavaScript
 - Less and less activity
 - Dead end?!
- Runs on Tomcat
- Basic viewing and editing

The screenshot shows the WebAletheia application interface. At the top, there's a toolbar with various icons for zooming and navigating. Below it is a menu bar with 'Content' selected, followed by 'Regions', 'Text Lines', 'Words', 'Glyphs', 'Border', and 'Print Space'. To the right of the menu is a 'TextRegion' panel containing a table with columns for ID, type, align, production, primaryLanguage, secondaryLanguage, primaryScript, secondaryScript, readingOrientation, readingDirection, orientation, textColour, bgColour, fontType, fontName, and kerning. The main workspace displays a document page with several regions highlighted in blue. One region contains text about Aletheia being an advanced system for ground truthing. Another region shows a tree diagram. At the bottom of the workspace is the URL www.primaresearch.org/tools/Aletheia.



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

OCR Engine Setup

Engine: Select to prefill all fields with default values for a specific OCR engine

Name: Tesseract4 Description: This is an experimental integration of the Tesseract engine.

Executable: tesseract4\TesseractToPAGE.exe Select...

Languages / training data: English, Dutch, German, Finnish
New Remove Up Down

Caption: ID: Aletheia Separator for multiple languages: +

Command line calls: Page layout and OCR
Single region/block
Single text line
Single word
...
-inp-img [IMG] -page [PAGEINDEX] -out-xml [OUT] -rec-mode [DEPTH] -lang [LANG] -tessdata [DATA] [CFG] [RDORD] [POLY]

Placeholders and parameter values

Input image [IMG]	OCR output [OUT]	Language [LANG]	Analysis depth [DEPTH]	Config file [CFG]	Reading order [RDORD]	Improve polygons [POLY]	Clipping [CLIP]
<input type="radio"/> XML	<input type="radio"/> Plain text	Layout layout	Prefix -config	Yes -reading-ord	No -orig-outlines	<input checked="" type="radio"/> Pass polygon file	<input type="radio"/> Image cutout
Regions ocr-regions	File Select...	Lines ocr-lines					
Words ocr-words		Glyphs ocr-glyphs					

Generate and show example command line call...

Window positions Reset

OCR Engines Add / modify...

Advanced... 1 : Requires restart of Aletheia

Restore Defaults Close

Page Analysis and OCR

Engine: Tesseract 4 ...

This is an experimental integration of the Tesseract OCR engine for automatic page segmentation and text recognition. The correctness of the results cannot be guaranteed and needs to be checked by the user.

Use the document validation (Home - Quality) for assistance.

Language: English, Dutch, German, Finnish, French, Italian, Latin, Polish

Analysis Depth: Regions (layout only), Regions with text, Text lines and regions with text, Words, lines and regions with text, Glyphs, words, lines and regions with text

