eCodicology

Algorithms for the Automatic Tagging of Medieval Manuscripts

# Development of New Technologies for the Automatic Analysis of Medieval Manuscripts

Hannah Busch, Philipp Vanscheidt (University of Trier)
Swati Chandna (Karlsruhe Institute of Technology)
Celia Krause (Technical University of Darmstadt)

Bundesministerium
für Bildung
und Forschung

Kompetenzzentrum
für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften

KIT
Karlsruher Institut für Technologie

TECHNISCHE
UNIVERSITÄT
DARMSTADT

# Project and Motivation

# Project and Motivation

*What is eCodicology?*

A BMBF funded joint research project of

- Technical University Darmstadt
- Trier Center for Digital Humanities
- Karlsruhe Institute for Technology

# Project and Motivation

*What is eCodicology?*

A BMBF funded joint research project of

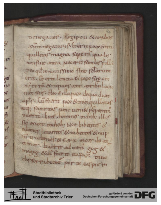
- Technical University Darmstadt
- Trier Center for Digital Humanities
- Karlsruhe Institute for Technology

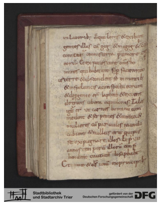
*What is eCodicology for?*

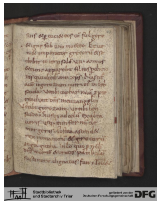
- Automatic identification of macro and micro structural layout elements
- Quantitative Codicology: Statistical analysis of reproducible features
- To identify hidden relationships

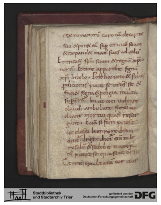
# How to Publish a Library?

Virtual Scriptorium St. Matthias



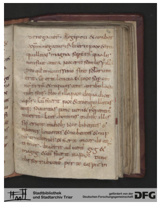


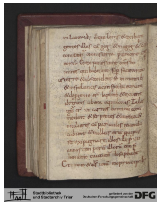


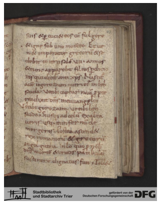


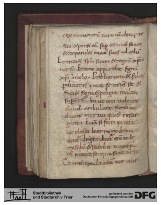
# How to Publish a Library?

Virtual Scriptorium St. Matthias





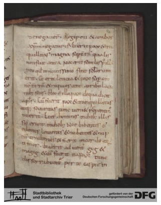


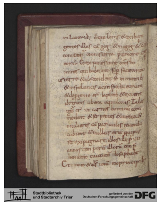


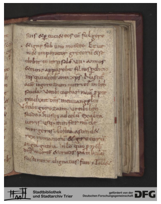
| | |
|---|---|
| Scope: | 440 codices online |
| Period: | 8th to 16th century |
| Contents: | Liturgica, Bible, patristics, mysticism, philosophy, poetry, law, charters, etc. |

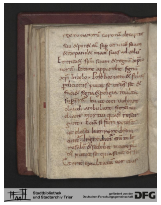
# How to Publish a Library?

## Virtual Scriptorium St. Matthias



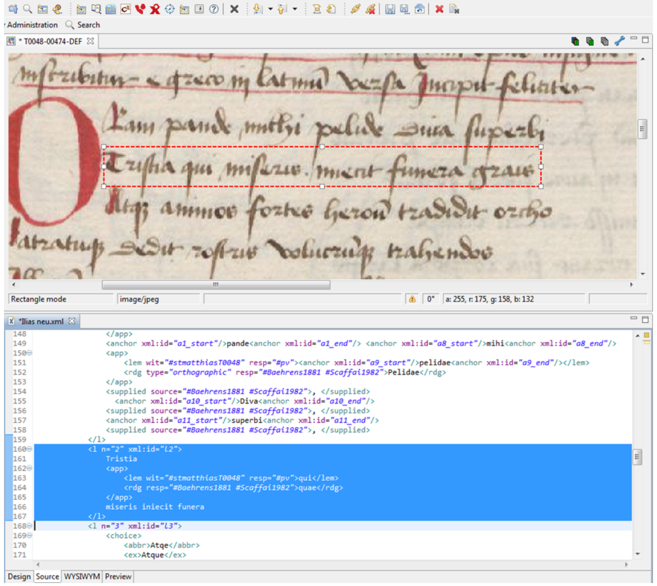






| | |
|---|---|
| Scope: | 440 codices online |
| Period: | 8th to 16th century |
| Contents: | Liturgica, Bible, patristics, mysticism, philosophy, poetry, law, charters, etc. |

## TextGrid

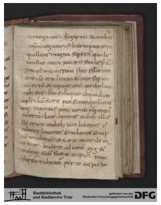


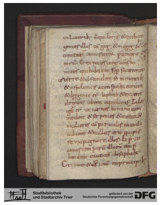
```
        </app>
        <anchor xml:id="a1_start"/>pande<anchor xml:id="a1_end"/> <anchor xml:id="a8_start"/>mihi<anchor xml:id="a8_end"/>
        <app>
            <lem wit="#stmatthiasT0048" resp="#pv"><anchor xml:id="a9_start"/>pelidae<anchor xml:id="a9_end"/></lem>
            <rdg type="orthographic" resp="#Baehrens1881 #Scaffai1982">Pelidae</rdg>
        </app>
        <supplied source="#Baehrens1881 #Scaffai1982">, </supplied>
          <anchor xml:id="a10_start"/>Diva<anchor xml:id="a10_end"/>
        <supplied source="#Baehrens1881 #Scaffai1982">, </supplied>
        <anchor xml:id="a11_start"/>superbi<anchor xml:id="a11_end"/>
        <supplied source="#Baehrens1881 #Scaffai1982">, </supplied>
    </l>
    <l n="2" xml:id="l2">
        Tristia
        <app>
            <lem wit="#stmatthiasT0048" resp="#pv">qui</lem>
            <rdg resp="#Baehrens1881 #Scaffai1982">quae</rdg>
        </app>
        miseris iniecit funera
    </l>
    <l n="3" xml:id="l3">
        <choice>
            <abbr>Atqe</abbr>
            <ex>Atque</ex>
```

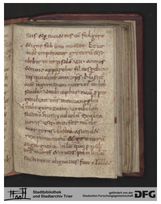
# How to Publish a Library?

