Universität Zürich^{UZH}

Institut für Computerlinguistik
Institut für Schweizerische Reformationsgeschichte

Crowd Correction Initiative zur Digitalisierung von Bullingers Briefwechsel

Projektdokumentation

Version 1

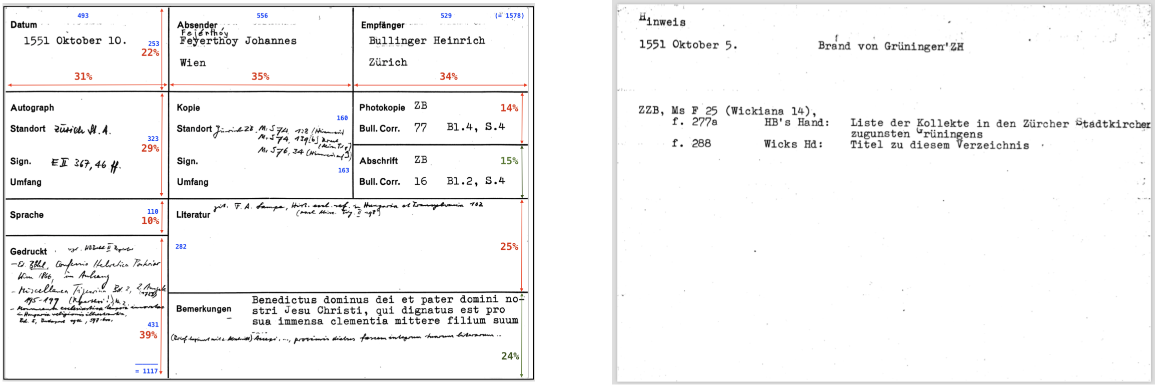
11. Oktober 2019

Inhaltsverzeichnis

1 Anforderungsspezifikation

1.1 IST-Zustand

1.1.1 Schema der Karteikarten



(a) `Karteikarten_HBBW_1551_100, S.13/99`

Datum
1551 Oktober 10.

Absender
Fejérthóy
Feyerthoy Johannes
Wien

Empfänger
Bullinger Heinrich
Zürich

Autograph
Standort Zürich St. A.
Sign. E II 367, 46 ff.
Umfang

Kopie
Standort Zürich ZB. Ms S 74, 138 (Himmeisen) Ms S 74, 139[b] Druck (Misc. Tig.) Ms S 76, 34 (Himmeisen [illegible])
Sign.
Umfang

Photokopie ZB
Bull. Corr. 77 Bl.4, S.4

Abschrift ZB
Bull. Corr. 16 Bl.2, S.4

Sprache

Literatur zit. F. A. Lampe, Hist. eccl. ref. in Hungaria et Transylvania 102 (nach Misc. Tig. II 198)

Gedruckt vgl. HBBibl II Register
– Ed. Böhl, Confessio Helvetica Posterior Wien 1866, im Anhang
– Miscellanea Tigurina Bd. 2, 2. Ausgabe (1723) 195–199 (Register!) Nr. 2
– Monumenta ecclesiastica tempore innovatae in Hungaria religionis illustrantia, Bd. 5, Budapest 1912, 597–600,

Bemerkungen
Benedictus dominus dei et pater domini no-
stri Jesu Christi, qui dignatus est pro
sua immensa clementia mittere filium suum
(Brief beginnt mit e. Abschnitt) Accepi ..., proximis diebus fasciculum integrum tuarum literarum ..

(b) `Karteikarten_HBBW_1551_100, S.5/99`

Hinweis

1551 Oktober 5. Brand von Grüningen'ZH

ZZB, Ms F 25 (Wickiana 14),
f. 277a HB's Hand: Liste der Kollekte in den Zürcher Stadtkircher zugunsten Grüningens
f. 288 Wicks Hd: Titel zu diesem Verzeichnis

Abbildung 1: typische Karteikarte (links) und Spezialfall (rechts)

# Felder	# Attribute	Attribute
7×	1	Datum; Absender; Empfänger; Sprache; Literatur; Gedruckt; Bemerkung
2×	2	[Photokopie, Bull. Corr.]; [Abschrift, Bull. Corr.]
2×	4	[Autograph, Standort, Sign., Umfang]; [Kopie, Standort, Sign. Umfang]
11×	19	$\sum$

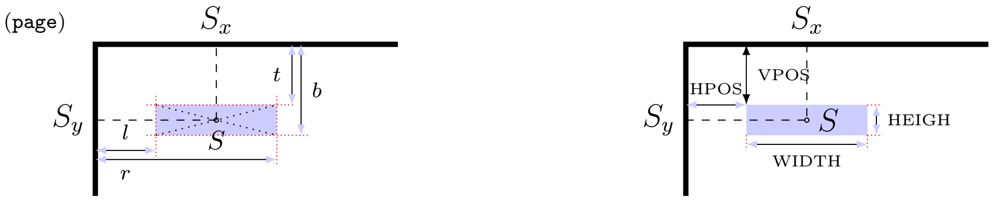
Abbildung 2: Felder und Attribute

Die Attributnamen «`Standort`», «`Sign.`», «`Umfang`» und «`Bull. Corr.`» sind auf den Karteikarten doppelt enthalten.

1.1.2 Schemata nach OCR

- Version 1: `http://www.abbyy.com/FineReader_xml/FineReader10-schema-v1.xml`
- Version 2: `http://www.loc.gov/standards/alto/alto-v2.0.xsd`

Positionsangaben:



$$S_{xy} = \left(\frac{r+l}{2}, \frac{t+b}{2}\right) \qquad S_{xy} = \left(\text{HPOS} + \frac{\text{WIDTH}}{2}, \text{VPOS} + \frac{\text{HEIGHT}}{2}\right)$$

Abbildung 3: Schwerpunktskoordinaten S_{xy} von Elementen in Version 1 (links) und 2.

1.2 SOLL-Zustand

1.2.1 Schema der Daten

Wir sollten die Daten so differenziert wie möglich erfassen (Datum → [Jahr, Monat, Tag]), Redundantes entfernen (Bull. Corr. → [Bull. Corr., Blatt, Seite]), und Attributwerte normieren (Jan., 1., 01., etc. → Januar).

Karteikarte (original)	Schema (neu)
Datum	Datum(Tag, Monat, Jahr)
Absender	Absender(Nachname, Vorname, Ort, Zusatz)
Empfänger	Empfänger(Nachname, Vorname, Ort, Zusatz)
Autograph	Autograph(Nachname, Vorname, Ort, Zusatz)
Standort A/B	Standort(Allgemein, Spezifisch, Zusatz)
Sign. A/B	Signatur(Allgemein, Spezifisch, Zusatz)
Umfang A/B	Umfang(Wert, Bemerkung)
Kopie	Kopie(Name, Bemerkung)
Photokopie	Photokopie(Name, Bemerkung)
Bull. Corr A/B	BullCorr(Blatt, Seite)
Abschrift	Abschrift(Name, Bemerkung)
Sprache	Sprache(Name, Zusatz)
Literatur	Literatur(Primär, Sekundär)
Gedruckt	Gedruckt(*Referenzen)
Bemerkungen	Bemerkung

Tabelle 1: Felder (Attribute), bzw. Schema der Daten original (links) und neu (rechts)

1.3 Anforderungsanalyse

1.3.1 Anwendungsfälle (User Stories)

Die folgenden Anwendungsszenarien dienen zur Analyse der Softwareanforderungen und so als Basis für die Formulierung der funktionalen Anforderungen.

Als Besucher/Anwender der Website/-applikation möchte ich...

- die Webseite über einen Link erreichen,
- allgemeine Informationen über Sinn/Zweck der Initiative erhalten,
-

1.3.2 Funktionale Anforderungen

Webapplikation

FRONT-END (CLIENT: WEBBROWSER)

- Erreichbare Website
- Benutzer Authentifizierung (Anmeldung)
-

BACK-END (Server)

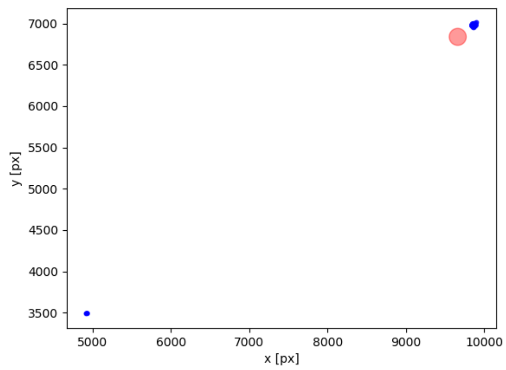
-

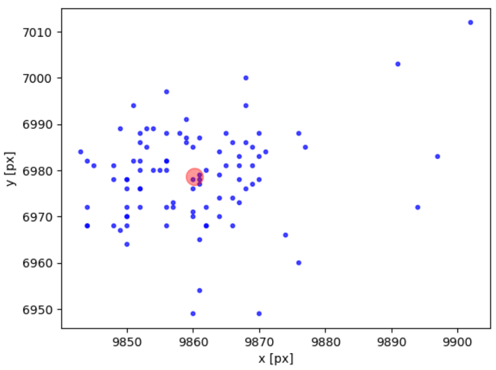
2 Implementation

2.1 Datenextraktion

2.1.1 Karteikartengrösse

Die Seitengrössen beider OCR-Versionen stimmen exakt überein.





Seite	μ_a	σ_a	min	max
Breite	9661	975	4922	9902
Höhe	6837	690	3488	7012

Tabelle 2: Dimensionen einer Karteikarte

Seite	μ_b	σ_b	min	max
Breite	**9860**	11	9843	9902
Höhe	**6978**	10	6949	7012

Tabelle 3: ohne Ausreisser

Durch Ausreisser verursachter Fehler:

$$\begin{aligned} \Delta\mu_x^{\%} &= 1 - \frac{9661}{9860} = 2.0183\% \\ \Delta\mu_y^{\%} &= 1 - \frac{6837}{6978} = 2.0206\% \end{aligned}$$

Ausreisser: $4/99 \approx 4\%$

- $(4922, 3488)$ `Karteikarten_HBBW_1551_1000010.xml`
- $(4923, 3489)$ `Karteikarten_HBBW_1551_1000041.xml`
- $(4935, 3492)$ `Karteikarten_HBBW_1551_1000069.xml`
- $(4937, 3489)$ `Karteikarten_HBBW_1551_1000095.xml`

Skalierung für Seitenlängen $(x, y)^T < \vec{\mu_b} - 4\vec{\sigma_b}$:

$$\mu_{bx} = \alpha\mu_{ax} \quad\Leftrightarrow\quad \alpha = \frac{\mu_{bx}}{\mu_{ax}} \quad\Rightarrow\quad \alpha(\mu_{ax}) = \frac{9860}{\mu_{ax}} \tag{1}$$

$$\mu_{by} = \beta\mu_{ay} \quad\Leftrightarrow\quad \beta = \frac{\mu_{by}}{\mu_{ay}} \quad\Rightarrow\quad \beta(\mu_{ay}) = \frac{6978}{\mu_{ay}} \tag{2}$$

