

## Bibliothèque nationale de Luxembourg

# Digitizing newspapers a case study

Carlo Blum - Head of IT Yves Maurer - IT

# **Project Goals**

- Preservation of old newspapers
  - Limit physical access to volumes
  - Replace microfilm with better access method
- Fulltext search
- Decompose images into articles
  - Take advantage of layout information
- Internet-based presentation system
- Use international standards (Mets/Alto)

# ALTO - Detailed OCR data

- Stores coordinates of words in image:
  - → Allows highlighting of search hits
- Groups words into Textblocks:
  - → Allows highlighting of articles
- Grouping Textblocks into articles is done in logical structure of METS

# METS - Complex objects

- Links to external files:
  - Pages in TIFF
  - Pages in PDF
  - ALTO XML
- Metadata about:
  - Newspaper itself (date, publisher...)
  - Articles
  - Images

# **METS - Layout**

- Logical structure describes article tree
- Links to blocks in ALTO OCR data

```
<div ID="DIVL24" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE1" LABEL="Zur Automobilsteuer. " ORDER="1">
<div ID="DIVL25" TYPE="HEADING">
  <div ID="DIVL26" TYPE="PARAGRAPH" ORDER="1">
   <div ID="DIVL27" TYPE="TEXT">
    <fptr>
    | <area FILEID="ALTO00001" BEGIN="BLOCK8"/>
    </fptr>
   </div>
  </div>
</div>
<div ID="DIVL28" TYPE="BODY">
  <div ID="DIVL29" TYPE="BODY CONTENT">
   <div ID="DIVL30" TYPE="PARAGRAPH" ORDER="1">
    <div ID="DIVL31" TYPE="TEXT">
     <fptr>
      <area FILEID="ALTO00001" BEGIN="BLOCK9"/>
     </fptr>
    </div>
   </div>
```

# Digitization workflow

- Public tenders (~150K pages per tender)
- Prepare examples for tender document
- QA on recieved batches

Ruseigen:
(pulvumerando substar)
Srit. 0.25 bie empolitige
Detityelle ob. beren tluam
Reklausen Srit. 0.50, uni.
Schalmadrichen Sr. 1.25
bie öelle, Jahrenmysigen
nni; Ilberenkunft.
Derlas; Gider Druskerel
unit Seftmasserie, Dersentanden.

Paul Schroell, Chi-fily.

### Escher Tageblatt Demokratisches Organ für die Interessen des Kantons Eich

Erscheint wochentäglich

Erfter Jahrgang

Wirffames Anzeigeblatt

#### Zur Minenkonzessionsvorlage.

Dem Kanimerburean gest ums der Betiel der Abuytiffenen gelebert verben foll, armsgeab it Gentreflektion über der Wincenbergeflomerer verein von der Prontinsatiere die Luferm 1882f. – Die Jentralisteins derfielt ans dem Ger-ses Briteauer, Wennen. Wortfins Suff. Kurzen. teuter, Bellet und hemmer. Berichterftatter

Senter Der Berner Berchere einhalten der Seine der Seine

m emfelle.

En millen uns hier, mie er rierigins oor de Ennemen net in de Stendenungen in de Stendenungen der Stendenungen

fich beine Maffinnen über die Comissinheiten die eine einheitliche Zermel bei der Anselderung mit fich bringte zu follährt jedoch, daß die die erreffende Ledingung unt fo ficherer einzehalter werde, als es an genügseb geschulten ingembur-

Cagestundical.

France in a describer.

The Undergrange and Secretaire.

The Undergrange and the Secretaire.

The Undergrange and the Secretaire.

The Undergrange and the Secretaire and State of the Secretaire.

The Undergrange and 

oudieboergaufer in Both de Judinia om, det er gien politiger Begefikung aborte date. Der Konig von Angland light fernet d jut "Ich felliche nich von ganzem derzieh Ob-Werficherungen beraltiger freundsfindt an." Gesono in Hets.

Die Handelsbarmer von Ronipelliet und Randon se hillen die Bertreitrimien der Benger Gibbrucker des einfimmig Befähligkei-träge ungenammen, fin benen fie fich im diehblich auf des geglante framelitäs fernitäte handelsaß-iorunen anfaltet.

