

# «Crowdsourcing»-Initiative zur Digitalisierung von Heinrich Bullingers Briefwechsel

Institut für Computerlinguistik  
Institut für Schweizerische Reformationsgeschichte

M. Volk, B. Schroffenegger

Universität Zürich

25. November 2019

# Total 10'093 Karteikarten ...

Datum 1547 Okt. 5.	Absender Bullinger Heinrich Zürich	Empfänger Myconius Oswald Basel
Autograph Standort <i>Genève, H.D.</i> Sign. <i>Genève 842.178</i> Umfang	Kopie Standort 228, Ms. 565, 47 Sign. Umfang	Photokopie ZB Bull. Corr. 72 Bl. 1, S. 1 Abschrift ZB Bull. Corr. 12 Bl. 1, S. 1
Sprache	Literatur	
Gedruckt		
		Bemerkungen Commendo tibi ... Myconi, hunc D. Lavateri consulis filium

## Abschätzungen:

- Arbeitsaufwand:  $3 \pm 2$  min/Karte  $\hat{=}$   $63 \pm 42$  Arbeitstage
- Kosten:  $11'102 \pm 7'402$  CHF (22 CHF/h)

- 1 Datenextraktion
  - OCR
  - Analysen
    - Karteikartengrößen
    - Dateigrößen
    - Partitionierung
  - Aufbereitung der Daten
    - Heuristiken
    - Silbentrennung
- 2 Datenbank
  - Anforderungen/Designprinzipien
  - Relationen
- 3 Interface
  - Live Demo
- 4 Tools

# OCR (= Optical Character Recognition)

Automatische Texterkennung (Optische Zeichenerkennung):

- digitaler Text  $\subseteq$  digitales Bilder  $\rightarrow$  Format-Transformation (Extrakt)
- Mustererkennung:
  - Formate: PDF  $\rightarrow$  XML
  - Atome: Pixel (Raster-/Vektorgraphik)  $\rightarrow$  Zeichen (Textformat)
  - Eigenschaften: Position/Farbe  $\rightarrow$  Position/Symbol
- Software: ABBYY FineReader
  - unterstützt insbesondere Latein & Althochdeutsch
  - 3 Problemerkarteikarten
  - proprietär (0.01 CHF/Seite)
- Typische Annotationen (Elemente/Attribute/Attributwerte):
  - Wort (Grösse/Position)
  - Textzeile (Grösse/Position)
  - Textblöcke, Seiten, ... (!)

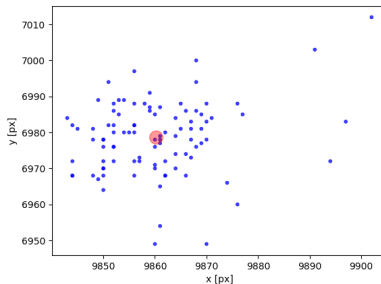
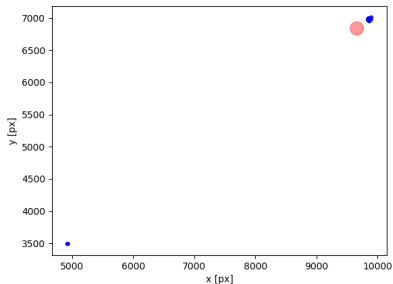
# OCR-Output (FineReader ABBYY)