Additional Options

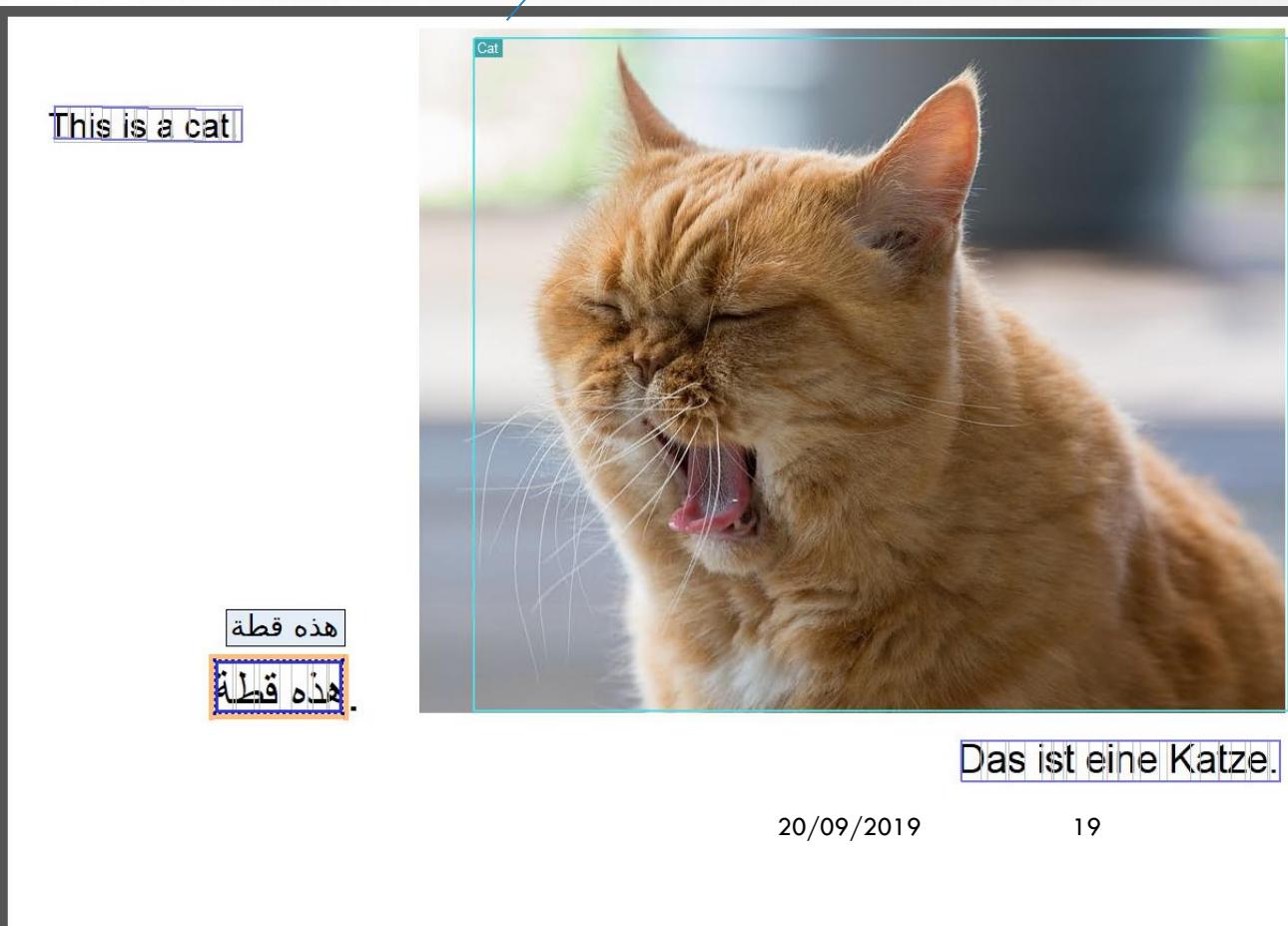
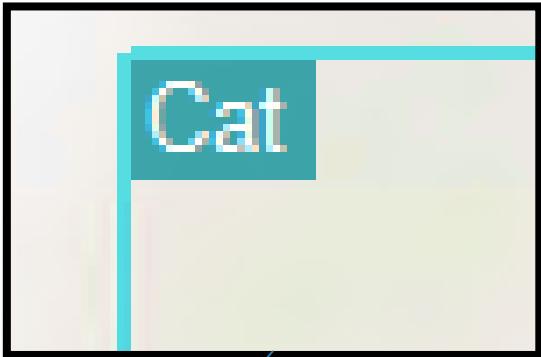
Image to use: Colour / greyscale Black-and-white

Create reading order Improve polygons Delete existing layout objects

20/09/2019 18

ents tool
outline tools
on tools ('Shrinkers')
plit tool
t tools
detection tool
and OCR tools
rt object tools

GOOGLE CLOUD VISION



This is a cat.

هذه قطة

هذه قطة

Das ist eine Katze.