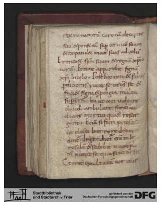
Virtual Scriptorium St. Matthias

TextGrid





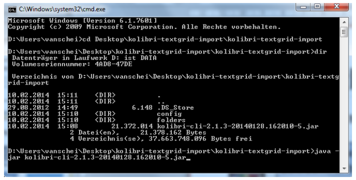




Scope: 440 codices online

Period: 8th to 16th century

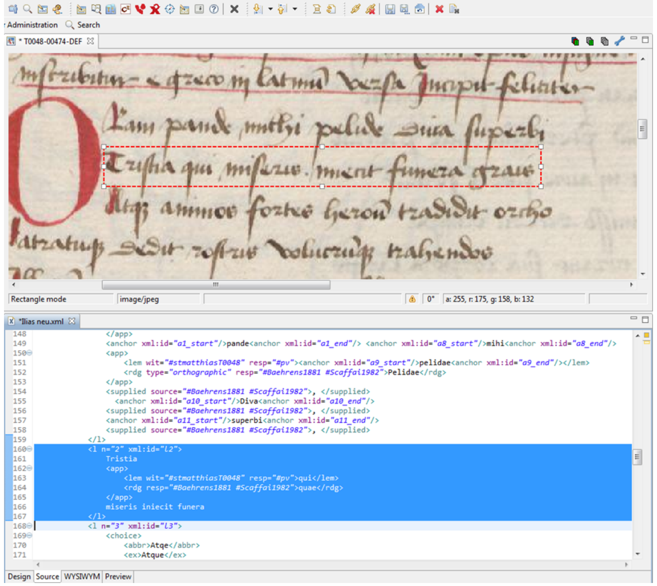
Contents: Liturgica, Bible, patristics, mysticism, philosophy, poetry, law, charters, etc.



```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Alle Rechte vorbehalten.

D:\Users\vanschei>cd Desktop\kolibri-textgrid-import\kolibri-textgrid-import

D:\Users\vanschei\Desktop\kolibri-textgrid-import\kolibri-textgrid-import>dir
 Datenträger in Laufwerk D: ist DATA
 Volumeseriennummer: 4AD8-47DE

 Verzeichnis von D:\Users\vanschei\Desktop\kolibri-textgrid-import\kolibri-textg
rid-import

10.02.2014  15:11    <DIR>          .
10.02.2014  15:11    <DIR>          ..
29.08.2012  14:49             6.148 .DS_Store
10.02.2014  15:10    <DIR>          config
10.02.2014  15:10    <DIR>          folders
10.02.2014  15:08        21.372.014 kolibri-cli-2.1.3-20140128.162010-5.jar
               2 Datei(en),     21.378.162 Bytes
               4 Verzeichnis(se), 37.663.748.096 Bytes frei

D:\Users\vanschei\Desktop\kolibri-textgrid-import\kolibri-textgrid-import>java -
jar kolibri-cli-2.1.3-20140128.162010-5.jar_
```

KoLibRi

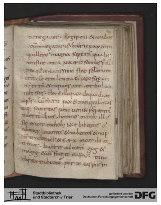
Mass Ingest

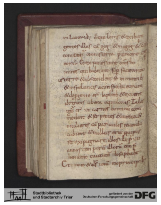


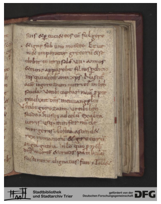
```
        </app>
        <anchor xml:id="a1_start"/>pande<anchor xml:id="a1_end"/> <anchor xml:id="a8_start"/>mihi<anchor xml:id="a8_end"/>
        <app>
            <lem wit="#stmatthiasT0048" resp="#pv"><anchor xml:id="a9_start"/>pelidae<anchor xml:id="a9_end"/></lem>
            <rdg type="orthographic" resp="#Baehrens1881 #Scaffai1982">Pelidae</rdg>
        </app>
        <supplied source="#Baehrens1881 #Scaffai1982">, </supplied>
          <anchor xml:id="a10_start"/>Diva<anchor xml:id="a10_end"/>
        <supplied source="#Baehrens1881 #Scaffai1982">, </supplied>
        <anchor xml:id="a11_start"/>superbi<anchor xml:id="a11_end"/>
        <supplied source="#Baehrens1881 #Scaffai1982">, </supplied>
    </l>
    <l n="2" xml:id="l2">
        Tristia
        <app>
            <lem wit="#stmatthiasT0048" resp="#pv">qui</lem>
            <rdg resp="#Baehrens1881 #Scaffai1982">quae</rdg>
        </app>
        miseris iniecit funera
    </l>
    <l n="3" xml:id="l3">
        <choice>
            <abbr>Atqe</abbr>
            <ex>Atque</ex>
```

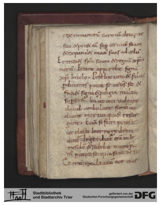
# How to Publish a Library?

## Virtual Scriptorium St. Matthias

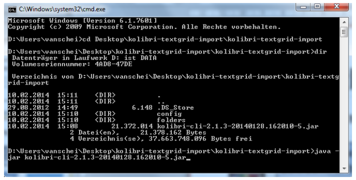








| | |
|---|---|
| Scope: | 440 codices online |
| Period: | 8th to 16th century |
| Contents: | Liturgica, Bible, patristics, mysticism, philosophy, poetry, law, charters, etc. |

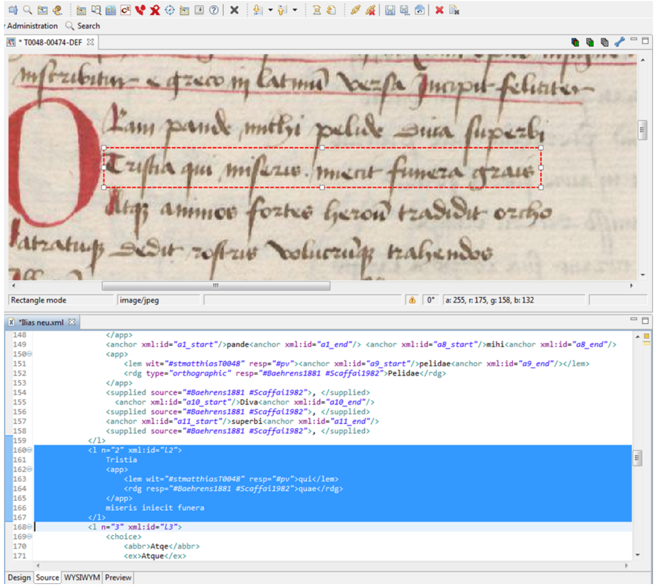


```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Alle Rechte vorbehalten.

D:\Users\vanschei>cd Desktop\kolibri-textgrid-import\kolibri-textgrid-import

D:\Users\vanschei\Desktop\kolibri-textgrid-import\kolibri-textgrid-import>dir
 Datenträger in Laufwerk D: ist DATA
 Volumeseriennummer: 4AD8-47DE

 Verzeichnis von D:\Users\vanschei\Desktop\kolibri-textgrid-import\kolibri-textg
rid-import

10.02.2014  15:11    <DIR>          .
10.02.2014  15:11    <DIR>          ..
29.08.2012  14:49             6.148 .DS_Store
10.02.2014  15:10    <DIR>          config
10.02.2014  15:10    <DIR>          folders
10.02.2014  15:08        21.372.014 kolibri-cli-2.1.3-20140128.162010-5.jar
               2 Datei(en),     21.378.162 Bytes
               4 Verzeichnis(se), 37.663.748.096 Bytes frei

D:\Users\vanschei\Desktop\kolibri-textgrid-import\kolibri-textgrid-import>java -
jar kolibri-cli-2.1.3-20140128.162010-5.jar_
```

