A DIGITAL APPROACH TO HANDWRITTEN DOCUMENTS

B.I.T. - Bureau Ingénieur Tomasi

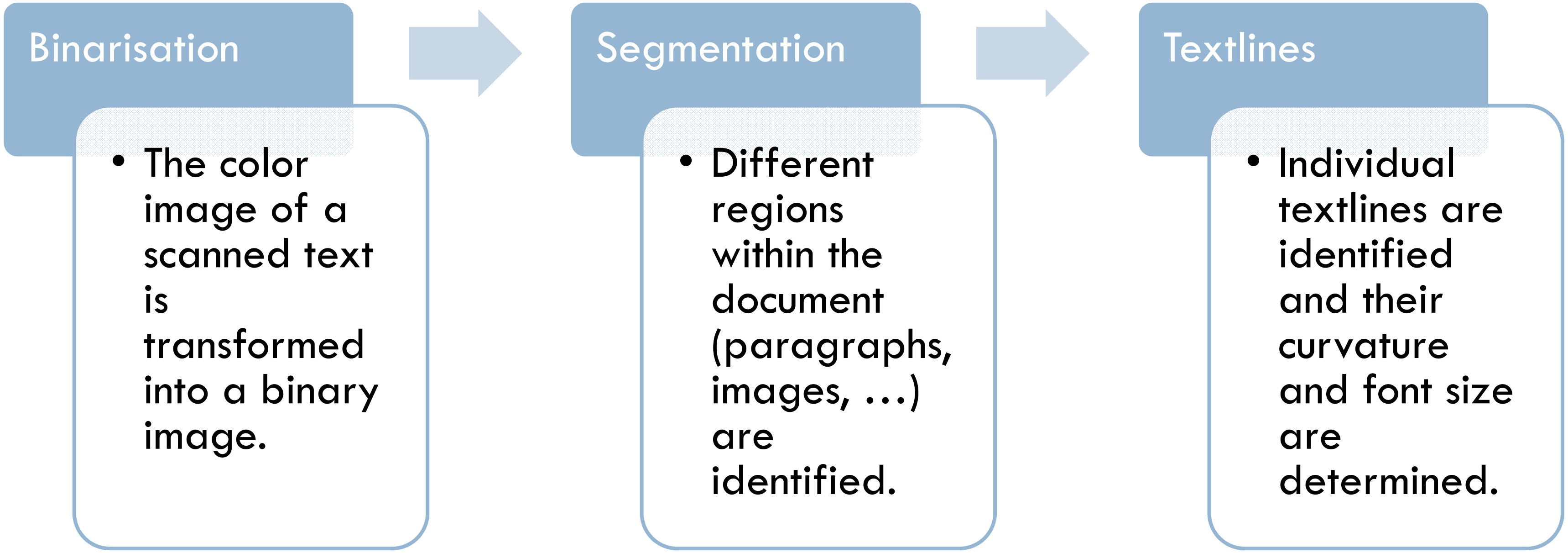
Introduction

- Handwritten documents can for the most part not be read by computers today.
- Our technology such as it has been implemented in the OCR software BIT-Alpha may be a starting point for the development of handwriting analysis tools.
- This digital approach to handwritten documents shall be presented in the following.

Content capture

- Binarisation
- Segmentation
- Textlines

Content capture



Binarisation

- The color image of a scanned text is transformed into a binary image.

Segmentation

- Different regions within the document (paragraphs, images, …) are identified.

Textlines

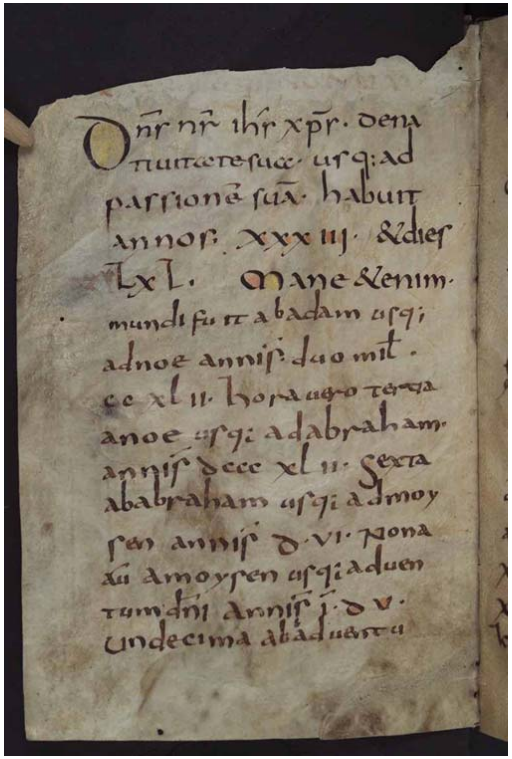
- Individual textlines are identified and their curvature and font size are determined.

Binarisation

- Archived documents are usually scanned color images (300-600dpi).
- For OCR a binary image of 300-400dpi is required.
- Binarisation is a non-trivial task:
  - Details of the characters are to be preserved.
  - Luminosity and contrast can vary across a single page.
  - The text on the backside of the page may shine through.