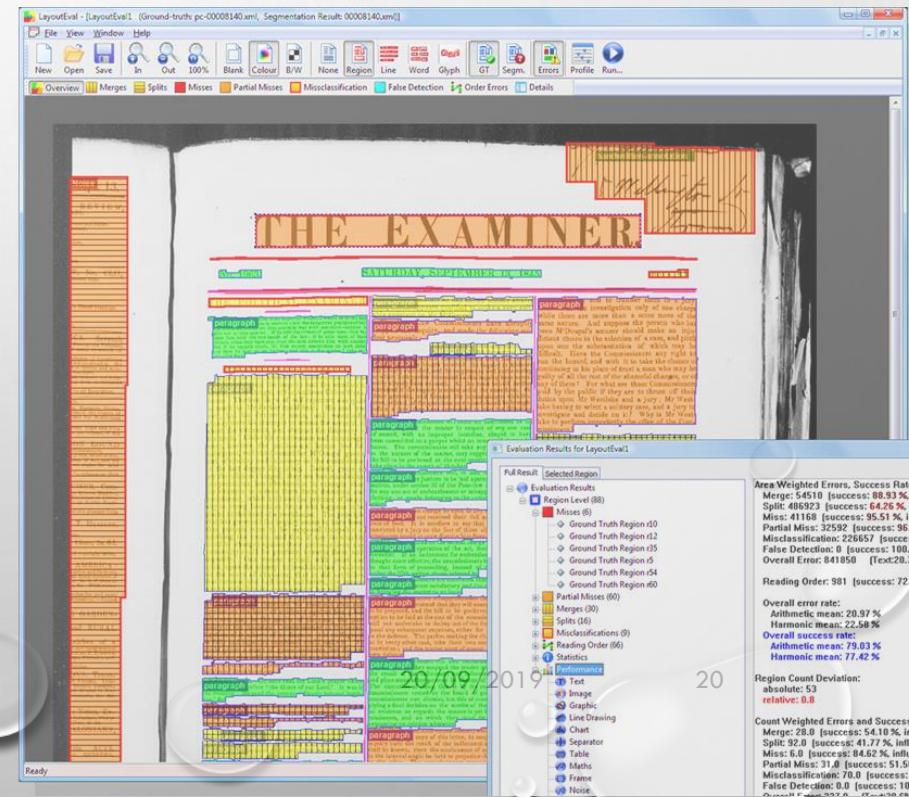
PERFORMANCE EVALUATION

- 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

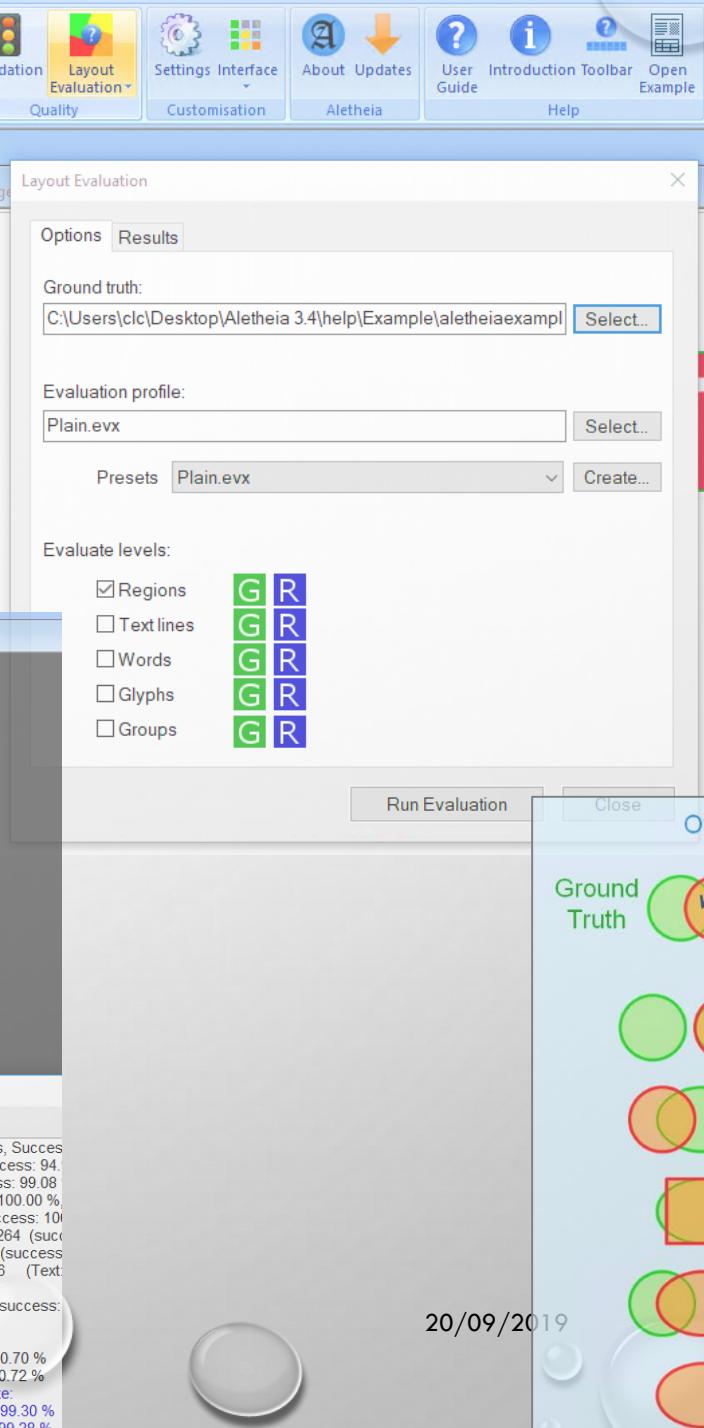