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
...
<Page ID="Page1" [...] HEIGHT="6946" WIDTH="9872">
...
<TextLine BASELINE="348" HEIGHT="136" WIDTH="1756" VPOS="214" HPOS="3246">
...
  <String CONTENT="Absender" HEIGHT="136" WIDTH="756" VPOS="214" HPOS="3246"/>
...
</TextLine>
<TextLine BASELINE="643" HEIGHT="156" WIDTH="1976" VPOS="494" HPOS="3306">
...
  <String CONTENT="Micronius" HEIGHT="156" WIDTH="1100" VPOS="494" HPOS="3306"/>
...
  <String CONTENT="Martin" HEIGHT="152" WIDTH="748" VPOS="494" HPOS="4534"/>
...
</TextLine>
<TextLine BASELINE="1056" HEIGHT="144" WIDTH="736" VPOS="918" HPOS="3314">
...
  <String CONTENT="London" HEIGHT="144" WIDTH="736" VPOS="918" HPOS="3314"/>
...
</TextLine>
...
```

Parsing: mit Python (xml.sax/pandas) → ... → Datenbank (SQLAlchemy)

# Karteikartengrösse (Länge $\times$ Breite)

Stichprobe: 100 Karteikarten mit/ohne Ausreisser (4%)

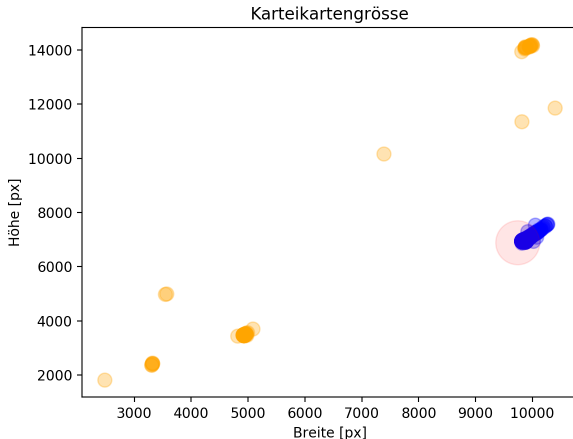


Seite	$\mu_a$	$\sigma_a$	min	max
Breite	9661	975	4922	9902
Höhe	6837	690	3488	7012

Seite	$\mu_b$	$\sigma_b$	min	max
Breite	9860	11	9843	9902
Höhe	6978	10	6949	7012

# Karteikartengrösse (Länge $\times$ Breite)

Allgemein: total 10'093 Dateien mit 259 Ausreissern ( $\approx 2.6\%$ , orange)



Mittelwert (rot):  $(x_m, y_m) = (9736, 6894)$

# Karteikartengrösse (Ausreisser)

## File-IDs:

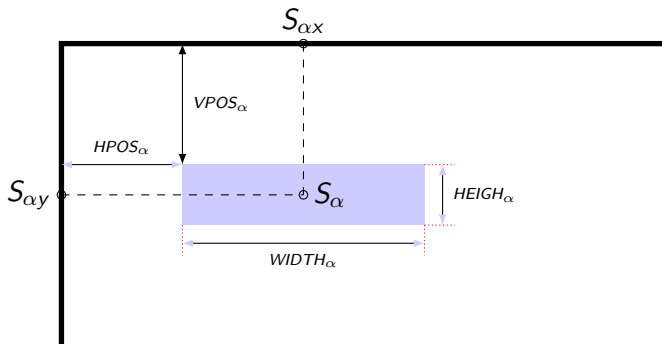
1, 16, 32, 33, 38, 50, 58, 63, 76, 80, 87, 92, 111, 116, 118, 128, 130, 138, 158, 161, 162, 170, 173, 186, 228, 237, 238, 246, 254, 255, 263, 266, 313, 315, 331, 332, 336, 354, 355, 356, 366, 368, 378, 401, 402, 426, 432, 439, 444, 448, 458, 469, 484, 490, 495, 502, 509, 512, 535, 541, 543, 553, 565, 569, 591, 593, 609, 618, 632, 638, 657, 658, 665, 666, 671, 679, 680, 694, 696, 698, 707, 713, 728, 746, 762, 763, 827, 828, 832, 833, 858, 860, 861, 867, 874, 889, 916, 918, 945, 965, 985, 1004, 1009, 1034, 1042, 1045, 1048, 1054, 1079, 1089, 1091, 1097, 1099, 1101, 1102, 1125, 1128, 1129, 1143, 1150, 1169, 1176, 1209, 1213, 1238, 1240, 1245, 1267, 1268, 1273, 1274, 1275, 1276, 1278, 1281, 1282, 1283, 1284, 1289, 1292, 1309, 1329, 1333, 1334, 1339, 1346, 1354, 1365, 1372, 1373, 1374, 1375, 1388, 1398, 1404, 1433, 1439, 1446, 1469, 1488, 1489, 1490, 1493, 1494, 1496, 1510, 1541, 1569, 1595, 1609, 1612, 1618, 1620, 1624, 1628, 1636, 1640, 1642, 1655, 1662, 1680, 1681, 1682, 1697, 1715, 1720, 1728, 1752, 1759, 1764, 1766, 1772, 1777, 1786, 1787, 1789, 1798, 1809, 1827, 1832, 1840, 1853, 1854, 1855, 1859, 1860, 1864, 1865, 1903, 1928, 1936, 1940, 1941, 1948, 1955, 1974, 1975, 1980, 1988, 1991, 2010, 2026, 2038, 2070, 2071, 2072, 2096, 2109, 2126, 2141, 2148, 2149, 2176, 2495, 3152, 4247, 4249, 4250, 4685, 4721, 5198, 5570, 5649, 5828, 6071, 6234, 6426, 6680, 6681, 6708, 6865, 7054, 7432, 7806, 7961, 7966, 8200, 8479, 8537, 8818, 9227, 9297, 9298, 9708, 9709, 9710, 9810, 9942, 10002