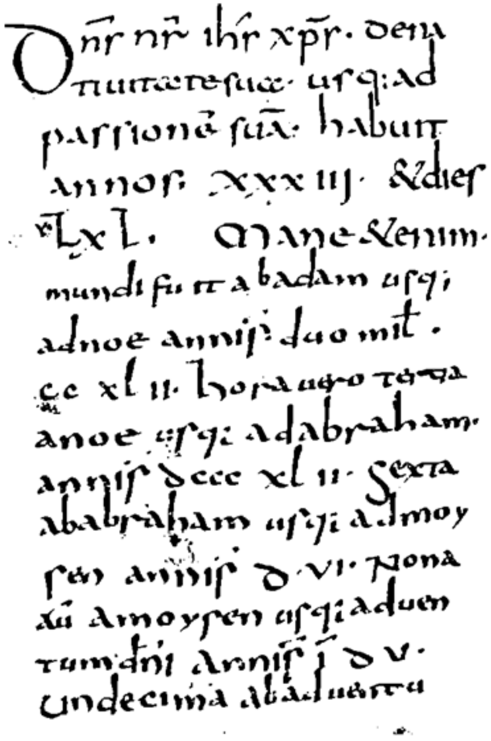
Binarisation

Scanned image

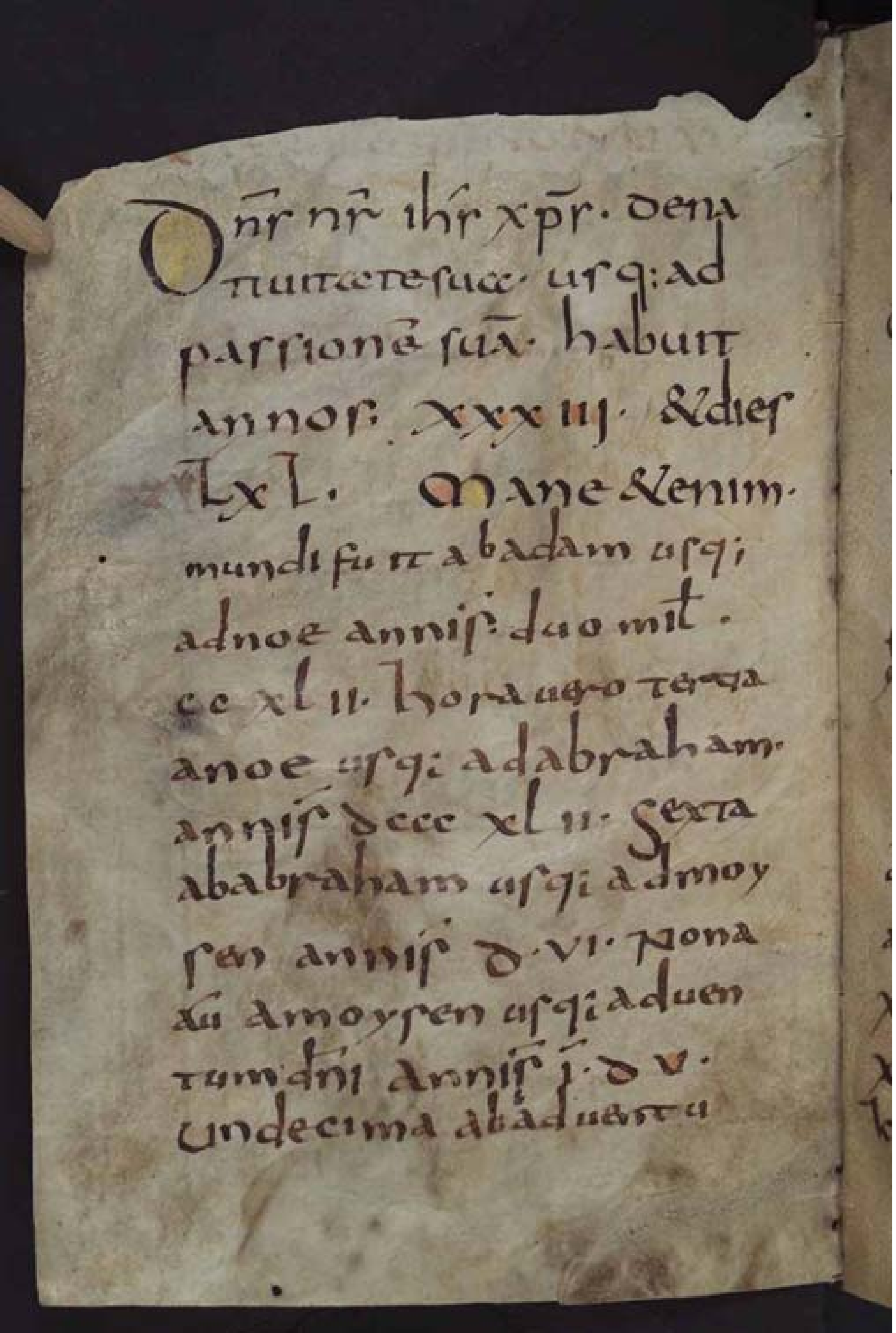


D nr̄ nr̄ ihr̄ xp̄r· dein
tiuitate suae· usq: ad
passionē suā· habuit
annos· xxxiiij· &dies
xl· Mane &enim·
mundi fuit abadam usq;
adnoe annis· duo mil·
cc xl ii· hora autem tertia
anoe usq: adabraham·
annis dccc xl ii· Sexta
ababraham usq: admoy
sen annis d·vi· Nona
au amoysen usq: aduen
tum dni annis i d v·
undecima abaduentu

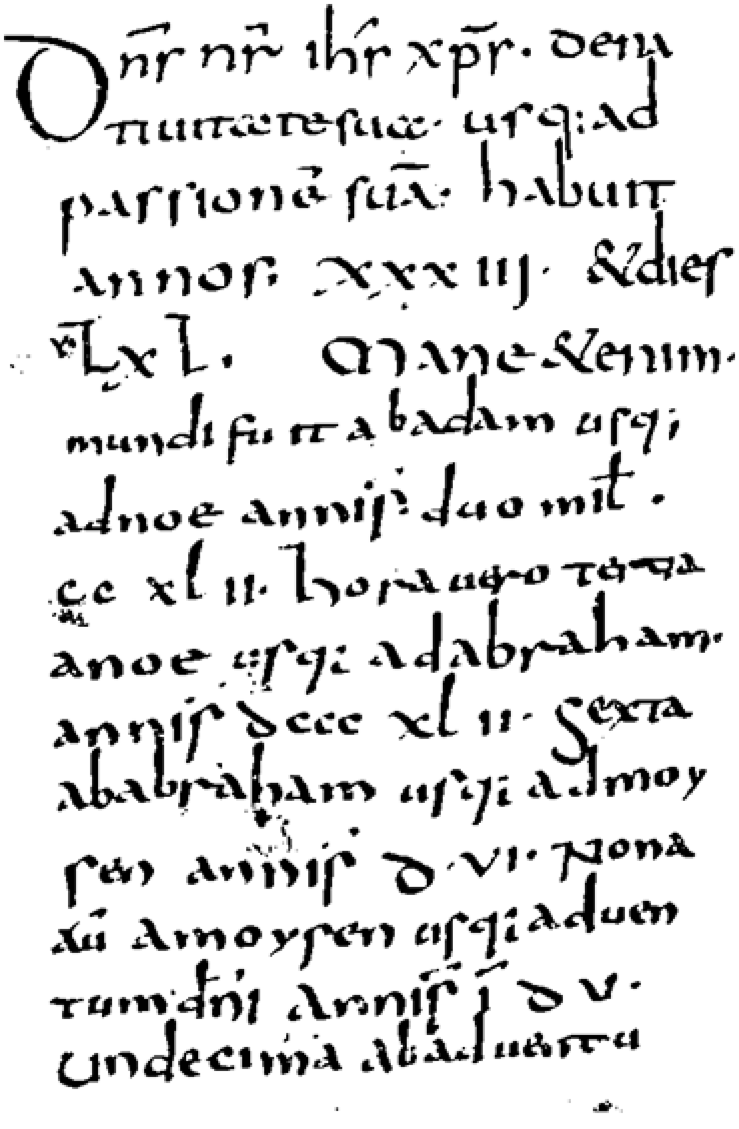
Binary image (mod. Niblack)



D nr̄ nr̄ ihr̄ xp̄r· dein
tiuitate suae· usq: ad
passionē suā· habuit
annos· xxxiiij· &dies
xl· Mane &enim·
mundi fuit abadam usq;
adnoe annis· duo mil·
cc xl ii· hora autem tertia
anoe usq: adabraham·
annis dccc xl ii· Sexta
ababraham usq: admoy
sen annis d·vi· Nona
au amoysen usq: aduen
tum dni annis i d v·
undecima abaduentu



Dn̄s n̄r ih̄s xp̄s. dena
tiuitate suæ· uſq; ad
paſſionē ſuā· habuit
annoſ· XXXIII· & dieſ
LXL· Mane & enim·
mundi fuit ab adam uſq;
ad noe anniſ duo mil·
CCXLII· hora uero tertia
a noe uſq; ad abraham·
anniſ DCCCXLII· Sexta
ab abraham uſq; ad moy
ſen anniſ DVI· Nona
xii a moyſen uſq; ad uen
tum dn̄i anniſ I DV·
undecima ab aduentu



Segmentation

- After binarisation different zones within the document are to be identified:
  - Text regions:
    - Titles
    - Paragraphs
  - Graphics regions:
    - Images/Figures
    - Lines

Segmentation

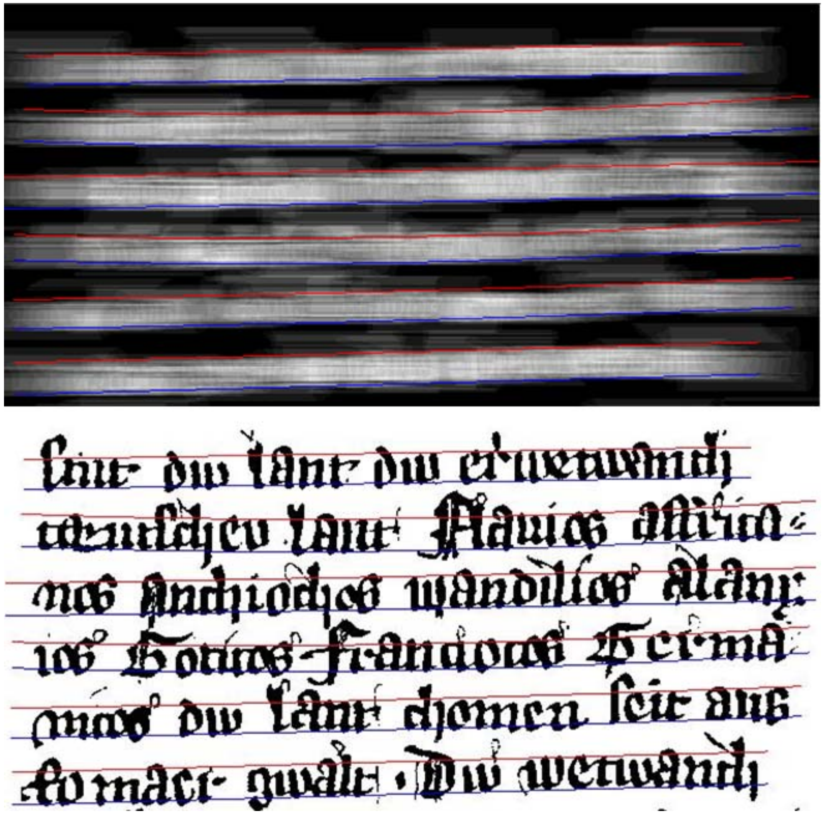
- Segmentation is often combined with a rotation/deskewing of the document in order to obtain horizontal textlines.
- For handwritten documents this can be very difficult: B.I.T. plans to incorporate a bidirectional variance method based on the work of Franck Le Bourgeois (INSA Lyon).