2.1.2 Datenfelder

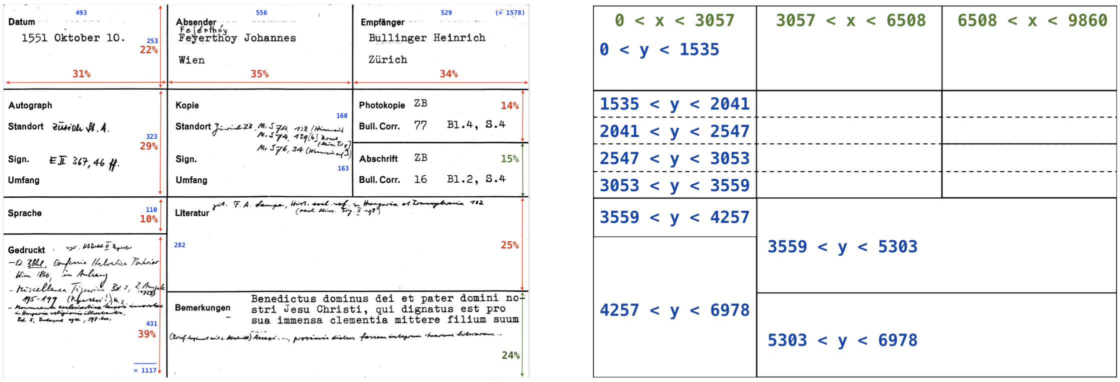
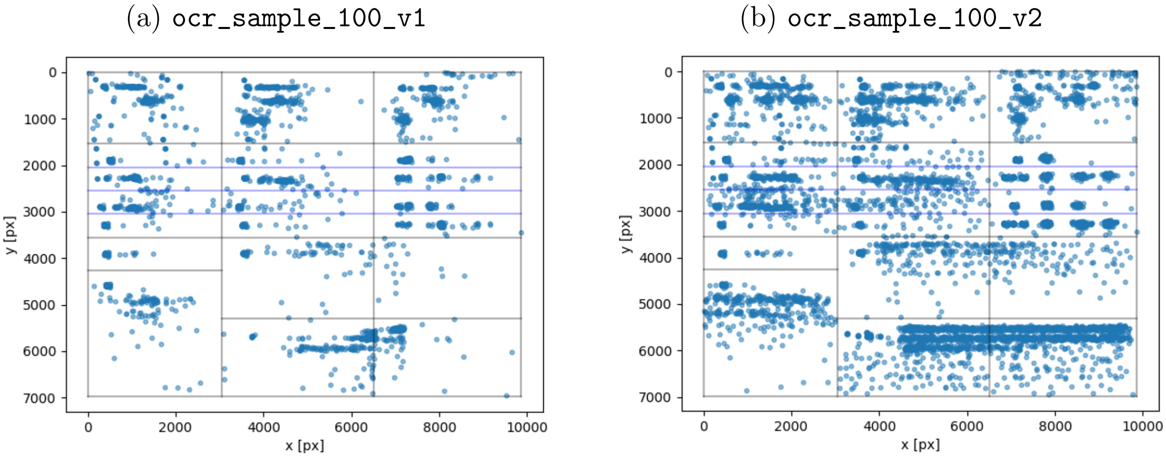


Abbildung 4: Manuelle Vermessung/Partitionierung einer Karteikarte

```
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]
```

2.1.3 OCR-Text


(a) ocr_sample_100_v1

(b) ocr_sample_100_v2

Abbildung 5: Schwerpunkte von OCR-Text-Elementen von jeweils 100 Karteikarten

2.1.4 OCR-Attribute

Attribute total: 1728. Gefiltert (FP): 1660

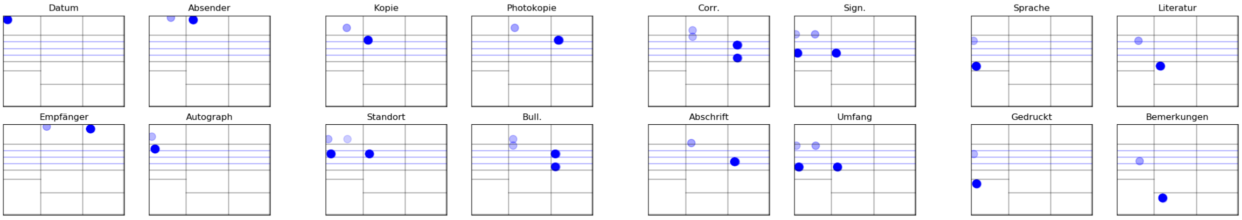


Abbildung 6: Verteilung einzelner Attributnamen

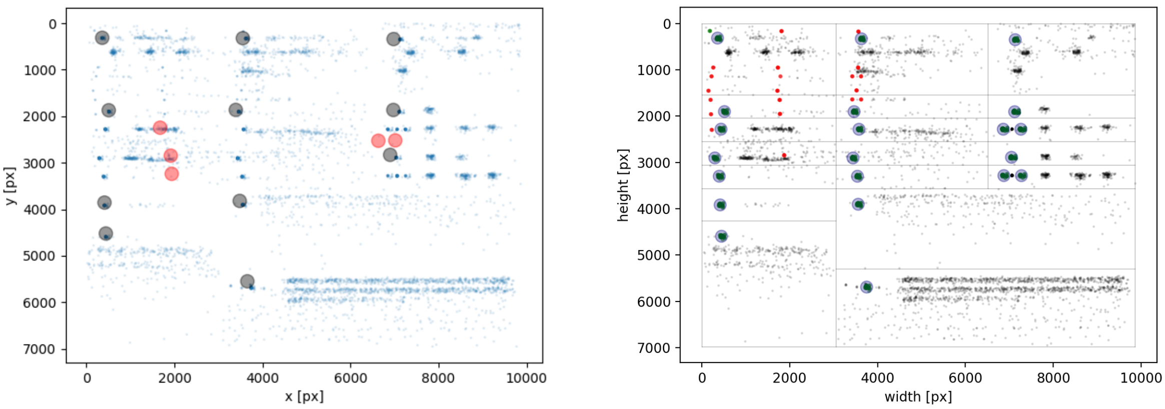


Abbildung 7: Durchschnittliche Attributpositionen (korrigiert/unkorrigiert)

Lineare Separierung:

$$\begin{aligned}
y[\text{Standort}_1] &\approx y[\text{Standort}_2] & x[\text{Standort}_1] &< x[\text{Standort}_2]\\
y[\text{Sign.}_1] &\approx y[\text{Sign.}_2] & x[\text{Sign.}_1] &< x[\text{Sign.}_2]\\
y[\text{Umfang}_1] &\approx y[\text{Umfang}_2] & x[\text{Umfang}_1] &< x[\text{Umfang}_2]\\
x[\text{Bull. Corr.}_1] &\approx x[\text{Bull. Corr.}_2] & y[\text{Bull. Corr.}_1] &< y[\text{Bull. Corr.}_2]
\end{aligned}$$

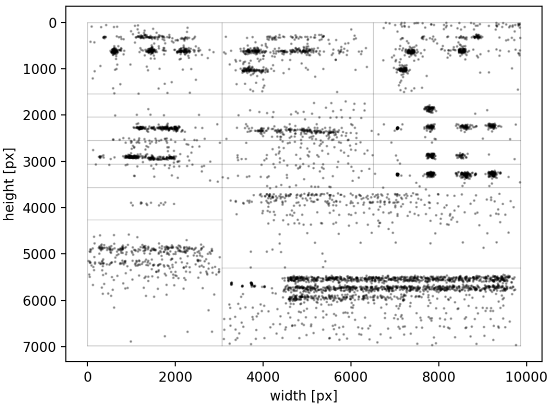


Abbildung 8: Attributwerte

2.1.5 Algorithmus

A Anhang

A.1 Python Code

A.1.1 LaTeX-Compiler

```python
#!/anaconda3/bin/python3.7
# -*- coding: utf-8 -*-
# Latex.py
# Bernard Schroffenegger
# 23th of September, 2019


""" .tex --> .pdf (pdflatex)
    - arguments: tex-file <str>, target-directory <str> (opt), (number of compilations <int>)
    - usage: $ python Latex.py example.tex ../ 3 """

import os, sys, subprocess
import webbrowser
from pathlib import Path


def main():

    # simple arg-parser
    runs = int(sys.argv[3]) if len(sys.argv) == 4 else 1
    output_dir = sys.argv[2] if len(sys.argv) > 2 else ’’
    if len(sys.argv) > 1:
        compile(sys.argv[1], output_dir, runs)


def compile(input_file, output_dir, runs, cleanup=[".log", ".aux", ".out"]):
    """ transforms *.tex into *.pdf, delete pdflatex-doc-files, and open the output
    :param input_file: <str> Path (document)
    :param output_dir: <str> Path (folder)
    :param runs: <int> number of compilations
    :param cleanup: list(str) delete by-product files
    :return: - """

    for _ in range(runs):
        if os.path.exists(input_file):
            pdf = ’-output-directory=’ + output_dir if output_dir else ’’
            cmd = [’pdflatex’, ’-halt-on-error’, ’-interaction’, ’batchmode’, pdf, input_file]
            subprocess.Popen(cmd).communicate()

    for ext in cleanup:
        os.unlink(output_dir + ’/’ + input_file.split(’.’)[0] + ext)

    webbrowser.open_new("file://"+str(Path(output_dir+’’+input_file.split(’.’)[0]+’.pdf’).absolute()))


if __name__ == ’__main__’:

    main()
```

