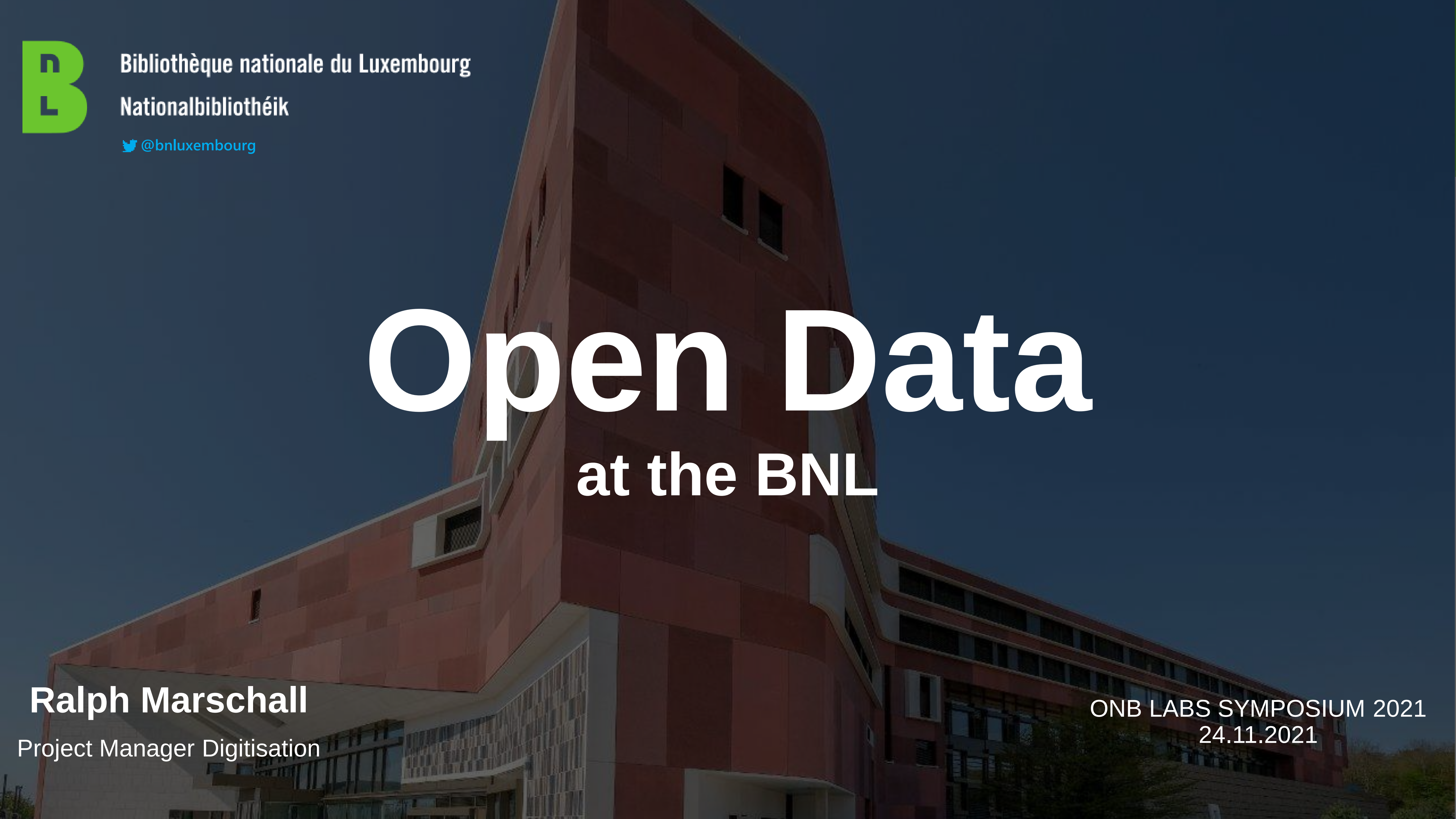
Bibliothèque nationale du Luxembourg

Nationalbibliothéik

@bnluxembourg



Open Data

at the BNL

Ralph Marschall

Project Manager Digitisation

ONB LABS SYMPOSIUM 2021
24.11.2021

Quick overview

- Digitization @ National Library of Luxembourg (BnL)
- Open Data & Tools
- A.I. / Machine Learning Project

Digitization at National Library of Luxembourg (BnL)

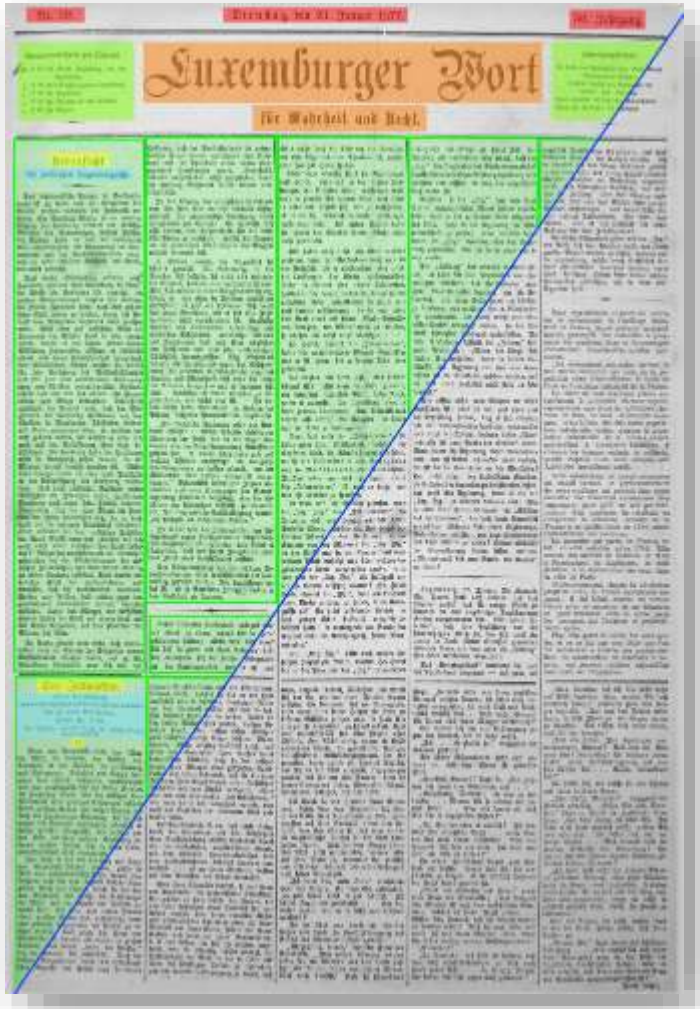
Digitization @ BnL

- **Preserve** old documents
  - **Restrict** access to physical volumes (scan once, minimize physical usage)
  - Some documents are **unique** (1 copy)
  - Replace microfilms with **improved digital access**
- **Search** Full Text – OCR (optical text recognition)
- Decompose images into **articles** (semantic units) – OLR (optical layout recognition)
- **Present** digital results online: eluxemburgensia.lu, a-z.lu

Digitization timeline

2003 to 2006	• Image only • Newspapers, postcards, books, manuscripts
2006+	• Luxembourg newspapers • Image & rich metadata (METS/ALTO)
2015	• 15 medieval manuscripts with metadata • 441 books with METS/ALTO
2016+	• Newspapers, books, postcards etc.

Standards



TIFF

- **300 DPI** Images
- **Uncompressed**
- 8 bit
- 256 levels of grey or full color

PDF

- **PDF/A** for online download
- With **Full Text** and **TOC**

XML

- **METS:** Container Format
- **ALTO:** OCR Format

For each issue / book:

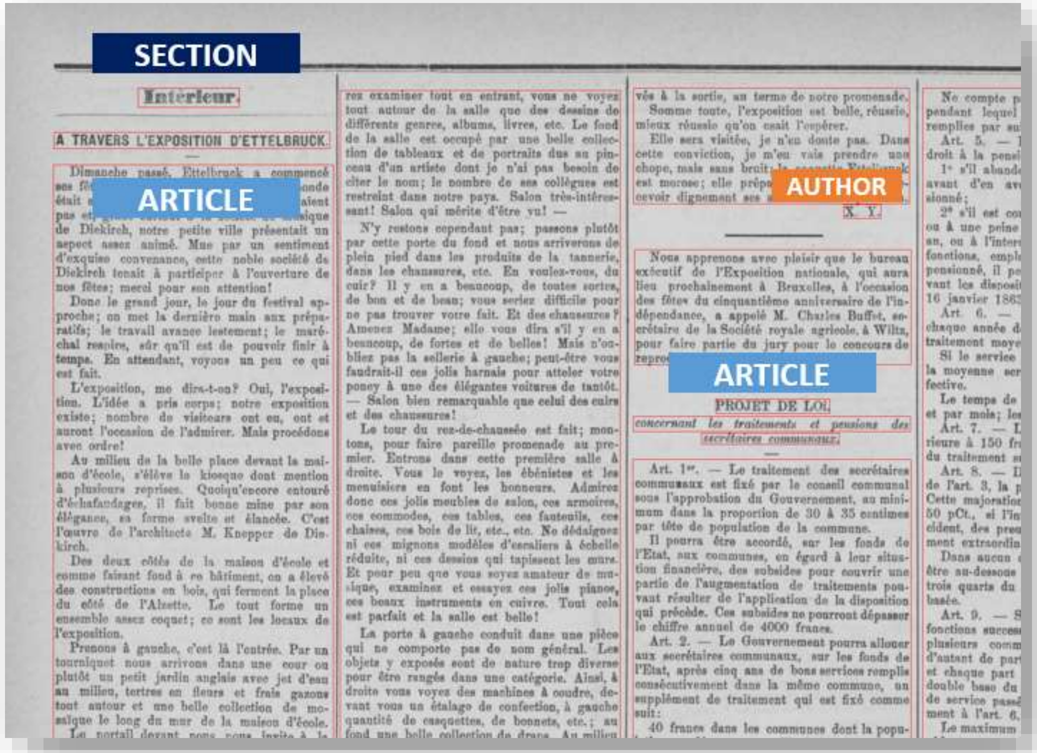
- 1 METS file
- 1 PDF file

For each page:

- 1 TIFF
- 1 PNG (B&W)
- 1 PDF
- 1 ALTO

METS / ALTO

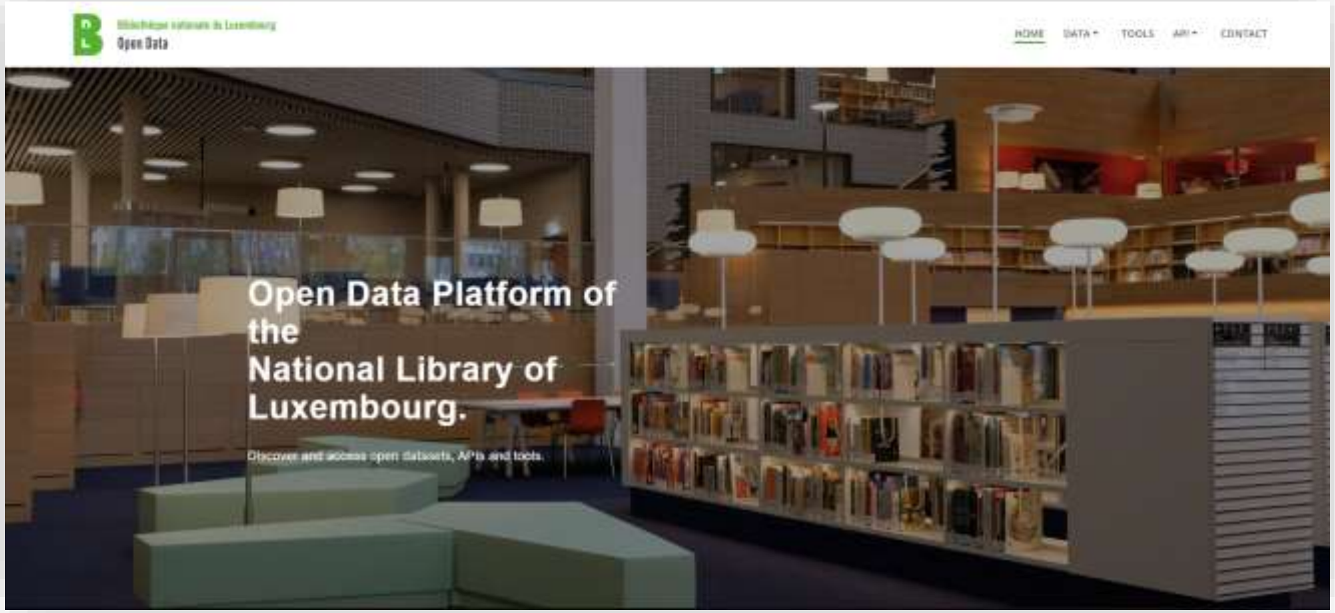
- **METS:**
  - 1 METS per document
  - Describes the logical and physical structure
  - Contains all the metadata (technical, descriptive, access rights etc.)
  - Links to ALTO blocks
- **ALTO:**
  - 1 ALTO file per page
  - OCR results with text coordinates
  - Organized in Text Blocks, Composed Blocks, Lines and Strings elements


SECTION
Intérieur
A TRAVERS L'EXPOSITION D'ETTELBRUCK
ARTICLE
AUTHOR
X. Y.
ARTICLE
PROJET DE LOI
concernant les traitements et pensions des secrétaires communaux.

