ALTO, PAGE, & Co.
Formate für Volltexte

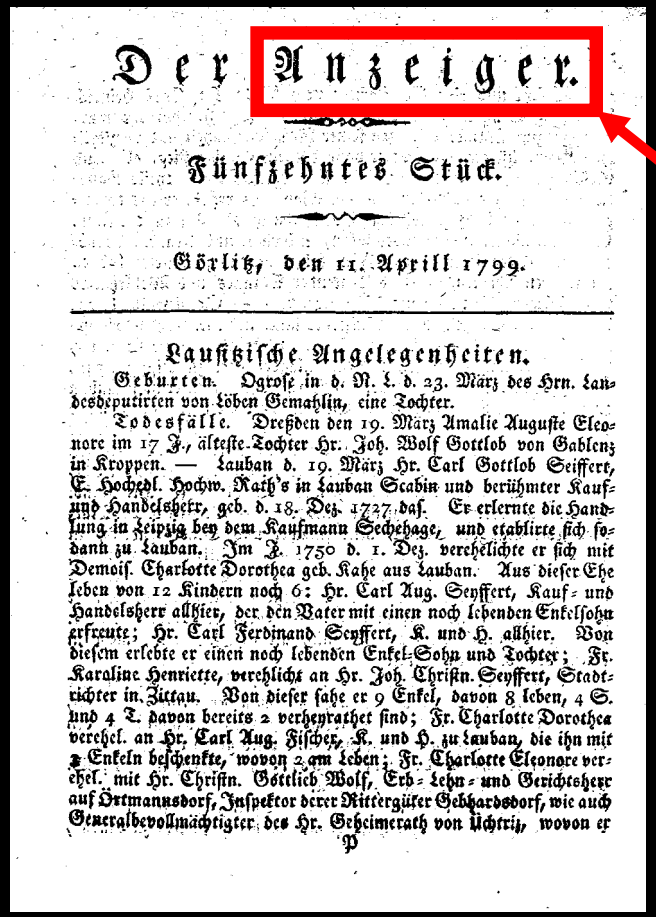
Clemens Neudecker (@cneudecker)

DINI AG KIM Workshop `21

27.04.2021

Wozu OCR?

- Volltextindexierung für Suche
- Trefferanzeige im Viewer mit Highlighting (Koordinaten!)
- Metadatenanreicherung
- Erstellung von Ebooks
- Text- und Datenmining, Digital Humanities, usw.



Der Anzeiger.

Fünfzehntes Stück.

Görlitz, den 11. April 1799.

Lausitzische Angelegenheiten.

Geburten. Ogrose in d. N. L. d. 23. März des Hrn. Landesdeputirten von Löben Gemahlin, eine Tochter.

Todesfälle. Dreßden den 19. März Amalie Auguste Eleonore im 17 J., älteste Tochter Hr. Joh. Wolf Gottlob von Gablenz in Kroppen. — Lauban d. 19. März Hr. Carl Gottlob Seiffert, E. Hochedl. Hochw. Rath's in Lauban Scabin und berühmter Kauf- und Handelsherr, geb. d. 18. Dez. 1727 daf. Er erlernte die Handlung in Leipzig bey dem Kaufmann Sechehage, und etablirte sich sodann zu Lauban. Im J. 1750 d. 1. Dez. verehelichte er sich mit Demois. Charlotte Dorothea geb. Kahe aus Lauban. Aus dieser Ehe leben von 12 Kindern noch 6: Hr. Carl Aug. Seyffert, Kauf- und Handelsherr allhier, der den Vater mit einen noch lebenden Enkelsohn erfreute; Hr. Carl Ferdinand Seyffert, K. und H. allhier. Von diesem erlebte er einen noch lebenden Enkel-Sohn und Tochter; Fr. Karoline Henriette, verehlicht an Hr. Joh. Christn. Seyffert, Stadtrichter in Zittau. Von dieser sahe er 9 Enkel, davon 8 leben, 4 S. und 4 T. davon bereits 2 verheyrathet sind; Fr. Charlotte Dorothea verehel. an Hr. Carl Aug. Fischer, K. und H. zu Lauban, die ihn mit 2 Enkeln beschenkte, wovon 2 am Leben; Fr. Charlotte Eleonore verehel. mit Hr. Christn. Göttlieb Wolf, Erb- Lehn- und Gerichtsherr auf Örtmannsdorf, Inspektor derer Rittergüter Gebhardsdorf, wie auch Generalbevollmächtigter des Hr. Geheimerath von Uchtritz, wovon er