Textlines

- Capturing textlines is a major difficulty to treating handwritten documents:
  - Textlines may overlap.
  - Characters within a textline can be very close or even touch each other.
  - Textlines may not follow a straight line and be curbed.
- Standard methods for printed documents are not suited to handwritten documents.

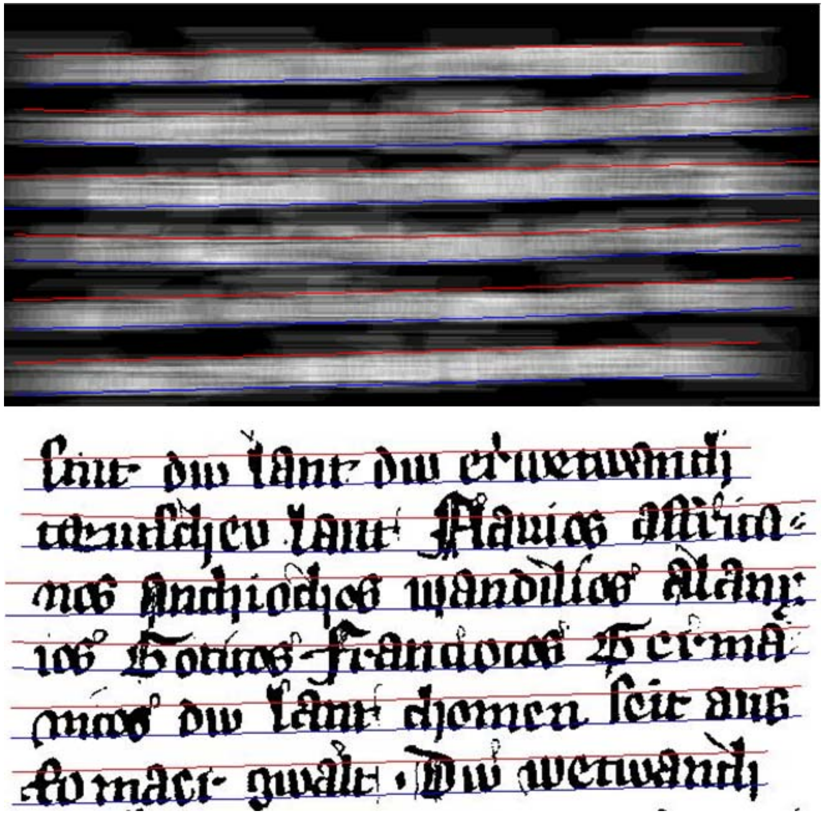
Textlines

- BIT-Alpha implements a method based on the research of the L.I.R.I.S. laboratory surrounding prof. Emptoz in Lyon: a one-dimensional variation analysis is performed to find zones that constitute textlines.



Textlines

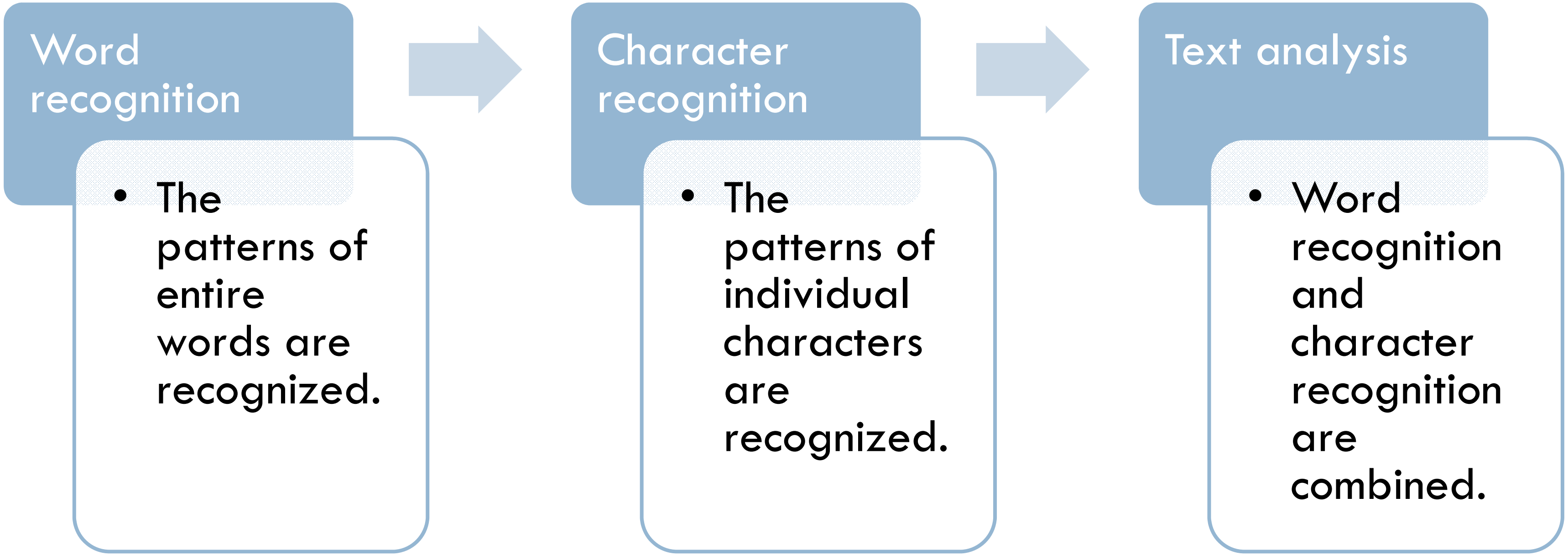
- After identifying zones of textlines mathematical functions have to be found that fit them.
- This allows to define a „topline“(red) and a „baseline“(blue) which also provides a measure for the font size.



Content analysis

- Training and recognition
- Word recognition
- Character recognition
- Text analysis

Content analysis



Word recognition

- The patterns of entire words are recognized.