Bibliothèque nationale du Luxembourg
Nationalbibliothéik

Open Data & Open Tools

2019: data.bnl.lu



Open Data Platform of the
National Library of Luxembourg.

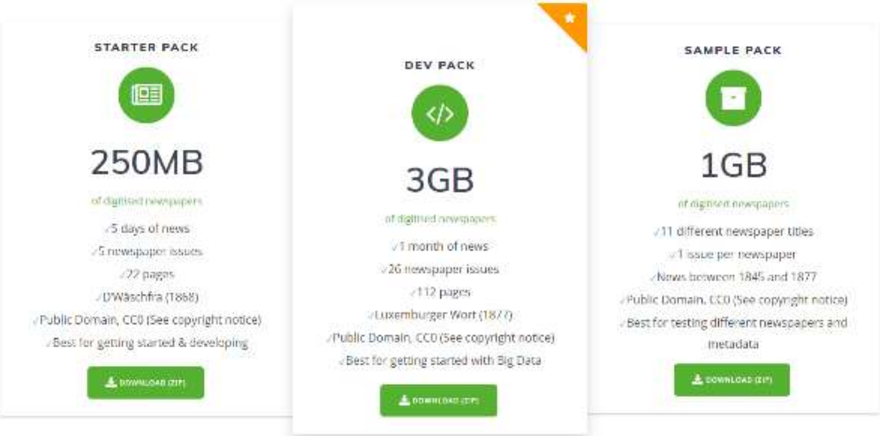
Discover and access open datasets, APIs and tools.

Data **Tools** **APIs**

data.bnl.lu

- Organised in Packs
- Download what you need

- Raw or Processed Data
- **Technical Specifications**



STARTER PACK

250MB

of digitised newspapers

✓5 days of news
✓5 newspaper issues
✓22 pages
✓D'Wäschfra (1868)
✓Public Domain, CC0 (See copyright notice)
✓Best for getting started & developing

DEV PACK

3GB

of digitised newspapers

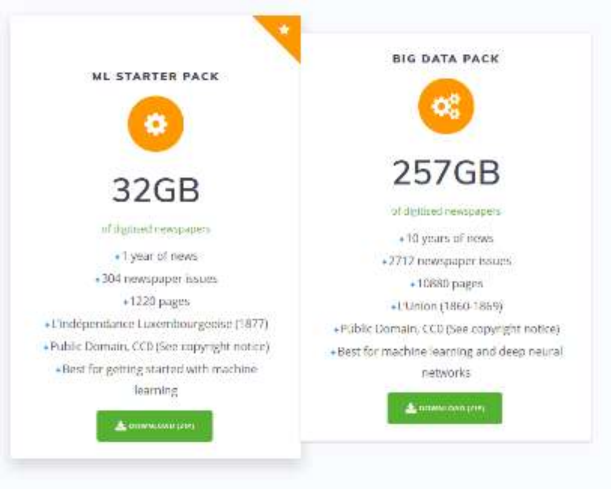
✓1 month of news
✓26 newspaper issues
✓112 pages
✓Luxemburger Wort (1877)
✓Public Domain, CC0 (See copyright notice)
✓Best for getting started with Big Data

SAMPLE PACK

1GB

of digitised newspapers

✓11 different newspaper titles
✓1 issue per newspaper
✓News between 1845 and 1877
✓Public Domain, CC0 (See copyright notice)
✓Best for testing different newspapers and metadata



ML STARTER PACK

32GB

of digitised newspapers

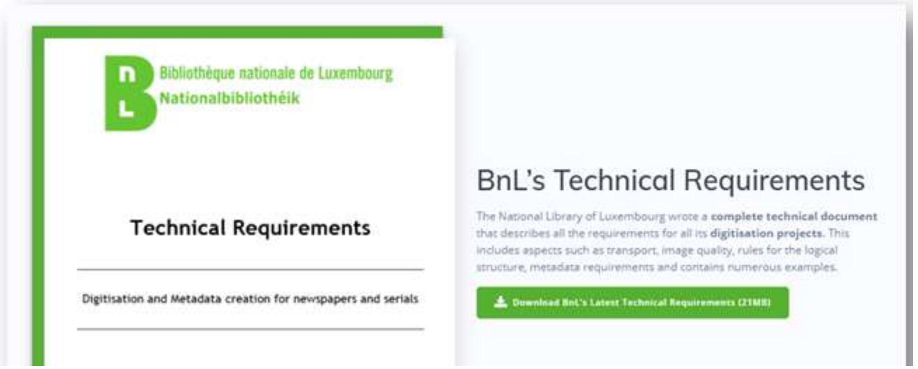
- 1 year of news
- 304 newspaper issues
- 1220 pages
- L'indépendance Luxembourgeoise (1877)
- Public Domain, CC0 (See copyright notice)
- Best for getting started with machine learning

BIG DATA PACK

257GB

of digitised newspapers

- 10 years of news
- 2712 newspaper issues
- 10880 pages
- L'Union (1860-1869)
- Public Domain, CC0 (See copyright notice)
- Best for machine learning and deep neural networks



Bibliothèque nationale de Luxembourg
Nationalbibliothéik

Technical Requirements

Digitisation and Metadata creation for newspapers and serials

BnL's Technical Requirements

The National Library of Luxembourg wrote a **complete technical document** that describes all the requirements for all its **digitisation projects.** This includes aspects such as transport, image quality, rules for the logical structure, metadata requirements and contains numerous examples.

Hackathon – Open Data Challenge

- Create novel interfaces / tools ideally using machine learning techniques
- METS/ALTO is not user-friendly when you have 2 days to learn it & use it

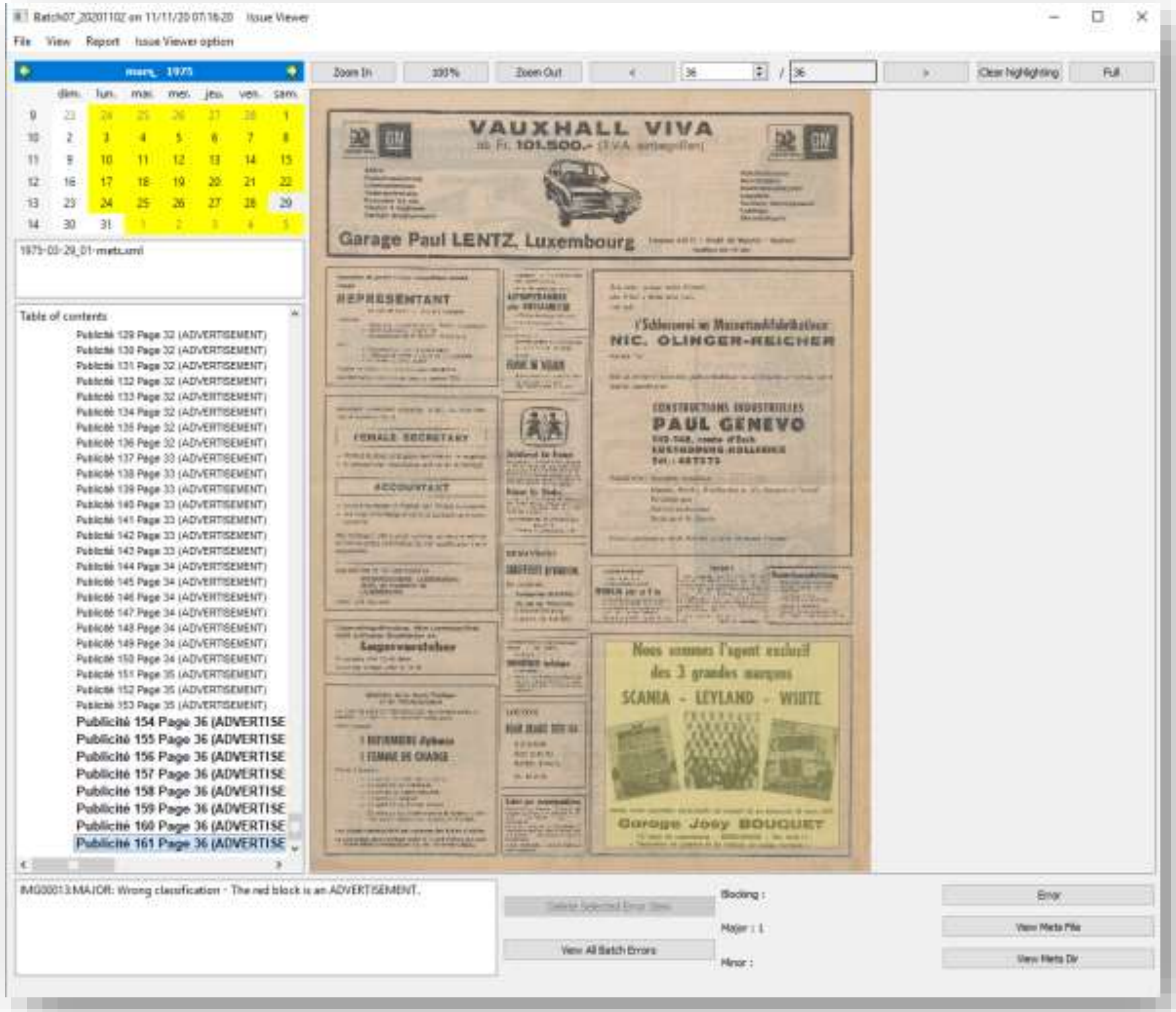
GAME OF [HACK ATHON] CODE
digital luxembourg
Bibliothèque nationale de Luxembourg
AWARDS CEREMONY

Quality Assurance Tool

Namalysator

- Checks METS/ALTO files
- Measures accuracy on the “titles to check manually”
- Generates Report (custom error codes)
- Open Source

https://github.com/natliblux/Namalysator



	dim.	lun.	mar.	mer.	jeu.	ven.	sam.
9	23	24	25	26	27	28	1
10	2	3	4	5	6	7	8
11	9	10	11	12	13	14	15
12	16	17	18	19	20	21	22
13	23	24	25	26	27	28	29
14	30	31	1	2	3	4	5