#### Die zweite Frau.

mevoringe auch noch gute Dienfte lelften könne.

Including and the series of th

tiede meine denementer wer eine die eine der Verlegerichte geber der Verlegerichte der Ver

# Digitization - QA

- Receive batches of 500GB 1TB
- Quick, thorough quality control needed
- Custom developed software for
  - Integrity (METS, ALTO, TIFF)
  - Tender conformance (naming schemes etc.)
  - Sampling of segmentation, title correction
  - Reasonableness
  - Completeness

## Infrastructure

- Sun v490 2 dual core Sparc 1.5 20GB RAM
- SAN duplicated on two physical sites:
  - 250 GB IBM ES8300 storage for Database
  - Sun StorageTek (12TB) 6140 (8 +1)
- ITSM 3584 Tape Backup

→ Migration to Solaris x86 in future

# Demo: eluxemburgensia

### Content:

- 100,000 pages / 800,000 articles
- 15500 issues

### Focus on:

- Different ways to acces newspapers
- New developments (Exlibris / BnL)

http://www.eluxemburgensia.lu

※ ■

HOME

**ABOUT US** 

CONTACT

COPYRIGHT



Previous searches

Advanced search



Find an issue



Please note: from May 10, 1940 to September 10, 1944 the national press was controlled by the Nazi



Search



- ◆ Luxemburger Wort
- Tageblatt
- \* More



In the spotlight

#### Findel

21st July 1946: Opening of the airport Luxembourg-Findel.



### Luxemburgs Flughafen

"Miches ist gewaltiger als der Mensch" einem Gefälle von 1.3 Grad die Bahn, mit unter dem Erdgelschaft die Garage, stang der griechliche Dichter Sopholder, dienn gerähet auf dem Mose in gefüngelten, sauscheiden Häusern, die Lauft in gefüggetten Häusern, die Lauft in Lauft in Lauft in gefüggetten Häusern, die Lauft in L

#### Retrospective Robespierre

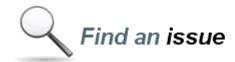
28 July 1944: Retrospective of the notorious revolutionary on the 150th anniversary of his execution.

View the issue

### Robespierre wird hingerichtet

Das Ende des "Schlächters" der französischen Revolution

Home | About us | Contact | Copyright @ 2009 by BnL



### By Date





Мо	Tu	We	Th	Fr	Sa	Su
						1
	<u>3</u>	4	<u>5</u>	<u>6</u>		8
<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>	<u>14</u>	15
<u>16</u>	<u>17</u>	<u>18</u>	<u>19</u>	<u>20</u>	<u>21</u>	22
<u>23</u>	24	<u>25</u>	26	<u>27</u>	<u>28</u>	29
-	0.4					

Please note: from May 10, 1940 to September 10, 1944 the national press was controlled by the Nazi occupier.

### By Newspaper



- Luxemburger Wort
- Tageblatt
- ◆ More ...





#### Proklamierung des westdeutschen Grundgesetzes

Der Parlamentarische Rat aufgelöst — Die Wahlvorbereitung soll demnächst beginnen

le Mri augulossi — Die Wahlworbereitung soll Bonn, 24. Mis. Die (eiersches Schiultstrung des Parlamentarischen Betes in 
orr dus Grundingstett Dr. Wesdundsteinschaft der Leiner der Leiner des 
nachmitigs kunz nach ist. Die 1. Anweise 
senheit der Generalle Haus (Vir. Slusten). Me Lean (Grochertamien) und 
preindenten und der Lendingspracionsten der 11 westdoctschen Länder. 
Präsidert Andernier wiest in einer Andieser Strang his, die "eine neue Acts 
der deutschen Coestlekte ernighet. 
Nachden er betatt haut, daß alle Linzuch lakten jelfende er einen deringerdier. 
Appell an die allierten Behörden ongen
zuch beiten jelfende er einen deringerdier. 
Appell an die allierten Behörden ongen
seinen, die die demokratische Entweck
lung Doutschland zu gelfürden debonen. 