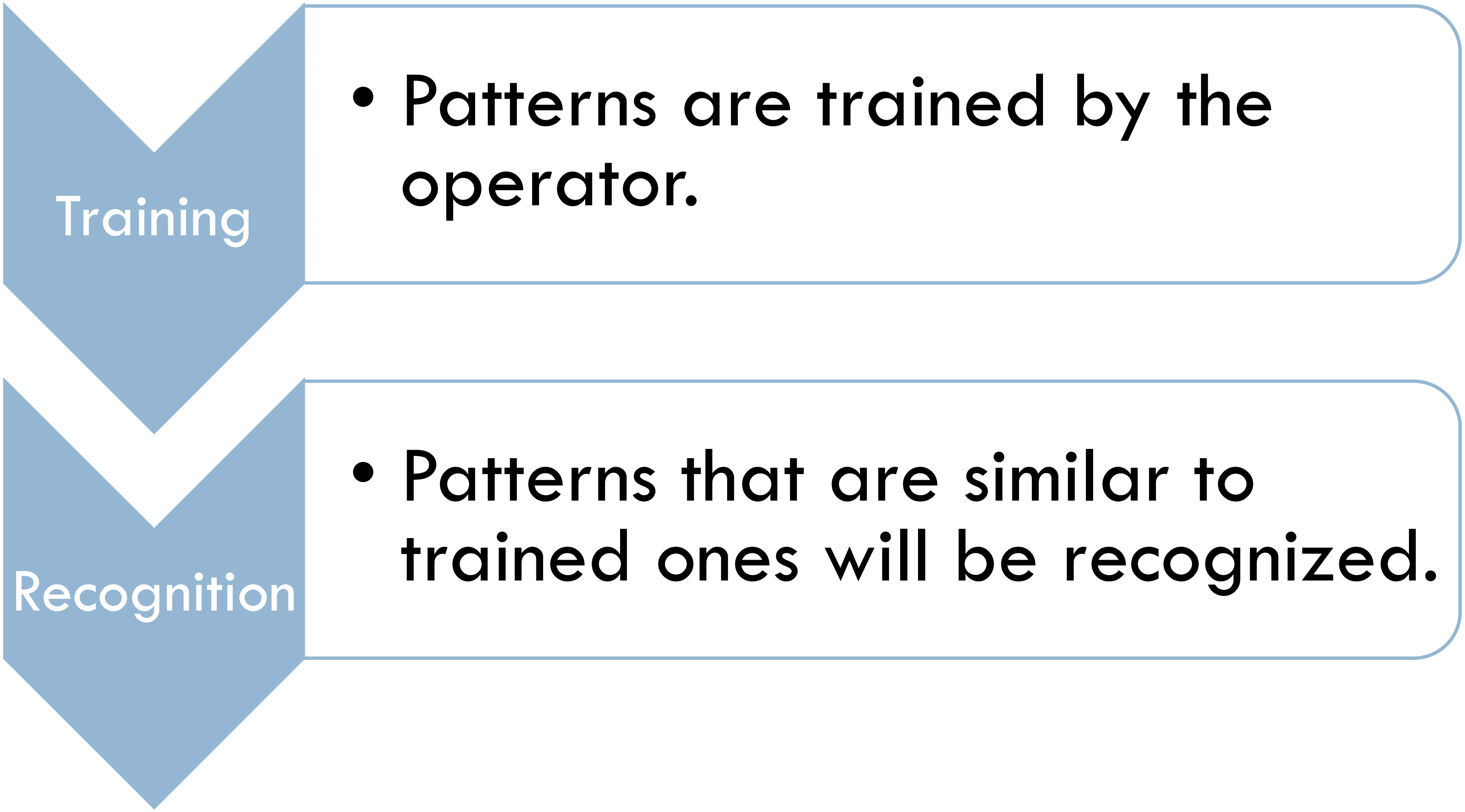
Character recognition

- The patterns of individual characters are recognized.

Text analysis

- Word recognition and character recognition are combined.

Training and recognition



Training

- Patterns are trained by the operator.

Recognition

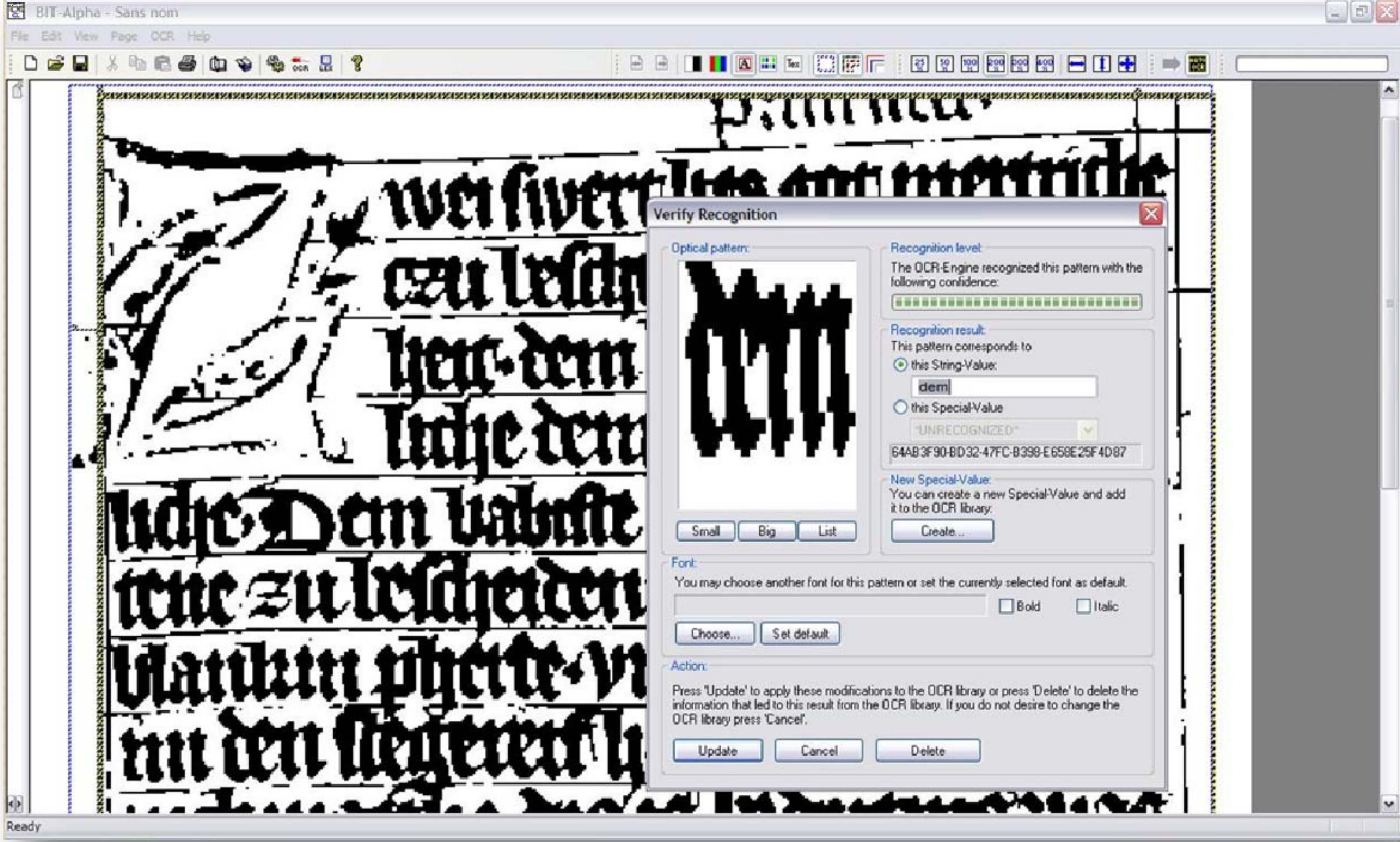
- Patterns that are similar to trained ones will be recognized.

Word recognition

- In handwritten documents individual characters are often joint together and have a high graphical variance
- Identifying entire words can be a natural approach in such cases.
- If not all characters within a word are linked, analyzing the distance between them is required to distinguish gaps between words from gaps between characters.

Word recognition

- BIT-Alpha has been designed for printed documents of the Renaissance where the distance between words is subject to high variance and not more homogeneous than in handwritten texts.
- The analysis of the distance between characters/words is done on a line-by-line basis in BIT-Alpha.



Verify Recognition

Optical pattern:

Recognition level:

The OCR-Engine recognized this pattern with the following confidence:

Recognition result:

This pattern corresponds to

● this String-Value: dem

○ this Special-Value: "UNRECOGNIZED"

64AB3F90-BD32-47FC-B398-E658E25F4D87

New Special-Value:

You can create a new Special-Value and add it to the OCR library:

Font:

You may choose another font for this pattern or set the currently selected font as default.