compiler.py

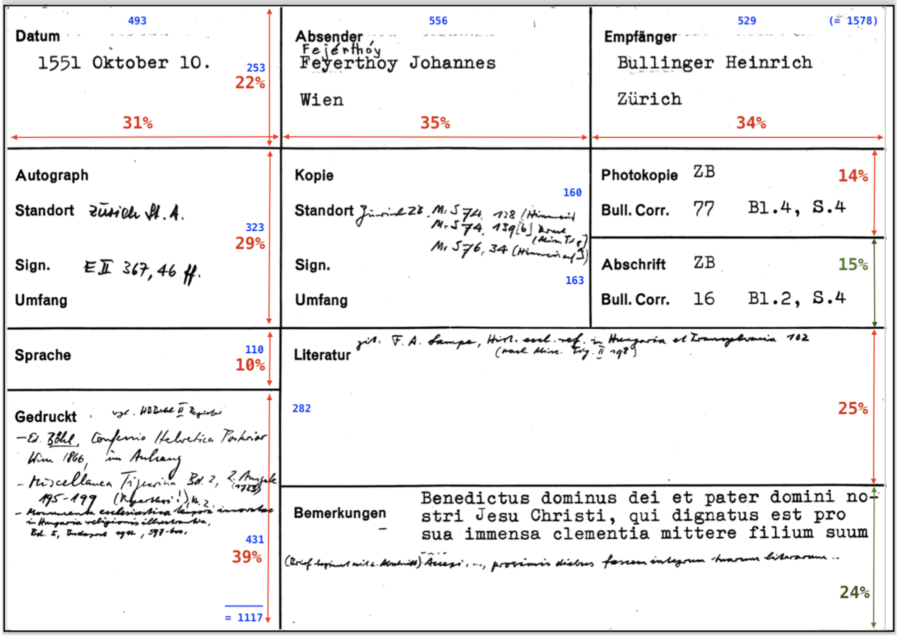
A.1.2 XML-Analysen

```python
#!/anaconda3/bin/python3.7
# -*- coding: utf-8 -*-
# xml.py
# Bernard Schroffenegger
# 24th of September, 2019

""" analyzing XML files """

import statistics
import xml.sax
import pandas as pd
import matplotlib.pyplot as plt

from xml.sax.handler import ContentHandler
from Tools.FileSystem import FileSystem
from Tools.Dictionaries import CountDict
from Tools.Dictionaries import ListDict


class Analyzer4XML:

    PRECISION = 3  # rounding
    SCHEMA = ["Attribut", "Mittelwert", "Standardabweichung"]
    SCATTER_V1, SCATTER_V2 = "scatter_v1.png", "scatter_v2.png"
    X_DISTRIBUTION_V1, Y_DISTRIBUTION_V1 = "x_distribution_v1.png", "y_distribution_v1.png"
    X_DISTRIBUTION_V2, Y_DISTRIBUTION_V2 = "x_distribution_v2.png", "y_distribution_v2.png"
    FIELDS = "fields.png"

    ATTRIBUTES = ["Datum", "Absender", "Empfänger", "Autograph", "Kopie", "Photokopie",
                  "Standort", "Bull.", "Corr.", "Sign.", "Abschrift", "Umfang", "Sprache",
                  "Literatur", "Gedruckt", "Bemerkungen"]

    def __init__(self):
        pass

    @staticmethod
    def compute_avg_page_dim(dir_path, dir_out, filter=False):
        """ mean/std-dev of page height & width
            :param dir_path: <string>
            :param dir_out: <string>
            :return: <pd.DataFrame> """
        out = dir_out+"page_size_adjusted.png" if filter else dir_out+"page_size.png"
        d = ListDict()
        x_max, y_max, x_min, y_min = -1, -1, 10**4, 10**4
        for path in FileSystem.get_file_paths(dir_path):
            dims = BullingerPage.get_dimensions(path)
            if filter and dims[’x’][0] < 9000 and dims[’x’][0] < 6000:
                print("Found:", dims[’x’][0], dims[’y’][0], path)
                continue
            x_max = dims[’x’][0] if dims[’x’][0] > x_max else x_max
            y_max = dims[’y’][0] if dims[’y’][0] > y_max else y_max
            x_min = dims[’x’][0] if dims[’x’][0] < x_min else x_min
            y_min = dims[’y’][0] if dims[’y’][0] < y_min else y_min
            d = ListDict.combine([dims, d])
        data = Analyzer4XML.compute_stats(d)
        data[’Minimum’] = [y_min, x_min]
        data[’Maximum’] = [y_max, x_max]
        data.set_index(’Mittelwert’)
        # print(data.to_latex(index=False))
        plt.scatter(d[’x’], d[’y’], alpha=0.7, s=len(d[’y’]) * [10], color="blue")
        plt.scatter(list(data[’Mittelwert’])[1], list(data[’Mittelwert’])[0], alpha=0.4, s=len(d[’y’]) * [200], color="red")
        plt.xlabel(’x [px]’)
        plt.ylabel(’y [px]’)
        fig = plt.gcf()  # get current figure
        if dir_out:  # write to file
            plt.draw()
            fig.savefig(out, dpi=100)
        plt.show()

    @staticmethod
    def create_plots_for_attributes(dir_path, out_path=None):
        for j in range(0, 4):
            fig = plt.figure()
            for i, attribute in enumerate(Analyzer4XML.ATTRIBUTES[0+4*j:4+4*j]):
                ld = ListDict()
                plt.subplot(2, 2, i + 1)
                for path in FileSystem.get_file_paths(dir_path):
                    ld = ListDict.combine([ld, BullingerAttributes.get_attribute_coordinates(path, attribute)])
                plt.scatter(ld[’x’], ld[’y’], alpha=0.1, s=len(ld[’x’]) * [100], color=’blue’)
                Analyzer4XML.draw_fields(plt)
                fig.add_subplot(2, 2, i + 1)
                # plt.ylabel("y [px]")
                # plt.xlabel("x [px]")
                plt.xticks([])
                plt.yticks([])
                axes = plt.gca()
                axes.set_xlim([0, 9903])
                axes.set_ylim([0, 7013])
                plt.ylim(plt.ylim()[::-1])  # reverse y-axis
                plt.title(attribute)
            if out_path:
                plt.draw()
                fig.savefig(out_path+"attributes_"+str(j), dpi=100)
            plt.show()

    @staticmethod
    def get_text_coordinates(dir_path_in, version=1):
        parser = BPV1 if version is 1 else BPV2
        data = pd.DataFrame({’x’: [], ’y’: []})
        for path in FileSystem.get_file_paths(dir_path_in):
            df = parser.get_coordinates(path)
            data = pd.concat([data, df])
        return data

    @staticmethod
    def get_attribute_name(hpos_t, vpos_r, height_b, width_l, version=2):
        """ key: position --> value: (attribute name, index)
            :param hpos_t: <int>
            :param vpos_r: <int>
            :param height_b: <int>
            :param width_l: <int>
            :param version: <1|2>: (top/right/bottom/left) || (top/left, height/width) """
        if version is 2:
            mx, my = int(hpos_t + 0.5 * width_l), int(vpos_r + 0.5 * height_b)  # mass point
        else:  # version 1
            mx, my = int((hpos_t+height_b)/2), int((vpos_r+width_l)/2)
        if mx <= 3057:  # 1st column
            if my <= 1535: return "Datum", None
            elif my <= 2041: return "Autograph", None
            elif my <= 2547: return "Standort", ’A’
            elif my <= 3053: return "Sign.", ’A’
            elif my <= 3559: return "Umfang", ’A’
            elif my <= 4257: return "Sprache", None
            else: return "Gedruckt", None
        elif mx <= 6508:  # 2nd column
            if my <= 1535: return "Absender", None
            elif my <= 2041: return "Kopie", None
            elif my <= 2547: return "Standort", ’B’
            elif my <= 3053: return "Sign.", ’B’
            elif my <= 3559: return "Umfang", ’B’
            elif my <= 5303: return "Literatur", None
            else: return "Bemerkungen", None
        else:  # 3rd column
            if my <= 1535: return "Empfänger", None
            elif my <= 2041: return "Photokopie", None
            elif my <= 2547: return "Bull. Corr.", 'A'
            elif my <= 3053: return "Abschrift", None
            elif my <= 3559: return "Bull. Corr.", 'B'
            elif my <= 5303: return "Literatur", None
            else: return "Bemerkungen", None


    @staticmethod
    def calculate_element_stats(dir_path):
        """ computes mean & standard deviation (element counts) over multiple files
            :param dir_path: <string>. Directory with xml-files
            :return: <pd.DataFrame> """
        count_dicts = [ElementCounter.count(path) for path in FileSystem.get_file_paths(dir_path)]
        data = Analyzer4XML.compute_stats(ListDict.combine(count_dicts))
        print(data.to_latex())

    @staticmethod
    def compute_stats(list_dict):
        """ computes averages and standard deviations
            :param list_dict: key <string> (classifier) --> value <num-list> (data points)
            :return: <DataFrame> """
        s = Analyzer4XML.SCHEMA
        data = pd.DataFrame(columns=s)
        for key in list_dict:
            mean = round(sum(list_dict[key])/len(list_dict[key]), Analyzer4XML.PRECISION)
            std_dev = round(statistics.stdev(list_dict[key]), Analyzer4XML.PRECISION)
            data = pd.concat([data, pd.DataFrame({s[0]: [key], s[1]: [mean], s[2]: std_dev})])
        return data

    @staticmethod  # OCR-V1
    def calculate_focus_points_v1(dir_path_in, dir_path_out):
        """ OCR-mass points (x,y) of all xml-files in <dir_path>
            :param dir_path_in: <string>
            :param dir_path_out: <string>
            :return: show/save plot """
        data = Analyzer4XML.get_text_coordinates(dir_path_in, version=1)
        Analyzer4XML.draw_scatter_plot(
            data['x'].to_list(), data['y'].to_list(),
            out_dir=dir_path_out+Analyzer4XML.SCATTER_V1)
        Analyzer4XML.draw_histogram(data['x'], 'x', out_dir=dir_path_out + Analyzer4XML.X_DISTRIBUTION_V1)
        Analyzer4XML.draw_histogram(data['y'], 'y', out_dir=dir_path_out + Analyzer4XML.Y_DISTRIBUTION_V1)

    @staticmethod  # OCR-V2
    def calculate_focus_points_v2(dir_path_in, dir_path_out):
        data = Analyzer4XML.get_text_coordinates(dir_path_in, version=2)
        Analyzer4XML.draw_scatter_plot(
            data['x'].to_list(), data['y'].to_list(),
            out_dir=dir_path_out+Analyzer4XML.SCATTER_V2)
        Analyzer4XML.draw_histogram(data['x'], 'x', out_dir=dir_path_out + Analyzer4XML.X_DISTRIBUTION_V2)
        Analyzer4XML.draw_histogram(data['y'], 'y', out_dir=dir_path_out + Analyzer4XML.Y_DISTRIBUTION_V2)

    @staticmethod
    def plot_fields(dir_path_in, dir_path_out):
        data = Analyzer4XML.get_text_coordinates(dir_path_in, version=2)
        Analyzer4XML.draw_scatter_plot2(data['x'], data['y'], out_dir=dir_path_out + Analyzer4XML.FIELDS)

    @staticmethod
    def determine_gaps(dir_path_in, version=2, dir_out=None):
        data = Analyzer4XML.get_text_coordinates(dir_path_in, version=version)
        x = data['y'].to_list()
        df = pd.DataFrame(columns=['i', 'y'])
        ranges = list(range(200, 500))+list(range(500, 750))+list(range(750, 1000, 5))\
                 + list(range(1000, 1500, 10))+list(range(1500, 2000, 15))+list(range(2000, 2500, 20))
        for n_bins in ranges:
            print(n_bins)
            ns, bins, bars = plt.hist(x, n_bins)
            plt.close()
            for i, n in enumerate(ns):
                if int(n) is 0:
                    d = pd.DataFrame({'i': [n_bins], 'y': [int((bins[i]+bins[i+1])/2)]})
                    df = pd.concat([df, d])
                    # df = df.reset_index()
        plt.scatter(df['y'], df['i'], alpha=0.4, s=len(df['i']) * [1])  # corrected
        plt.xlabel('Koordinate y [px]')
        plt.ylabel('#Buckets [IN]')
        fig = plt.gcf()  # get current figure
        if dir_out:  # write to file
            plt.draw()
            fig.savefig(dir_out+"gaps_y.png", dpi=100)
        plt.show()

    @staticmethod
    def compute_average_attribute_coordinates(dir_in, dir_out=None):
        """ data/plots
            :param dir_in: <string>. Path
            :param dir_out: <string>. Path
            :return: 2x <DataFrame>, [Attributname, mean, stddev] (für x/y)"""

        # All OCR-text Elements
        c = Analyzer4XML.get_text_coordinates(dir_in, version=2)

        # Attribute Coordinates: Mean & Standard Deviation
        l_dicts = [BPV2Attributes.get_data(path) for path in FileSystem.get_file_paths(dir_in)]
        l_dict = ListDict.combine(l_dicts)
        x_dict = {key: tuple(zip(*l_dict[key]))[0] for key in l_dict}
        y_dict = {key: tuple(zip(*l_dict[key]))[1] for key in l_dict}
        x_stats, y_stats = Analyzer4XML.compute_stats(x_dict), Analyzer4XML.compute_stats(y_dict)
        x_e, y_e = pd.DataFrame(columns=Analyzer4XML.SCHEMA), pd.DataFrame(columns=Analyzer4XML.SCHEMA)

        # Biased Data
        for e in ['Standort', 'Sign.', 'Umfang', 'Bull.', 'Corr.']:
            x_e = pd.concat([x_e, x_stats[x_stats.Attribut == e]])
            y_e = pd.concat([y_e, y_stats[y_stats.Attribut == e]])
            x_stats = x_stats[x_stats.Attribut != e]
            y_stats = y_stats[y_stats.Attribut != e]

        # Linear Separation
        dict_lx, dict_ly, dict_rx, dict_ry = ListDict(), ListDict(), ListDict(), ListDict()
        for e in [’Standort’, ’Sign.’, ’Umfang’]:  # vertically (left/right)
            data_l = tuple(zip(*[pair for pair in l_dict[e] if pair[0] < 2000 < pair[1]]))
            data_r = tuple(zip(*[pair for pair in l_dict[e] if pair[0] >= 2000 and pair[1] > 2000]))
            dict_lx[e], dict_ly[e], dict_rx[e], dict_ry[e] = data_l[0], data_l[1], data_r[0], data_r[1]
        dict_tx, dict_ty, dict_bx, dict_by = ListDict(), ListDict(), ListDict(), ListDict()
        for e in [’Bull.’, ’Corr.’]:  # horizontally (top/bottom)
            data1 = tuple(zip(*[pair for pair in l_dict[e] if pair[1] < 3000 and pair[0] > 6000]))
            data2 = tuple(zip(*[pair for pair in l_dict[e] if pair[1] >= 3000 and pair[0] > 6000]))
            dict_tx[e], dict_ty[e], dict_bx[e], dict_by[e] = data1[0], data1[1], data2[0], data2[1]

        # Corrected Attributes: Mean & Standard Deviation
        lx, ly = Analyzer4XML.compute_stats(dict_lx), Analyzer4XML.compute_stats(dict_ly)
        rx, ry = Analyzer4XML.compute_stats(dict_rx), Analyzer4XML.compute_stats(dict_ry)
        tx, ty = Analyzer4XML.compute_stats(dict_tx), Analyzer4XML.compute_stats(dict_ty)
        bx, by = Analyzer4XML.compute_stats(dict_bx), Analyzer4XML.compute_stats(dict_by)

        # Plots
        m, c1, c2, c3, c4 = ’Mittelwert’, ’cornflowerblue’, ’slategrey’, ’black’, ’green’
        # Analyzer4XML.plot_attributes(coords[’x’], coords[’y’], x_stats, y_stats, x_e, y_e, dir_out=dir_out)