P

```
Der Anzeiger.
Fünfzehntes Stü.
Todesfälle. Dreßden den 19.
März Amalie Auguſte Eleo=
nore im 17 J., älteſte
To.ter Hr. Joh. Wolf
Gottlob von Gablenz
in Kroppen. – Lauban
d. 19. März Hr. Carl
Gottlob Seiffert,
E. Ho.edl. Ho.w.

[...]
```

ALTO

- ALTO = Analyzed Layout and Text Object
- Entwickelt 2004 im EU-Projekt „Metadata Engine“ (METAe)
  web.archive.org/web/20160318002552/http://meta-e.aib.uni-linz.ac.at/
- 2009: Standardisierung durch Library of Congress
- XML-Schema basiert
- Wird meist in Kombination mit METS verwendet

  loc.gov/standards/alto/

  github.com/altoxml
- Aktuelle Version: `4.2` (Released 2020-07-14)

ALTO

- Vom DFG-Viewer unterstützter Standard für Volltexte
- Seitenbasiert, d.h. 1x Seite = 1x ALTO Datei
- ALTO erfordert `<String>`, daher ist eine Wortsegmentierung bei der OCR immer zwingend erforderlich!
- OCR Software mit ALTO Unterstützung:
  - Tesseract (ab Version 4)
  - Kraken
  - ABBYY FineReader (Server/Engine) – aber nur bis ALTO-Version 3.1

```
<Page ID="Page1" PHYSICAL_IMG_NR="1" HEIGHT="2774" WIDTH="1951">
    <TopMargin HEIGHT="104" WIDTH="1951" VPOS="0" HPOS="0"></TopMargin>
    <LeftMargin HEIGHT="2501" WIDTH="167" VPOS="104" HPOS="0"></LeftMargin>
    <RightMargin HEIGHT="2501" WIDTH="83" VPOS="104" HPOS="1868"></RightMargin>
    <BottomMargin HEIGHT="169" WIDTH="1951" VPOS="2605" HPOS="0"></BottomMargin>
    <PrintSpace HEIGHT="2501" WIDTH="1701" VPOS="104" HPOS="167">
        <ComposedBlock ID="Page1_Block1" HEIGHT="447" WIDTH="1464" VPOS="104" HPOS="296" TYPE="container">
            <Shape>
                <Polygon POINTS="298,110 1760,110 1760,552 298,552 298,110"/>
            </Shape>
            <TextBlock ID="Page1_Block2" HEIGHT="169" WIDTH="1463" VPOS="104" HPOS="297" language="de" STYLEREFS="font5">
                <Shape>
                    <Polygon POINTS="298,110 1760,110 1760,274 298,274 298,110"/>
                </Shape>
                <TextLine HEIGHT="153" WIDTH="1429" VPOS="116" HPOS="313">
                    <String WC="0.3000000119" CONTENT="D" HEIGHT="125" WIDTH="125" VPOS="116" HPOS="313"/>
                    <SP WIDTH="71" VPOS="117" HPOS="438"/>
                    <String WC="0.3600000143" CONTENT="c" HEIGHT="92" WIDTH="40" VPOS="145" HPOS="509"/>
                    <SP WIDTH="70" VPOS="145" HPOS="550"/>
                    <String WC="0.3400000036" CONTENT="r" HEIGHT="88" WIDTH="48" VPOS="145" HPOS="621"/>
                    <SP WIDTH="143" VPOS="117" HPOS="670"/>
                    <String WC="0.3799999952" CONTENT="An " HEIGHT="137" WIDTH="329" VPOS="118" HPOS="813"/>
                    <SP WIDTH="70" VPOS="139" HPOS="1142"/>
                    <String WC="0.4900000095" CONTENT="c" HEIGHT="92" WIDTH="44" VPOS="147" HPOS="1213"/>
                    <SP WIDTH="63" VPOS="123" HPOS="1258"/>
                    <String WC="0.2399999946" CONTENT="t" HEIGHT="120" WIDTH="37" VPOS="124" HPOS="1321"/>
                    <SP WIDTH="71" VPOS="124" HPOS="1358"/>
                    <String WC="0.2599999905" CONTENT="g" HEIGHT="124" WIDTH="60" VPOS="144" HPOS="1429"/>
                    <SP WIDTH="66" VPOS="144" HPOS="1490"/>
                    <String WC="0.2599999905" CONTENT="k" HEIGHT="96" WIDTH="44" VPOS="145" HPOS="1557"/>
                    <SP WIDTH="66" VPOS="145" HPOS="1602"/>
                    <String WC="0.6549999714" CONTENT="r." HEIGHT="96" WIDTH="72" VPOS="149" HPOS="1669"/>
                </TextLine>
            </TextBlock>
```