Die Leinerseichnung.

Die Unterzeichnung

Die Unferreichnung
Die Abgeordneten wurden dem in
The Abgeordneten wurden dem in
The State Rechterlage aufgeruffen
The State Rechterlage aufgeruffen
ternet unter das Grundgestet. Nur die
seient Komministen Bemann und Rechter Leiten der Entsteheit in der State
Beiden Komministen Bemann und Rechter leiten die Untziehrlich in die State
Beiden Komministen Bemann und Rechter
The Vertreiter Weisberling, die als BeDie Abgehaften des Prafaments
des Bestehe des Prafaments
des Bestehe der Arteilen des Prafaments
des Bestehe des Grundgesteit geliefelblides Rechtellich vom den Ministerprafadere und der Unterzeitwang verfallender
Prafation Ademanne Feortien die Beitreitselaung der previserischen Verfanung
de "einen Beliefes ger Vereinstellichung
taug absolitieren die Bildigere die
Beantien Lieuter der Verlagen des Bildiger
Beantien answeretrauser Ferner Bildiger
Germins viewer Kommission en dieser Sitzing einen
Germins viewer Kommission en dieser Sitzing einen
Germins viewer Kommission en dieser Sitzing einen
Germins viewer den dieser Sitzing einen
Germins viewer Kommission en dieser Sitzing einen
Germins viewer den den den Schall aus S. 73

(Portsetruit, von Stell 1)

(Portsetruit, von Stell 1)

(Schliß auf S. 7.)

(Fortsetzung von Selfs 1)

Ausschuß, der damit besetzung wurde,

äuf eventuelle Fresen der wer Außerminister über des deutsche Problem zu
antworfen.

Der Parlamentarische Rat ist damit

den wurden um das Wahlgesetz des par-lamentarischen Bats zu prüfen. Da Rein vollständiges Einvernehmen erzielt wer-den könnte, unterbreiteiten als dies ihren Regierungen.

den Route, umer-teene al dies richt der Dernach hat eine sich ist Peris, Lier den und Wachtschen pepperwickig mit den und Wachtschen pepperwickig mit den Meistungswerzichtenen den Ansen, welche im Frankfurst ab Schwiereiten Stellen und der Schwiesen de

eluxemburgensia

HOME

ABOUT US

CONTACT

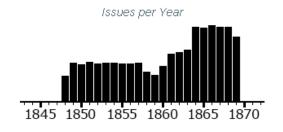
COPYRIGHT

Other newspapers

### Luxemburger Wort

Title(s) Luxemburger Wort Publisher i saint paul luxembourg S.A. Years published ! 1848/03/23 -Years scanned : 1848/03/23 - 1869/12/31

Link to bibnet.lu catalogue | Holding information



Search in this newspaper





#### Newspaper information

#### Daily founded in 1848

#### Publisher:

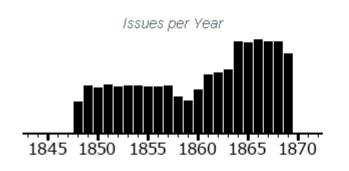
saint paul S.A. 2, rue Christophe Plantin L-2988 Luxembourg Tel: +352 49 93 - 1

Fax: +352 49 93 - 726 Website: www.wort.lu

Please note: The Nazi occupation of Luxembourg began on May 10, 1940 and ended on September 10, 1944. In the beginning, the press was censored but later the occupier confiscated the publishing houses and printing offices, gaining complete control over the publications, which kept their pre-war titles.