        plt.scatter(c[’x’], c[’y’], alpha=0.1, s=len(c[’x’].to_list())*[1])  # all
        plt.scatter(x_stats[m], y_stats[m], alpha=0.7, s=len(x_stats[m])*[100], color=c3)  # mean
        # plt.scatter(x_e[m], y_e[m], alpha=0.4, s=len(y_e[m])*[100], color=’red’)  # errors
        plt.scatter(lx[m], ly[m], alpha=0.7, s=len(lx[m])*[100], color=c4)  # corrected
        plt.scatter(rx[m], ry[m], alpha=0.7, s=len(rx[m])*[100], color=c4)
        plt.scatter(tx[m], ty[m], alpha=0.7, s=len(tx[m])*[100], color=c4)
        plt.scatter(bx[m], by[m], alpha=0.7, s=len(bx[m])*[100], color=c4)

        # Output
        plt.xlabel(’x [px]’)
        plt.ylabel(’y [px]’)
        plt.ylim(plt.ylim()[::-1])  # reverse y-axis
        fig = plt.gcf()  # get current figure
        if dir_out:  # write to file
            plt.draw()
            fig.savefig(dir_out+"ocr_attributes_2.png", dpi=100)
        plt.show()

        # Final Data
        # print(pd.concat([x_stats, lx, rx, tx, bx]).to_latex(index=None))
        # print(pd.concat([y_stats, ly, ry, ty, by]).to_latex(index=None))
        return pd.concat([x_stats, lx, rx, tx, bx]), pd.concat([y_stats, ly, ry, ty, by])

    @staticmethod
    def plot_attributes(x, y, x_stats, y_stats, x_e, y_e, dir_out=None):
        plt.scatter(x, y, alpha=0.2, s=len(x.to_list())*[3])
        plt.scatter(x_stats[’Mittelwert’], y_stats[’Mittelwert’], alpha=1, s=len(x_stats[’Mittelwert’]) * [100], color=’black’
     )
        plt.scatter(x_e[’Mittelwert’], y_e[’Mittelwert’], alpha=0.4, s=len(y_e[’Mittelwert’]) * [100], color=’red’)
        plt.ylim(plt.ylim()[::-1])  # reverse y-axis
        plt.xlabel(’x [px]’)
        plt.ylabel(’y [px]’)
        fig = plt.gcf()  # get current figure
        if dir_out:
            plt.draw()
            fig.savefig(dir_out+"ocr_attributes_1.png", dpi=100)
        plt.show()

    @staticmethod
    def draw_scatter_plot(x, y, out_dir=None):
        plt.scatter(x, y, alpha=0.5, s=len(x)*[10])
        plt.ylim(plt.ylim()[::-1])  # reverse y-axis
        plt.xlabel(’x [px]’)
        plt.ylabel(’y [px]’)
        Analyzer4XML.draw_fields(plt)
        fig = plt.gcf()  # get current figure
        if out_dir:
            plt.draw()
            fig.savefig(out_dir, dpi=100)
        plt.show()

    @staticmethod
    def draw_scatter_plot2(x, y, out_dir=None):
        plt.scatter(x, y, alpha=0.5, s=len(x)*[10])
        plt.ylim(plt.ylim()[::-1])  # reverse y-axis
        plt.xlabel(’x [px]’)
        plt.ylabel(’y [px]’)
        Analyzer4XML.draw_fields(plt)
        fig = plt.gcf()  # get current figure
        if out_dir:
            plt.draw()
            fig.savefig(out_dir+Analyzer4XML.FIELDS, dpi=100)
        plt.show()

    @staticmethod
    def draw_fields(plt):
        x0, x1, x2, x3 = 0, 3057, 6508, 9860
        y0, y1, y2, y3, y4, y5, y6, y7, y8 = 0, 1535, 2041, 2547, 3053, 3559, 4257, 5303, 6978

        # Vertical Lines
        plt.plot((x0, x0), (y0, y8), ’k-’, alpha=0.3)
        plt.plot((x1, x1), (y0, y8), ’k-’, alpha=0.3)
        plt.plot((x2, x2), (y0, y8), ’k-’, alpha=0.3)
        plt.plot((x3, x3), (y0, y8), ’k-’, alpha=0.3)

        # Horizontal Lines
        plt.plot((x0, x3), (y0, y0), ’k-’, alpha=0.3)  # top
        plt.plot((x0, x3), (y1, y1), ’k-’, alpha=0.3)
        plt.plot((x0, x3), (y2, y2), ’blue’, alpha=0.3)
        plt.plot((x0, x3), (y3, y3), ’blue’, alpha=0.3)
        plt.plot((x0, x3), (y4, y4), ’blue’, alpha=0.3)
        plt.plot((x0, x3), (y5, y5), ’k-’, alpha=0.3)
        plt.plot((x0, x1), (y6, y6), ’k-’, alpha=0.3)
        plt.plot((x1, x3), (y7, y7), ’k-’, alpha=0.3)
        plt.plot((x0, x3), (y8, y8), ’k-’, alpha=0.3)

    @staticmethod
    def draw_histogram(x, x_name, out_dir=None):
        fig = plt.figure()
        for i, n_bins in enumerate([10**i for i in range(1, 5)]):
            plt.subplot(2, 2, i+1)
            plt.hist(x, n_bins, facecolor=’green’, alpha=0.5)
            p = fig.add_subplot(2, 2, i+1)
            p.title.set_text(str(n_bins)+" Buckets")
            if i < 2:
                plt.xticks([])
            if i is 0 or i is 2:
                plt.ylabel("Frequency")
            if i is 2 or i is 3:
                plt.xlabel(x_name + " [px]")
        if out_dir:
            plt.draw()
            fig.savefig(out_dir, dpi=100)
        plt.show()


class ElementCounter(ContentHandler):

    """ counts XML-elements """

    def __init__(self):
        super(ElementCounter, self).__init__()
        self.elements = CountDict()
        # self.attributes = CountDict()
        # self.values = CountDict()

    def startElement(self, name, attributes):
        self.elements.add(name)
        # for a in attributes:
        #    self.attributes[a] += 1
        #    self.values[attributes[a]] += 1

    @staticmethod
    def count(path):
        """ elements and their frequencies
            :param path: <string> (xml-file)
            :return: <CountingDict> """
        try:
            parser = xml.sax.make_parser()
            handler = ElementCounter()
            parser.setContentHandler(handler)
            parser.parse(path)
            return handler.elements
        except (AttributeError, TypeError):
            print("Warning: Parser failed on", path)
            return None


# Bullinger Parser  V2
class BPV2(ContentHandler):

    """ Elements:
            <String CONTENT="Johannes" HEIGHT="152" WIDTH="960" VPOS="554" HPOS="4526"/>
        --> mass points (x, y) """

    def __init__(self):
        super(BPV2, self).__init__()
        self.data = pd.DataFrame({’x’: [], ’y’: []})

    def startElement(self, name, attributes):
        if name == "String" and "STYLE" not in attributes.getNames():
            hpos, vpos, height, width = 0, 0, 0, 0
            for a in attributes.getNames():
                if a == "HPOS":
                    hpos = int(attributes.getValue(a))
                elif a == "VPOS":
                    vpos = int(attributes.getValue(a))
                elif a == "HEIGHT":
                    height = int(attributes.getValue(a))
                elif a == "WIDTH":
                    width = int(attributes.getValue(a))
            x, y = int(hpos + 0.5*width), int(vpos + 0.5*height)
            data = pd.DataFrame({’x’: [x], ’y’: [y]})
            self.data = pd.concat([self.data, data])

    @staticmethod
    def get_coordinates(path):
        try:
            parser = xml.sax.make_parser()
            counter = BPV2()
            parser.setContentHandler(counter)
            parser.parse(path)
            return counter.data
        except (AttributeError, TypeError):
            print("Warning: xml-sax-parser failed on", path)
            return None


class BPV1(ContentHandler):

    def __init__(self):
        super(BPV1, self).__init__()
        self.data = pd.DataFrame({’x’: [], ’y’: []})
        self._charBuffer = []
        self._result = []
        self.bool = False

    def startElement(self, name, attributes):
        if name == "line":
            self.t, self.l, self.r, self.b = 0, 0, 0, 0
            for a in attributes.getNames():
                if a == "t":
                    self.t = int(attributes.getValue(a))
                elif a == "r":
                    self.r = int(attributes.getValue(a))
                elif a == "b":
                    self.b = int(attributes.getValue(a))
                elif a == "l":
                    self.l = int(attributes.getValue(a))

    def endElement(self, name):
        if name == ’line’:
            data = pd.DataFrame({’x’: [int((self.r+self.l)/2)], ’y’: [int((self.b+self.t)/2)]})
            self.data = pd.concat([self.data, data])

    def _getCharacterData(self):
        data = ’’.join(self._charBuffer).strip()
        self._charBuffer = []
        return data.strip()

    def characters(self, data):
        self._charBuffer.append(data)

    @staticmethod
    def get_coordinates(path):
        try:
            parser = xml.sax.make_parser()
            counter = BPV1()
            parser.setContentHandler(counter)
            parser.parse(path)
            return counter.data
        except (AttributeError, TypeError):
            print("Warning: parser failed on", path)
            return None


class BPV2Attributes(ContentHandler):

    """ Elements:
            <String CONTENT="Johannes" HEIGHT="152" WIDTH="960" VPOS="554" HPOS="4526"/>
        --> avg. (x, y) = f(attribute_name)  """

    NAMES = ["Datum", "Absender", "Empfänger", "Autograph", "Kopie", "Photokopie",
             "Standort", "Bull.", "Corr.", "Sign.", "Abschrift", "Umfang", "Sprache",
             "Literatur", "Gedruckt", "Bemerkungen"]

    def __init__(self):
        super(BPV2Attributes, self).__init__()
        self.l_dict = ListDict()

    def startElement(self, name, attributes):
        if name == "String" and "STYLE" not in attributes.getNames():
            key, value, hpos, vpos, height, width = None, None, 0, 0, 0, 0
            for a in attributes.getNames():
                key = attributes.getValue(a)
                if a == "CONTENT" and key in self.NAMES:
                    value = str(key)
                elif a == "HPOS":
                    hpos = int(key)
                elif a == "VPOS":
                    vpos = int(key)
                elif a == "HEIGHT":
                    height = int(key)
                elif a == "WIDTH":
                    width = int(key)
            if key is not None and value is not None:
                x, y = BPV2Attributes.get_mass_point(hpos, vpos, width, height)
                self.l_dict.add(value, (x, y))

    @staticmethod
    def get_mass_point(hpos, vpos, width, height):
        return int(hpos + 0.5*width), int(vpos + 0.5*height)

    @staticmethod
    def get_data(path):
        try:
            parser = xml.sax.make_parser()
            counter = BPV2Attributes()
            parser.setContentHandler(counter)
            parser.parse(path)
            return counter.l_dict
        except (AttributeError, TypeError):
            print("Warning: parser failed on", path)
            return None


class BullingerPage(ContentHandler):

    """ Computes avg page dimensions (x_may, y_may) [px] """

    def __init__(self, path):
        super(BullingerPage, self).__init__()
        self.l_dict = ListDict()  # x, y
        self.path = path

    def startElement(self, name, attributes):
        if name == "Page":
            for a in attributes.getNames():
                if a == "WIDTH":
                    self.l_dict.add(’x’, int(attributes.getValue(a)))
                if a == "HEIGHT":
                    self.l_dict.add(’y’, int(attributes.getValue(a)))
                    if int(attributes.getValue(a)) == 3488:
                        print(self.path)

    @staticmethod
    def get_dimensions(path):
        try:
            parser = xml.sax.make_parser()
            counter = BullingerPage(path)
            parser.setContentHandler(counter)
            parser.parse(path)
            return counter.l_dict
        except (AttributeError, TypeError):
            print("Warning: parser failed on", path)
            return None


class BullingerAttributes(ContentHandler):

    """ Computes avg page dimensions (x_may, y_may) [px] """

    def __init__(self, path, attr):
        super(BullingerAttributes, self).__init__()
        self.l_dict = ListDict()  # x, y
        self.path = path
        self.attr = attr

    def startElement(self, name, attributes):
        if name == "String":
            key, value, hpos, vpos, height, width = None, None, 0, 0, 0, 0
            for a in attributes.getNames():
                key = attributes.getValue(a)
                if a == "CONTENT" and key == self.attr:
                    value = str(key)
                elif a == "HPOS":
                    hpos = int(key)
                elif a == "VPOS":
                    vpos = int(key)
                elif a == "HEIGHT":
                    height = int(key)
                elif a == "WIDTH":
                    width = int(key)
            if value:
                x, y = BPV2Attributes.get_mass_point(hpos, vpos, width, height)
                self.l_dict.add(’x’, x)
                self.l_dict.add(’y’, y)

    @staticmethod
    def get_attribute_coordinates(path, attr_name):
        try:
            parser = xml.sax.make_parser()
            counter = BullingerAttributes(path, attr_name)
            parser.setContentHandler(counter)
            parser.parse(path)
            return counter.l_dict
        except (AttributeError, TypeError):
            print("Warning: parser failed on", path)
            return None
```