1975-03-29_01-mets.xml

Table of contents

Publicité 129 Page 32 (ADVERTISEMENT)
Publicité 130 Page 32 (ADVERTISEMENT)
Publicité 131 Page 32 (ADVERTISEMENT)
Publicité 132 Page 32 (ADVERTISEMENT)
Publicité 133 Page 32 (ADVERTISEMENT)
Publicité 134 Page 32 (ADVERTISEMENT)
Publicité 135 Page 32 (ADVERTISEMENT)
Publicité 136 Page 32 (ADVERTISEMENT)
Publicité 137 Page 33 (ADVERTISEMENT)
Publicité 138 Page 33 (ADVERTISEMENT)
Publicité 139 Page 33 (ADVERTISEMENT)
Publicité 140 Page 33 (ADVERTISEMENT)
Publicité 141 Page 33 (ADVERTISEMENT)
Publicité 142 Page 33 (ADVERTISEMENT)
Publicité 143 Page 33 (ADVERTISEMENT)
Publicité 144 Page 34 (ADVERTISEMENT)
Publicité 145 Page 34 (ADVERTISEMENT)
Publicité 146 Page 34 (ADVERTISEMENT)
Publicité 147 Page 34 (ADVERTISEMENT)
Publicité 148 Page 34 (ADVERTISEMENT)
Publicité 149 Page 34 (ADVERTISEMENT)
Publicité 150 Page 34 (ADVERTISEMENT)
Publicité 151 Page 35 (ADVERTISEMENT)
Publicité 152 Page 35 (ADVERTISEMENT)
Publicité 153 Page 35 (ADVERTISEMENT)
Publicité 154 Page 36 (ADVERTISE
Publicité 155 Page 36 (ADVERTISE
Publicité 156 Page 36 (ADVERTISE
Publicité 157 Page 36 (ADVERTISE
Publicité 158 Page 36 (ADVERTISE
Publicité 159 Page 36 (ADVERTISE
Publicité 160 Page 36 (ADVERTISE
Publicité 161 Page 36 (ADVERTISE

IMG00013:MAJOR: Wrong classification - The red block is an ADVERTISEMENT.

Blocking :
Major : 1
Minor :

METS / ALTO Viewer

OpenSeadragon

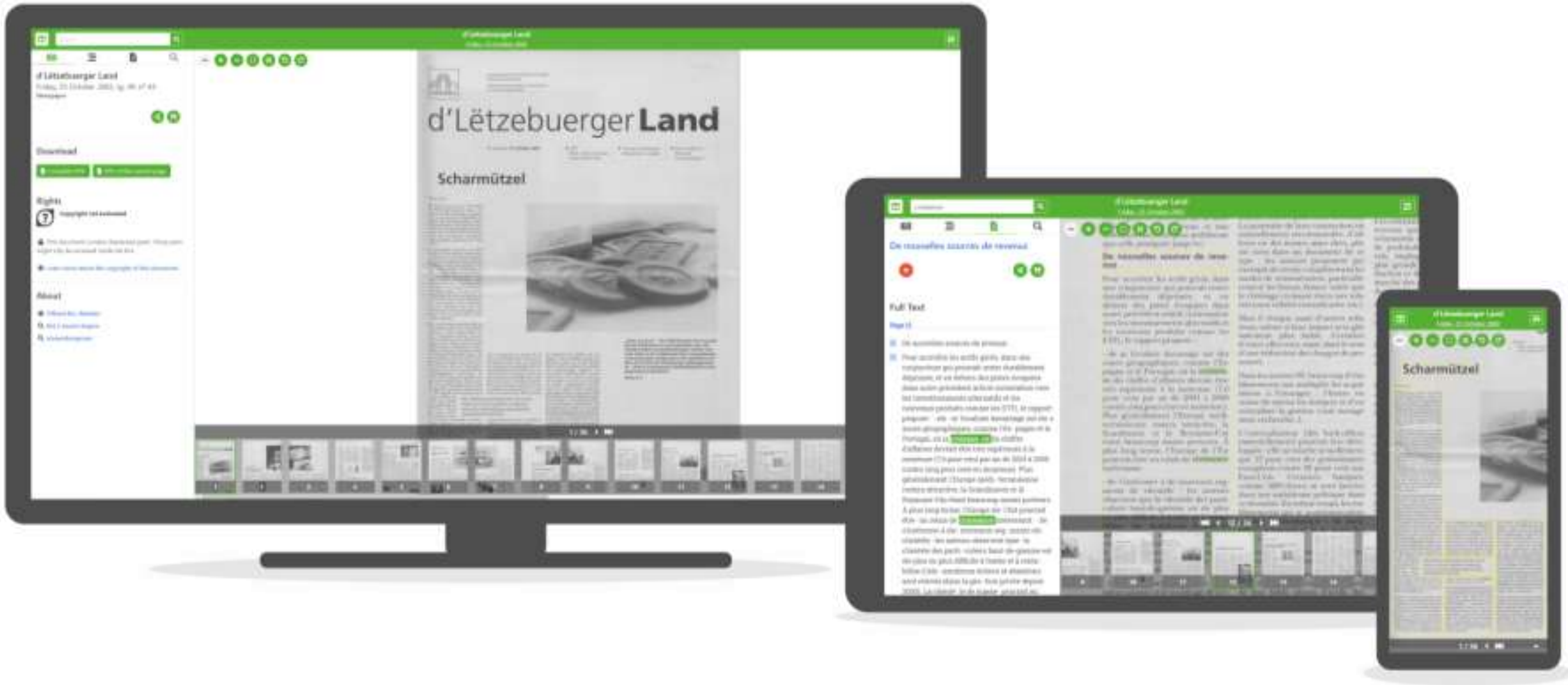
Cantaloupe
High-performance dynamic image server in Java

International Image Interoperability Framework

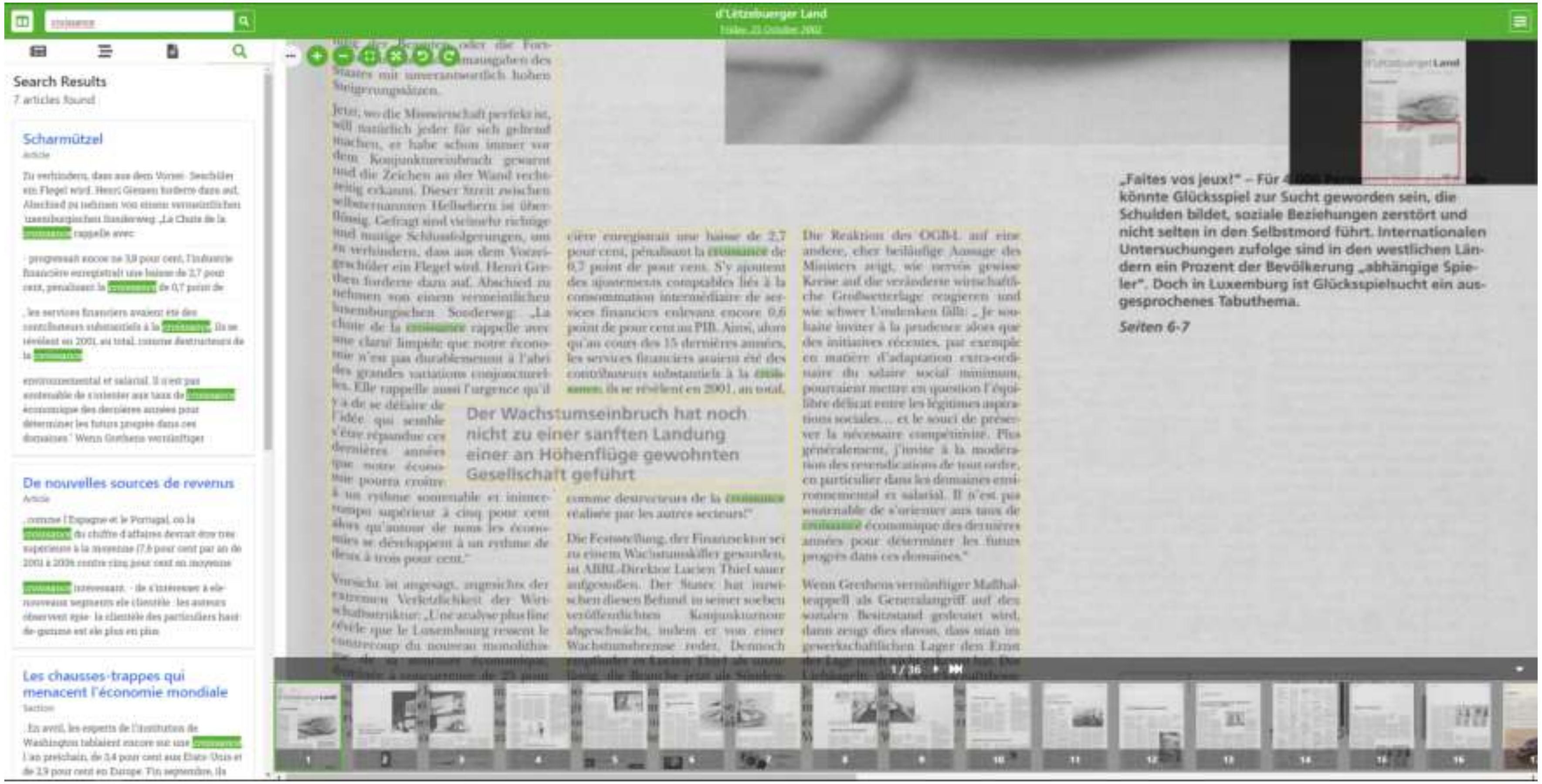
Vue.js

spring boot

HAPROXY



METS / ALTO Viewer



Scharmützel

Article

Zu verhindern, dass aus dem Vorzei- Seschüler ein Flegel wird. Henri Grethen forderte dazu auf, Abschied zu nehmen von einem vermeintlichen luxemburgischen Sonderweg. „La Chute de la croissance rappelle avec

· progressait encore de 3,8 pour cent, l'industrie financière enregistrait une baisse de 2,7 pour cent, penalisant la croissance de 0,7 point de

, les services financiers avaient été des contributeurs substantiels à la croissance, ils se révèlent en 2001 au total comme destructeurs de la croissance

environnemental et salarial. Il n'est pas soutenable de s'orienter aux taux de croissance économique des dernières années pour déterminer les futurs progrès dans ces domaines." Wenn Grethens vernünftiger

De nouvelles sources de revenus

Article

, comme l'Espagne et le Portugal, où la croissance du chiffre d'affaires devrait être très supérieure à la moyenne (7,6 pour cent par an de 2001 à 2006 contre cinq pour cent en moyenne

croissance intéressant. · de s'intéresser à de nouveaux segments de clientèle . les auteurs observent que · la clientèle des particuliers haut-de-gamme est de plus en plus

Les chausses-trappes qui menacent l'économie mondiale

Section

. En avril, les experts de l'institution de Washington tablaient encore sur une croissance l'an prochain, de 3,4 pour cent aux Etats-Unis et de 2,9 pour cent en Europe. Fin septembre, ils

Der Wachstumseinbruch hat noch nicht zu einer sanften Landung einer an Höhenflüge gewohnten Gesellschaft geführt

„Faites vos jeux!" – Für 4 000 Personen ... könnte Glücksspiel zur Sucht geworden sein, die Schulden bildet, soziale Beziehungen zerstört und nicht selten in den Selbstmord führt. Internationalen Untersuchungen zufolge sind in den westlichen Ländern ein Prozent der Bevölkerung „abhängige Spieler". Doch in Luxemburg ist Glücksspielsucht ein ausgesprochenes Tabuthema.