# Karteikartengrösse (Skalierung)

Normierung aller Karteikarten  $i \in \{1, \dots, 10'093\}$  auf Einheitsgrösse:


















- Seitenbreite  $B_i \propto S_{i\alpha x}(\text{WIDTH}_{i\alpha x}, \text{HPOS}_{i\alpha x})$
- Seitenhöhe  $H_i \propto S_{i\alpha y}(\text{HEIGHT}_{i\alpha y}, \text{VPOS}_{i\alpha y})$



Skalierungsfaktoren:  $x_s(B_i) = \frac{\hat{B}}{B_i}$ ,  $y_s(H_i) = \frac{\hat{H}}{H_i}$  ( $\hat{B} = 9860$ ,  $\hat{H} = 6978$ ).

# Dateigrößen [MB]

Verdächtig: 10 (?) Briefe (0.99‰)

	HBBW_Karteikarte_00632.png	10.11.2019, 21:24	38.5 MB	PNG-Bild
	HBBW_Karteikarte_00354.png	10.11.2019, 21:12	34.1 MB	PNG-Bild
	HBBW_Karteikarte_00355.png	10.11.2019, 21:13	31.2 MB	PNG-Bild
	HBBW_Karteikarte_02071.png	10.11.2019, 22:17	16.5 MB	PNG-Bild
	HBBW_Karteikarte_00426.png	10.11.2019, 21:15	15.6 MB	PNG-Bild
	HBBW_Karteikarte_00432.png	10.11.2019, 21:16	15.6 MB	PNG-Bild
	HBBW_Karteikarte_04249.png	10.11.2019, 23:34	7.3 MB	PNG-Bild
	HBBW_Karteikarte_01273.png	10.11.2019, 21:48	5.9 MB	PNG-Bild
	HBBW_Karteikarte_01129.png	10.11.2019, 21:43	5.7 MB	PNG-Bild
	HBBW_Karteikarte_01697.png	10.11.2019, 22:03	5.2 MB	PNG-Bild
	HBBW_Karteikarte_01128.png	10.11.2019, 21:43	4.9 MB	PNG-Bild
	HBBW_Karteikarte_00092.png	10.11.2019, 21:02	4.6 MB	PNG-Bild
	HBBW_Karteikarte_01854.png	10.11.2019, 22:09	4.3 MB	PNG-Bild
	HBBW_Karteikarte_00591.png	10.11.2019, 21:22	4.3 MB	PNG-Bild
	HBBW_Karteikarte_01489.png	10.11.2019, 21:56	4.2 MB	PNG-Bild
	HBBW_Karteikarte_00565.png	10.11.2019, 21:21	4.1 MB	PNG-Bild
	HBBW_Karteikarte_06680.png	11.11.2019, 00:59	4.1 MB	PNG-Bild

# Attribute/Werte (Position)

Datum 1551 Oktober 10.	Absender Feldmay Feyerthoy Johannes Wien	Empfänger Bullinger Heinrich Zürich
31%	35%	34%
Autograph	Kopie	Photokopie ZB
Standort Zürich, A. A.	Standort Zürich, A. A.	Bull. Corr. 77 Bl. 4, S. 4
Sign. W. I. 267, 46 ff.	Sign.	Abschrift ZB
Umfang	Umfang	Bull. Corr. 16 Bl. 2, S. 4
Sprache	Literatur	
Gedruckt	Bemerkungen	

$0 < x < 3057$ $0 < y < 1535$	$3057 < x < 6508$	$6508 < x < 9860$
$1535 < y < 2041$		
$2041 < y < 2547$		
$2547 < y < 3053$		
$3053 < y < 3559$		
$3559 < y < 4257$	$3559 < y < 5303$	
$4257 < y < 6978$	$5303 < y < 6978$	

Abbildung: Messungsergebnisse (in [%] und [px])

## % Berechnung