Home | About us | Contact | Copyright @ 2009 by BnL

### Find an issue



Search

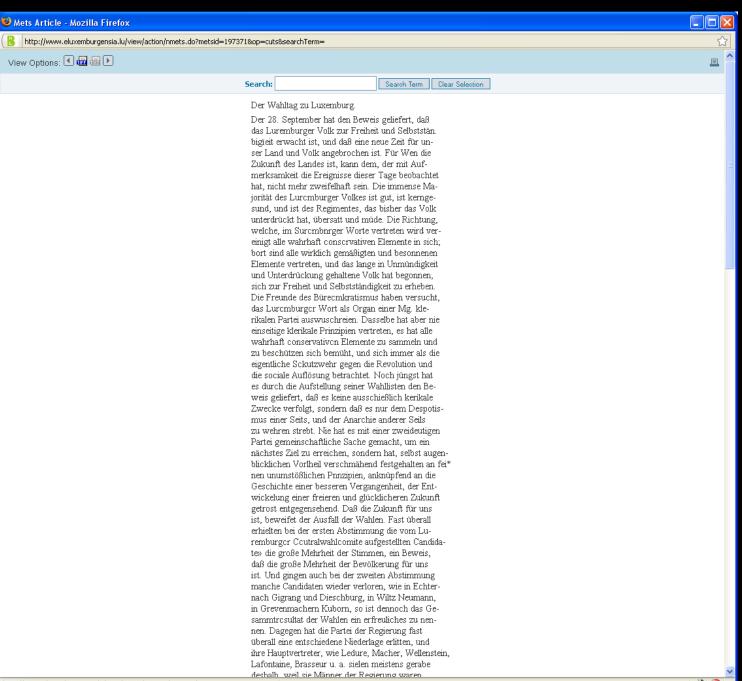






Der Wahltog ju Lugemburg.

Der 28. September hat ben Beweis geliefert, bag bas Luremburger Bolf gur Freiheit und Gelbftffanbigfeit erwacht ift, und daß eine neue Zeit für un= fer land und Bolf angebrochen ift. Für Ben bie Bufunft bes landes ift, fann bem, ber mit Aufmerffamfeit bie Ereigniffe biefer Tage beobachtet hat, nicht mehr zweifelhaft fein. Die immense Da= joritat bes Luxemburger Bolfes ift gut, ift fernge= fund, und ift bes Regimentes, bas bisher bas Bolf unterbrudt hat, überfatt und mube. Die Richtung, welche im Luremburger Worte vertreten wird; vereinigt alle wahrhaft conservativen Elemente in sich; bort find alle wirklich gemäßigten und besonnenen Elemente vertreten, und bas lange in Unmunbigfeit und Unterbrückung gehaltene Bolf hat begonnen, fich jur Freiheit und Gelbftftandigfeit zu erheben. Die Freunde bes Bureaufratismus haben verfucht, bas Luxemburger Wort als Drgan einer Ig. flerifalen Partei auszuschreien. Dasselbe hat aber nie einseitige flerifale Prinzivien vertreten, es bat alle





ACCUEIL A PROPOS

CONTACT DROITS



Todler

Recherche



🚖 Recherches précédentes



HOME

ABOUT US

CONTACT

COPYRIGHT

7 matches for "todler"



Table	e view 🔻 Recon	d view		Page 1 of 1	Sort by Ranking 💌
#	Ranking	Title	Date	Snippet	Object
1		<u>Luxemburger Wort</u>	1853/09/18	es wie wenn in jedem einzelnen Hause ein <b>Todler</b> wäre Alle brannten sich zu feiner Leich	<b>■ ■ </b>
2	-	<u>Luxemburger Wort</u>	1864/05/04	Gott wurde früher nicht feiten als ein todler ©Ott betrachtet oder vielmehr als ein Got	
3		<u>Luxemburger Wort</u>	1854/09/29	wozu als Mittel die Einziehung aller Güter todler Hand bezeichnet wird 9 Aufmunterung der Anlegun	🛅 🏗 💾 🤣
4	-	Escher Tageblatt	1944/09/30	Verwandten EscÄ Uelzecht Woltz Differdeng an Todler D Begriefnes fönd statt zu E s c h U e	<b>6 8</b> 4 9
5	-	Escher Tageblatt	1915/06/24	frühern von dem schweizerischen Hauptmann Todler und einem Leutnant Luxemburg 21 «Juni Wanderfahr	
6		Escher Tageblatt	1937/07/05	dem nördlichen Teil des Landes nieder Aus T <b>odler</b> Ringler und Umgegend werden große Schäde	
7		<u>Luxemburger Wort</u>	1855/05/04	Gesetze verkäuflich erklärten Gütern in t <b>odle</b> r Hand Alle übrigen Bestimmungen des Gesetz	🛅 🌃 💾 🧇