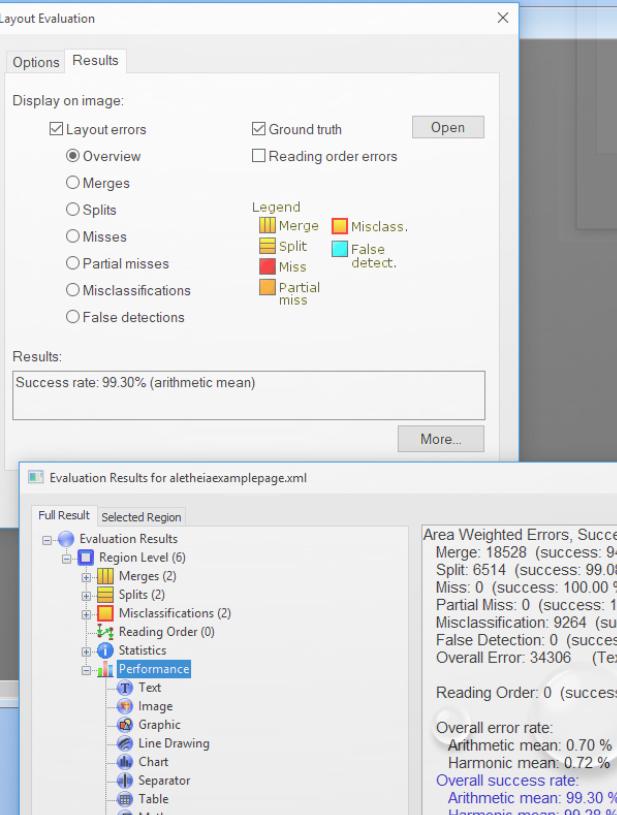
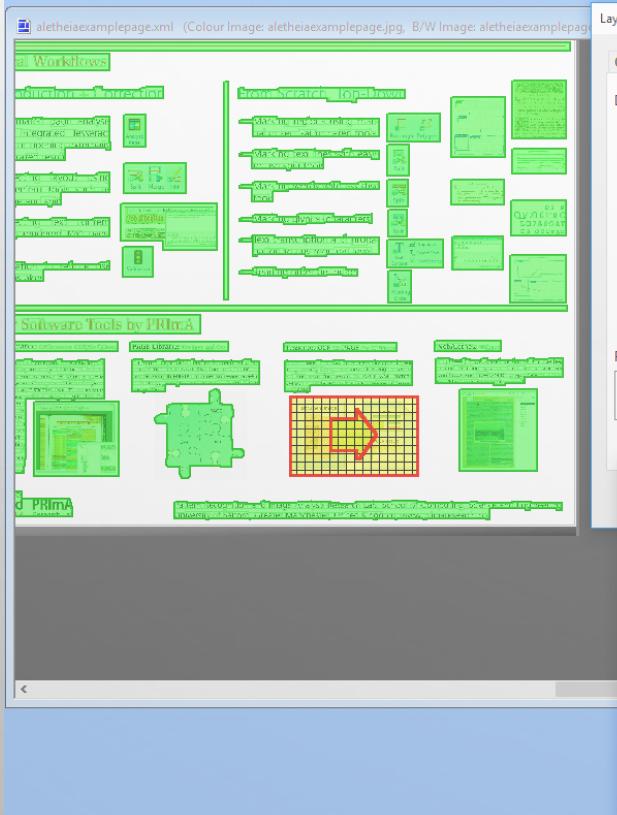
github.com/PRImA-Research-Lab/prima-text