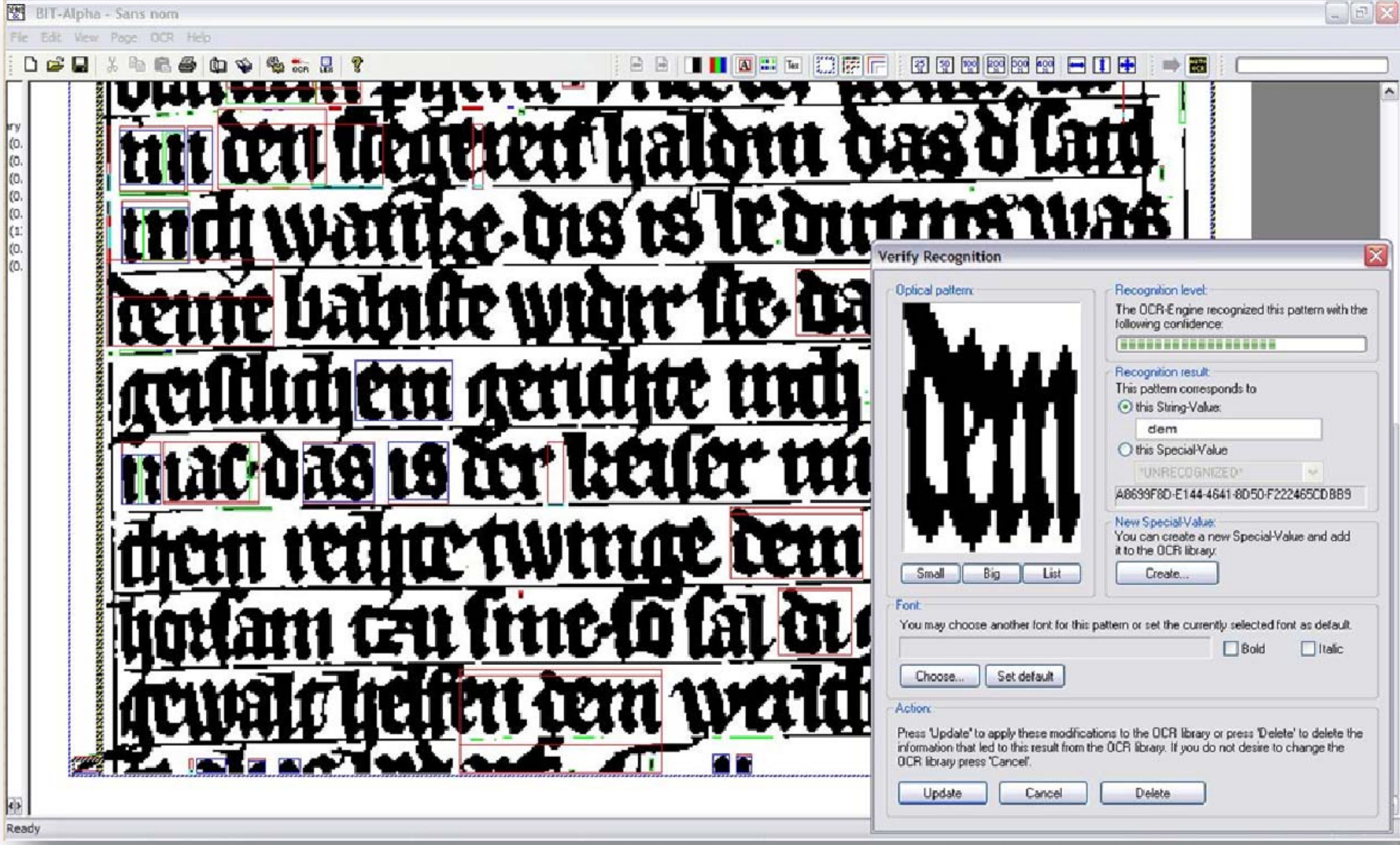
□ Bold □ Italic

Action:

Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

Training of words

The pattern of the word „dem“ is trained in this example.



Verify Recognition

Optical pattern:

Recognition level:

The OCR-Engine recognized this pattern with the following confidence:

Recognition result:

This pattern corresponds to

● this String-Value: dem

○ this Special-Value: *UNRECOGNIZED*

A8699F8D-E144-4641-8D50-F222465CDBB9

New Special-Value:

You can create a new Special-Value and add it to the OCR library.

Font:

You may choose another font for this pattern or set the currently selected font as default.

□ Bold □ Italic

Action:

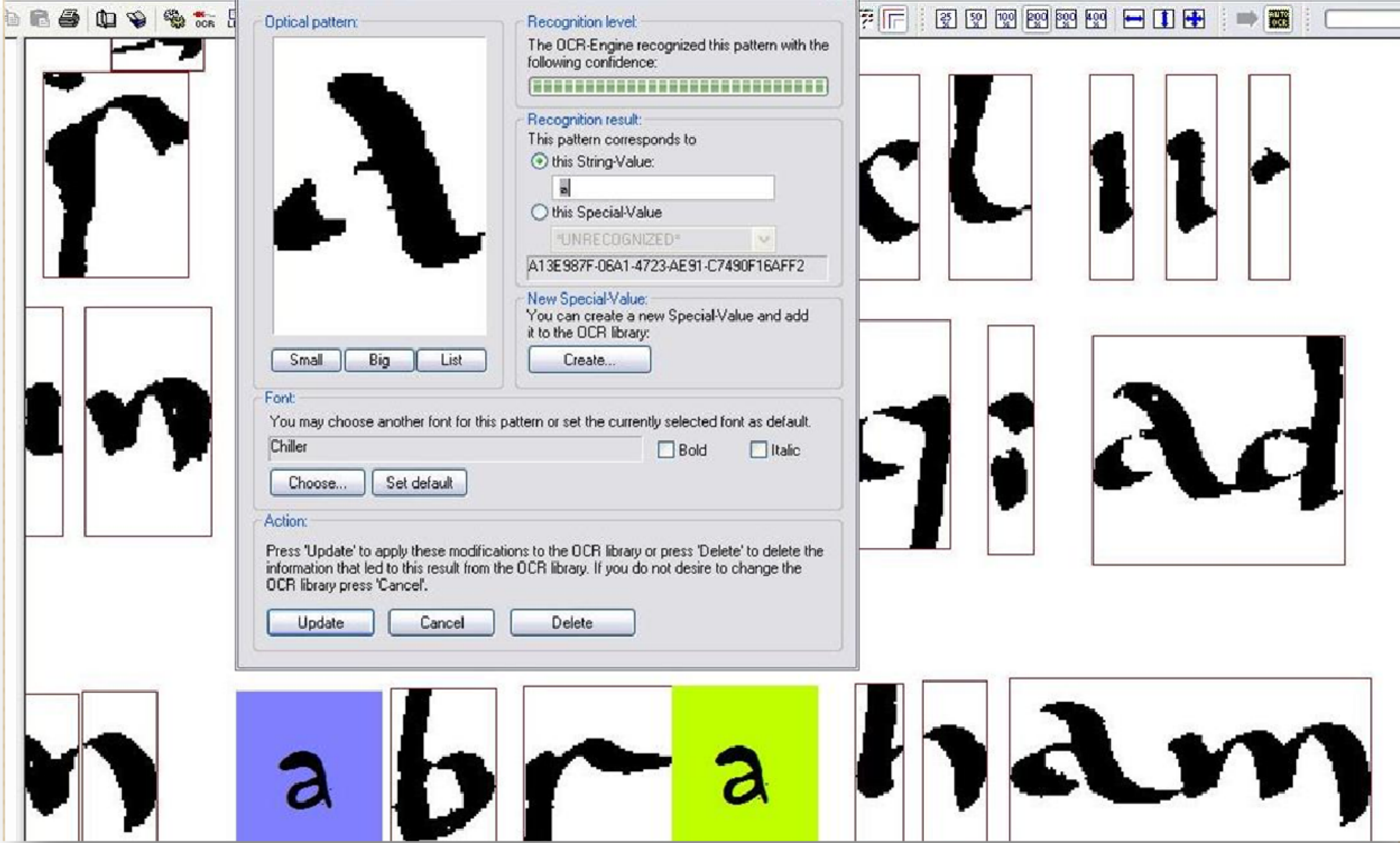
Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

Recognition of words

A new occurrence of a pattern similar to the previously trained one is recognized.

Character recognition

- In order to obtain a reliable recognition only identifying the patterns of entire words is not enough.
- The isolation of individual characters within word patterns is therefore necessary.
- The approach taken with BIT-Alpha is to recognize sub patterns of individual characters within the patterns of words. We call this process „disjunction“.



Recognition level:

The OCR-Engine recognized this pattern with the following confidence:

Recognition result:

This pattern corresponds to

● this String-Value: a

○ this Special-Value: *UNRECOGNIZED*

A13E987F-06A1-4723-AE91-C7490F16AFF2

New Special-Value:

You can create a new Special-Value and add it to the OCR library:

Font:

You may choose another font for this pattern or set the currently selected font as default.