PAGE

- PAGE = Page Analysis and Ground Truth Environment
- Entwickelt von PRImA (Pattern Recognition and Image Analysis Lab), Universität Salford, Greater Manchester, UK (primaresearch.org)
- De-facto Standard für Ground Truth
- XML-Schema basiert
- Seitenbasiert, d.h. 1x Seite = 1x PAGE Datei
- Verwendung in wissenschaftlichen Publikationen und Wettbewerben

  github.com/PRImA-Research-Lab/PAGE-XML
- Letzte Version: `2019-07-15` (Released: 2019-07-15)

PAGE

- Granulare Elemente/Attribute für bspw. Regionentypen
- PAGE-spezifische Konzepte:
  - Reading Order
  - Layers (z-Level)
  - AlternativeImage

- OCR Software mit PAGE Unterstützung:
  - OCR-D
  - OCR4all
  - Transkribus
  - Aletheia

```
<TextRegion id="r7" type="heading">
    <Coords points="343,928 705,928 705,930 835,930 835,931 1000,931
        1000,932 1256,932 1256,933 1593,933 1593,932 1649,932 1649,933
        1708,933 1708,934 1853,934 1853,944 1708,944 1708,942 1256,942
        1256,941 1000,941 1000,940 870,940 870,941 835,941 835,940 705,940
        705,938 344,938 344,939 272,939 272,940 225,940 225,941 190,941
        190,940 188,940 188,939 187,939 187,937 184,937 184,933 189,933
        189,932 207,932 207,931 224,931 224,930 272,930 272,929 343,929"/>
    <TextEquiv>
        <Unicode>Der Anzeiger.</Unicode>
    </TextEquiv>
</TextRegion>
```

hOCR

- hOCR = Google OCR Format
- Ursprünglich entwickelt von Thomas Breuel für die OCR Software OCRopus | ocropy
- HTML-basiert
- Zumeist 1x hOCR Datei für gesamtes Dokument

  kba.cloud/hocr-spec/1.2/
- Letzte Version: `1.2` (Released: 2020-02-06)

hOCR

- OCR Software mit hOCR Unterstützung:
  - OCRopus | ocropy
  - Kraken
  - Tesseract

```
<div class='ocr_page' id='page_1' title='image "T:\ENP_GT\00675515.tif";
  bbox 0 0 1951 2774; ppageno 0'>
 <div class='ocr_carea' id='block_1_1' title="bbox 185 19 1924 482">
  <p class='ocr_par' id='par_1_1' lang='Fraktur' title="bbox 185 19 1940 482">
   <span class='ocr_line' id='line_1_1' title="bbox 185 19 1924 482; baseline 0.009 -245.101;
    x_size 139.41451; x_descenders 16.414507; x_ascenders 27">
   <span class='ocrx_word' id='word_1_1' title='bbox 185 19 532 338; x_wconf 84'>Di</span>
   <span class='ocrx_word' id='word_1_2' title='bbox 471 56 595 341; x_wconf 55'>e</em></span>
   <span class='ocrx_word' id='word_1_3' title='bbox 587 50 757 325; x_wconf 79'>r</span>
   <span class='ocrx_word' id='word_1_4' title='bbox 730 27 956 325; x_wconf 85'>A</span>
   <span class='ocrx_word' id='word_1_5' title='bbox 936 49 1077 326; x_wconf 76'>u</span>
   <span class='ocrx_word' id='word_1_6' title='bbox 1013 35 1161 330; x_wconf 6'>E</span>
   <span class='ocrx_word' id='word_1_7' title='bbox 1184 50 1276 320; x_wconf 56'>e</span>
   <span class='ocrx_word' id='word_1_8' title='bbox 1286 86 1402 324; x_wconf 86'>i</span>
   <span class='ocrx_word' id='word_1_9' title='bbox 1328 82 1569 392; x_wconf 57'>ig</span>
   <span class='ocrx_word' id='word_1_10' title='bbox 1551 97 1643 341; x_wconf 68'>e</span>
   <span class='ocrx_word' id='word_1_11' title='bbox 1574 69 1924 482; x_wconf 22'>en</span>
   </span>
  </p>
 </div>
```