```
f = lambda l: [sum(l[:i+1]) for i, _ in enumerate(l)] # Partialsummen
f([b*9860 for b in [0.31, 0.35, 0.34]]) # [3057, 6508, 9860]
f([h*6978 for h in [0.22, 0.29, 0.1, 0.39]]) # [1535, 3559, 4257, 6978]
f([h*6978 for h in [0.22, 0.14, 0.15, 0.25, 0.24]]) # [1535, 2512, 3559, 5303, 6978]
[1535+i*0.29*6978/4 for i in range(1,5)] # [2041, 2547, 3053, 3559]
```

# Attribute/Werte (Position)

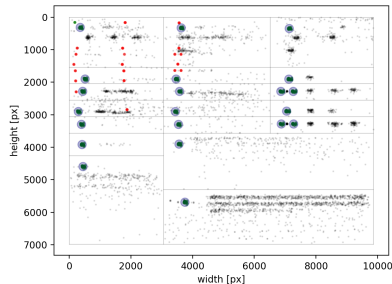
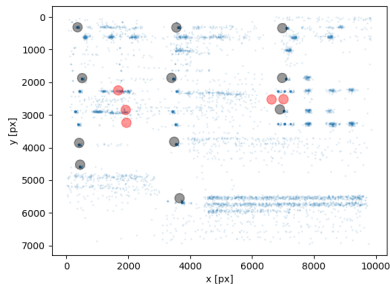
Typische Werte (Mittelwerte/Standardabweichungen):

Value	x	y	$\sigma_x$	$\sigma_y$
Datum	364.0	320.0	20.0	13.0
Absender	3631.0	329.0	17.0	12.0
Empfänger	7134.0	349.0	18.0	12.0
Autograph	515.0	1901.0	19.0	13.0
Kopie	3465.0	1903.0	17.0	12.0
Photokopie	7120.0	1904.0	26.0	12.0
Standort	437.0	2281.0	20.0	12.0
Standort	3577.0	2283.0	19.0	12.0
Bull.	6861.0	2282.0	31.0	13.0
Bull.	6872.0	3284.0	37.0	18.0
Corr.	7261.0	2284.0	31.0	13.0
Corr.	7273.0	3285.0	36.0	17.0
Sign.	298.0	2903.0	18.0	12.0
Sign.	3442.0	2903.0	16.0	12.0
Abschrift	7049.0	2890.0	25.0	11.0
Umfang	397.0	3297.0	21.0	13.0
Umfang	3543.0	3299.0	17.0	12.0
Sprache	417.0	3917.0	22.0	13.0
Literatur	3560.0	3904.0	20.0	13.0
Gedruckt	448.0	4594.0	21.0	13.0
Bemerkungen	3752.0	5691.0	19.0	15.0

# Attribute/Werte (Position)

Plots (100 KK): Attributnamen, Attributwerte & Trennlinien zw. Feldern

- doppelte Attributnamen
  - Standort, Sign., Umfang (Autograph/Kopie)
  - Bull. Corr. (Photokopie/Abschrift)
- korrupte Daten: durchschnittliche Position eines Attributnamens...
  - links: undifferenziert zwischen Namen und Position
  - rechts (unskaliert): separate Darstellung der Ausreisser



# Funktionen

Beziehungen zw. Attributen/Werten und Position derselben:

- $(x, y) \rightarrow$  Zugehörigkeit
- $(\text{Attributname}, x, y) \rightarrow$  Wert oder Bezeichnung?