../../Tools/xml.py

A.2 Screenshots

A.2.1 Karteikarte (Original)



Datum
1551 Oktober 10.

Absender
Fejérthóy
Feyerthoy Johannes
Wien

Empfänger
Bullinger Heinrich
Zürich

Autograph
Standort Zürich St.A.
Sign. E II 367, 46 ff.
Umfang

Kopie
Standort Zürich ZB, Ms S 74, 138 (Hinweis)
Ms S 74, 139[b] Druck (Misc. Tig.)
Ms S 76, 34 (Hinweis auf 3)
Sign.
Umfang

Photokopie ZB
Bull. Corr. 77 Bl.4, S.4

Abschrift ZB
Bull. Corr. 16 Bl.2, S.4

Sprache

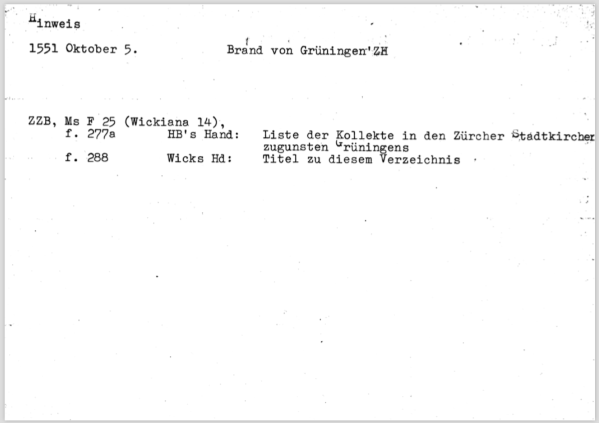
Literatur zit. F. A. Lampe, Hist. eccl. ref. in Hungaria et Transylvania 102 (nach Misc. Tig. II 198)

Gedruckt vgl. HBBibl II Register
– Ed. Böhl, Confessio Helvetica Posterior Wien 1866, im Anhang
– Miscellanea Tigurina Bd. 2, 2. Ausgabe (1723) 195–199 (Fejerthoi!) Nr. 2
– Monumenta ecclesiastica tempora innovatae in Hungaria religionis illustrantia, Bd. 5, Budapest 1912, 597–600,

Bemerkungen Benedictus dominus dei et pater domini nostri Jesu Christi, qui dignatus est pro sua immensa clementia mittere filium suum
(Brief beginnt mit d. Abschnitt) Accepi, ..., proximis diebus formam integram tuarum literarum ..

Abbildung 9: Sammlung von Karteikarten (Bilder) im pdf-Format (HBBW_1551_100), S. 13/99

A.2.2 Karteikarte (Spezialfall)



Hinweis

1551 Oktober 5. Brand von Grüningen'ZH

```
ZZB, Ms F 25 (Wickiana 14),
     f. 277a      HB's Hand:   Liste der Kollekte in den Zürcher Stadtkircher
                               zugunsten Grüningens
     f. 288       Wicks Hd:    Titel zu diesem Verzeichnis
```

Abbildung 10: Sammlung von Karteikarten (Bilder) im pdf-Format (HBBW_1551_100), S. 13/99