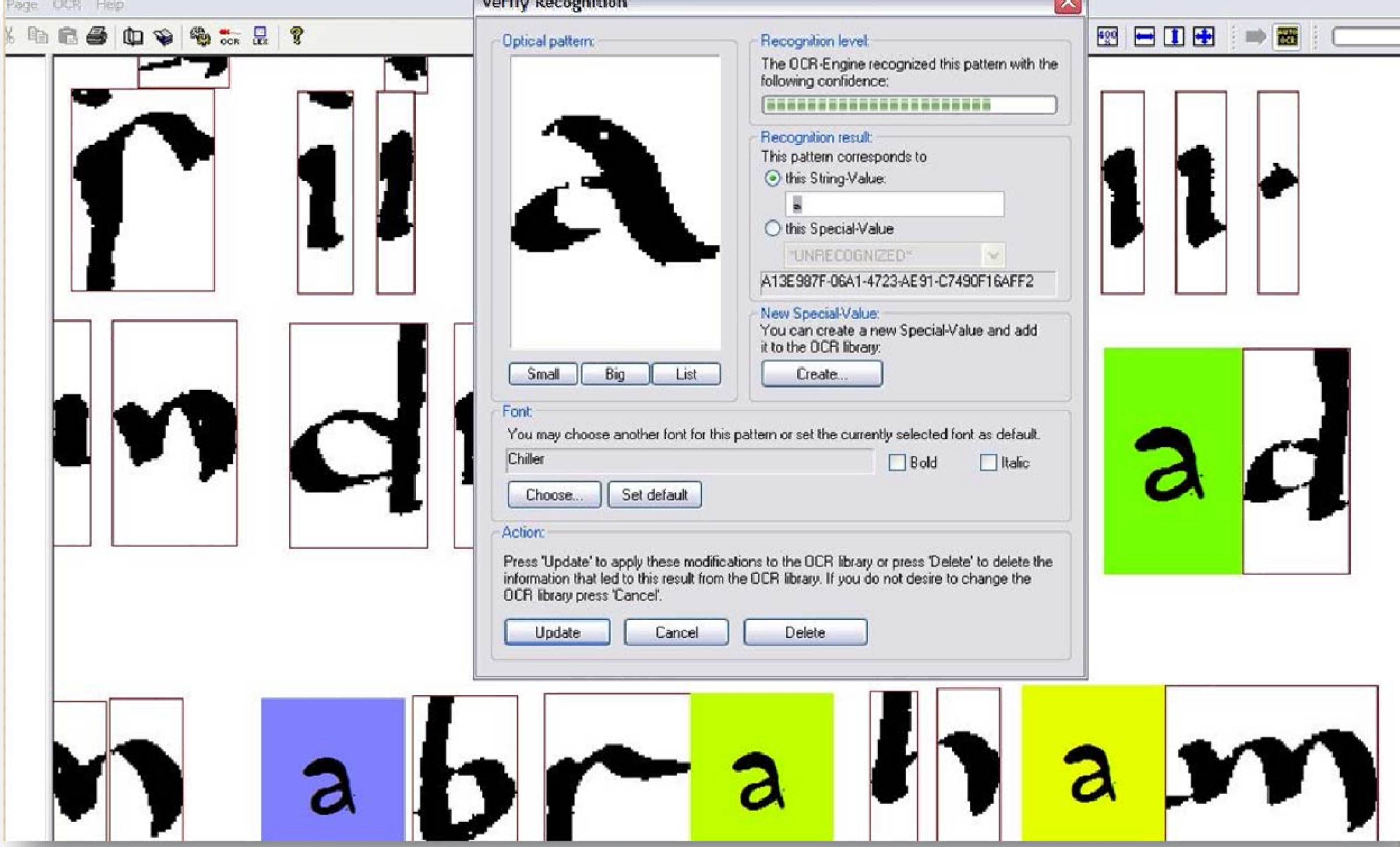
Chiller □ Bold □ Italic

Action:

Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

Recognition of characters

This screenshot shows the recognition of multiple patterns for the character „a“. Note that a blue background indicates a pattern which is identical to a trained pattern, whereas a green background indicates a very similar but not identical pattern.



Optical pattern:

Recognition level:
The OCR-Engine recognized this pattern with the following confidence:

Recognition result:
This pattern corresponds to
● this String-Value: a
○ this Special-Value "UNRECOGNIZED"
A13E987F-06A1-4723-AE91-C7490F16AFF2

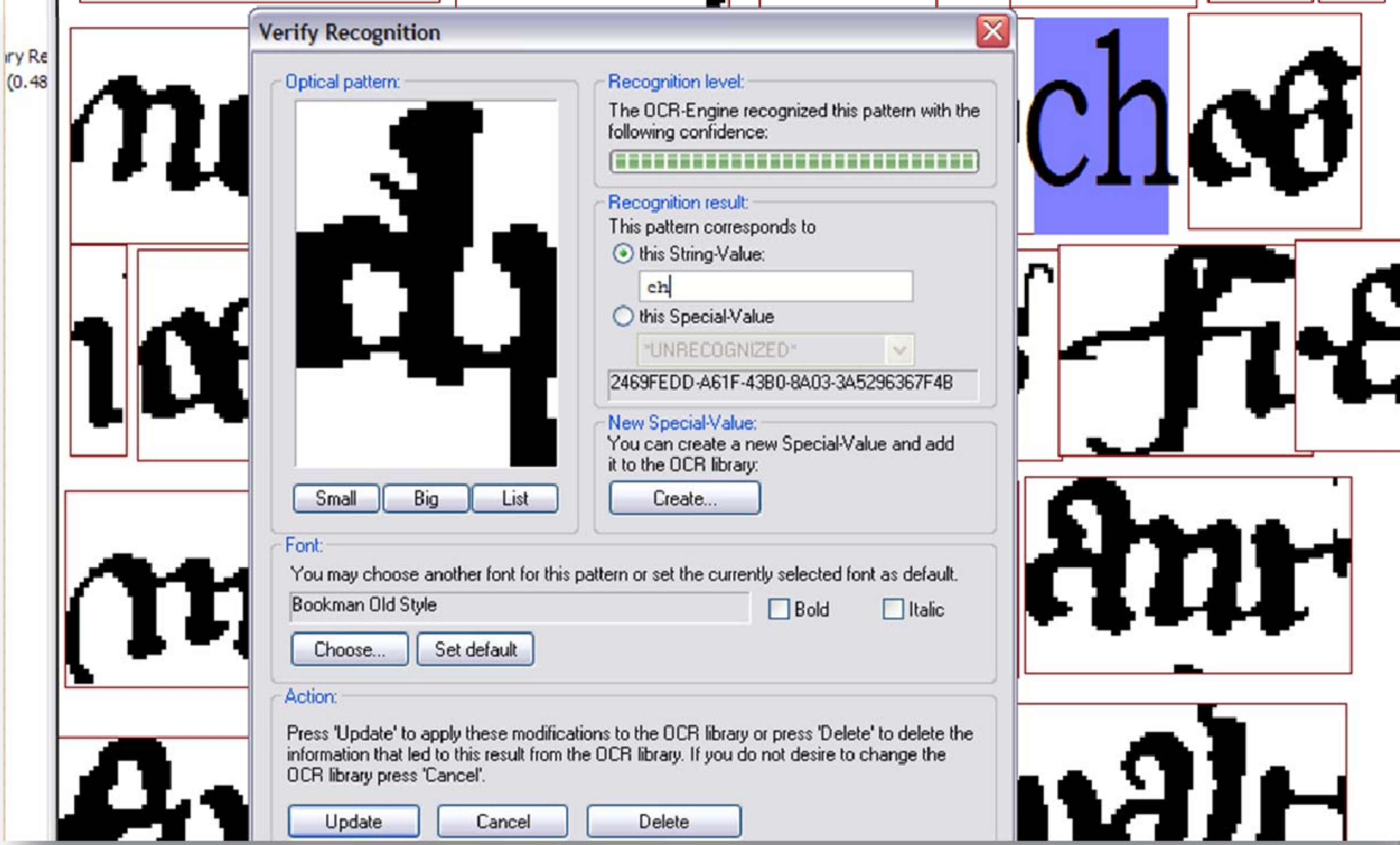
New Special-Value:
You can create a new Special-Value and add it to the OCR library:

Font:
You may choose another font for this pattern or set the currently selected font as default.
Chiller □ Bold □ Italic

Action:
Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

Recognition of characters and disjunction

Once the disjunction has been activated other patterns for the character „a“ are recognized as sub patterns within words.



Verify Recognition

Optical pattern:

Recognition level:

The OCR-Engine recognized this pattern with the following confidence:

Recognition result:

This pattern corresponds to

● this String-Value: ch

○ this Special-Value *UNRECOGNIZED*

2469FEDD-A61F-43B0-8A03-3A5296367F4B

New Special-Value:

You can create a new Special-Value and add it to the OCR library:

Font:

You may choose another font for this pattern or set the currently selected font as default.

Bookman Old Style □ Bold □ Italic

Action:

Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

Ligatures

Ligatures can be trained and recognized as a whole.

Text analysis

- Word recognition proposes strings for each word by comparing the patterns of entire words.
- Character recognition proposes strings for each word by comparing the patterns of constituent characters.
- The correlation of both provides a first clue to the transcription of the text.
- All information concerning the transcription comes essentially from pattern recognition so far.

Content valorization

- Text transcription
- Text export
- Scribe identification

Text transcription

- The results from text analysis have to be confronted with linguistic considerations.
- Assuming the language of the text is known the words that have been suggested by text analysis can be matched against a database which contains the words of the language.
- A distance measure for strings is needed to find the best match.