```
@staticmethod
def get_attribute_name(x, y):
    if x <= 3057: # 1st column
        if y <= 1535: return "Datum", None
        elif y <= 2041: return "Autograph", None
        elif y <= 2547: return "Standort", 'A'
        elif y <= 3053: return "Sign", 'A'
        elif y <= 3559: return "Umfang", 'A'
        elif y <= 4257: return "Sprache", None
        else: return "Gedruckt", None
    elif x <= 6508: # 2nd column
        if y <= 1535: return "Absender", None
        elif y <= 2041: return "Kopie", None
        elif y <= 2547: return "Standort", 'B'
        elif y <= 3053: return "Sign", 'B'
        elif y <= 3559: return "Umfang", 'B'
        elif y <= 5303: return "Literatur", None
        else: return "Bemerkungen", None
    else: # 3rd column
        if y <= 1535: return "Empfänger", None
        elif y <= 2041: return "Photokopie", None
        elif y <= 2547: return "Bull. Corr.", 'A'
        elif y <= 3053: return "Abschrift", None
        elif y <= 3559: return "Bull. Corr.", 'B'
        elif y <= 5303: return "Literatur", None
        else: return "Bemerkungen", None
```

```
@staticmethod
def is_attribute(attribute_name, x_coord, y_coord, deviation=3):
    d = deviation # tolerance
    return {
        "Datum": lambda x, y: True if 364-d*x20 < x < 364+d*x20 and 320-d*x13 < y < 320+d*x13 else False,
        "Absender": lambda x, y: True if 3631-d*x17 < x < 3631+d*x17 and 329-d*x12 < y < 329+d*x12 else False,
        "Empfänger": lambda x, y: True if 7134-d*x18 < x < 7174+d*x18 and 349-d*x12 < y < 349+d*x12 else False,
        "Autograph": lambda x, y: True if 515-d*x19 < x < 515+d*x19 and 1901-d*x13 < y < 1901+d*x13 else False,
        "Standort A": lambda x, y: True if 437-d*x20 < x < 437+d*x20 and 2291-d*x12 < y < 2281+d*x12 else False,
        "Standort B": lambda x, y: True if 3577-d*x19 < x < 3577+d*x19 and 2283-d*x12 < y < 2283+d*x12 else False,
        "Sign. A": lambda x, y: True if 298-d*x18 < x < 298+d*x18 and 2903-d*x12 < y < 2903+d*x12 else False,
        "Sign. B": lambda x, y: True if 3442-d*x18 < x < 3442+d*x18 and 2903-d*x12 < y < 2903+d*x12 else False,
        "Umfang A": lambda x, y: True if 397-d*x21 < x < 397+d*x21 and 3297-d*x13 < y < 3297+d*x13 else False,
        "Umfang B": lambda x, y: True if 3543-d*x17 < x < 3543+d*x17 and 3297-d*x13 < y < 3297+d*x13 else False,
        "Photokopie": lambda x, y: True if 7120-d*x26 < x < 7120+d*x26 and 1904-d*x12 < y < 1904+d*x12 else False,
        "Bull. Corr. A": lambda x, y: True if 6865-d*x35 < x < 7273+d*x37 and 2282-d*x15 < y < 2282+d*x15 else False,