KoLibRi

Mass Ingest

## TextGrid

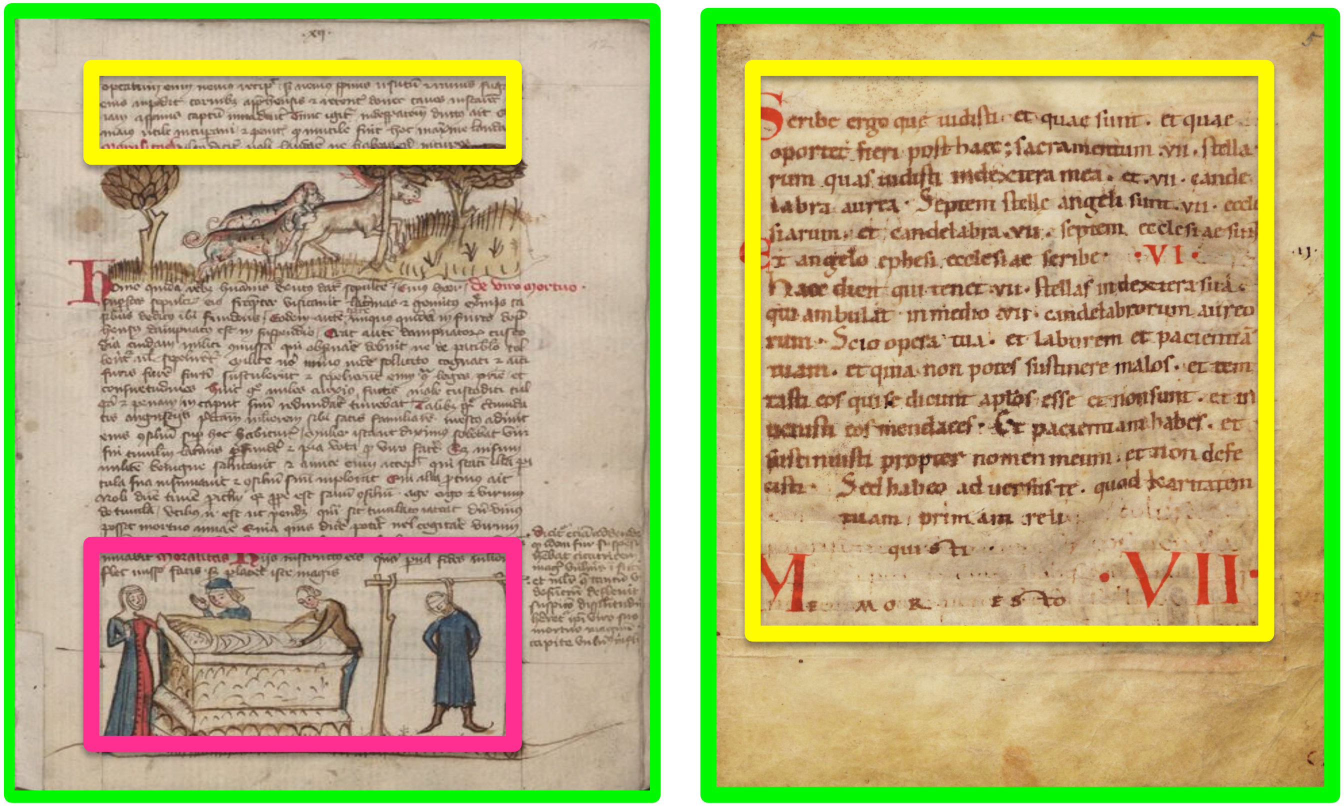


```
                </app>
                <anchor xml:id="a1_start"/>pande<anchor xml:id="a1_end"/> <anchor xml:id="a8_start"/>mihi<anchor xml:id="a8_end"/>
                <app>
                    <lem wit="#stmatthiasT0048" resp="#pv"><anchor xml:id="a9_start"/>pelidae<anchor xml:id="a9_end"/></lem>
                    <rdg type="orthographic" resp="#Baehrens1881 #Scaffai1982">Pelidae</rdg>
                </app>
                <supplied source="#Baehrens1881 #Scaffai1982">, </supplied>
                  <anchor xml:id="a10_start"/>Diva<anchor xml:id="a10_end"/>
                <supplied source="#Baehrens1881 #Scaffai1982">, </supplied>
                <anchor xml:id="a11_start"/>superbi<anchor xml:id="a11_end"/>
                <supplied source="#Baehrens1881 #Scaffai1982">, </supplied>
            </l>
            <l n="2" xml:id="l2">
                Tristia
                <app>
                    <lem wit="#stmatthiasT0048" resp="#pv">qui</lem>
                    <rdg resp="#Baehrens1881 #Scaffai1982">quae</rdg>
                </app>
                miseris iniecit funera
            </l>
            <l n="3" xml:id="l3">
                <choice>
                    <abbr>Atqe</abbr>
                    <ex>Atque</ex>
```

- PID Handling by GWDG
- Exploration of material by human researchers, eg. Critical editions