Displaying records 1 to 7 of 7

Home | About us | Contact | Copyright

HOME

**ABOUT US** 

CONTACT

COPYRIGHT

### 7 matches for "todler"



Table view

Record view ▼



Result 3 | 4 | 5 | 6 of 7



Sort by Ranking 🗸

### Escher Tageblatt 1937/07/05







Object

Object

Tageblatt - PDF Document (45 M)

000092338

MEWSPAPER: Tageblatt 1937-07-05 01 - Mets (147 K)

Record number

Title

**Escher Tageblatt** 

Publisher

Editpress S.A.

Scanned by

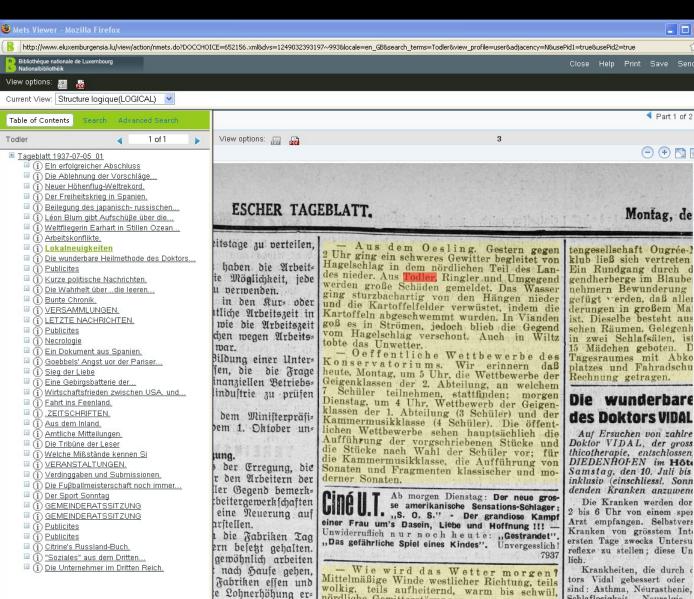
Bibliothèque nationale de Luxembourg

Date

1937/07/05

Snippet

dem nördlichen Teil des Landes nieder Aus Todler Ringler und Umgegend werden große Schäde



Serald Times" mel=

feure von Philadel=

find. 25 000 Chauf=

nördliche Gewitterstörungen. Mit großer Ungeduid erwartet!

Schifflingen, 5. Juli. Aus dem Gemeinderat. Am Samstag Nachmittag fand hier eine Gemeinderatssitzung statt, die aber ohne n des Gemüsemarks einen einzigen Artikal der Tagasardnung zu

15 Mädchen geboten. D Tagesraumes mit Abko platzes und Fahradschu Rechnung getragen.

Part 1 of 2

(-) (+) D

#### Die wunderbare des Doktors VIDAL

Auf Ersuchen von zahlre Doktor VIDAL, der gross thicotherapie, entschlossen DIEDENHOFEN im Hôte Samstag, den 10. Juli bis inklusiv (einschliessl. Sonn denden Kranken anzuwene

Die Kranken werden dor 2 bis 6 Uhr von einem spez Arzt empfangen. Selbstvers Kranken von grösstem Inte ersten Tage zwecks Untersu reflexe zu stellen; diese Un

Krankheiten, die durch tors Vidal gebessert oder sind: Asthma, Neurasthenie, Schlaflosigkeit, Neuralgie, störungen, Störungen des Ne erscheinungen des kritisch Lähmungen.

Die Heilmethode besteht chen Schmerz das sympathis stimmten Punkten der Nase diese Anregung ruft die wol .