        "Bull. Corr. B": lambda x, y: True if 6865-d*x35 < x < 7273+d*x37 and 3284-d*x15 < y < 3284+d*x15 else False,
        "Abschrift": lambda x, y: True if 7049-d*x25 < x < 7049+d*x25 and 2890-d*x11 < y < 2890+d*x11 else False,
        "Sprache": lambda x, y: True if 417-d*x22 < x < 417+d*x22 and 3917-d*x13 < y < 3917+d*x13 else False,
        "Gedruckt": lambda x, y: True if 448-d*x21 < x < 448+d*x21 and 4594-d*x13 < y < 4594+d*x13 else False,
        "Literatur": lambda x, y: True if 3560-d*x20 < x < 3560+d*x20 and 3904-d*x13 < y < 3904+d*x13 else False,
        "Bemerkungen": lambda x, y: True if 3752-d*x19 < x < 3752+d*x19 and 5691-d*x15 < y < 5691+d*x15 else False,
        "Kopie": lambda x, y: True if 3465-d*x17 < x < 3465+d*x17 and 1903-d*x12 < y < 1903+d*x12 else False,
    }[attribute_name](x_coord, y_coord)
```

# Aufbereitung der Daten

## Attributnamen entfernen

→ Position und «physische» Erscheinung bekannt

## Beseitigung von OCR-Fehlern...

- Datum: Karteikartenordnung ausnutzen (Jahre/Monate)
- Personen: Bullinger immer präsent (immer?); Ähnlichkeitsvergleiche (N-Gramme)
- ((Ortschaften: Ähnlichkeitsvergleiche gegen bekannte (Schweizer-) Städtenamen))
- Sprache: Deutsch, Lateinisch, Griechisch. Spracherkennungstool «*langid*»
- Referenzen (Photokopie/Abschrift): Einschränkungen auf numerische Zeichen
- Standort: Präsenz bestimmter «keywords» (Schreibmaschine?)
- Bemerkung: insbesondere Bindestriche entfernen
- Allgemein: exotische Zeichen entfernen; nach Zahlen/Buchstaben filtern, typische Fehler korrigieren. Beispiele:  
{0, O, o, °}, {B, 8, β, g}, {l, I, i, 1, |}, {ii, ü}, {S, \$},  
{7, ?, y}, {A, 4}, {£, E, €}, ...

# Heuristiken

- Schreibmaschinen-Schlüsselwörter
  1. StA, StB, Ms., Nr., Hr., (ZB), ... (?)
  2. Zürich, Chur, Basel, St.Gallen, Genf, London, Den Haag, ...
- Autograph/Kopie
  - Beispiele (Standort/Signatur)
    - Zürich StA, E II 360,85
    - Zürich ZB (ZZB), Ms S 67, 84
  - Regeln:
    - $\{E, \text{£}, \text{€}\} \in \text{Signatur}[0:4] \Rightarrow \text{Standort: «Zürich StA»}$
    - $\{A, 4\} \in \text{Standort}[-4:-1] \Rightarrow \text{Standort: «Zürich StA»}$
    - $\{22, 2Z, Z2, 22\} \in \text{Standort}[0:4] \Rightarrow \text{«Zürich ZB» (ZZB)}$
    - $\{Z, 2, 7, ^\wedge\} \in \text{Standort}[0:4] \Rightarrow \text{«Zürich»}$
  - typische Fehler:
    - $\{U, T, X\}$  statt II
    - Kombinationen [Zürich StA EU 365,757]
- Junk?
  - Buchstaben/Zeichen-Verhältnis; Durchschnittliche Wortlänge; Einzelzeichen; ...



# Silbentrennungen

Viele Bindestriche: Bspw. «Baselines» [['Ein', 'Bei-'], ['spiel']]