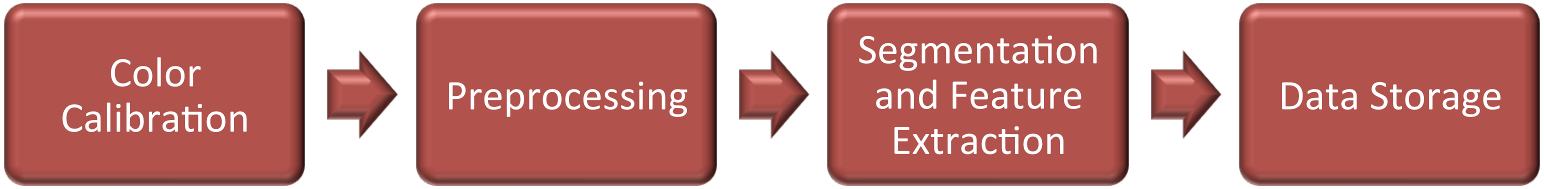
# Feature Extraction

- **Feature:** Individual measurable property of phenomenon observed



🟩 page size, 🟨 written size, 🟥 pictorial space

# How to Extract Layout Features?

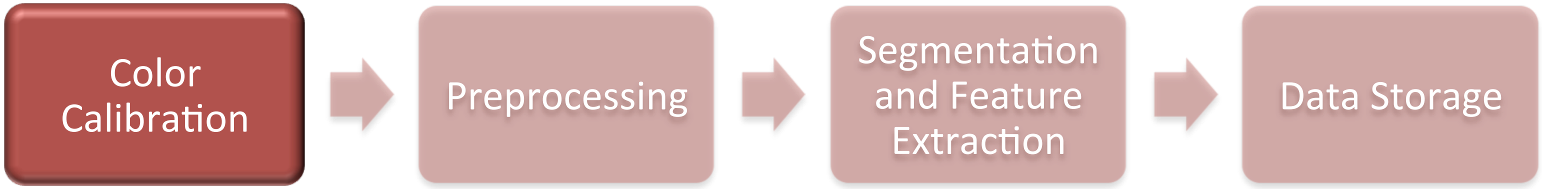


Color Calibration

Preprocessing

Segmentation and Feature Extraction

Data Storage



Color Calibration

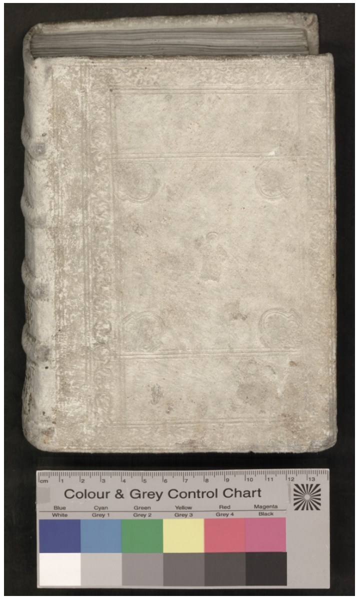
Preprocessing

Segmentation and Feature Extraction

Data Storage

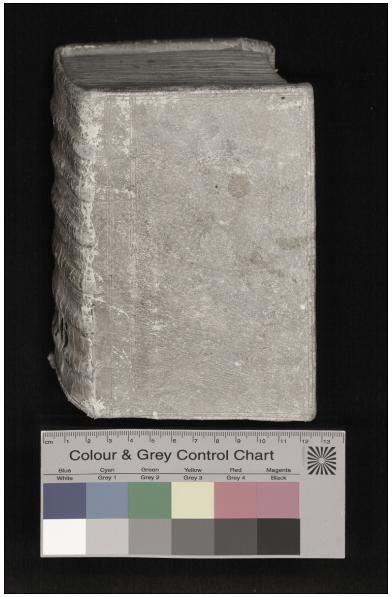
# Color Calibration

- Adjusting different color spaces to similar color space according to a standard color chart



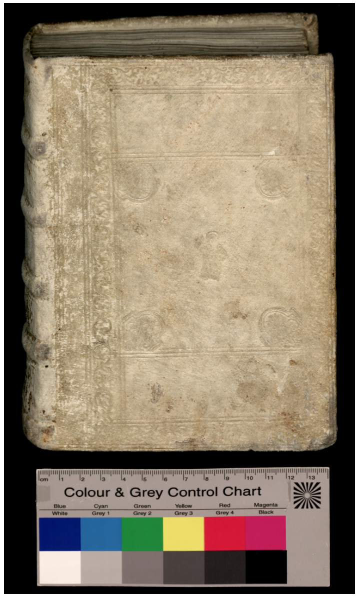
Before Calibration

Scanner 1



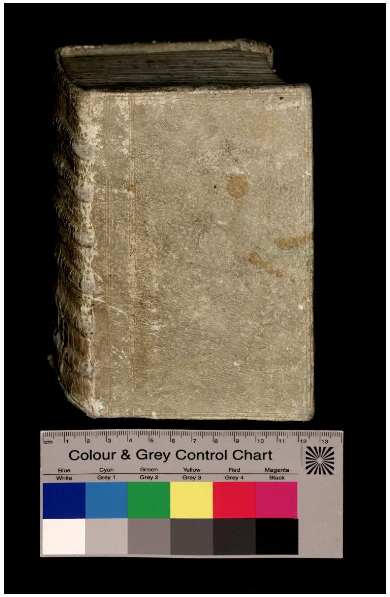
Before Calibration

Scanner 2



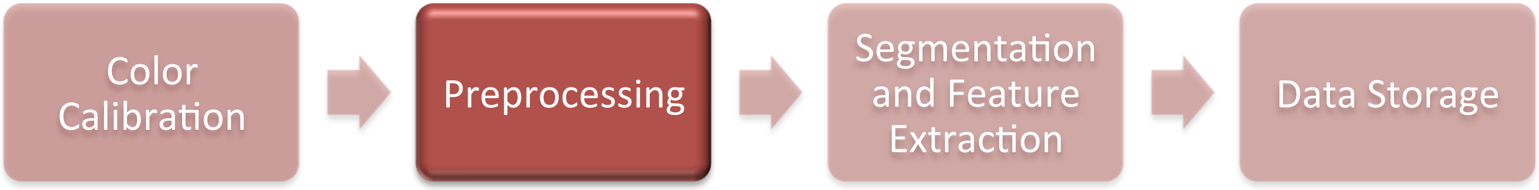
After Calibration

Scanner 1



After Calibration

Scanner 2



Color Calibration

Preprocessing

Segmentation and Feature Extraction

Data Storage

# Preprocessing

- To increase accuracy of the digital data for feature extraction
  - Spatial Calibration - Adjusting different resolutions to same resolution
  - Filtering - Removing low frequency background noise
  - Scaling and Duplication – Scaling images to different resolutions