A.3 OCR-Output

A.3.1 Version 1

Beispiel: Karteikarten_HBBW_1551_1000012.xml
Schema: https://fr7.abbyy.com/FineReader_xml/FineReader10-schema-v1.xml
Formatter: https://www.freeformatter.com/html-formatter.html#ad-output

```
<document xmlns="http://www.abbyy.com/FineReader_xml/FineReader10-schema-v1.xml" version="1.0" producer=""
    languages="" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.abbyy
    .com/FineReader_xml/FineReader10-schema-v1.xml http://www.abbyy.com/FineReader_xml/FineReader10-schema-
    v1.xml">
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<documentData>
  <sections>
    <section>
```

```
         <stream role="text" beginPage="0">
            <mainText columnCount="1"/>
            <elemId id="{5602AE7C-9916-4028-A87C-D0C8717F904B}"/>
         </stream>
      </section>
   </sections>
</documentData>
<page width="9851" height="6994" resolution="1200">
   <block blockType="Table" pageElemId="{5602AE7C-9916-4028-A87C-D0C8717F904B}" l="4" t="0" r="9848" b="6994">
      <region>
         <rect l="4" t="0" r="9848" b="6994"/>
      </region>
      <row>
         <cell leftBorder="White" topBorder="White" width="3068" height="1592">
            <text id="{50FDE2C2-8C1C-4D1A-9E6D-D245E6B9DA23}">
               <par align="Justified" lineSpacing="1360">
                  <line baseline="388" l="110" t="258" r="2042" b="394">
                     <formatting lang="GermanNewSpelling">
                        Datum.....................
                     </formatting>
                  </line>
               </par>
               <par leftIndent="2100" lineSpacing="1360">
                  <line baseline="702" l="358" t="550" r="2270" b="722">
                     <formatting lang="GermanNewSpelling">
                        1551 Oktober 10.
                     </formatting>
                  </line>
               </par>
            </text>
         </cell>
         <cell topBorder="White" width="3476" height="1592">
            <text id="{435055F0-FC4E-4199-8D97-6216683C70B8}">
               <par align="Justified" leftIndent="1100" lineSpacing="2720">
                  <line baseline="560" l="3244" t="172" r="5940" b="588">
                     <formatting lang="GermanNewSpelling">
                        ttWfRAT ......    apos;
                     </formatting>
                  </line>
               </par>
               <par align="Justified" leftIndent="1100" lineSpacing="1360">
                  <line baseline="704" l="3290" t="550" r="5486" b="746">
                     <formatting lang="GermanNewSpelling">
                        Feyerthoy Johannes
                     </formatting>
                  </line>
               </par>
               <par align="Justified" leftIndent="1100" lineSpacing="1360">
                  <line baseline="1112" l="3290" t="970" r="3782" b="1114">
                     <formatting lang="GermanNewSpelling">
                        Wien
                     </formatting>
                  </line>
               </par>
            </text>
         </cell>
         <cell topBorder="White" rightBorder="White" width="3300" height="1592">
            <text id="{A2C724D8-7590-4EAC-9EB3-039FEC098F5B}">
               <par align="Justified" leftIndent="800" lineSpacing="1360">
                  <line baseline="400" l="6714" t="270" r="8898" b="434">
                     <formatting lang="GermanNewSpelling">
                        Empfanger ...
                     </formatting>
                  </line>
               </par>
               <par leftIndent="1900" lineSpacing="1360">
                  <line baseline="700" l="6858" t="554" r="9046" b="742">
                     <formatting lang="GermanNewSpelling">Bullinger Heinrich</formatting>
                  </line>
               </par>
               <par leftIndent="1900" lineSpacing="1360">
                  <line baseline="1108" l="6858" t="962" r="7578" b="1110">
                     <formatting lang="GermanNewSpelling">Zuerich</formatting>
                  </line>
               </par>
            </text>
         </cell>
      </row>
      <row>
         <cell rowSpan="2" leftBorder="White" width="3068" height="2012">
            <text id="{25EBF6D7-8027-4493-90BC-7A297F4743EE}">
               <par align="Justified" lineSpacing="1120">
                  <line baseline="1948" l="98" t="1818" r="914" b="1982">
                     <formatting lang="GermanNewSpelling">Autograph</formatting>
                  </line>
               </par>
               <par align="Justified" lineSpacing="1250">
                  <line baseline="2347" l="98" t="2202" r="1986" b="2402">
                     <formatting lang="GermanNewSpelling">Standort . 4.</formatting>
            </line>
         </par>
         <par align="Justified" lineSpacing="1360">
            <line baseline="2972" l="94" t="2826" r="2134" b="3138">
               <formatting lang="GermanNewSpelling">Sign.    C J    </formatting>
            </line>
         </par>
         <par align="Justified" lineSpacing="1120">
            <line baseline="3348" l="98" t="3218" r="674" b="3382">
               <formatting lang="GermanNewSpelling">Umfang</formatting>
            </line>
         </par>
      </text>
   </cell>
   <cell rowSpan="2" width="3476" height="2012">
      <text id="{9DDAB928-A90D-49C0-A802-D03967F6C2B7}">
         <par align="Justified" leftIndent="1100" lineSpacing="1120">
            <line baseline="1952" l="3242" t="1822" r="3666" b="1986">
               <formatting lang="GermanNewSpelling">Kopie</formatting>
            </line>
         </par>
         <par align="Justified" leftIndent="1100" lineSpacing="1360">
            <line baseline="2346" l="3234" t="2214" r="5086" b="2350">
               <formatting lang="GermanNewSpelling">Standort    J</formatting>
            </line>
         </par>
         <par align="Justified" leftIndent="5100" lineSpacing="1360">
            <line baseline="2522" l="3918" t="2358" r="6274" b="2594">
               <formatting lang="GermanNewSpelling">lt;7    -fif/ij gt;w</formatting>
            </line>
         </par>
         <par align="Justified" leftIndent="1100" lineSpacing="1360">
            <line baseline="2956" l="3238" t="2798" r="6514" b="2986">
               <formatting lang="GermanNewSpelling">Sign.    apos;</formatting>
            </line>
         </par>
         <par align="Justified" leftIndent="1100" lineSpacing="1120">
            <line baseline="3352" l="3242" t="3218" r="3818" b="3382">
               <formatting lang="GermanNewSpelling">Umfang</formatting>
            </line>
         </par>
      </text>
   </cell>
   <cell rightBorder="White" width="3300" height="1000">
      <text id="{43A62CAC-F01E-4F13-8868-8F448A23D4D4}">
         <par align="Justified" leftIndent="800" lineSpacing="1120">
            <line baseline="1950" l="6678" t="1782" r="7946" b="1986">
               <formatting lang="GermanNewSpelling">Photokopie ZB</formatting>
            </line>
         </par>
         <par align="Justified" leftIndent="800" lineSpacing="1360">
            <line baseline="2339" l="6678" t="2178" r="9418" b="2366">
               <formatting lang="GermanNewSpelling">Bull.Corr. 77    Bl.4, S.4</formatting>
            </line>
         </par>
      </text>
   </cell>
</row>
<row>
   <cell rightBorder="White" width="3300" height="1012">
      <text id="{9E6CEBDE-D503-435E-977F-A050AACD92E4}">
         <par align="Justified" leftIndent="800" lineSpacing="1360">
            <line baseline="2953" l="6674" t="2802" r="7946" b="2962">
               <formatting lang="GermanNewSpelling">Abschrift ZB</formatting>
            </line>
         </par>
         <par align="Justified" leftIndent="800" lineSpacing="1120">
            <line baseline="3347" l="6678" t="3194" r="9418" b="3382">
               <formatting lang="GermanNewSpelling">Bull.Corr. 16    Bl.2, S.4</formatting>
            </line>
         </par>
      </text>
   </cell>
</row>
<row>
   <cell rowSpan="2" leftBorder="White" bottomBorder="White" width="3068" height="3390">
      <text id="{2BA9D218-5AE0-443D-9DDD-3EC756237D37}">
         <par align="Justified" lineSpacing="1120">
            <line baseline="3964" l="90" t="3834" r="718" b="3998">
               <formatting lang="GermanNewSpelling">Sprache</formatting>
            </line>
         </par>
         <par align="Justified" lineSpacing="1120">
            <line baseline="4656" l="90" t="4526" r="950" b="4662">
               <formatting lang="GermanNewSpelling">Gedruckt</formatting>
            </line>
         </par>
         <par align="Justified" leftIndent="1500" lineSpacing="1250">
            <line baseline="4889" l="230" t="4746" r="2314" b="4906">
               <formatting lang="GermanNewSpelling">amp;.    c //d*rlt;-Arcc*</formatting>
```

```
                  </line>
               </par>
               <par align="Justified" lineSpacing="1360">
                  <line baseline="5383" l="114" t="5114" r="2370" b="5414">
                     <formatting lang="GermanNewSpelling">˜    /iw/ amp;*,</formatting>
                  </line>
               </par>
               <par leftIndent="11400" lineSpacing="1250">
                  <line baseline="5568" l="1907" t="5405" r="2259" b="5617">
                     <formatting lang="GermanNewSpelling">!h </formatting>
                  </line>
               </par>
               <par align="Justified" lineSpacing="510">
                  <line baseline="5695" l="90" t="5592" r="990" b="5705">
                     <formatting lang="GermanNewSpelling">ArK4Wuy4</formatting>
                  </line>
               </par>
               <par align="Justified" leftIndent="1500" lineSpacing="1360">
                  <line baseline="5934" l="267" t="5837" r="2019" b="5965">
                     <formatting lang="GermanNewSpelling"> xxx </formatting>
                  </line>
               </par>
            </text>
         </cell>
         <cell colSpan="2" rightBorder="White" width="6776" height="1768">
            <text id="{F58B7F07-4841-4ECD-9F23-11493F5565E4}">
               <par align="Justified" leftIndent="900" lineSpacing="1360">
                  <line baseline="3949" l="3229" t="3787" r="7059" b="3969">
                     <formatting lang="GermanNewSpelling">Literatur    r    o****-    **japos;H*i* j</formatting>
                  </line>
               </par>
            </text>
         </cell>
      </row>
      <row>
         <cell colSpan="2" rightBorder="White" bottomBorder="White" width="6776" height="1622">
            <text id="{9DF6A937-B519-4E09-BFB6-8E17A19B0B96}">
               <par leftIndent="900" startIndent="8500" lineSpacing="1134">
                  <line baseline="5572" l="4646" t="5422" r="9778" b="5610">
                     <formatting lang="GermanNewSpelling">Benedictus dominus dei et pater domini </formatting>
                     <formatting lang="GermanStandard">no</formatting>
                     <formatting lang="GermanNewSpelling"></formatting>
                  </line>
                  <line baseline="5764" l="3226" t="5610" r="9410" b="5814">
                     <formatting lang="GermanStandard">Bemerkungen</formatting>
                     <formatting lang="GermanNewSpelling">    stri Jesu Christi, qui dignatus est pro</formatting>
                  </line>
               </par>
               <par align="Justified" leftIndent="6600" lineSpacing="1360">
                  <line baseline="5973" l="4166" t="5826" r="9666" b="5986">
                     <formatting lang="GermanNewSpelling">sua immensa clementia mittere filium suum</formatting>
                  </line>
               </par>
               <par align="Justified" leftIndent="12500" lineSpacing="1360">
                  <line baseline="6231" l="5154" t="6134" r="9202" b="6318">
                     <formatting lang="GermanNewSpelling">1 r..yA..LPX1fr7imiL</formatting>
                  </line>
               </par>
            </text>
         </cell>
      </row>
   </block>
   <block blockType="Separator" l="4" t="1568" r="9844" b="1616">
      <region>
         <rect l="2232" t="1568" r="2908" b="1572"/>
         <rect l="4" t="1572" r="3244" b="1576"/>
         <rect l="4" t="1576" r="6464" b="1580"/>
         <rect l="4" t="1580" r="9568" b="1588"/>
         <rect l="4" t="1588" r="9844" b="1596"/>
         <rect l="4" t="1596" r="9844" b="1600"/>
         <rect l="3244" t="1600" r="9844" b="1604"/>
         <rect l="6464" t="1604" r="9844" b="1612"/>
         <rect l="9568" t="1612" r="9844" b="1616"/>
      </region>
      <separator type="Black" thickness="7">
         <start x="4" y="1592"/>
         <end x="9844" y="1592"/>
      </separator>
   </block>
   <block blockType="Separator" l="6532" t="2576" r="9844" b="2612">
      <region>
         <rect l="6532" t="2576" r="9548" b="2580"/>
         <rect l="6532" t="2580" r="9844" b="2608"/>
         <rect l="9548" t="2608" r="9844" b="2612"/>
      </region>
      <separator type="Black" thickness="8">
```

```
         <start x="6532" y="2594"/>
         <end x="9844" y="2594"/>
      </separator>
   </block>
   <block blockType="Separator" l="4" t="3588" r="9844" b="3624">
      <region>
         <rect l="3060" t="3588" r="8116" b="3592"/>
         <rect l="4" t="3592" r="9844" b="3620"/>
         <rect l="4" t="3620" r="3084" b="3624"/>
      </region>
      <separator type="Black" thickness="7">
         <start x="4" y="3606"/>
         <end x="9844" y="3606"/>
      </separator>
   </block>
   <block blockType="Separator" l="7756" t="3724" r="8624" b="3756">
      <region>
         <rect l="7756" t="3724" r="8516" b="3728"/>
         <rect l="7756" t="3728" r="8516" b="3736"/>
         <rect l="7816" t="3736" r="8516" b="3740"/>
         <rect l="7816" t="3740" r="8624" b="3744"/>
         <rect l="7944" t="3744" r="8624" b="3748"/>
         <rect l="8496" t="3748" r="8624" b="3752"/>
         <rect l="8520" t="3752" r="8624" b="3756"/>
      </region>
      <separator type="Black" thickness="5">
         <start x="7756" y="3740"/>
         <end x="8624" y="3740"/>
      </separator>
   </block>
   <block blockType="Separator" l="4724" t="3748" r="5252" b="3784">
      <region>
         <rect l="4724" t="3748" r="4880" b="3760"/>
         <rect l="4724" t="3760" r="5252" b="3764"/>
         <rect l="4724" t="3764" r="5252" b="3772"/>
         <rect l="4820" t="3772" r="5252" b="3776"/>
         <rect l="4820" t="3776" r="5224" b="3780"/>
         <rect l="5076" t="3780" r="5224" b="3784"/>
      </region>
      <separator type="Black" thickness="5">
         <start x="4724" y="3766"/>
         <end x="5252" y="3766"/>
      </separator>
   </block>
   <block blockType="Separator" l="4" t="4280" r="3080" b="4312">
      <region>
         <rect l="536" t="4280" r="3080" b="4284"/>
         <rect l="4" t="4284" r="3080" b="4308"/>
         <rect l="4" t="4308" r="3080" b="4312"/>
      </region>
      <separator type="Black" thickness="7">
         <start x="4" y="4296"/>
         <end x="3080" y="4296"/>
      </separator>
   </block>
   <block blockType="Separator" l="3052" t="5356" r="9840" b="5392">
      <region>
         <rect l="8988" t="5356" r="9568" b="5360"/>
         <rect l="4780" t="5360" r="9840" b="5364"/>
         <rect l="3052" t="5364" r="9840" b="5384"/>
         <rect l="3052" t="5384" r="9840" b="5388"/>
         <rect l="3052" t="5388" r="4780" b="5392"/>
      </region>
      <separator type="Black" thickness="7">
         <start x="3052" y="5374"/>
         <end x="9840" y="5374"/>
      </separator>
   </block>
   <block blockType="Separator" l="668" t="5904" r="1104" b="5928">
      <region>
         <rect l="668" t="5904" r="980" b="5908"/>
         <rect l="668" t="5908" r="1104" b="5920"/>
         <rect l="796" t="5920" r="1104" b="5928"/>
      </region>
      <separator type="Black" thickness="5">
         <start x="668" y="5916"/>
         <end x="1104" y="5916"/>
      </separator>
   </block>
   <block blockType="Separator" l="3048" t="4" r="3096" b="6984">
      <region>
         <rect l="3068" t="4" r="3096" b="424"/>
         <rect l="3064" t="424" r="3092" b="2844"/>
         <rect l="3060" t="2844" r="3088" b="3592"/>
         <rect l="3060" t="3592" r="3084" b="3616"/>
         <rect l="3052" t="3616" r="3084" b="4520"/>
         <rect l="3048" t="4520" r="3076" b="6984"/>
      </region>
      <separator type="Black" thickness="7">
         <start x="3072" y="4"/>
         <end x="3072" y="6984"/>
      </separator>
   </block>
   <block blockType="Separator" l="6532" t="8" r="6568" b="3620">
      <region>
         <rect l="6536" t="8" r="6568" b="744"/>
         <rect l="6536" t="744" r="6564" b="1580"/>
         <rect l="6532" t="1580" r="6568" b="1608"/>
         <rect l="6532" t="1608" r="6564" b="3620"/>
      </region>
      <separator type="Black" thickness="7">
         <start x="6550" y="8"/>
         <end x="6550" y="3620"/>
      </separator>
   </block>
   <block blockType="Separator" l="7708" t="320" r="7868" b="332">
      <region>
         <rect l="7708" t="320" r="7868" b="332"/>
      </region>
      <separator type="Dotted" thickness="2">
         <start x="7708" y="326"/>
         <end x="7868" y="326"/>
      </separator>
   </block>
</page>
</document>
```