1 1



eluxemburgensia beta

洲 ■

Mathias MONGENAST Contains Exact Search terms Date from Date to mm 🗸 dd 🗸 ууууу Restrict search to selected titles Luxemburger Wort Tageblatt Restrict your search to Article headings Image captions Article body Advertisements Obituaries

Home | About us | Contact | Copyright

器 🔳 🚎

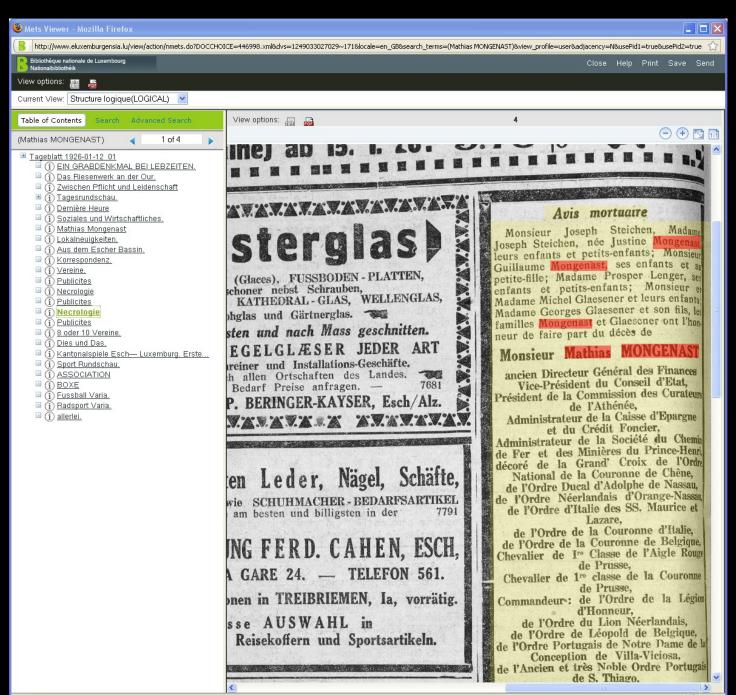
CONTAC

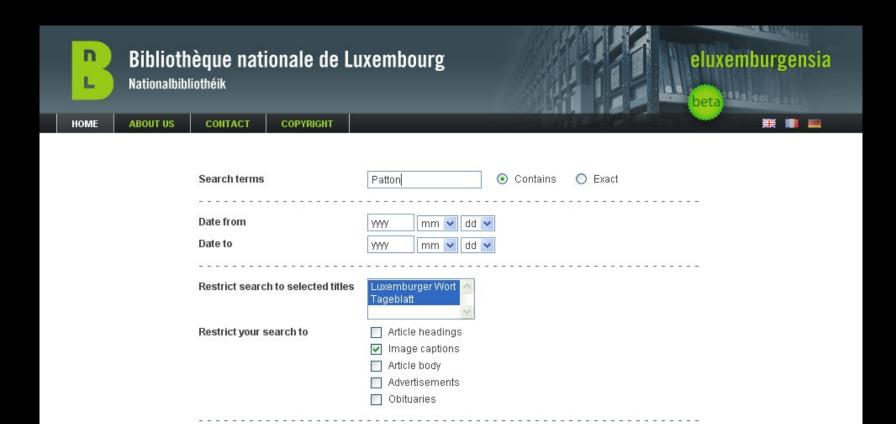
COPYRIGHT

7 matches for "( mathias mongenast )"



Table	view ▼ Record	view		Page 1 of 1	Sort by Ranking 💌
#	Ranking	Title	Date	Snippet	Object
1		Escher Tageblatt	1923/05/15	allgemeinen Kenntnis gebracht dass 1 d Bergmann <b>Mathias</b> Gorges wohnhaft in Fell vorher in Esch Alz	
2		Escher Tageblatt	1925/06/20	alt ehelos aus Bonneweg Peter alt Peter Mathias 11 Maria Sophie alt ohne Stand ehelos aus	POP 🚻 💾 🧐
3		Escher Tageblatt	1925/06/23	Grund 2 Rudolph I Grund 3 Viktor Feiten Mathias Kremer Jean I Kuhn Jean Karl 49 8 Kirchen	POP 🚻 💾 🧐
4		Escher Tageblatt	1926/01/11	Gestern starb im Alter von 83 Jahren Herr Mathias Mo genast langjähriger General Direktor der	
5		Escher Tageblatt	1926/01/12	familles <b>Mongenast</b> et Glaesener ont I hon neur de faire part du décès de Monsieur <b>Mathias</b> <b>MONGENAST</b> ancien Directeu	
6		Escher Tageblatt	1926/02/23	qui leur ont été à du décès de Monsieur <b>Mathias MONGENAST</b> de la part de leurs amis et connaissances	
7		Escher Tageblatt	1933/07/15	Prüfung Folgende haben bestanden Antony <b>Mathias</b> Rümelingen Bastian Georges Lux Rauler Marcel	