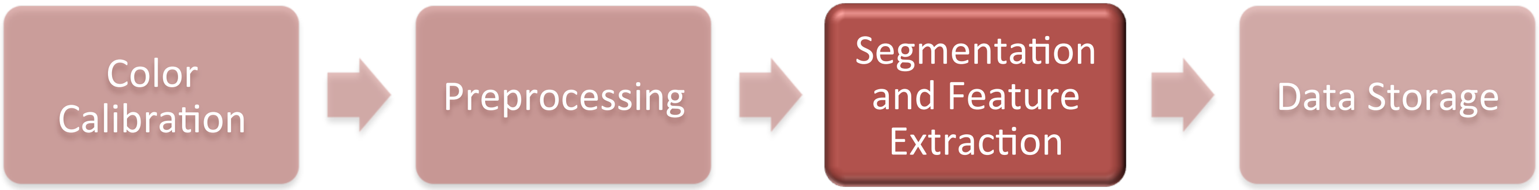
300 dpi

400 dpi

Different Resolutions

Spatial Calibrated Image

Filtered Image



Color Calibration

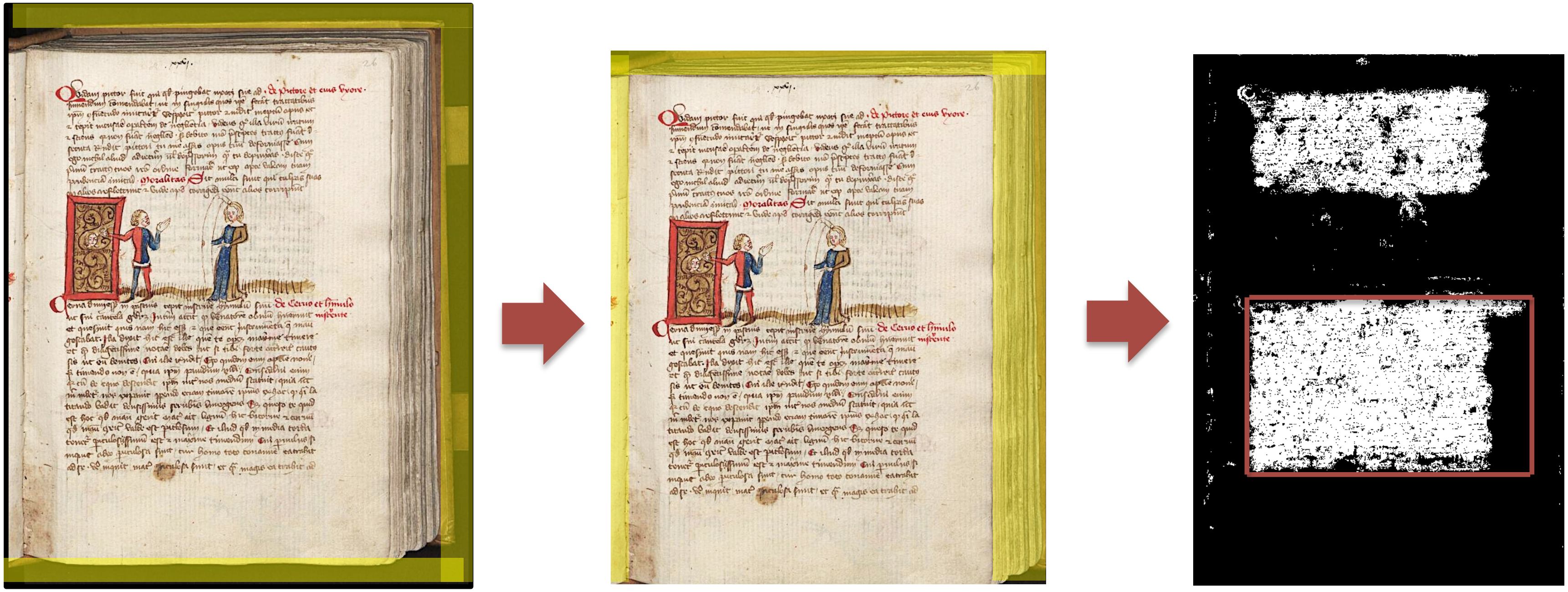
Preprocessing

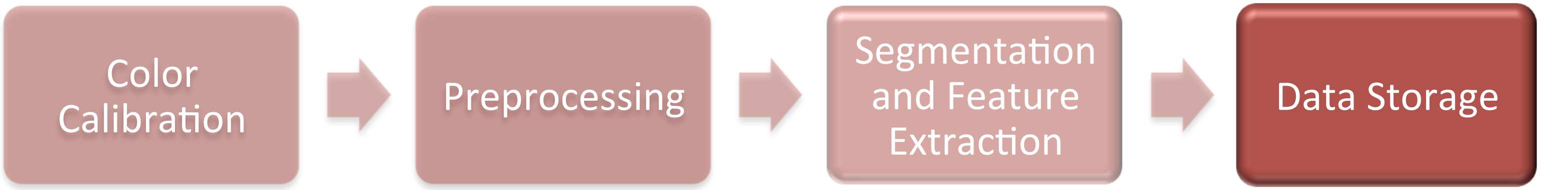
Segmentation and Feature Extraction

Data Storage

# Segmentation

- Divide image into regions with similar properties
- Determine region of interest