A.3.2 Version 2

Beispiel: Karteikarten_HBBW_1551_1000012.xml

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<alto xmlns="http://www.loc.gov/standards/alto/ns-v2#" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi=
    "http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.loc.gov/standards/alto/ns-v2
    # http://www.loc.gov/standards/alto/alto-v2.0.xsd">
   <Description>
      <MeasurementUnit>pixel</MeasurementUnit>
      <OCRProcessing ID="IdOcr">
         <ocrProcessingStep>
            <processingDateTime>2019-09-23</processingDateTime>
            <processingSoftware>
               <softwareCreator>ABBYY</softwareCreator>
               <softwareName>ABBYY Recognition Server</softwareName>
               <softwareVersion>4.0</softwareVersion>
            </processingSoftware>
         </ocrProcessingStep>
      </OCRProcessing>
   </Description>
   <Styles>
      <ParagraphStyle ID="StyleId-FFFFFFFF-FFFF-FFFF-FFFF-FFFFFFFFFFFF-" ALIGN="Left" LEFT="0." RIGHT="0."
    FIRSTLINE="0."/>
      <ParagraphStyle ID="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-" ALIGN="Left" LEFT="0." RIGHT="0."
    FIRSTLINE="0."/>
   </Styles>
   <Layout>
      <Page ID="Page1" PHYSICAL_IMG_NR="1" HEIGHT="6982" WIDTH="9856">
         <PrintSpace HEIGHT="6982" WIDTH="9856" VPOS="0" HPOS="0">
            <ComposedBlock ID="Page1_Block1" HEIGHT="6982" WIDTH="9856" VPOS="0" HPOS="0" TYPE="table">
               <TextBlock ID="Page1_Block2" HEIGHT="500" WIDTH="3064" VPOS="0" HPOS="0" language="de"
    STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
                  <TextLine BASELINE="380" HEIGHT="136" WIDTH="1712" VPOS="250" HPOS="102">
                     <String CONTENT="Datum" HEIGHT="136" WIDTH="488" VPOS="250" HPOS="102"/>
                     <SP WIDTH="728" VPOS="342" HPOS="690"/>
                     <String CONTENT="-" HEIGHT="20" WIDTH="48" VPOS="338" HPOS="1514"/>
                     <SP WIDTH="228" VPOS="338" HPOS="1586"/>
                  </TextLine>
               </TextBlock>
               <TextBlock ID="Page1_Block3" HEIGHT="500" WIDTH="3480" VPOS="0" HPOS="3064" language="de"
    STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
                  <TextLine BASELINE="388" HEIGHT="136" WIDTH="2852" VPOS="258" HPOS="3234">
                     <String CONTENT="Absender" HEIGHT="136" WIDTH="756" VPOS="258" HPOS="3234"/>
                     <SP WIDTH="180" VPOS="354" HPOS="4002"/>
                     <String CONTENT="-" HEIGHT="28" WIDTH="48" VPOS="342" HPOS="4186"/>
                     <SP WIDTH="572" VPOS="350" HPOS="4282"/>
                     <String CONTENT="-" HEIGHT="20" WIDTH="264" VPOS="350" HPOS="4890"/>
                     <SP WIDTH="788" VPOS="350" HPOS="5218"/>
                     <String CONTENT="-" HEIGHT="16" WIDTH="36" VPOS="354" HPOS="6050"/>
                  </TextLine>
               </TextBlock>
               <TextBlock ID="Page1_Block4" HEIGHT="500" WIDTH="3312" VPOS="0" HPOS="6544" language="de"
    STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
                  <TextLine BASELINE="392" HEIGHT="164" WIDTH="2320" VPOS="266" HPOS="6710">
                     <String CONTENT="Empfaenger" HEIGHT="164" WIDTH="808" VPOS="266" HPOS="6710"/>
                     <SP WIDTH="94" VPOS="298" HPOS="7519"/>
```

```xml
                  <String CONTENT="-" HEIGHT="24" WIDTH="60" VPOS="346" HPOS="7614"/>
                  <SP WIDTH="104" VPOS="354" HPOS="7710"/>
                  <String CONTENT="-" HEIGHT="12" WIDTH="60" VPOS="354" HPOS="7822"/>
                  <SP WIDTH="624" VPOS="346" HPOS="7914"/>
                  <String CONTENT="-" HEIGHT="24" WIDTH="60" VPOS="342" HPOS="8562"/>
                  <SP WIDTH="226" VPOS="342" HPOS="8623"/>
                  <String CONTENT="--" HEIGHT="16" WIDTH="100" VPOS="350" HPOS="8850"/>
                  <SP WIDTH="68" VPOS="354" HPOS="8962"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block5" HEIGHT="380" WIDTH="3064" VPOS="500" HPOS="0" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="740" HEIGHT="176" WIDTH="1784" VPOS="586" HPOS="362">
                  <String CONTENT="1551" HEIGHT="176" WIDTH="472" VPOS="586" HPOS="362"/>
                  <SP WIDTH="130" VPOS="586" HPOS="835"/>
                  <String CONTENT="Oktober" HEIGHT="148" WIDTH="856" VPOS="586" HPOS="966"/>
                  <SP WIDTH="118" VPOS="586" HPOS="1823"/>
                  <String CONTENT="8." HEIGHT="148" WIDTH="204" VPOS="586" HPOS="1942"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block6" HEIGHT="380" WIDTH="3480" VPOS="500" HPOS="3064" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="736" HEIGHT="196" WIDTH="2808" VPOS="590" HPOS="3290">
                  <String CONTENT="Vergerius" HEIGHT="196" WIDTH="1092" VPOS="590" HPOS="3290"/>
                  <SP WIDTH="92" VPOS="766" HPOS="4398"/>
                  <String CONTENT="Petrus" HEIGHT="148" WIDTH="728" VPOS="590" HPOS="4514"/>
                  <SP WIDTH="126" VPOS="594" HPOS="5243"/>
                  <String CONTENT="Paulus" HEIGHT="148" WIDTH="728" VPOS="594" HPOS="5370"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block7" HEIGHT="380" WIDTH="3312" VPOS="500" HPOS="6544" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="730" HEIGHT="188" WIDTH="2192" VPOS="586" HPOS="6854">
                  <String CONTENT="Bullinger" HEIGHT="184" WIDTH="1084" VPOS="590" HPOS="6854"/>
                  <SP WIDTH="130" VPOS="590" HPOS="7939"/>
                  <String CONTENT="Heinrich" HEIGHT="148" WIDTH="976" VPOS="586" HPOS="8070"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block8" HEIGHT="708" WIDTH="3480" VPOS="880" HPOS="3064" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="1148" HEIGHT="188" WIDTH="1332" VPOS="998" HPOS="3294">
                  <String CONTENT="Vicosoprano" HEIGHT="188" WIDTH="1332" VPOS="998" HPOS="3294"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block9" HEIGHT="708" WIDTH="3312" VPOS="880" HPOS="6544" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="1136" HEIGHT="152" WIDTH="720" VPOS="994" HPOS="6858">
                  <String CONTENT="Zuerich" HEIGHT="152" WIDTH="720" VPOS="994" HPOS="6858"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block10" HEIGHT="496" WIDTH="3064" VPOS="1588" HPOS="0" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="1944" HEIGHT="164" WIDTH="812" VPOS="1814" HPOS="94">
                  <String CONTENT="Autograph" HEIGHT="164" WIDTH="812" VPOS="1814" HPOS="94"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block11" HEIGHT="496" WIDTH="3480" VPOS="1588" HPOS="3064" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="1948" HEIGHT="160" WIDTH="424" VPOS="1818" HPOS="3234">
                  <String CONTENT="Kopie" HEIGHT="160" WIDTH="424" VPOS="1818" HPOS="3234"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block12" HEIGHT="496" WIDTH="3312" VPOS="1588" HPOS="6544" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="1948" HEIGHT="160" WIDTH="1144" VPOS="1818" HPOS="6674">
                  <String CONTENT="Photokopie" HEIGHT="160" WIDTH="852" VPOS="1818" HPOS="6674"/>
                  <SP WIDTH="186" VPOS="1854" HPOS="7527"/>
                  <String CONTENT="-" HEIGHT="28" WIDTH="104" VPOS="1886" HPOS="7714"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block13" HEIGHT="500" WIDTH="3064" VPOS="2084" HPOS="0" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="2345" HEIGHT="192" WIDTH="2000" VPOS="2174" HPOS="90">
                  <String CONTENT="Standort" HEIGHT="140" WIDTH="660" VPOS="2206" HPOS="90"/>
                  <SP WIDTH="150" VPOS="2202" HPOS="751"/>
                  <String CONTENT="SueWelt" HEIGHT="168" WIDTH="600" VPOS="2178" HPOS="902"/>
                  <SP WIDTH="130" VPOS="2250" HPOS="1503"/>
                  <String CONTENT="&apos;U" HEIGHT="184" WIDTH="164" VPOS="2174" HPOS="1634"/>
                  <SP WIDTH="102" VPOS="2174" HPOS="1799"/>
                  <String CONTENT="A," HEIGHT="176" WIDTH="188" VPOS="2190" HPOS="1902"/>
               </TextLine>
            </TextBlock>
            <TextBlock ID="Page1_Block14" HEIGHT="1024" WIDTH="3480" VPOS="2084" HPOS="3064" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
               <TextLine BASELINE="2353" HEIGHT="172" WIDTH="1412" VPOS="2210" HPOS="3226">
                  <String CONTENT="Standort" HEIGHT="136" WIDTH="660" VPOS="2210" HPOS="3226"/>
                  <SP WIDTH="158" VPOS="2218" HPOS="3887"/>
                  <String CONTENT="L-^ze" HEIGHT="132" WIDTH="592" VPOS="2250" HPOS="4046"/>
               </TextLine>
               <TextLine BASELINE="2948" HEIGHT="160" WIDTH="388" VPOS="2818" HPOS="3230">
            <String CONTENT="Sign." HEIGHT="160" WIDTH="388" VPOS="2818" HPOS="3230"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="Page1_Block15" HEIGHT="500" WIDTH="3312" VPOS="2084" HPOS="6544" language="de"
 STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
          <TextLine BASELINE="2340" HEIGHT="136" WIDTH="752" VPOS="2210" HPOS="6670">
            <String CONTENT="Bull." HEIGHT="128" WIDTH="332" VPOS="2214" HPOS="6670"/>
            <SP WIDTH="46" VPOS="2210" HPOS="7003"/>
            <String CONTENT="Corr." HEIGHT="136" WIDTH="372" VPOS="2210" HPOS="7050"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="Page1_Block16" HEIGHT="524" WIDTH="3064" VPOS="2584" HPOS="0" language="de"
STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
          <TextLine BASELINE="2960" HEIGHT="184" WIDTH="2000" VPOS="2814" HPOS="90">
            <String CONTENT="Sign." HEIGHT="160" WIDTH="384" VPOS="2818" HPOS="90"/>
            <SP WIDTH="370" VPOS="2818" HPOS="475"/>
            <String CONTENT="E" HEIGHT="156" WIDTH="136" VPOS="2818" HPOS="846"/>
            <SP WIDTH="66" VPOS="2818" HPOS="983"/>
            <String CONTENT="TT" HEIGHT="156" WIDTH="172" VPOS="2814" HPOS="1050"/>
            <SP WIDTH="550" VPOS="2814" HPOS="1223"/>
            <String CONTENT="4Jz" HEIGHT="160" WIDTH="316" VPOS="2838" HPOS="1774"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="Page1_Block17" HEIGHT="524" WIDTH="3312" VPOS="2584" HPOS="6544" language="de"
 STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
          <TextLine BASELINE="2957" HEIGHT="200" WIDTH="2232" VPOS="2802" HPOS="6666">
            <String CONTENT="Abschrift" HEIGHT="132" WIDTH="720" VPOS="2818" HPOS="6666"/>
            <SP WIDTH="326" VPOS="2826" HPOS="7387"/>
            <String CONTENT="ZB" HEIGHT="140" WIDTH="232" VPOS="2834" HPOS="7714"/>
            <SP WIDTH="150" VPOS="2806" HPOS="7947"/>
            <String CONTENT="(&apos;Druck)" HEIGHT="200" WIDTH="800" VPOS="2802" HPOS="8098"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="Page1_Block18" HEIGHT="488" WIDTH="3064" VPOS="3108" HPOS="0" language="de"
STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
          <TextLine BASELINE="3344" HEIGHT="160" WIDTH="580" VPOS="3214" HPOS="90">
            <String CONTENT="Umfang" HEIGHT="160" WIDTH="580" VPOS="3214" HPOS="90"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="Page1_Block19" HEIGHT="488" WIDTH="3480" VPOS="3108" HPOS="3064" language="de"
 STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
          <TextLine BASELINE="3344" HEIGHT="164" WIDTH="576" VPOS="3214" HPOS="3238">
            <String CONTENT="Umfang" HEIGHT="164" WIDTH="576" VPOS="3214" HPOS="3238"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="Page1_Block20" HEIGHT="488" WIDTH="3312" VPOS="3108" HPOS="6544" language="de"
 STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
          <TextLine BASELINE="3359" HEIGHT="208" WIDTH="2740" VPOS="3210" HPOS="6670">
            <String CONTENT="Bull.Corr." HEIGHT="136" WIDTH="756" VPOS="3210" HPOS="6670"/>
            <SP WIDTH="290" VPOS="3246" HPOS="7427"/>
            <String CONTENT="iB" HEIGHT="152" WIDTH="224" VPOS="3238" HPOS="7718"/>
            <SP WIDTH="378" VPOS="3234" HPOS="7943"/>
            <String CONTENT="Bl.l," HEIGHT="184" WIDTH="576" VPOS="3234" HPOS="8322"/>
            <SP WIDTH="146" VPOS="3230" HPOS="8899"/>
            <String CONTENT="S.l" HEIGHT="148" WIDTH="364" VPOS="3230" HPOS="9046"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="Page1_Block21" HEIGHT="696" WIDTH="3064" VPOS="3596" HPOS="0" language="de"
STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
          <TextLine BASELINE="3960" HEIGHT="164" WIDTH="628" VPOS="3830" HPOS="82">
            <String CONTENT="Sprache" HEIGHT="164" WIDTH="628" VPOS="3830" HPOS="82"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="Page1_Block22" HEIGHT="696" WIDTH="6792" VPOS="3596" HPOS="3064" language="de"
 STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
          <TextLine BASELINE="3787" HEIGHT="184" WIDTH="5204" VPOS="3618" HPOS="3970">
            <String CONTENT="*&gt;?:" HEIGHT="132" WIDTH="360" VPOS="3650" HPOS="3970"/>
            <SP WIDTH="598" VPOS="3618" HPOS="4331"/>
            <String CONTENT="Htult," HEIGHT="184" WIDTH="560" VPOS="3618" HPOS="4930"/>
            <SP WIDTH="78" VPOS="3618" HPOS="5491"/>
            <String CONTENT="%&apos;e" HEIGHT="152" WIDTH="268" VPOS="3618" HPOS="5570"/>
            <SP WIDTH="78" VPOS="3646" HPOS="5839"/>
            <String CONTENT="evaua" HEIGHT="156" WIDTH="540" VPOS="3646" HPOS="5918"/>
            <SP WIDTH="82" VPOS="3646" HPOS="6459"/>
            <String CONTENT="6rtotteM*h," HEIGHT="160" WIDTH="944" VPOS="3642" HPOS="6542"/>
            <SP WIDTH="70" VPOS="3678" HPOS="7487"/>
            <String CONTENT="M.L6Ctit%t" HEIGHT="168" WIDTH="860" VPOS="3634" HPOS="7558"/>
            <SP WIDTH="98" VPOS="3670" HPOS="8419"/>
            <String CONTENT="TL*A" HEIGHT="132" WIDTH="656" VPOS="3670" HPOS="8518"/>
          </TextLine>
          <TextLine BASELINE="3943" HEIGHT="140" WIDTH="6428" VPOS="3823" HPOS="3223">
            <String CONTENT="Literatur" HEIGHT="132" WIDTH="630" VPOS="3831" HPOS="3223"/>
            <SP WIDTH="528" VPOS="3823" HPOS="3854"/>
            <String CONTENT="^" HEIGHT="100" WIDTH="110" VPOS="3823" HPOS="4383"/>
            <SP WIDTH="636" VPOS="3823" HPOS="4494"/>
            <String CONTENT="v" HEIGHT="58" WIDTH="68" VPOS="3845" HPOS="5131"/>
            <String STYLE="subscript" CONTENT="v" HEIGHT="72" WIDTH="74" VPOS="3855" HPOS="5203"/>
            <String CONTENT="\y^" HEIGHT="136" WIDTH="3384" VPOS="3825" HPOS="5303"/>
            <SP WIDTH="552" VPOS="3825" HPOS="8688"/>
            <String CONTENT="*" HEIGHT="64" WIDTH="60" VPOS="3835" HPOS="9241"/>
```