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



University of
Salford
MANCHESTER

- Effort to reimplement the layout evaluation in python
- OCR-D is role model project for openness
- Talk to Clemens Neudecker or colleagues

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



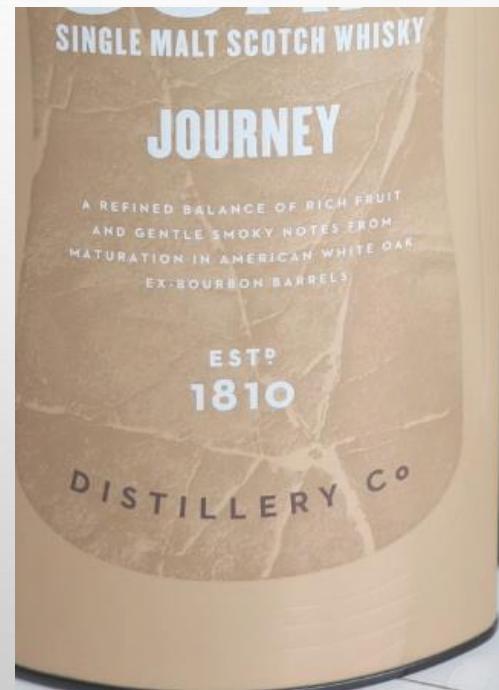
University of
Salford
MANCHESTER

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

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

SEMANTIC LABELLING

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



Ph.D. Thesis



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

The screenshot shows the Aletheia software interface. At the top, there's a toolbar with various icons and a dropdown menu. A sub-menu is open under the 'Competitions...' option, listing 'Page layout evaluation (single page)', 'Page layout evaluation (multiple pages)', 'Open evaluation profile...', and 'Competitions...'. Below this, there's a section titled 'PRIMA Research Competitions' with a sub-section for 'Competition on Recognition of Historical Arabic Scientific Manuscripts (RASM2018)'. It includes a thumbnail image of a manuscript page with green and blue annotations, a brief description, and a note about the challenge. Another section for 'ICDAR2017 Competition on Recognition of Documents with Complex Layouts - RDCL 2017' is shown with its own description and a thumbnail of a document page with complex layout analysis. The bottom section is for 'Competition on Historical Book Recognition (HBR2013)', also with its own description and a thumbnail of a historical book page.

Evaluate results in Aletheia

Validation Layout Evaluation Settings Interface About Updates User Guide Introduction Toolbar Open Example Help

Page layout evaluation (single page)
Page layout evaluation (multiple pages)
Open evaluation profile...
Competitions...

PRIMA Research Competitions

Measure the **Competitions** Evaluation for competition datasets

Select a competition:

Competition on Recognition of Historical Arabic Scientific Manuscripts (RASM2018)

Challenge focussing on finding an optimal solution for accurately and automatically transcribing the British Library's vast and growing digital archive of historical Arabic scientific handwritten manuscripts within the Qatar Digital Library.

ICDAR2017 Competition on Recognition of Documents with Complex Layouts - RDCL 2017

The competition presents challenges for page segmentation, region classification, and text recognition in an end-to-end scenario. The dataset contains scanned pages from contemporary magazines and technical articles. Participants will be provided with know-how and tools that aid the development or extension of their page analysis systems.