Color Calibration

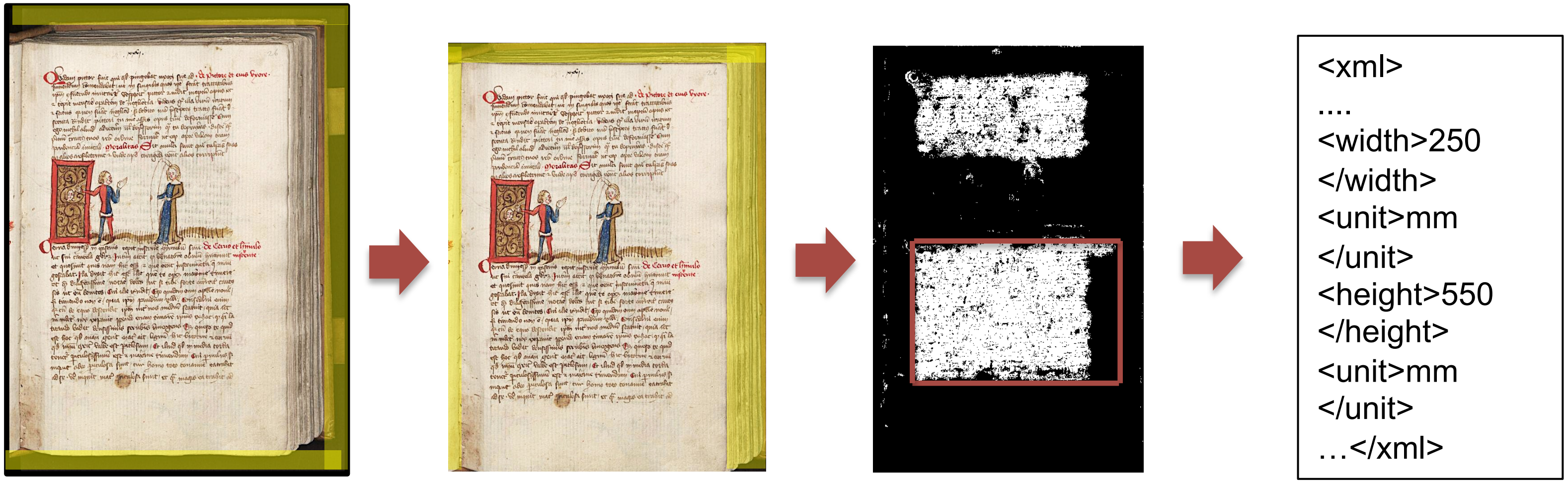
Preprocessing

Segmentation and Feature Extraction

Data Storage

# Data Storage

- Extracted features are stored in XML file



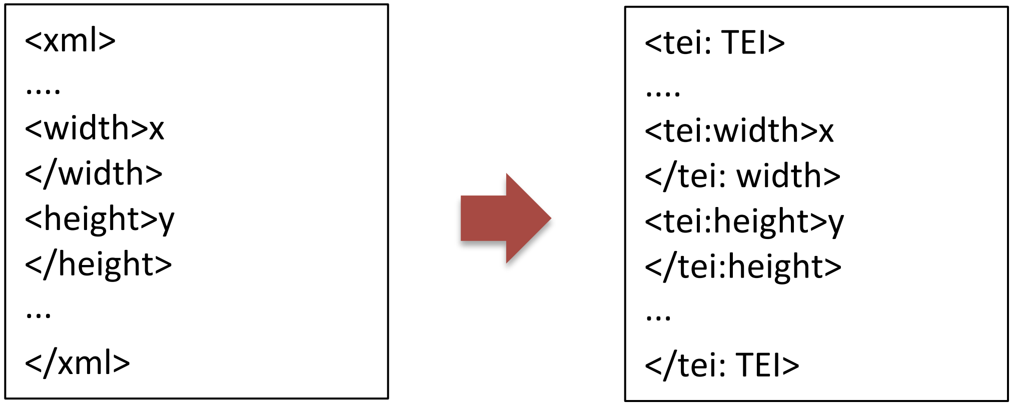
```
<xml>
....
<width>250
</width>
<unit>mm
</unit>
<height>550
</height>
<unit>mm
</unit>
…</xml>
```

# Feature Extraction….In a nutshell

- Automatic layout analysis of manuscript images is challenging due to different color spaces, heavy noise etc.
- Suitable preprocessing and segmentation methodology are very essential steps for feature extraction
- Accuracy and convergence rate of such techniques must be significantly high in order to ensure the success of subsequent steps

# Metadata Management

- Customization of TEI, using TEI ODD – One Document Does It All
- Preferable schema language for output: RelaxNG + Schematron
- Special focus on measurements, written and pictorial spaces, e.g. marginalia, pictorial elements



```
<xml>
....
<width>x
</width>
<height>y
</height>
...
</xml>
```

```
<tei: TEI>
....
<tei:width>x
</tei: width>
<tei:height>y
</tei:height>
...
</tei: TEI>
```

# Metadata Management

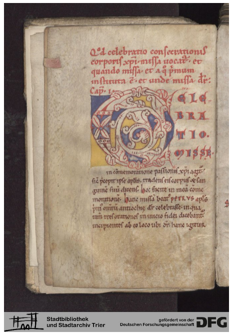
TEI XML Example for Manuscript Description:

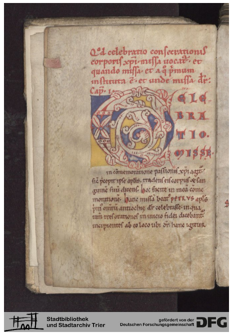
```
<tei:TEI>
    <tei:teiHeader>
    …
        <tei:msDesc>
        …
            <tei:supportDesc>

                <tei:extent>
                    <tei:measure type="leaves">44</tei:measure>
                    <tei:measure type="format">8°</tei:measure>
                </tei:extent>
                …
            </tei:supportDesc>
            <tei:layoutDesc>
                <tei:layout>
                    <tei:dimensions type="written" corresp="#written1">
                        <tei:height quantity="250" unit="mm" min="249" max="251" confidence="0.8">250mm</tei:height>
                        <tei:width quantity="100" unit="mm" min="98" max="101" confidence="0.77">100mm</tei:width>
                    </tei:dimensions>
                    <tei:dimensions type="image" facs="#image1"/>
                </tei:layout>
                <tei:layout columns="1" writtenLines="26 30"/>
            </tei:layoutDesc>
            </tei:objectDesc>
            …
            <tei:additions>
                <tei:note type="gloss" place="gutter"/>
            </tei:additions>
            …
        </tei:msDesc>
        …
    </tei:teiHeader>
    <tei:facsimile>
        <tei:surface type="leaf" ulx="0" uly="0" lrx="100" lry="250">
            <tei:graphic url="0099-00005.jpg"/>
            <tei:zone type="image" ulx="20" uly="20" lrx="70" lry="70" xml:id="image1"/>
            <tei:zone  type="written" ulx="20" uly="20" lrx="90" lry="90" xml:id="written1"/>
        </tei:surface>
    </tei:facsimile>
    <tei:body>…</tei:body>
</tei:TEI>
```

# How to analyze the results?

+

=

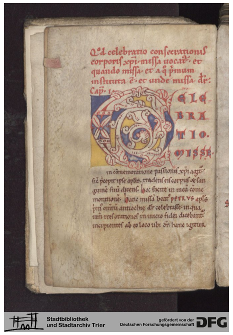

Stadtbibliothek und Stadtarchiv Trier
gefördert von der Deutschen Forschungsgemeinschaft DFG


Stadtbibliothek und Stadtarchiv Trier
gefördert von der Deutschen Forschungsgemeinschaft DFG