```
                     <SP WIDTH="316" VPOS="3835" HPOS="9302"/>
                     <String CONTENT="*" HEIGHT="18" WIDTH="32" VPOS="3853" HPOS="9619"/>
                  </TextLine>
               </TextBlock>
               <TextBlock ID="Page1_Block23" HEIGHT="460" WIDTH="3064" VPOS="4292" HPOS="0" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
                  <TextLine BASELINE="4652" HEIGHT="136" WIDTH="692" VPOS="4522" HPOS="86">
                     <String CONTENT="Gedruckt" HEIGHT="136" WIDTH="692" VPOS="4522" HPOS="86"/>
                  </TextLine>
               </TextBlock>
               <TextBlock ID="Page1_Block24" HEIGHT="616" WIDTH="3064" VPOS="4752" HPOS="0" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
                  <TextLine BASELINE="5066" HEIGHT="336" WIDTH="1648" VPOS="4862" HPOS="1398">
                     <String CONTENT="%kJl" HEIGHT="224" WIDTH="456" VPOS="4862" HPOS="1398"/>
                     <String STYLE="subscript" CONTENT=";" HEIGHT="80" WIDTH="56" VPOS="5078" HPOS="1822"/>
                     <SP WIDTH="54" VPOS="4998" HPOS="1879"/>
                     <String CONTENT="yf," HEIGHT="196" WIDTH="144" VPOS="4906" HPOS="1934"/>
                     <SP WIDTH="98" VPOS="4930" HPOS="2079"/>
                     <String CONTENT="Uff,!" HEIGHT="280" WIDTH="868" VPOS="4918" HPOS="2178"/>
                  </TextLine>
               </TextBlock>
               <TextBlock ID="Page1_Block25" HEIGHT="1614" WIDTH="6792" VPOS="5368" HPOS="3064" language="de" STYLEREFS="StyleId-173DA853-4BBF-4C24-B4D9-8E1AA430F9AC-">
                  <TextLine BASELINE="5611" HEIGHT="200" WIDTH="5296" VPOS="5458" HPOS="4530">
                     <String CONTENT="accepi" HEIGHT="180" WIDTH="720" VPOS="5478" HPOS="4530"/>
                     <SP WIDTH="130" VPOS="5478" HPOS="5251"/>
                     <String CONTENT="heri" HEIGHT="144" WIDTH="480" VPOS="5478" HPOS="5382"/>
                     <SP WIDTH="146" VPOS="5478" HPOS="5863"/>
                     <String CONTENT="literas" HEIGHT="144" WIDTH="832" VPOS="5478" HPOS="6010"/>
                     <SP WIDTH="142" VPOS="5478" HPOS="6843"/>
                     <String CONTENT="tuas," HEIGHT="172" WIDTH="568" VPOS="5478" HPOS="6986"/>
                     <SP WIDTH="170" VPOS="5510" HPOS="7555"/>
                     <String CONTENT="quibus" HEIGHT="184" WIDTH="704" VPOS="5466" HPOS="7726"/>
                     <SP WIDTH="134" VPOS="5498" HPOS="8431"/>
                     <String CONTENT="mihi" HEIGHT="148" WIDTH="480" VPOS="5462" HPOS="8566"/>
                     <SP WIDTH="138" VPOS="5458" HPOS="9047"/>
                     <String CONTENT="inter" HEIGHT="144" WIDTH="640" VPOS="5458" HPOS="9186"/>
                  </TextLine>
                  <TextLine BASELINE="5795" HEIGHT="244" WIDTH="6332" VPOS="5618" HPOS="3218">
                     <String CONTENT="Bemerkungen" HEIGHT="148" WIDTH="1028" VPOS="5618" HPOS="3218"/>
                     <SP WIDTH="278" VPOS="5642" HPOS="4247"/>
                     <String STYLE="subscript" CONTENT="r" HEIGHT="108" WIDTH="116" VPOS="5718" HPOS="4526"/>
                     <String CONTENT="@liqua" HEIGHT="184" WIDTH="720" VPOS="5678" HPOS="4658"/>
                     <SP WIDTH="142" VPOS="5714" HPOS="5379"/>
                     <String CONTENT="significabas" HEIGHT="180" WIDTH="1444" VPOS="5678" HPOS="5522"/>
                     <SP WIDTH="134" VPOS="5674" HPOS="6967"/>
                     <String CONTENT="de" HEIGHT="140" WIDTH="228" VPOS="5674" HPOS="7102"/>
                     <SP WIDTH="134" VPOS="5702" HPOS="7331"/>
                     <String CONTENT="nuptiis" HEIGHT="180" WIDTH="844" VPOS="5666" HPOS="7466"/>
                     <SP WIDTH="142" VPOS="5662" HPOS="8311"/>
                     <String CONTENT="Iosiae" HEIGHT="148" WIDTH="712" VPOS="5662" HPOS="8454"/>
                     <SP WIDTH="146" VPOS="5638" HPOS="9167"/>
                     <String CONTENT="ii:" HEIGHT="184" WIDTH="236" VPOS="5638" HPOS="9314"/>
                  </TextLine>
                  <TextLine BASELINE="6021" HEIGHT="156" WIDTH="1324" VPOS="5870" HPOS="4538">
                     <String CONTENT="et" HEIGHT="140" WIDTH="220" VPOS="5886" HPOS="4538"/>
                     <SP WIDTH="126" VPOS="5870" HPOS="4759"/>
                     <String CONTENT="Elisabet" HEIGHT="152" WIDTH="976" VPOS="5870" HPOS="4886"/>
                  </TextLine>
               </TextBlock>
            </ComposedBlock>
            <GraphicalElement ID="Page1_Block26" HEIGHT="40" WIDTH="9848" VPOS="1568" HPOS="0"/>
            <GraphicalElement ID="Page1_Block27" HEIGHT="32" WIDTH="3320" VPOS="2568" HPOS="6528"/>
            <GraphicalElement ID="Page1_Block28" HEIGHT="36" WIDTH="9844" VPOS="3580" HPOS="4"/>
            <GraphicalElement ID="Page1_Block29" HEIGHT="32" WIDTH="984" VPOS="3800" HPOS="4408"/>
            <GraphicalElement ID="Page1_Block30" HEIGHT="32" WIDTH="3068" VPOS="4276" HPOS="4"/>
            <GraphicalElement ID="Page1_Block31" HEIGHT="44" WIDTH="6808" VPOS="5348" HPOS="3044"/>
            <GraphicalElement ID="Page1_Block32" HEIGHT="6968" WIDTH="48" VPOS="8" HPOS="3040"/>
            <GraphicalElement ID="Page1_Block33" HEIGHT="3608" WIDTH="32" VPOS="4" HPOS="6528"/>
            <GraphicalElement ID="Page1_Block34" HEIGHT="12" WIDTH="160" VPOS="348" HPOS="1584"/>
         </PrintSpace>
      </Page>
   </Layout>
</alto>
```

A.3.3 Element Frequenzen Statistik

Tabelle 4: Mittelwert μ und Standardabweichung σ der Elementfrequenzen in den Dateien der Ordner `ocr_sample_100_v1` und `ocr_sample_100_v2`.

Element	μ	σ
document	1.0	0.0
documentData	1.0	0.0
sections	1.0	0.0
section	1.12	0.46
stream	1.11	0.43
mainText	1.11	0.43
elemId	1.42	1.34
page	1.0	0.0
block	10.0	3.9
region	10.0	3.9
rect	35.81	9.43
row	8.09	2.32
cell	21.96	10.38
text	21.29	10.43
par	30.76	6.52
line	28.88	4.35
formatting	29.56	4.74
separator	7.83	0.86
start	7.83	0.86
end	7.83	0.86

Abbildung 11: Version 1

Element	μ	σ
alto	1.0	0.0
Description	1.0	0.0
MeasurementUnit	1.0	0.0
OCRProcessing	1.0	0.0
ocrProcessingStep	1.0	0.0
processingDateTime	1.0	0.0
processingSoftware	1.0	0.0
softwareCreator	1.0	0.0
softwareName	1.0	0.0
softwareVersion	1.0	0.0
Styles	1.0	0.0
ParagraphStyle	2.64	1.68
Layout	1.0	0.0
Page	1.0	0.0
PrintSpace	1.0	0.0
ComposedBlock	1.09	0.32
TextBlock	17.71	6.64
TextLine	28.67	4.38
String	74.27	17.98
SP	45.06	13.08
GraphicalElement	7.78	0.93
HYP	1.07	0.26
TopMargin	1.0	0.0
LeftMargin	1.0	0.0
RightMargin	1.0	0.0
BottomMargin	1.0	0.0
Shape	1.67	0.58
Polygon	1.67	0.58
Illustration	Illustration	0

Abbildung 12: Version 2