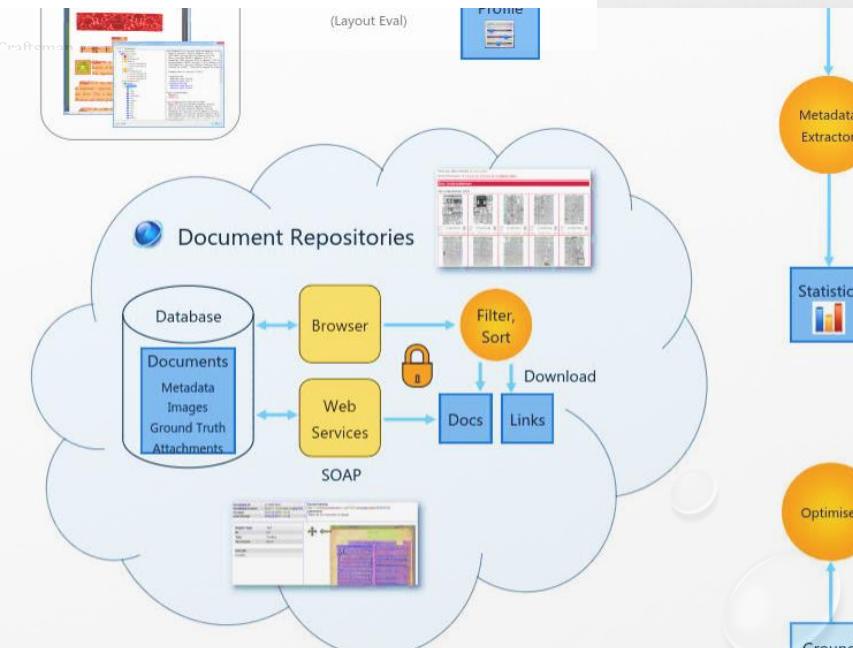
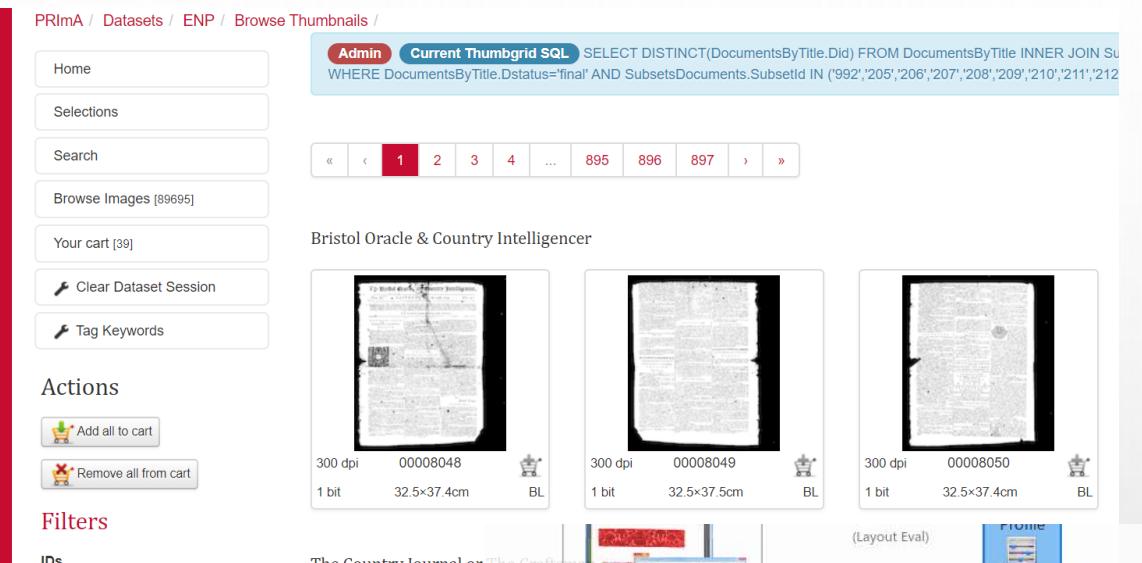
Competition on Historical Book Recognition (HBR2013)

Historical books represent a large proportion of libraries' holdings and continue to be the focus of large-scale digitisation projects. A number of distortions frequently manifest themselves in scans of historical books, hindering layout analysis and text recognition. The motivation of the competition is to evaluate existing approaches using a realistic dataset and an objective performance analysis system.