Home | About us | Contact | Copyright

CONTACT

COPYRIGHT

### 4 matches for "( patton )"



Table	e view ▼ Record	d view		Page 1 of 1	Sort by Ranking 💌
#	Ranking	Title	Date	Snippet	Object
1		Escher Tageblatt	1945/12/10	General George S PATTON stehend an der Westfront 1944	🔐 🖺 🤣
2		Escher Tageblatt	1945/12/28	Der Sarg von General <b>Patton</b> auf dem amerikanischen Soldatenfriedhof in	
3		Escher Tageblatt	1948/08/07	touristischen Briefmarken <b>PATTON</b> s REVANCHE Im 200 nahm Me <b>L Patton</b> Revanche für seine Niederlage übe	
4	-	Escher Tageblatt	1944/09/16	E shake hand dem <b>Patton</b> an dem Felix An der Mött de Diderich Photo	🗃 🌃 💾 🤣
				<del>-</del>	



E shake-hand töschent dem Generol Patton an dem Prönz Felix An der Mött

de Burgermêschter Diderich

gen neute unter russischem Arthierefeuer vom Ostufer der Weichsel. Die
Deutschen führen Zerstörungen durch
und haben zwei Weichselbrücken gesprengt. Am Westufer der Weichsel wird
schwer gekämpft. Hier versuchen die
Deutschen ihre Stellungen zu erweitern,
um dem russischen Angriff Widerstand
leisten zu können. Russische Flugzeuge
warfen für die in Warschau kämpfenden polnischen Befreiungstruppen Waffen und Lebensmittel ab. Das polnische
Befreiungskommuniqué richtete im Sender Lublin folgenden Aufruf an die in
Warschau kämpfenden Befreiungstruppen: Die bevorsfehende Befreiung Warschaus ist nahe. Die ersten polnischen
Truppen sind in Praga eingerückt, im
Verband mit der roten Armee. An den
Ufern der Weichsel wird die Entscheidungsschlacht geschlagen. Hilfe ist nahe,
der Sieg ist nahe! Kämpft weiter! Praga
soll, Meldungen zufolge, ebenso verwüsteine einzige Trümmerstätte. Es war die

stärkste Riegelstellung, die die Deutschen

im Osten errichtet hatten.

Photo Rouster

\* 1

# Non-visible changes

### Perfomance tuning:

- Harvest: 7 days -> 1 day (fix by Exlibris)
- Image conversion: 10s -> 1s
   (replacement of ImageMagick)

### **Additions:**

Scripts for batch loading of METS

# Why Digitool?

- Open, flexible system which allows diverse content types
- Small team, 1.5 FTE, Aleph experience
  - → small learning curve
- Well-defined update cycle
- At start of project (2006), Exlibris promised to implement the features we wanted.

### **Current Limitations**

- Cannot use logical bases
  - Newspaper functionality is on or off per silo
- RD Search not quite there yet:
  - 1 hit per newspaper (instead of 1 per article)
  - Simplistic ranking (only frequency)
  - No article title for hits
- METS Viewer not customizable enough
- Missing direct links to articles or pages

# Limitations (2)

- Many OCR errors
- Maybe correct entire Table of Contents
- Tenders, but preparation and control is work intensive:
  - High level of quality expected
  - Library wants completeness

# **Future**

- Better search engine (Lucene)
- New Viewer
- Ingest existing content:
  - 24 titles with 55,000 pages
- Add 200,000 pages in next year
- Other content: postcards, maps, CDs
- Aleph ⇔ Digitool synchronisation
- Persistent identifiers

# Questions?

# Thank you

www.eluxemburgensia.lu

- Carlo Blum
   carlo.blum@ctie.etat.lu
- Yves Maurer
   yves.maurer@bnl.etat.lu