Seiten 6-7

Article Segmentation

Full Text

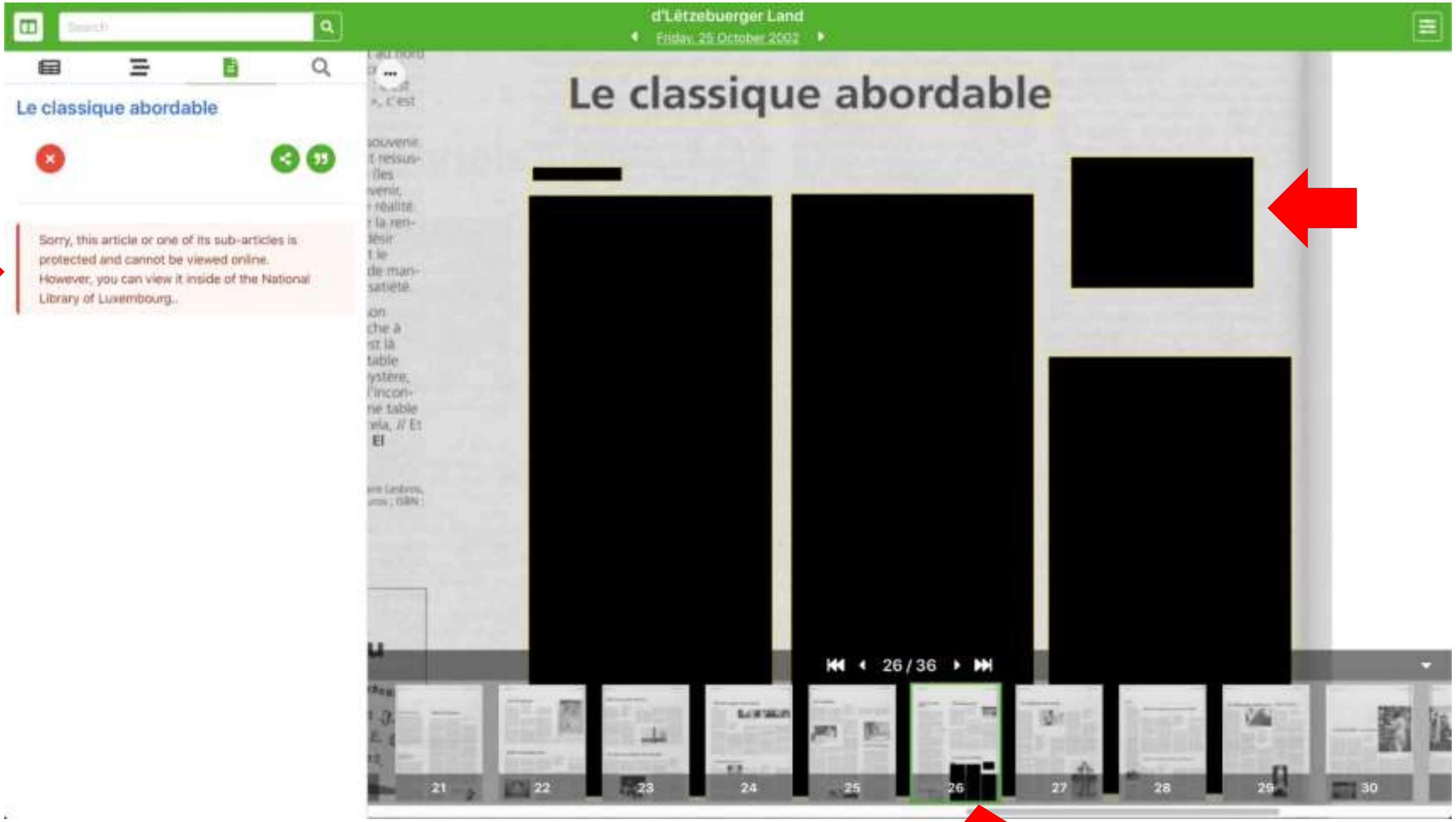
Block-level Links

Word-Highlight

Word-To-Image Link

Bibliothèque nationale du Luxembourg
Nationalbibliothéik

METS / ALTO Viewer



d'Lëtzebuerger Land

Friday, 25 October 2002

Le classique abordable

Sorry, this article or one of its sub-articles is protected and cannot be viewed online. However, you can view it inside of the National Library of Luxembourg..

Le classique abordable

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

26 / 36

- Metadata
- Table of Contents
- Calendar
- Search
- IIIF Images
- Redaction

A.I. / Machine Learning Projects

eLmA = eLuxemburgensia meets AI

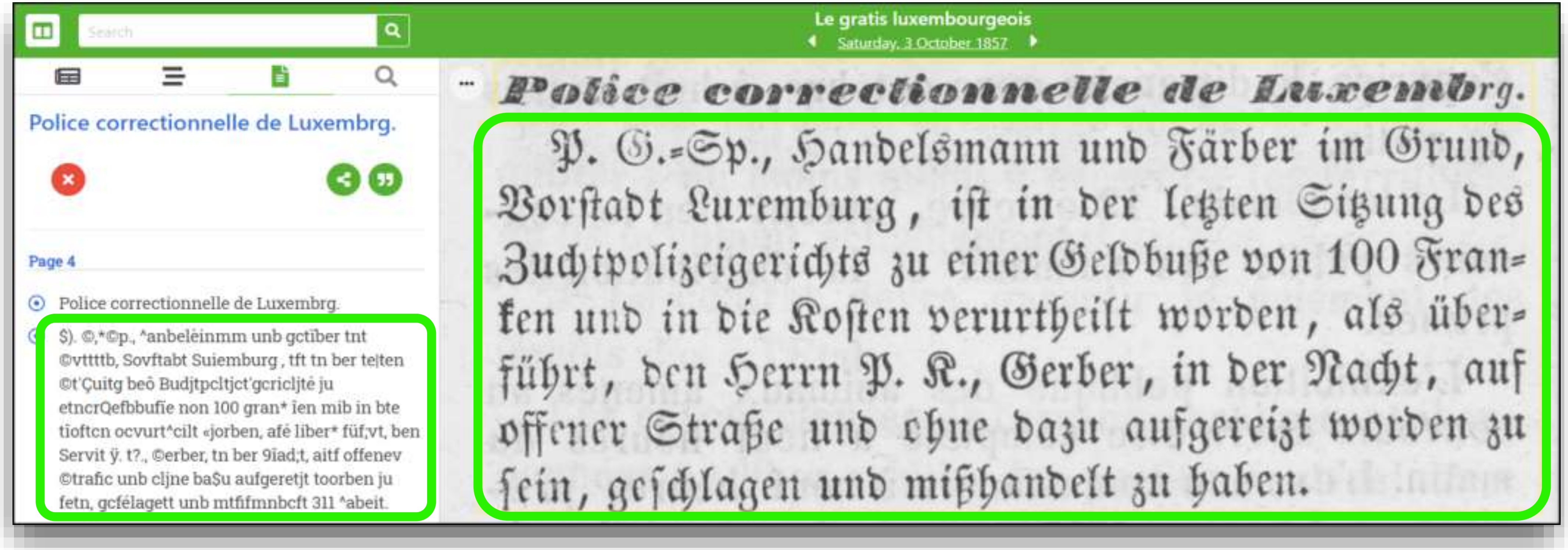
Improving OCR

- Ground Truth
- Engine Training
- Re-OCR

Article Recommendation

- Visual Exploration
- Named Entities

OCR



Le gratis luxembourgeois

Saturday, 3 October 1857

Police correctionnelle de Luxembrg.

Page 4

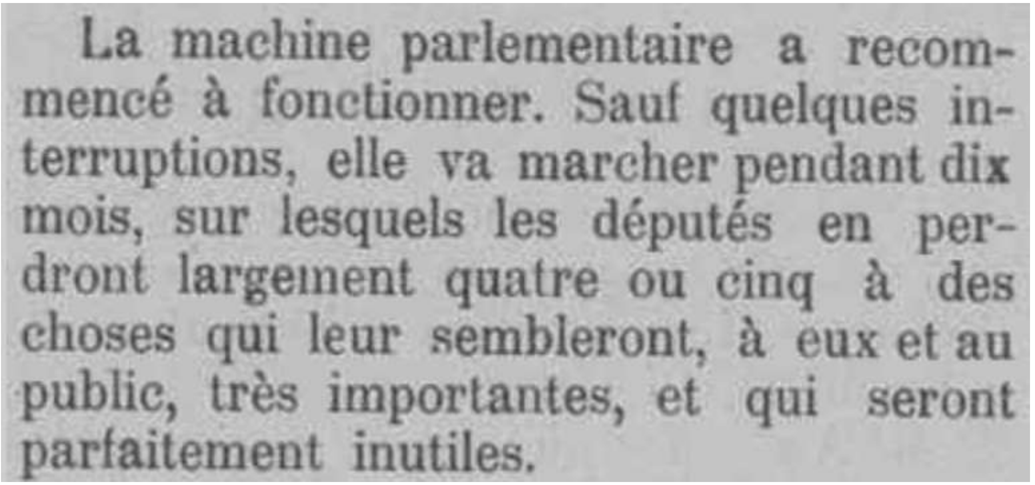
○ Police correctionnelle de Luxembrg.

○ $). ©,*©p., ^anbeléinmm unb gctíber tnt ©vttttb, Sovftabt Suiemburg , tft tn ber te|ten ©t'Çuitg beô Budjtpcltjct'gcricljté ju etncrQefbbufie non 100 gran* ïen mib in bte tïoftcn ocvurt^cilt «jorben, afé liber* füf;vt, ben Servit ỹ. t?., ©erber, tn ber 9ïad;t, aitf offenev ©trafic unb cljne ba$u aufgeretjt toorben ju fetn, gcfélagett unb mtfifmnbcft 311 ^abeit.

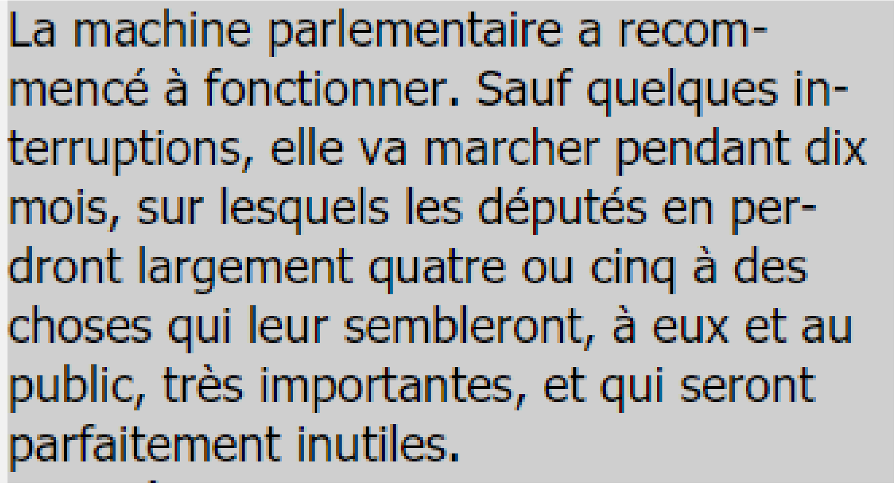
Police correctionnelle de Luxembrg.

P. G.=Sp., Handelsmann und Färber im Grund, Vorſtadt Luxemburg, iſt in der letzten Sitzung des Zuchtpolizeigerichts zu einer Geldbuße von 100 Franken und in die Koſten verurtheilt worden, als überführt, den Herrn P. K., Gerber, in der Nacht, auf offener Straße und ohne dazu aufgereizt worden zu ſein, geſchlagen und mißhandelt zu haben.