Close

DATASETS

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

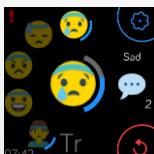
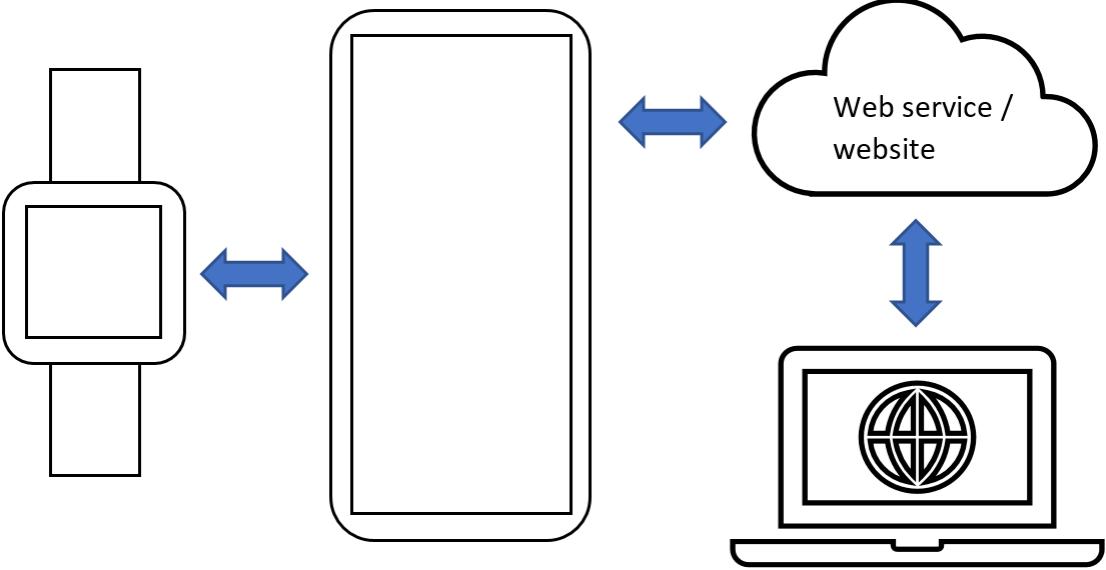


UNIVERSITY ENVIRONMENT

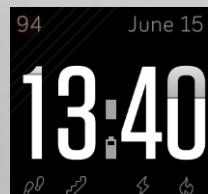
- People coming and going
 - Loss of know-how
 - E.g. website dev stalling
- University changes
 - Politics
 - Structure
- REF – Research Excellence Framework
 - Need impact case studies
 - Need to show impact (hence we ask for user information)