+

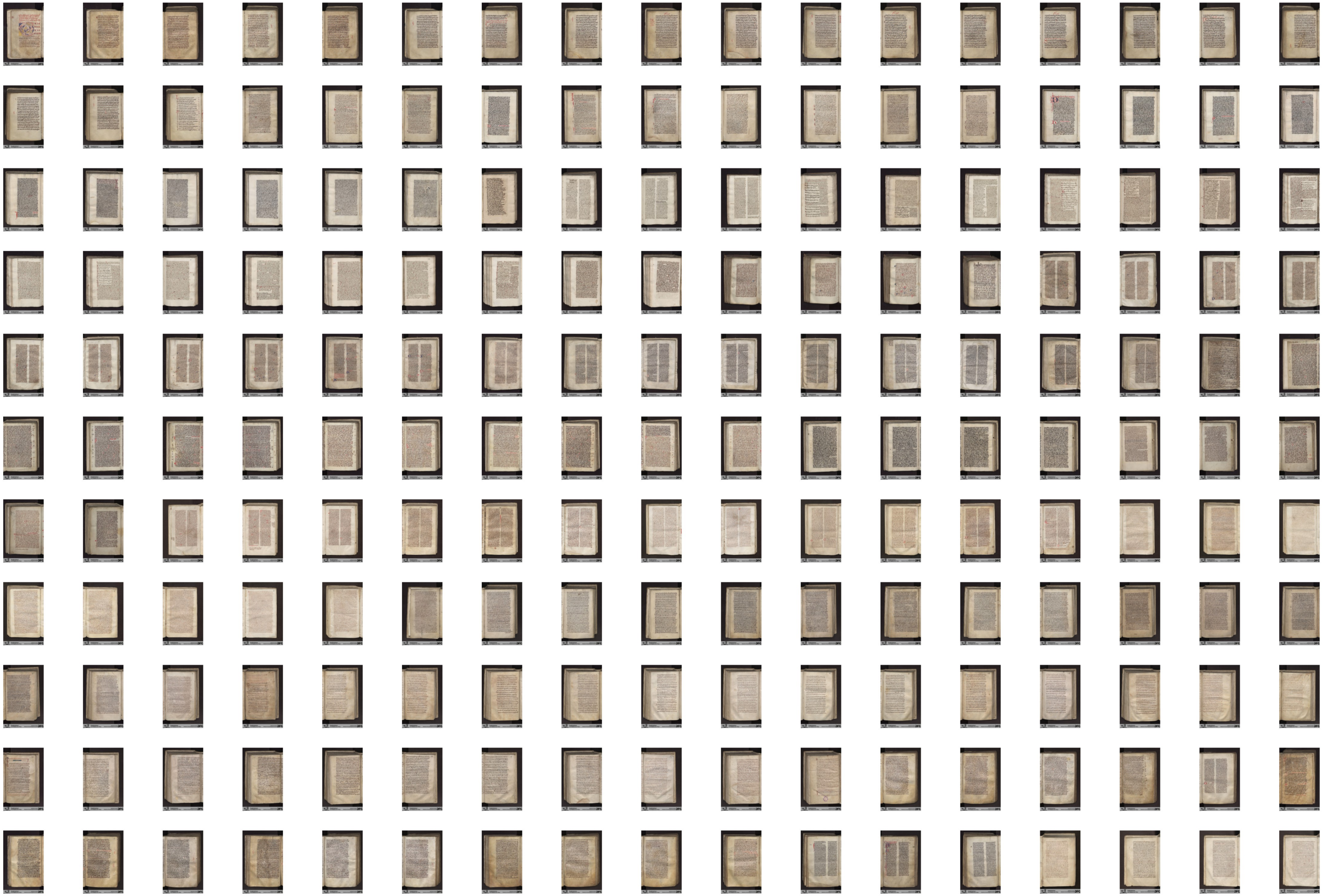
=

```
          <tei:TEI>
          <tei:teiHeader>
          <tei:msDesc>
          <tei:layoutDesc>
          <tei:layout>
          <tei:dimensions type="written" corresp="#written1">
          <tei:height quantity="250" unit="mm" min="249" max="251"
confidence="0.8">250mm</tei:height>
          <tei:width quantity="100" unit="mm" min="98" max="101"
confidence="0.77">100mm</tei:width>
          </tei:dimensions>
          <tei:dimensions type="image" facs="#image1"/>
</tei:layout>
          <tei:layout columns="1" writtenLines="26 30"/>
          </tei:layoutDesc>
          </tei:msDesc>
</tei:teiHeader>
          </tei:TEI>
```


Stadtbibliothek und Stadtarchiv Trier
gefördert von der Deutschen Forschungsgemeinschaft DFG

```
          <tei:TEI>
          <tei:teiHeader>
          <tei:msDesc>
          <tei:layoutDesc>
          <tei:layout>
          <tei:dimensions type="written" corresp="#written1">
          <tei:height quantity="250" unit="mm" min="249" max="251"
confidence="0.8">250mm</tei:height>
          <tei:width quantity="100" unit="mm" min="98" max="101"
confidence="0.77">100mm</tei:width>
          </tei:dimensions>
          <tei:dimensions type="image" facs="#image1"/>
</tei:layout>
          <tei:layout columns="1" writtenLines="26 30"/>
          </tei:layoutDesc>
          </tei:msDesc>
</tei:teiHeader>
          </tei:TEI>
```

# Quantitative Codicology



# Current Case Study (100 codices)

eCodicology

Numerical variables (absolute values)

A) page dimensions (height x width)