FR-XML

- FR-XML = ABBYY FineReader XML
  support.abbyy.com/hc/en-us/articles/360017336699-ABBYY-FineReader-Engine-XML-Export
- Letzte Version: `FineReader10-schema-v1` (Released: ???)
- Seitenbasiert, d.h. 1 Seite = 1 FR-XML Datei
- OCR Software mit FR-XML Unterstützung:
  - ABBYY FineReader
  - Kraken
  - Dienstleister

```
<line baseline="240" l="313" t="116" r="1742" b="269">
    <formatting lang="GermanStandard" ff="Arial" fs="40." spacing="150">
        <charParams l="313" t="116" r="438" b="241" wordStart="1" wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="30" serifProbability="255" wordPenalty="0" meanStrokeWidth="158">D</charParams>
        <charParams l="438" t="117" r="509" b="241"></charParams>
        <charParams l="509" t="145" r="549" b="237" wordStart="1" wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="36" serifProbability="255" wordPenalty="0" meanStrokeWidth="158">c</charParams>
        <charParams l="550" t="145" r="620" b="237"></charParams>
        <charParams l="621" t="145" r="669" b="233" wordStart="1" wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="34" serifProbability="255" wordPenalty="0" meanStrokeWidth="158">r</charParams>
        <charParams l="670" t="117" r="813" b="234"></charParams>
        <charParams l="813" t="118" r="898" b="234" wordStart="1" wordFromDictionary="1" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="40" serifProbability="255" wordPenalty="0" meanStrokeWidth="160">A</charParams>
        <charParams l="969" t="142" r="1029" b="231" wordStart="0" wordFromDictionary="1" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="48" serifProbability="255" wordPenalty="0" meanStrokeWidth="160">n</charParams>
        <charParams l="1101" t="139" r="1141" b="255" wordStart="0" wordFromDictionary="1" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="26" serifProbability="255" wordPenalty="0" meanStrokeWidth="160">;</charParams>
        <charParams l="1142" t="139" r="1212" b="255"></charParams>
        <charParams l="1213" t="147" r="1257" b="239" wordStart="1" wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="49" serifProbability="255" wordPenalty="0" meanStrokeWidth="184">c</charParams>
        <charParams l="1258" t="123" r="1321" b="244"></charParams>
        <charParams l="1321" t="124" r="1358" b="244" wordStart="1" wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="24" serifProbability="255" wordPenalty="0" meanStrokeWidth="184">t</charParams>
        <charParams l="1358" t="124" r="1429" b="268"></charParams>
        <charParams l="1429" t="144" r="1489" b="268" wordStart="1" wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="26" serifProbability="255" wordPenalty="0" meanStrokeWidth="184">g</charParams>
        <charParams l="1490" t="144" r="1556" b="269"></charParams>
        <charParams l="1557" t="145" r="1601" b="241" wordStart="1" wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="26" serifProbability="255" wordPenalty="5" meanStrokeWidth="184">k</charParams>
        <charParams l="1602" t="145" r="1668" b="245"></charParams>
        <charParams l="1669" t="149" r="1717" b="245" wordStart="1" wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="31" serifProbability="255" wordPenalty="0" meanStrokeWidth="184">r</charParams>
        <charParams l="1717" t="217" r="1741" b="245" wordStart="0" wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
        charConfidence="100" serifProbability="255" wordPenalty="0" meanStrokeWidth="184">.</charParams>
    </formatting>
</line>
```