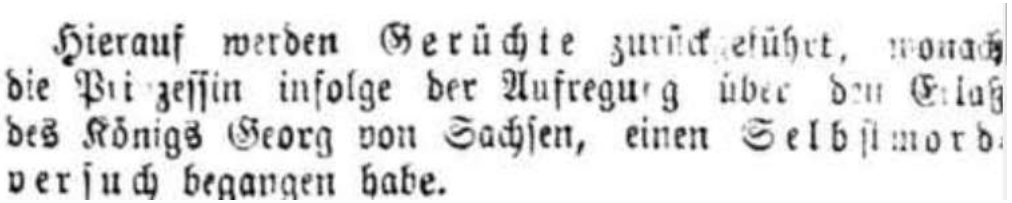
Varying quality for OCR



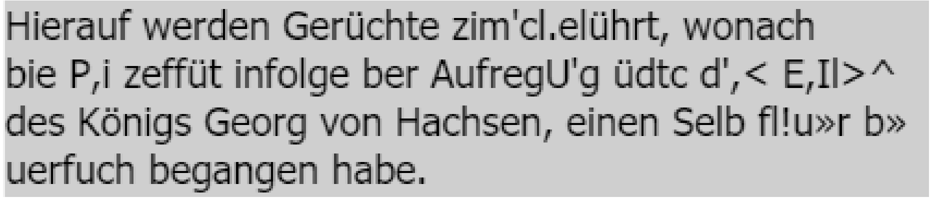
La machine parlementaire a recommencé à fonctionner. Sauf quelques interruptions, elle va marcher pendant dix mois, sur lesquels les députés en perdront largement quatre ou cinq à des choses qui leur sembleront, à eux et au public, très importantes, et qui seront parfaitement inutiles.



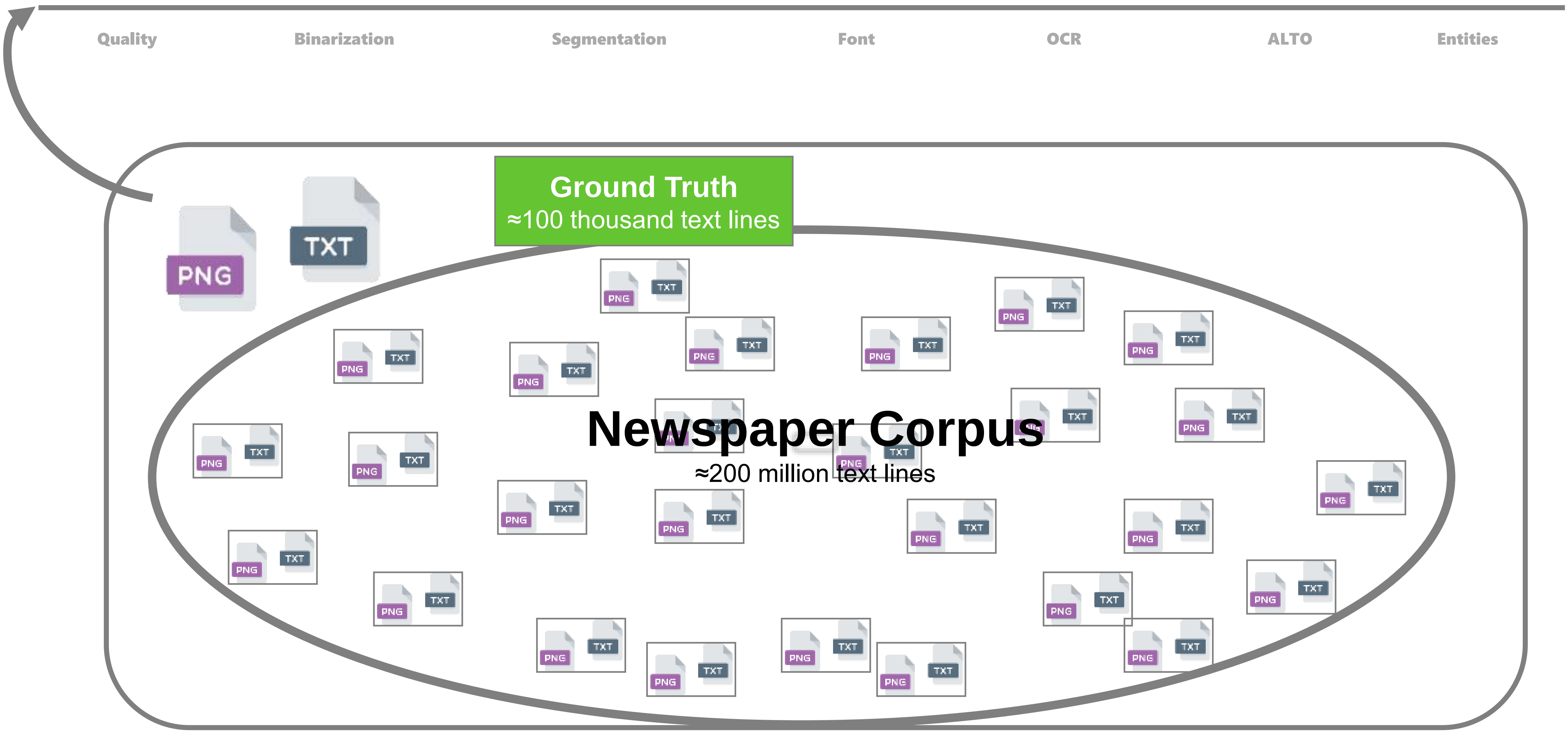
La machine parlementaire a recommencé à fonctionner. Sauf quelques interruptions, elle va marcher pendant dix mois, sur lesquels les députés en perdront largement quatre ou cinq à des choses qui leur sembleront, à eux et au public, très importantes, et qui seront parfaitement inutiles.



Hierauf werden Gerüchte zurückgeführt, wonach die Prinzeſſin infolge der Aufregung über den Erlaß des Königs Georg von Sachſen, einen Selbſtmordverſuch begangen habe.



Hierauf werden Gerüchte zim'cl.elührt, wonach
bie P,i zeffüt infolge ber AufregU'g üdtc d',< E,Il>^
des Königs Georg von Hachsen, einen Selb fl!u»r b»
uerfuch begangen habe.



Quality

Binarization

Segmentation

Font

OCR

ALTO

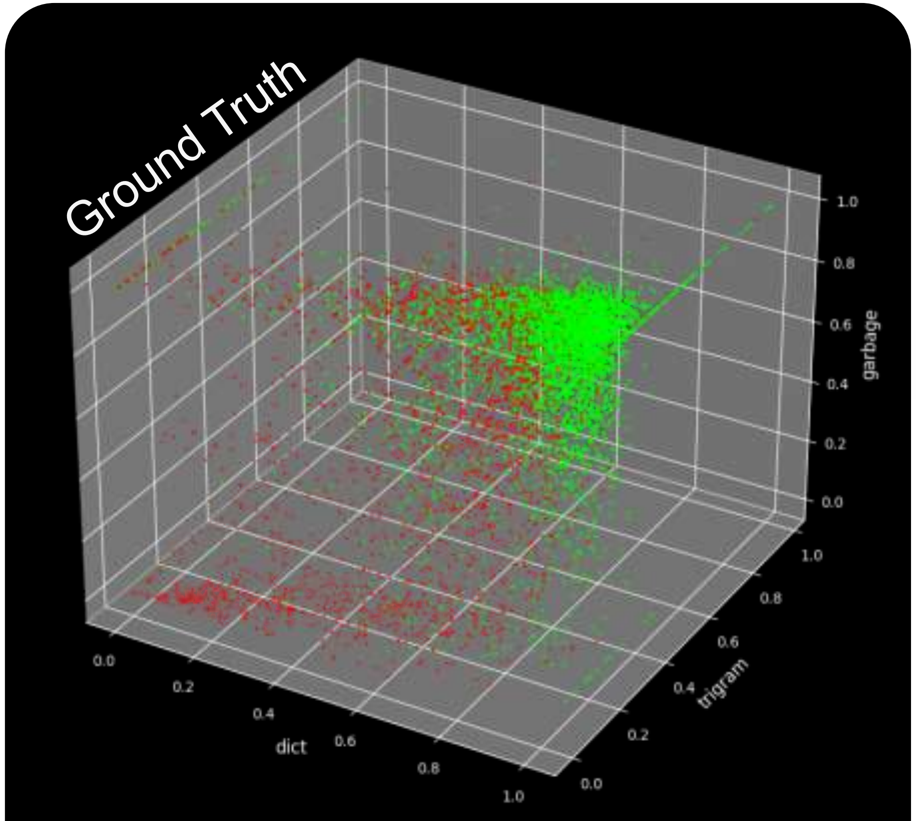
Entities

Ground Truth

≈100 thousand text lines

Newspaper Corpus

≈200 million text lines



unsupervised quality evaluation • kNN algorithm

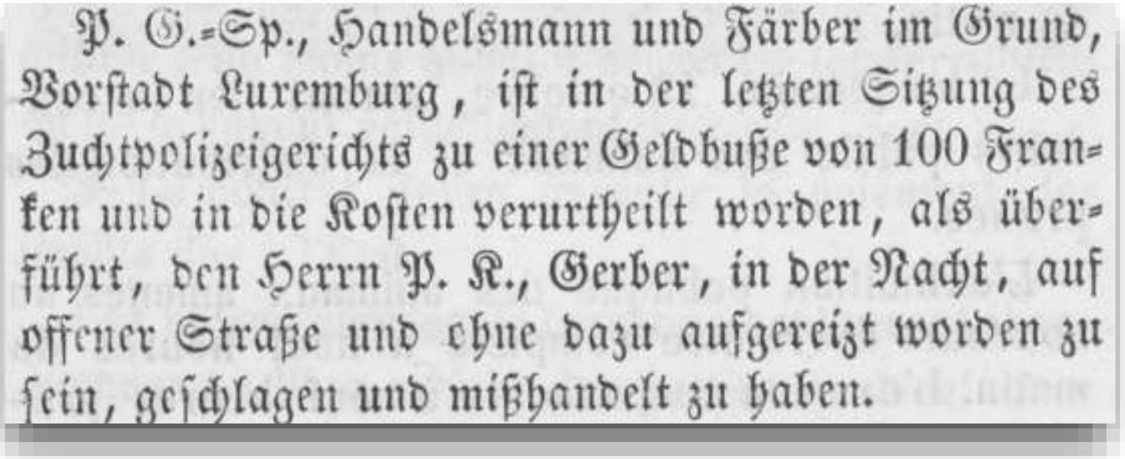
● score > 0.95

● else

score = 1 - (editDistance / |chars|)

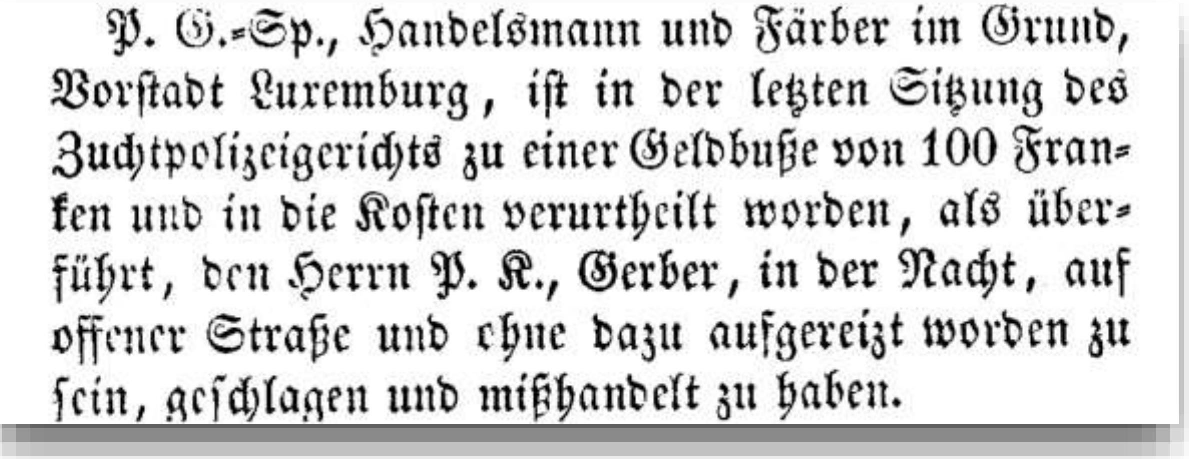
85% Accuracy on test set

Quality **Binarization** Segmentation Font OCR ALTO Entities



P. G.-Sp., Handelsmann und Färber im Grund, Vorſtadt Luxemburg, iſt in der letzten Sitzung des Zuchtpolizeigerichts zu einer Geldbuße von 100 Franken und in die Koſten verurtheilt worden, als überführt, den Herrn P. K., Gerber, in der Nacht, auf offener Straße und ohne dazu aufgereizt worden zu ſein, geſchlagen und mißhandelt zu haben.