Text transcription

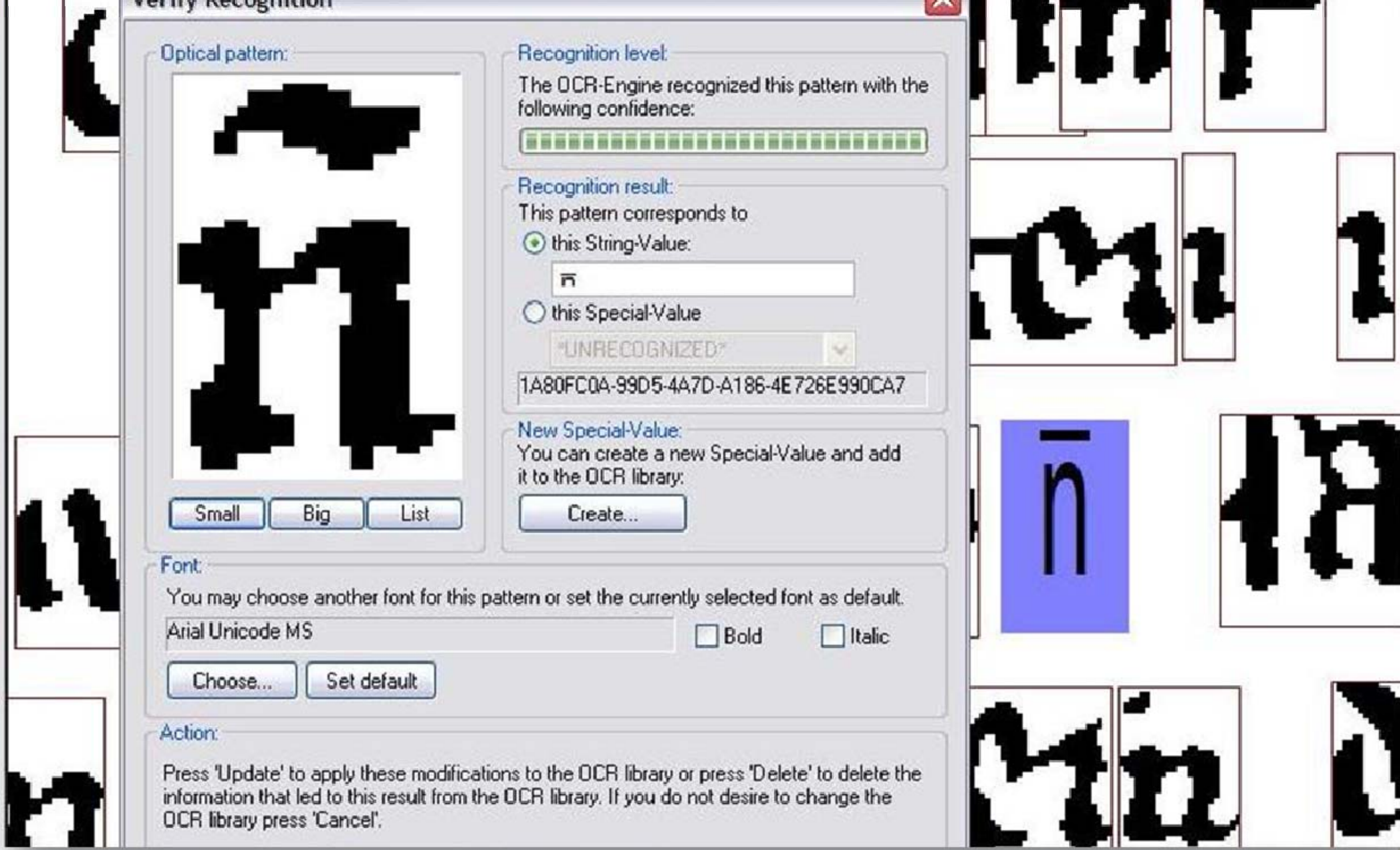
- The number of edit operations (replacement, insertion, deletion, …) required to transform one string into another gives a distance measure („edit distance“).

Text transcription

- The number of edit operations (replacement, insertion, deletion, …) required to transform one string into another gives a distance measure („edit distance“).
- Errors stemming from OCR-lecture are not necessarily comparable to editing errors (typos). The concept of edit distance has to be extended and adapted for this purpose.

Text transcription

- In handwritten documents abbreviations have been frequently used.
- We suggest to use Unicode symbols for abbreviations in conformance with the MUFI (Medieval Unicode Font Initiative) recommendations.
- In order to obtain a text that is readily readable by non-specialists these abbreviations have to be expanded.



Optical pattern:

Recognition level:

The OCR-Engine recognized this pattern with the following confidence:

Recognition result:

This pattern corresponds to

● this String-Value: n̄

○ this Special-Value: *UNRECOGNIZED*

1A80FC0A-99D5-4A7D-A186-4E726E990CA7

New Special-Value:

You can create a new Special-Value and add it to the OCR library:

Font:

You may choose another font for this pattern or set the currently selected font as default.

Arial Unicode MS □ Bold □ Italic

Action:

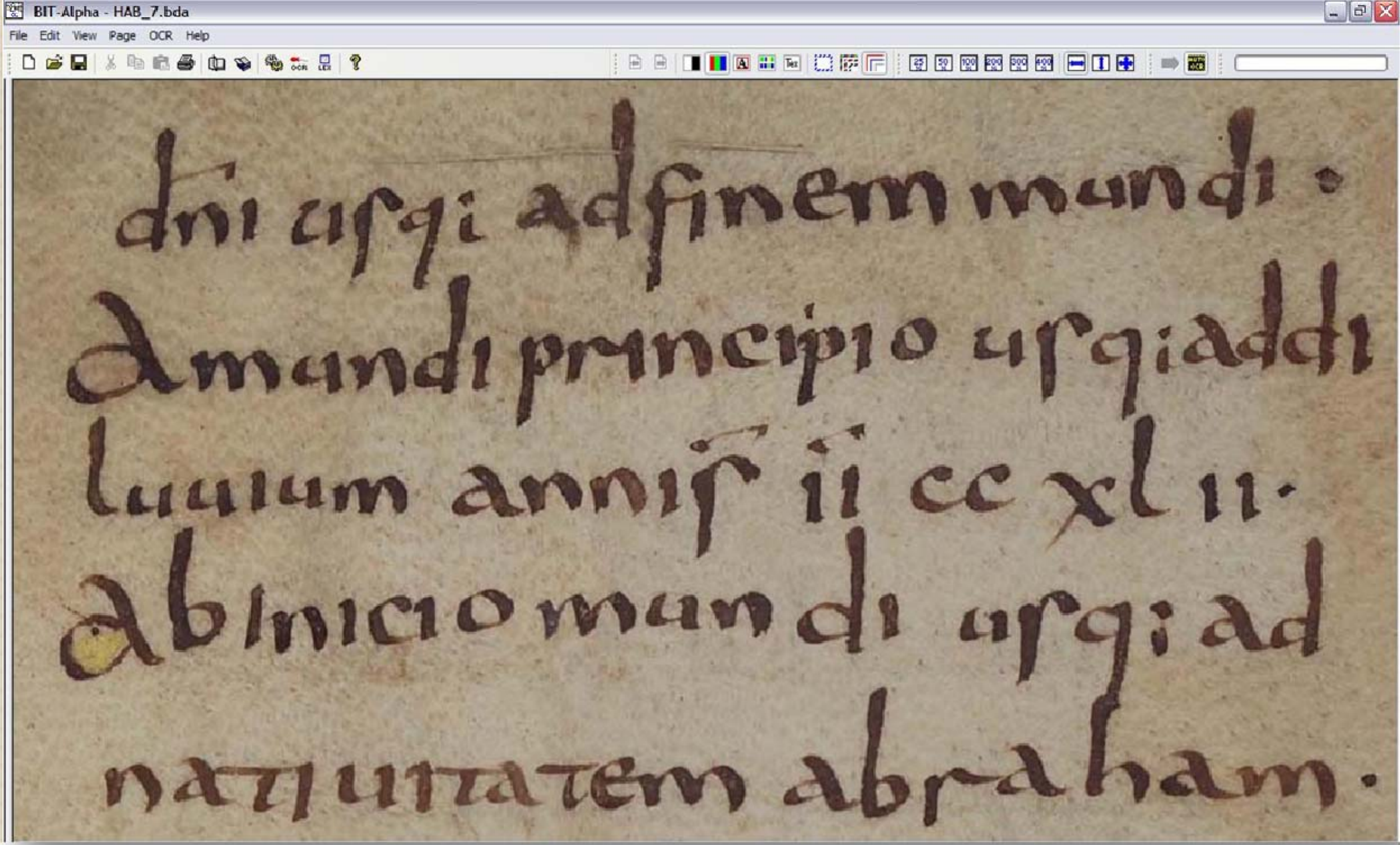
Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

Unicode abbreviation

Abbreviations can be trained and recognized as a Unicode symbols.