TEI

- TEI = Text Encoding Initiative
- Standard für Transkriptionen, digitale Editionen
- Breite Verwendung in den Digital Humanities

  tei-c.org
- Dokumentenbasiert, d.h. 1x gesamtes Dokument = 1x TEI-Datei
- Kein XML-Schema sondern flexible „Toolbox“ von Elementen/Attributen
- DTABf = Deutsches TextArchiv Basisformat
  deutschestextarchiv.de/doku/basisformat/

```
<ab facs="#fPage1_Block2" rendition="#font5" type="TextBlock" xml:id="Page1_Block2">
    <s>
        <w facs="#f32" xml:id="e32">D</w>
        <w facs="#f34" xml:id="e34">c</w>
        <w facs="#f36" xml:id="e36">r</w>
        <w facs="#f38" xml:id="e38">An /w>
        <pc type="post" xml:id="pc">;</pc>
        <w facs="#f40" xml:id="e40">c</w>
        <w facs="#f42" xml:id="e42">t</w>
        <w facs="#f44" xml:id="e44">g</w>
        <w facs="#f46" xml:id="e46">k</w>
        <w facs="#f48" xml:id="e48">r</w>
        <pc type="post" xml:id="pc">.</pc>
    </s>
    <lb facs="#f31" xml:id="e31"/>
```

Text

- Plain Text (ohne Markup/Strukturierung)
- Encoding (UTF8 vs. ASCII)
  - Historische Sonderzeichen!
    Vgl. ocr-d.de/en/gt-guidelines/trans/ocr_d_koordinationsgremium_codierung.html
- Zeilenumbrüche (`<CR>` vs `<LF>`)
- Aber:
  - Sehr geringe Dateigröße
  - Häufiges Eingabeformat für bspw. TDM

Der Anzeiger.
Fünfzehntes Stü.
Todesfälle. Dreßden den 19. März Amalie Auguſte Eleo=
nore im 17 J., älteſte To ter Hr. Joh. Wolf Gottlob von Gablenz
in Kroppen. — Lauban d. 19. März Hr. Carl Gottlob Seiffert,
E. Ho edl. Ho w. Rath's in Lauban Scabin und berühmter Kauf=
und Handelsherr, geb. d. 18. Dez. 1727 daſ. Er erlernte die Hand=
lung in Leipzig bey dem Kaufmann Se ehage, und etablirte ſo=
dann zu Lauban. Im J. 1750 d. I. Dez. verehli te er mit
Demoiſ. Charlotte Dorothea geb. Kahe aus Lauban. Aus dieſer Ehe
leben von 12 Kindern no 6 : Hr. Carl Aug. Seyffert, Kauf= und
Handelsherr a hier, der den Vater mit einen no lebenden Enkelſohn
erfreute ; Hr. Carl Ferdinand Seyffert, K. und H. allhier. Von
dieſem erlebte er einen no lebenden Enkel=Sohn und To ter ; Fr.
Karoline Henriette, verehli t an Hr. Joh. Chriſtn. Seyffert, Stadt=
ri ter in Zittau. Von dieſer ſahe er 9 Enkel, davon 8 leben, 4 S.
und 4 T. davon bereits 2 verheyrathet nd ; Fr. Charlotte Dorothea
verehel. an Hr. Carl Aug. Fiſ er, K. und H. zu Lauban, die ihn mit
3 Enkeln beſ enkte, wovon 2 am Leben ; Fr. Charlotte Eleonore ver=
ehel. mit Hr. Chri n. Gottlieb Wolf, Erb= Lehn= und Geri tsherr
auf Örtmannsdorf, Inſpektor derer Rittergüter Gebhardsdorf, wie au
Generalbevo mä tiger des Hr. Geheimerath von Ü triz, wovon er

OCR-D Spezifikationen

- METS als Containerformat ocr-d.de/en/spec/mets
  - Erfordert unique ID `<mods:identifier>` (purl, urn, handle, url)
  - Stellt vordefinierte FileGroups bereit `<mets:fileGrp USE=„...“>`
  - Prozessinformationen werden dokumentiert via `<mets:agent>`

- PAGE als OCR Ausgabeformat ocr-d.de/en/spec/page
  - Polygone für Regionentypen, Unicode für Text
  - Informationen zu `<AlternativeImage>` (binarized, cropped, deskewed, etc.)
  - Angabe der `<ReadingOrder>`