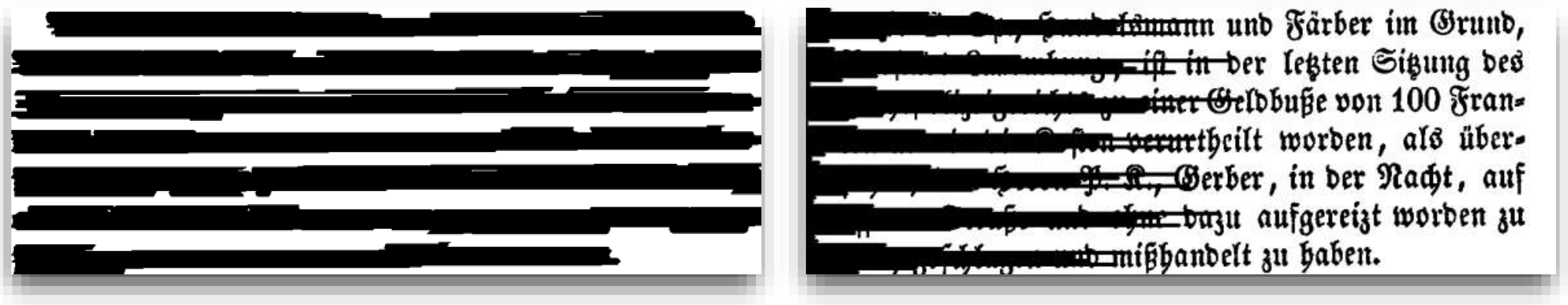
Original Text Block

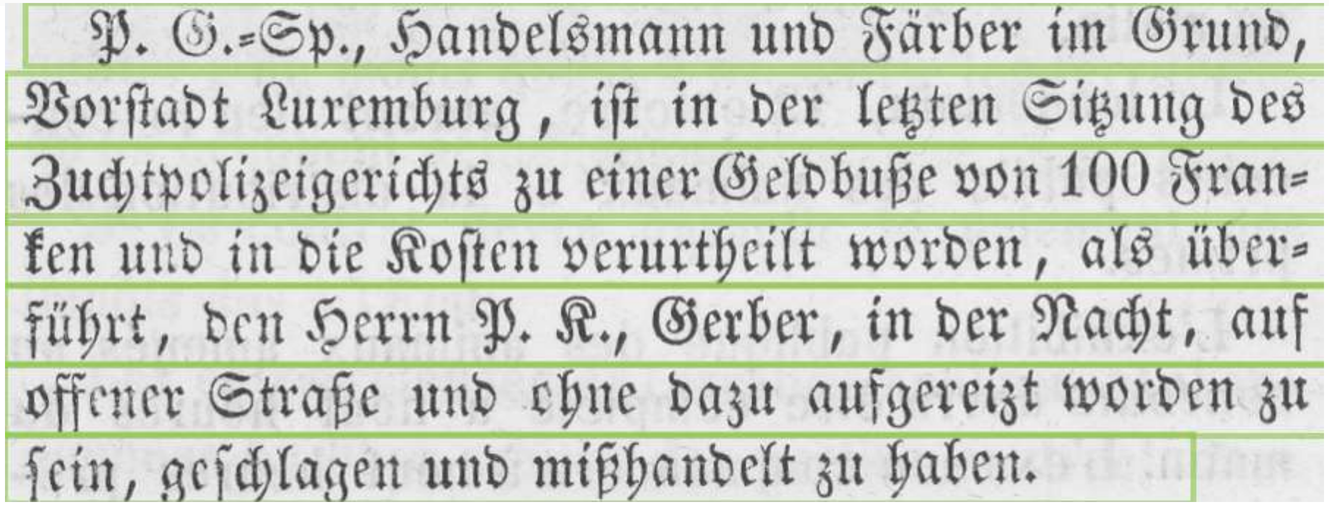


P. G.-Sp., Handelsmann und Färber im Grund, Vorſtadt Luxemburg, iſt in der letzten Sitzung des Zuchtpolizeigerichts zu einer Geldbuße von 100 Franken und in die Koſten verurtheilt worden, als überführt, den Herrn P. K., Gerber, in der Nacht, auf offener Straße und ohne dazu aufgereizt worden zu ſein, geſchlagen und mißhandelt zu haben.

Binarized Text Block

cleaning • dilation • padding • inversion • white on black detection

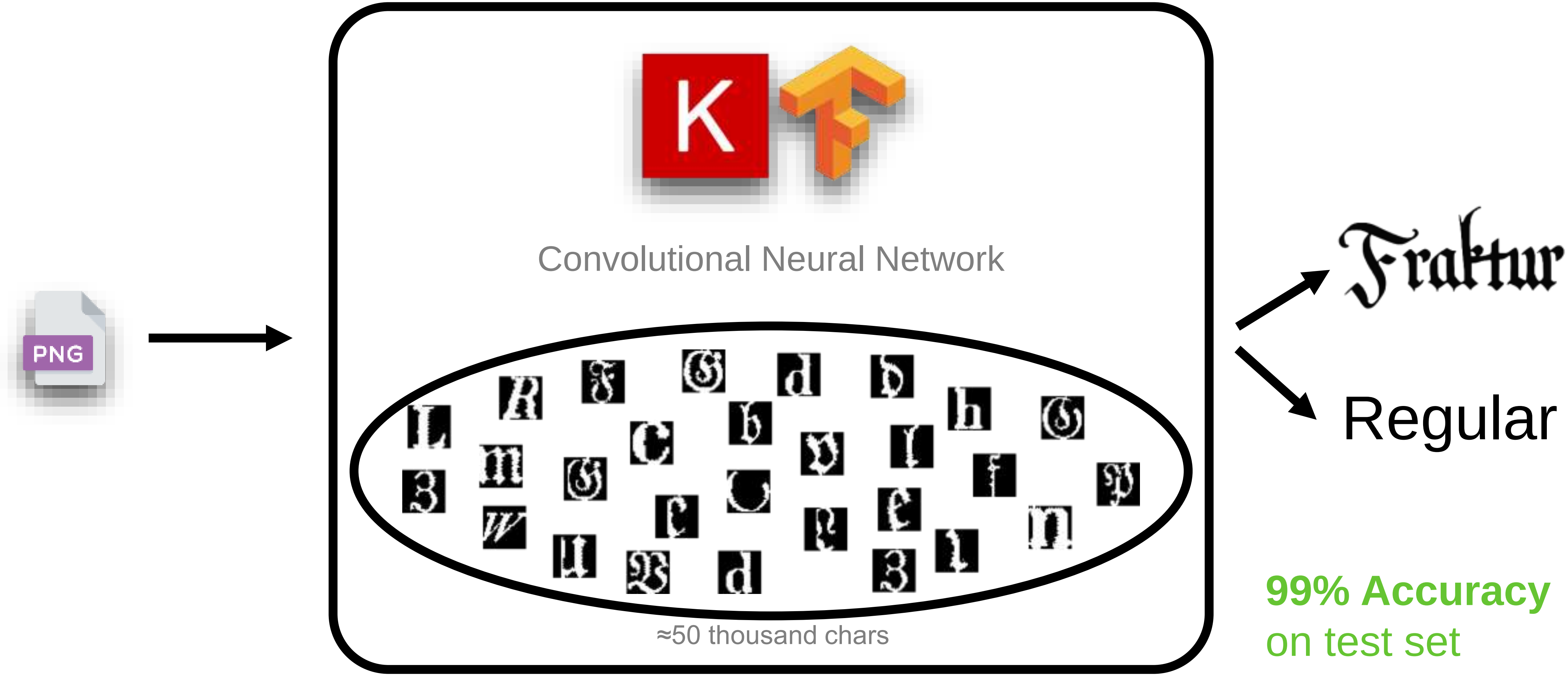




P. G.-Sp., Handelsmann und Färber im Grund, Vorſtadt Luxemburg, iſt in der letzten Sitzung des Zuchtpolizeigerichts zu einer Geldbuße von 100 Franken und in die Koſten verurtheilt worden, als überführt, den Herrn P. K., Gerber, in der Nacht, auf offener Straße und ohne dazu aufgereizt worden zu ſein, geſchlagen und mißhandelt zu haben.