B) number of columns

} 20 randomly selected pages

C) number of folios

D) number of blank pages within a codex

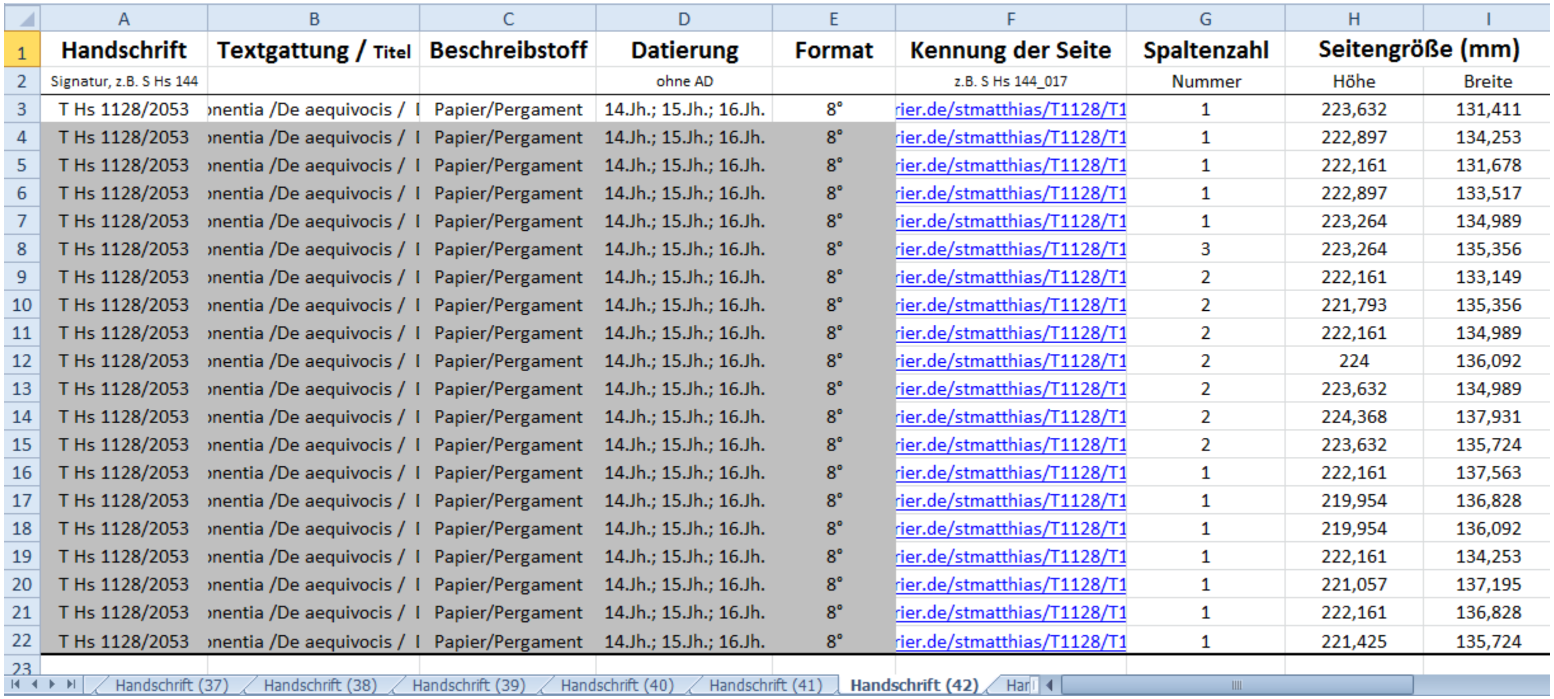
Categorial variables

E) codex format (2°; 4°; 8°)

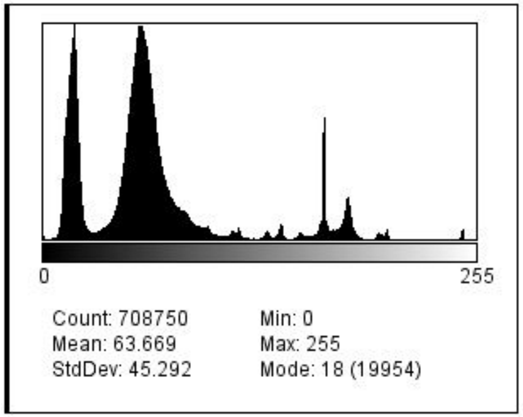
F) dating

G) text genre

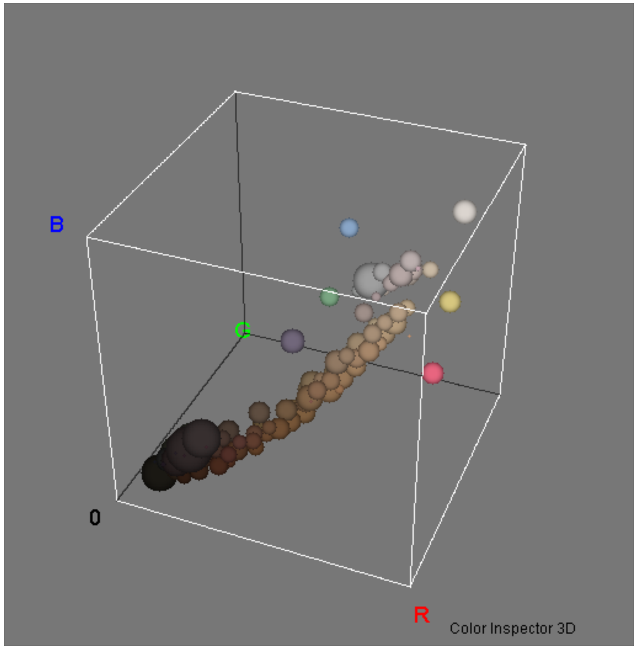
H) writing material (parchment; paper; both) eCodicology



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Handschrift | Textgattung / Titel | Beschreibstoff | Datierung | Format | Kennung der Seite | Spaltenzahl | Seitengröße (mm) | |
| 2 | Signatur, z.B. S Hs 144 | | | ohne AD | | z.B. S Hs 144_017 | Nummer | Höhe | Breite |
| 3 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 223,632 | 131,411 |
| 4 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 222,897 | 134,253 |
| 5 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 222,161 | 131,678 |
| 6 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 222,897 | 133,517 |
| 7 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 223,264 | 134,989 |
| 8 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 3 | 223,264 | 135,356 |
| 9 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 2 | 222,161 | 133,149 |
| 10 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 2 | 221,793 | 135,356 |
| 11 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 2 | 222,161 | 134,989 |
| 12 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 2 | 224 | 136,092 |
| 13 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 2 | 223,632 | 134,989 |
| 14 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 2 | 224,368 | 137,931 |
| 15 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 2 | 223,632 | 135,724 |
| 16 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 222,161 | 137,563 |
| 17 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 219,954 | 136,828 |
| 18 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 219,954 | 136,092 |
| 19 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 222,161 | 134,253 |
| 20 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 221,057 | 137,195 |
| 21 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 222,161 | 136,828 |
| 22 | T Hs 1128/2053 | ›nentia /De aequivocis / I | Papier/Pergament | 14.Jh.; 15.Jh.; 16.Jh. | 8° | rier.de/stmatthias/T1128/T1 | 1 | 221,425 | 135,724 |
| 23 | | | | | | | | | |



Count: 708750
Mean: 63.669
StdDev: 45.292
Min: 0
Max: 255
Mode: 18 (19954)



# Conclusion

Our goals are …

to build correlations between different parameters within groups of **codices with common features** (for ex. dating / text genre / writing material)

to make comparisons between groups of **codices with different similarities** (for ex. codices made of parchment vs. codices made of paper or handwritten codices vs. incunables)

The eCodicology project is associated to

DARIAH-DE