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





University of
Salford
MANCHESTER

youtube.com/channel/UCtxCXg-UvSnTKPOzLH4wJaQ

Search for “Coding Tech” on YouTube

CODING TECH

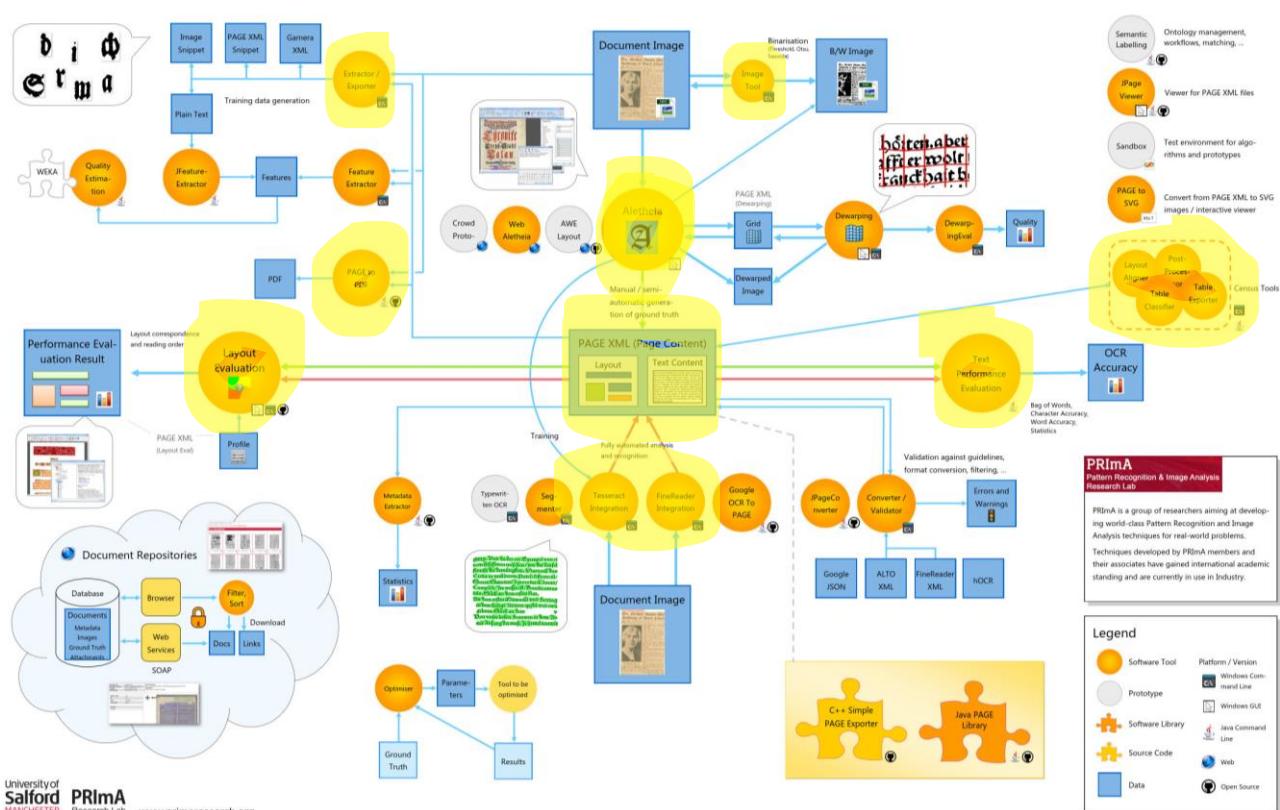
- Great source of information
- Very diverse content, something for everyone
- Talks, tutorials, ...

The screenshot shows the 'PLAYLISTS' tab of the Coding Tech YouTube channel. The channel has 481K subscribers. It displays a grid of 10 playlists:

Playlist Name	Thumbnail	Last Updated	View Full Playlist
Security	[Icon: Headphones]	Updated today	VIEW FULL PLAYLIST
IoT	[Icon: USB port]	Updated today	VIEW FULL PLAYLIST
Performance and Testing	[Icon: CPU and chart]	Updated yesterday	VIEW FULL PLAYLIST
Python	[Icon: Python logo]	Updated 2 days ago	VIEW FULL PLAYLIST
Design	[Icon: UX icons]	Updated 4 days ago	VIEW FULL PLAYLIST
Soft Skills	[Icon: Person with speech bubbles]	Updated 7 days ago	VIEW FULL PLAYLIST
THE REACT PRODUCTIVITY REVOLUTION	[Icon: React logo]	Updated 7 days ago	VIEW FULL PLAYLIST
Software Development	[Icon: Code editor]	Updated 7 days ago	VIEW FULL PLAYLIST
React	[Icon: Airbnb logo]	20/09/2019	VIEW FULL PLAYLIST
Web Development	[Icon: HTTP/HTTPS]	28	VIEW FULL PLAYLIST

CENSUS

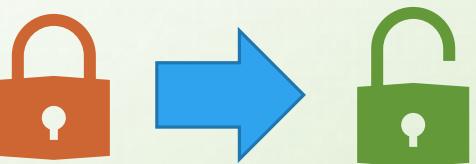
- Used many of the shown tools + some new ones





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