99% Accuracy
on test set

morphology • connected components • horizontal histogram projection



99% Accuracy
on test set

font recognition • binary classifier

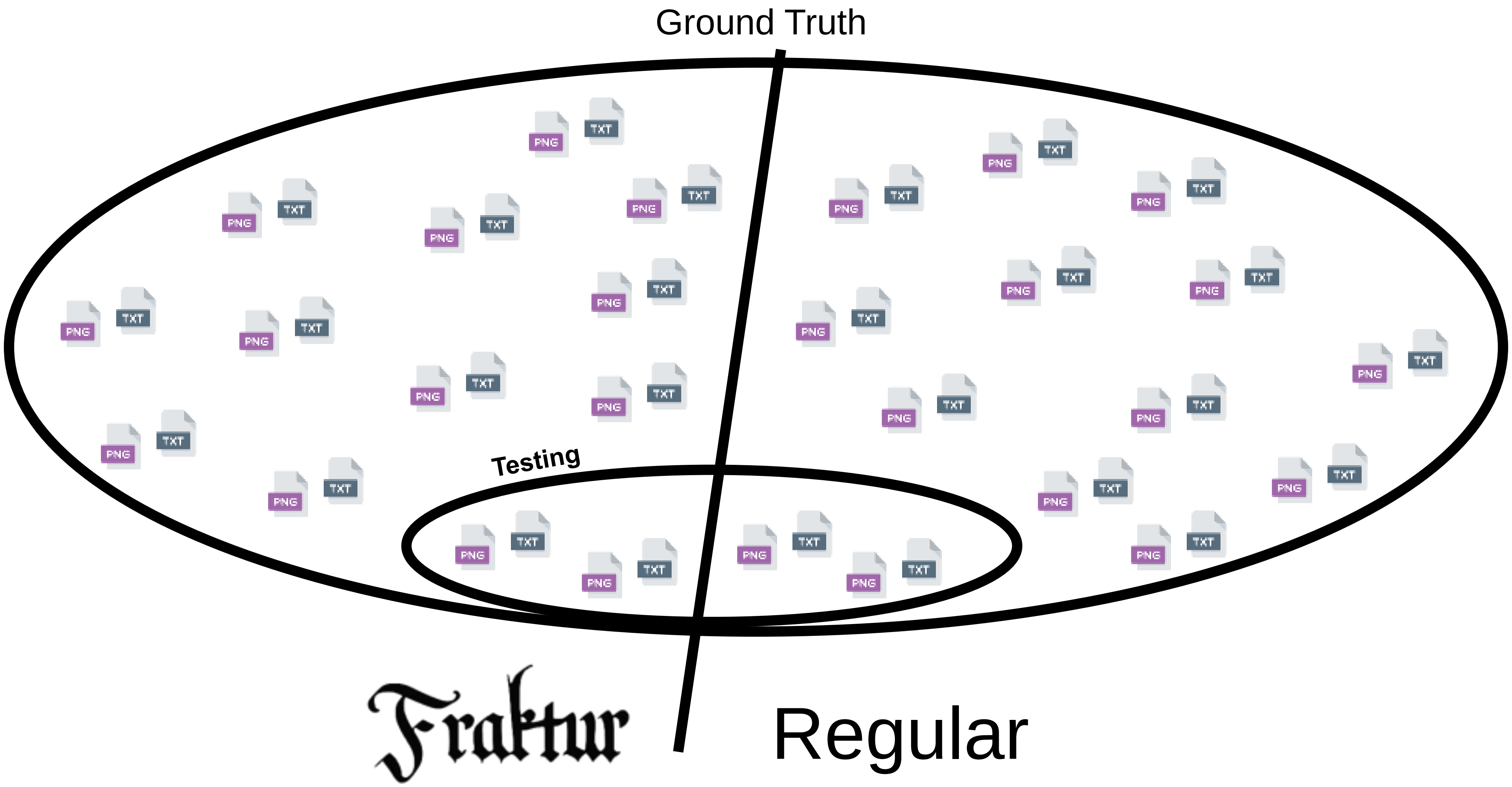
kraken

Tesseract

Calamari

96% Accuracy
on test set

open source software • custom model training

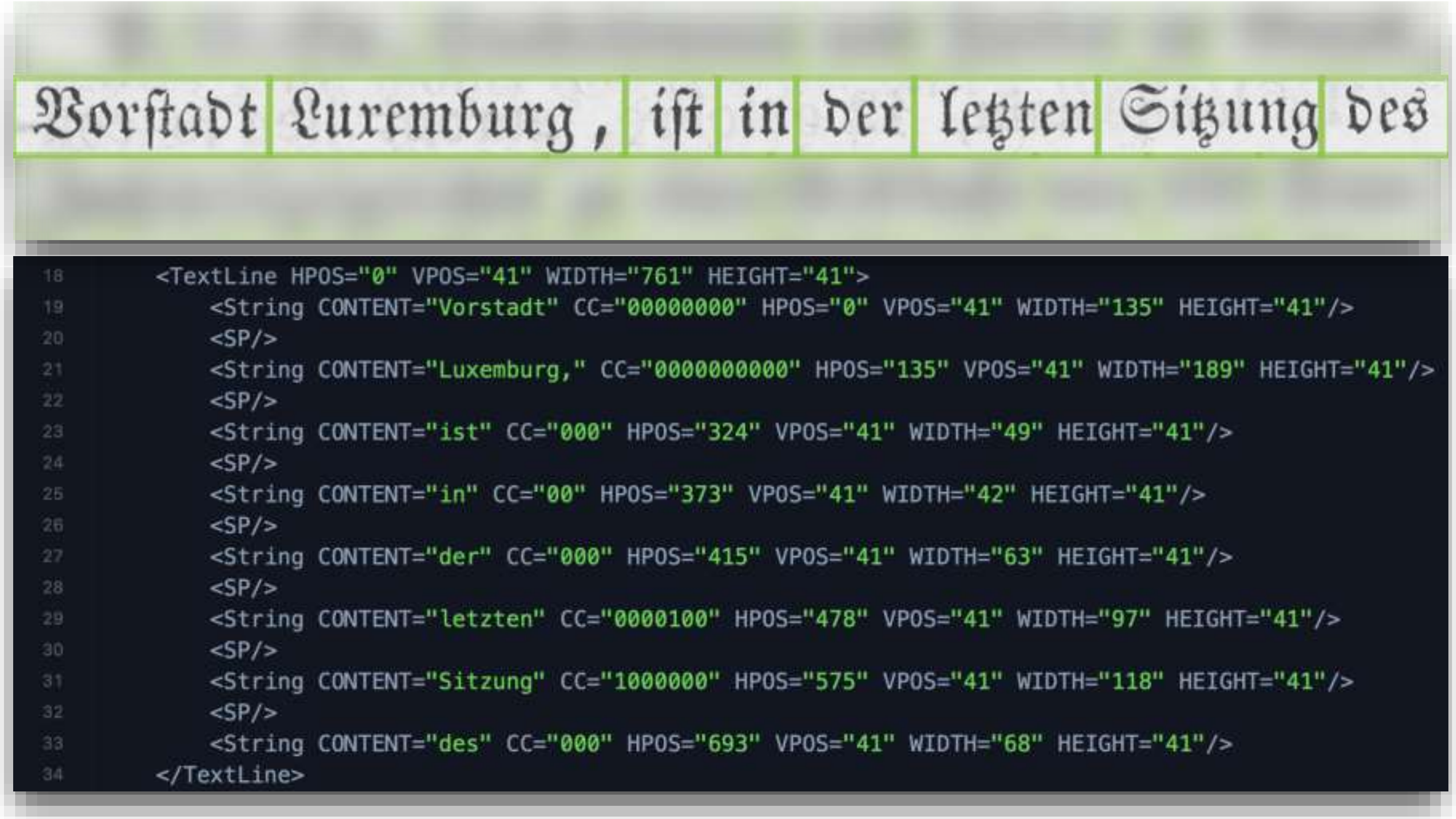


Ground Truth

Testing

Fraktur

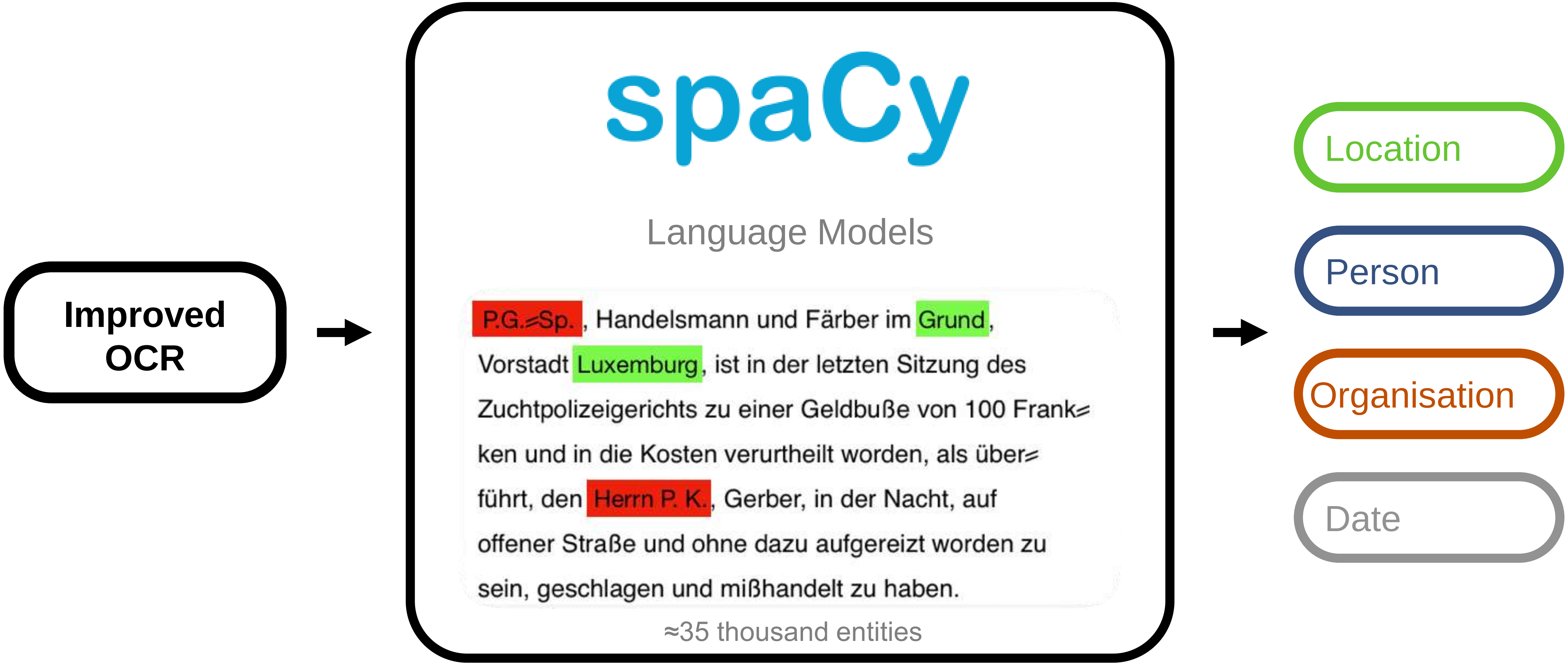
Regular



Vorſtadt Luxemburg, iſt in der letzten Sitzung des

```
<TextLine HPOS="0" VPOS="41" WIDTH="761" HEIGHT="41">
    <String CONTENT="Vorstadt" CC="00000000" HPOS="0" VPOS="41" WIDTH="135" HEIGHT="41"/>
    <SP/>
    <String CONTENT="Luxemburg," CC="0000000000" HPOS="135" VPOS="41" WIDTH="189" HEIGHT="41"/>
    <SP/>
    <String CONTENT="ist" CC="000" HPOS="324" VPOS="41" WIDTH="49" HEIGHT="41"/>
    <SP/>
    <String CONTENT="in" CC="00" HPOS="373" VPOS="41" WIDTH="42" HEIGHT="41"/>
    <SP/>
    <String CONTENT="der" CC="000" HPOS="415" VPOS="41" WIDTH="63" HEIGHT="41"/>
    <SP/>
    <String CONTENT="letzten" CC="0000100" HPOS="478" VPOS="41" WIDTH="97" HEIGHT="41"/>
    <SP/>
    <String CONTENT="Sitzung" CC="1000000" HPOS="575" VPOS="41" WIDTH="118" HEIGHT="41"/>
    <SP/>
    <String CONTENT="des" CC="000" HPOS="693" VPOS="41" WIDTH="68" HEIGHT="41"/>
</TextLine>
```

word bounding boxes • ALTO XML schema



Improved OCR

spaCy

Language Models

P.G.⸗Sp., Handelsmann und Färber im Grund,
Vorstadt Luxemburg, ist in der letzten Sitzung des
Zuchtpolizeigerichts zu einer Geldbuße von 100 Frank⸗
ken und in die Kosten verurtheilt worden, als über⸗
führt, den Herrn P. K., Gerber, in der Nacht, auf
offener Straße und ohne dazu aufgereizt worden zu
sein, geschlagen und mißhandelt zu haben.