```
1 if (lines[i][-1][-1] == '-'):
2     lines[i][-1][-1] = ''
3     concat(lines[i][-1], lines[i+1][0])
```

→ Erfolgsquote bei ca. 40% (9:13 über 100)

```
***
<String CONTENT="omni" SUBS_TYPE="HypPart1" SUBS_CONTENT="omnipotente" .../>
<HYP CONTENT="-"/>
</TextLine>
<TextLine>
  <String CONTENT="potente" SUBS_TYPE="HypPart2" SUBS_CONTENT="omnipotente" .../>
***
```

Lösung (Parser):

```
1 d, v = dict(), None
2 for a in attr.getNames(): d[a] = attr.getValue(a)
3 if ("SUBS_TYPE" in d) and ("SUBS_CONTENT" in d):
4     if (d["SUBS_TYPE"] == "HypPart1"):
5         v = d["SUBS_CONTENT"]
6     else: v = d["CONTENT"]
```

# Datenbank

- Herausforderungen:
  - Vandalismus
  - Redundanz
- Lösungen:
  - Benutzerkonten
  - Normalisierung
- Normalformen
  1. keine mehrwertigen Attribute (Sortier-/Selektierbarkeit)
  2. keine partielle Abhängigkeiten von Nicht-Schlüsselattributen vom Gesamtschlüssel (weniger Redundanzen/Updateanomalien)
  3. kein Nichtschlüsselattribut hängt transitiv von einem Schlüsselkandidaten ab (logische, monothematische DB-Struktur)

# Relationen

## Primärschlüssel: Karten Nummer (ID)

- Benutzer(Name, E-Mail, Passwort-Hash)
- Kartei(ID, Rezensionen, Status, Pfad\_OCR, Pfad\_IMG, Pfad\_PDF)
- Datum(ID, J1, M1, D1, J2, M2, D2, Bemerkung, Benutzer, Zeit)
- Person(IDP, NN, VN, Ort, Titel, Benutzer, Zeit)
- Absender(ID, IDP, Bemerkung, Benutzer, Zeit)
- Empfänger(ID, IDP, Bemerkung, Benutzer, Zeit)
- Autograph(ID, Standort, Signatur, Umfang, Benutzer, Zeit)
- Kopie(ID, Standort, Signatur, Umfang, Benutzer, Zeit)
- Photokopie(ID, Standort, Bull. Corr, Blatt, Seite, Benutzer, Zeit)
- Abschrift(ID, Standort, Bull. Corr, Blatt, Seite, Benutzer, Zeit)
- Sprache(ID, Sprache, Benutzer, Zeit)
- Literatur(ID, Literatur, Benutzer, Zeit)
- Gedruckt(ID, Gedruckt, Benutzer, Zeit)
- Bemerkung(ID, Bemerkung, Benutzer, Zeit)

# Werkzeuge

- Frontend:
  - HTML, CSS
  - JS/JQuery
- Backend (Python):
  - Server (Flask/Jinja2/Werkzeug)
  - Datenbank (SQLAlchemy)
  - Parser (xml.sax)
  - Sprachidentifikation (LangID)
  - Tabellenmanagement (pandas)
  - Dictionaries (defaultdict)
  - Dateimanagement (os)
  - pdf2image

## Interface

← → ↻ 🏠  90% ... 📄 ⭐

📁 A 📁 M 📁 UZH 📁 Bullinger

🏛️ Universität Zürich | Allgemeines Übersicht KONTO 🌐

Übersicht / 1549 / Januar / 28. Rezeptionen: 0; Status: offen

[illegible]

Status abgeschlossen: ● / ungültig: ● / unklar: ●
 

Zurück

Speichern

Vorwärts

Datum	Absender	Empfänger
1549 <span style="float: right;">28</span>		
(max. Jahr)		
Bemerkung	John	Heinrich
	Strassburg	Zürich
	Bemerkung	Bemerkung
<b>Autograph</b>	<b>Kopie</b>	<b>Standort</b>
Zürich StA	Zürich ZB	Bull. Corr.
E I 343,40	Ms S 221,6,2	ZB (Druck)
Umfang	Umfang	13
Latinisch	Non possum satis admirari,... quo nandam vestratium negligenclam, ne dicam an exosam	
Gedruckt	Mbribiaamt Deorillo senis Bindi ad familiare... , amicum meum Joannem Hopperum lileras	

Institut für Schweizerische Reformationsgeschichte, Institut für Computergistik, Universität Zürich, 2020