Konvertierung

github.com/cneud/ocr-conversion-scripts

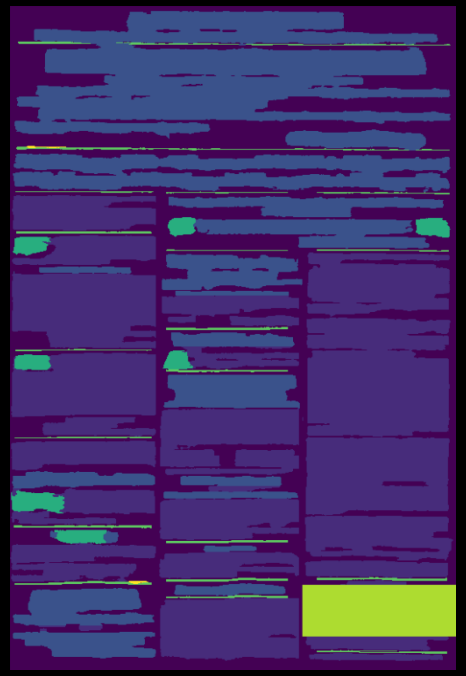
digi.bib.uni-mannheim.de/ocr-fileformat/

github.com/kba/page-to-alto

- Was fehlt noch?
  - ALTO → PAGE
  - METS/ALTO bzw. METS/PAGE → IIIF Manifest
  - METS/ALTO bzw. METS/PAGE → TEI (DTABf?)

Sonderfall Zeitungen

- Seitenübergreifende Strukturdaten („*Fortsetzung dieses Artikels auf S. 5*“)
- Artikelsegmentierung
- Lesereihenfolge („Reading Order“)




```
<UnorderedGroup>
  . . .
  <OrderedGroup>
    . . .
  </OrderedGroup>
  . . .
</UnorderedGroup>
```

Direct bij den Dam
[illegible]

Te Huur gevraagd,
[illegible]

Baarn.
[illegible]

Wijk aan Zee.
[illegible]

Te Huur.
[illegible]

TE HUUR te Baarn
[illegible]

Hilversum.
[illegible]

Gevraagd, te Huur of te Koop.
[illegible]

Zeist of Omstreken.
[illegible]

BAARN.
[illegible]

Limburg.
[illegible]

Arnhem. Sonsbeek.
[illegible]

Te Huur of te Koop:
[illegible]

Zandvoort.
[illegible]

Te Huur:
[illegible]

Wordt gevraagd:
[illegible]

TE HUUR.
[illegible]

Te Koop:
[illegible]

Te Koop of te Huur:
[illegible]

Posterijen.
[illegible]

Ter Overname gevraagd:
[illegible]

Zeldzaam aanbod.
[illegible]

Vergunning.
[illegible]

Een nette Zaak
[illegible]

Grutterswinkel
[illegible]

Velocipède.
[illegible]

Te Koop:
[illegible]

Manufacturen, Garen en Band.
[illegible]

NEDERLANDSCHE TYPEN EN SCHETSEN,
door C. F. VAN DUIJL,
MET 8 PLATEN VAN JOH. BRAAKENSIEK.
Een geestig Boek!

Op fraai papier en keurig gedrukt.
[illegible]

Volière
[illegible]

Te Koop te Haarlem,
[illegible]

BOEKENKAST
[illegible]

Te Koop.
[illegible]

IJZERGIETERIJ.
[illegible]

IJzergieterij.
[illegible]

Wegens afschaffing.
[illegible]

TE KOOP voor f 1200:
[illegible]

Uit de hand te Koop.
[illegible]

Ter Overname:
[illegible]

INSECTENPOEDER
[illegible]

C. BERNDT & Co., Leipzig.
[illegible]

Zomer- en Winterverblijf.
[illegible]

Amsteldijk, wijk K, No. 13.
[illegible]

18 AKERS.
[illegible]

Frederikplein No. 9.
[illegible]

Groote Wittenburgerstraat,
[illegible]

6 Aren 6 Centiaren
[illegible]

SEIJFFARDT's Boekhandel
[illegible]

33% Gasbezuiniging MET GARANTIE
[illegible]

Tweede Wedstrijd
[illegible]

Velocipeden-Fabriek.
[illegible]

KRAAM- en ZIEKENBEDDEN
[illegible]

Ter leen gevraagd.
[illegible]

Wandelkaart
[illegible]

STAATS EVERS,
[illegible]

Hollandsch Bier
[illegible]

JACHT.
[illegible]

Dames!
[illegible]

D. EICHHORN,
[illegible]

Vaderlandsche Geschiedenis.
[illegible]

HET NIEUWS VAN DEN DAG VAN DONDERDAG [illegible] MEI [illegible]. 2e BLAD.

Danke für die Aufmerksamkeit!
Fragen?

Clemens Neudecker (@cneudecker)

DINI AG KIM Workshop

27.04.2021