≈35 thousand entities

Location

Person

Organisation

Date

ground truth generation • named entity recognition (NER)

Results

1) Improve Optical Character Recognition

Accuracy
Improvement of ~30% text blocks

Production
OCR application on entire corpus

2) Improve Newspaper Exploration

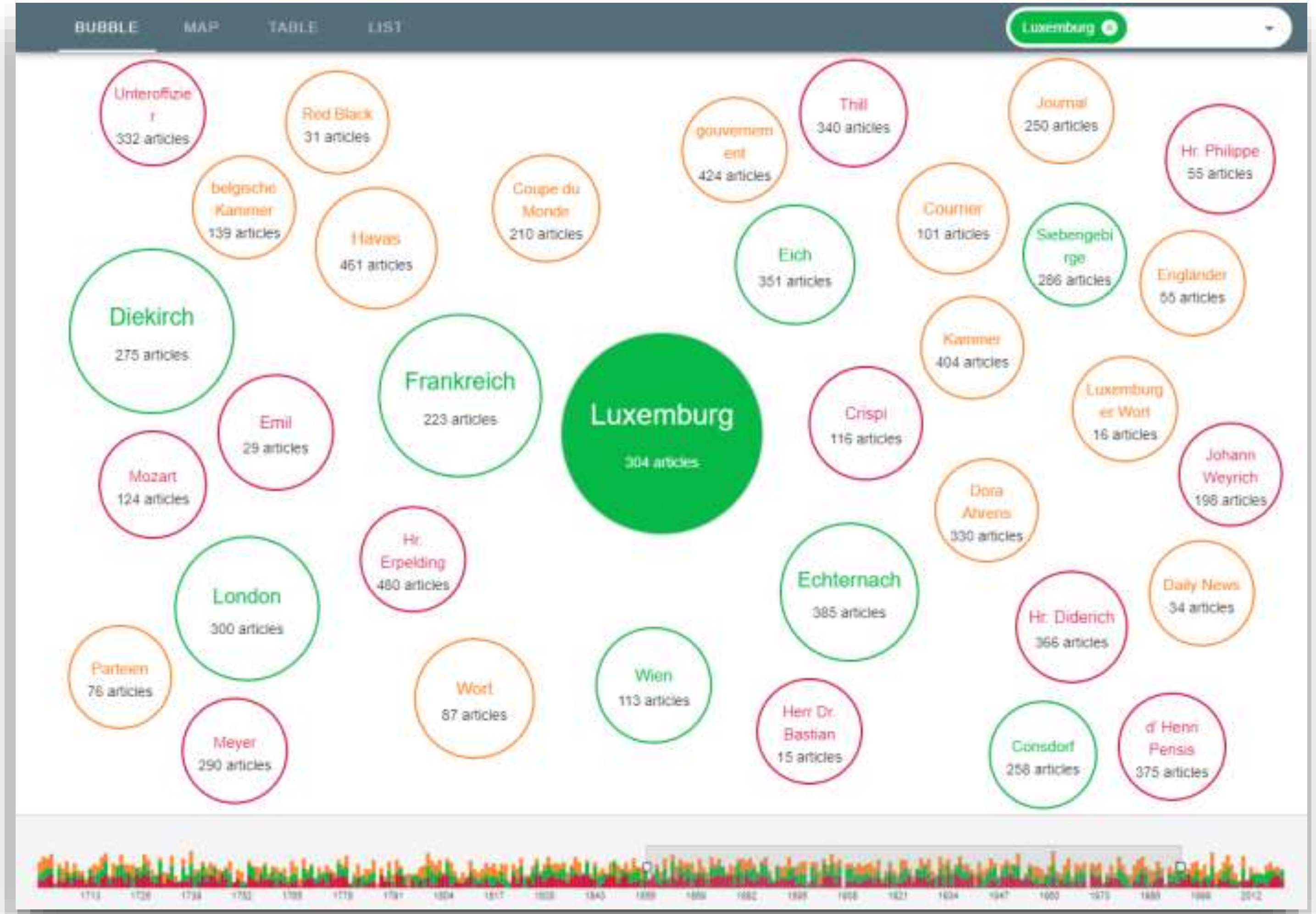
Named Entity Recognition
Finished ground truth - observing first results

New User Interface
Proof of concept

Bibliothèque nationale du Luxembourg
Open Data

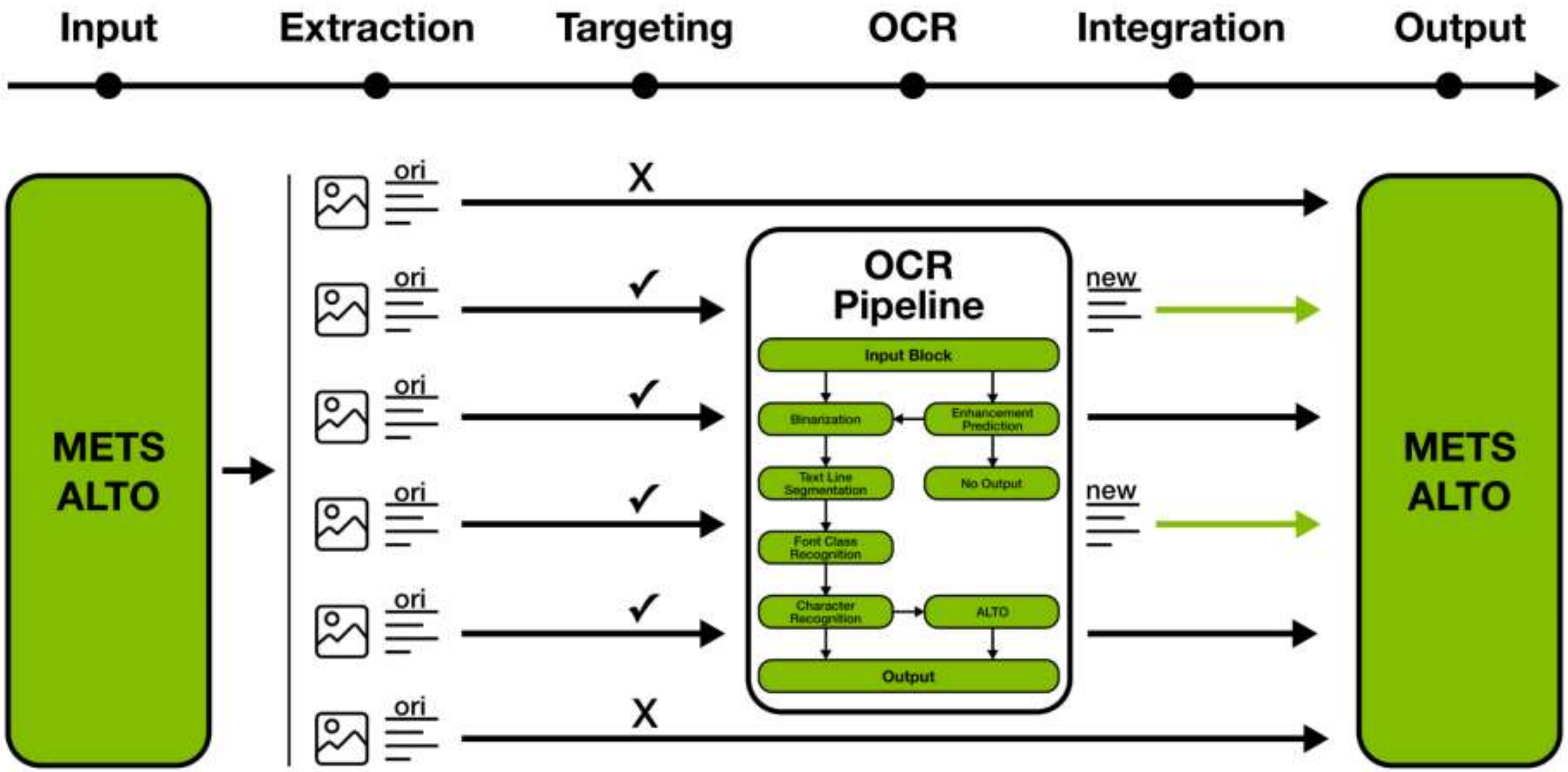
data.public.lu

POC



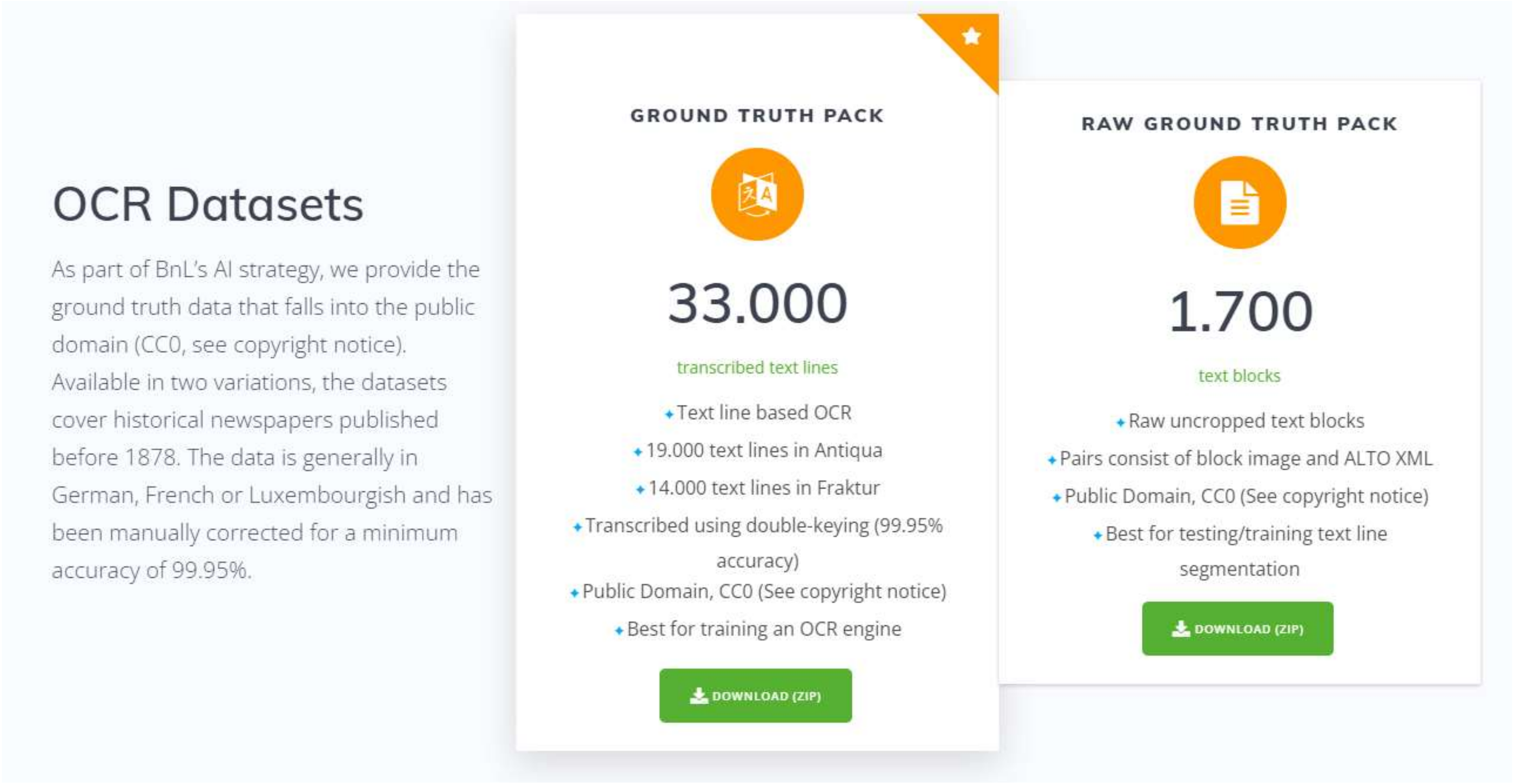
- Named Entities
- Entity Relations
- Timeline
- Maps
- Wikidata

Nautilus-OCR



https://github.com/natliblux/nautilusocr

data.bnl.lu



OCR Datasets

As part of BnL's AI strategy, we provide the ground truth data that falls into the public domain (CC0, see copyright notice). Available in two variations, the datasets cover historical newspapers published before 1878. The data is generally in German, French or Luxembourgish and has been manually corrected for a minimum accuracy of 99.95%.

GROUND TRUTH PACK

33.000

transcribed text lines

- Text line based OCR
- 19.000 text lines in Antiqua
- 14.000 text lines in Fraktur
- Transcribed using double-keying (99.95% accuracy)
- Public Domain, CC0 (See copyright notice)
- Best for training an OCR engine

RAW GROUND TRUTH PACK

1.700

text blocks

- Raw uncropped text blocks
- Pairs consist of block image and ALTO XML
- Public Domain, CC0 (See copyright notice)
- Best for testing/training text line segmentation

Thank you!
Questions?

Ralph Marschall

Ralph.Marschall@bnl.etat.lu

@RalphMarschall