Text export

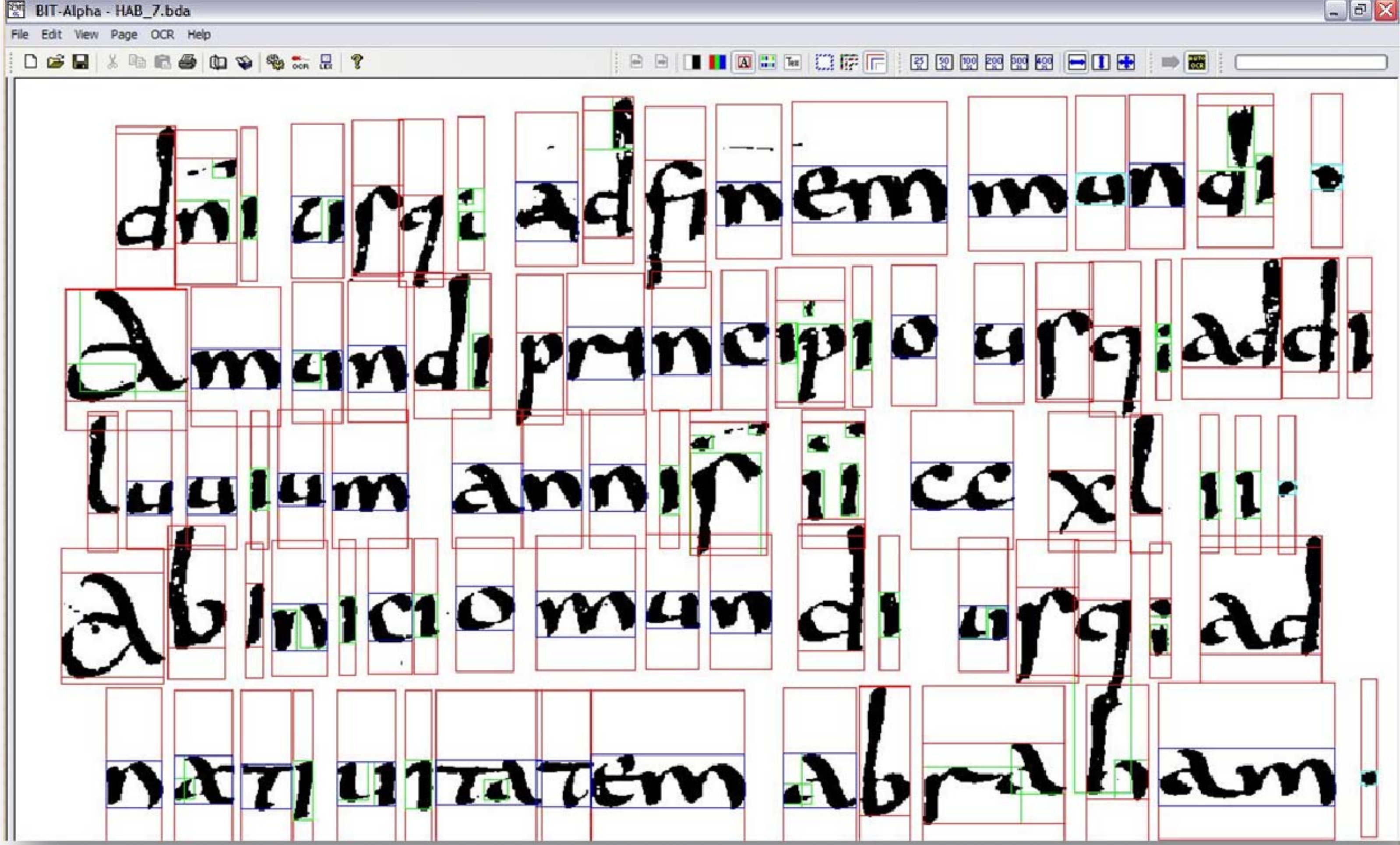
- BIT-Alpha offers a variety of formats for the export of the processed document, including:
  - PDF file containing a color-image of the document in the background and transparent text in the foreground.
  - PDF file containing a binarized image of the document in the background and transparent text in the foreground.
  - PDF file containing only text (opaque).
  - XML file according to the METS/ALTO standard.



dni uſqi adfinem mundi ·
amundi principio uſqi addi
luuium anniſ ii cc xl ii ·
abinicio mundi uſqi ad
natiuitatem abraham ·

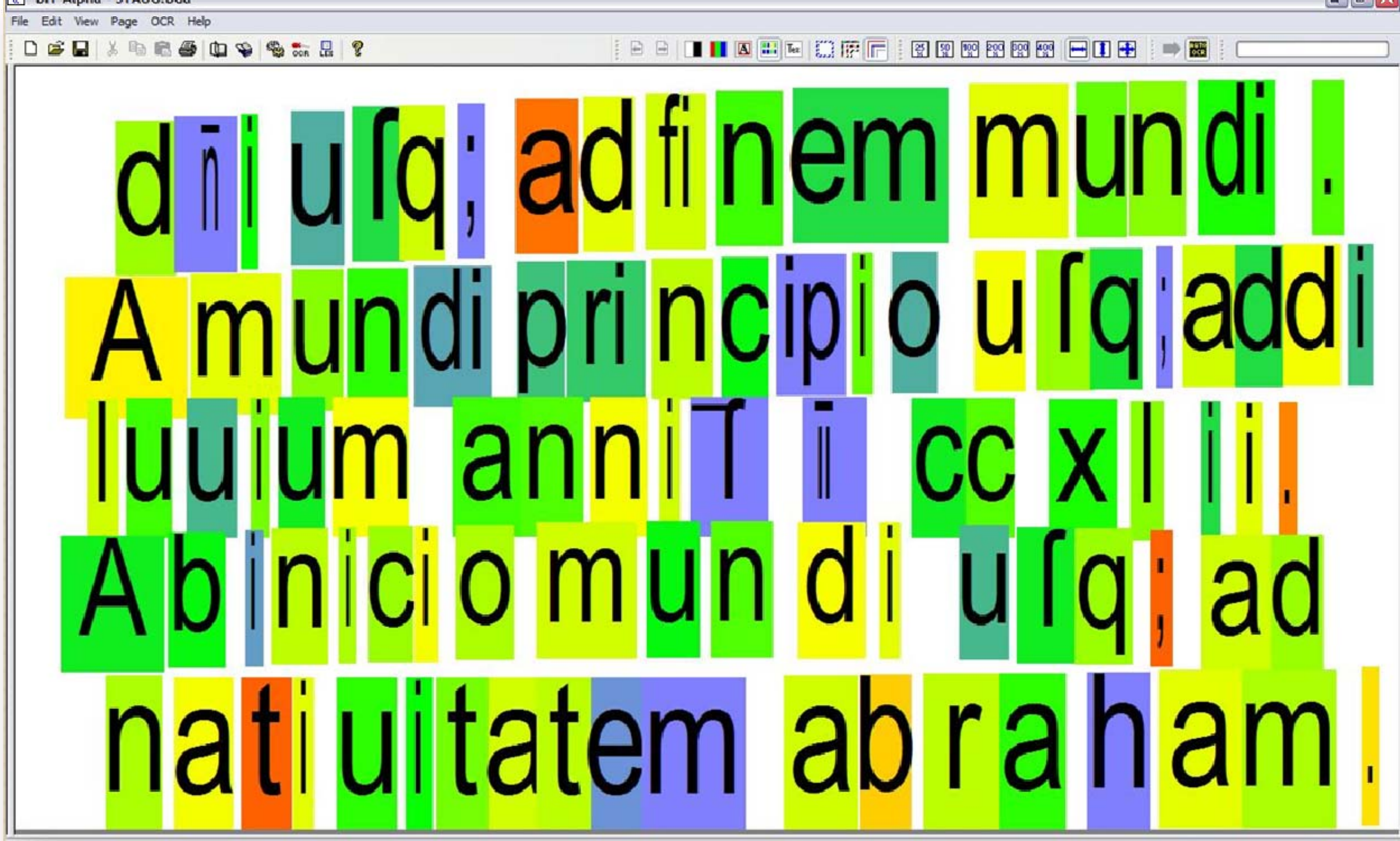
Scanned document

Screenshot of BIT-Alpha showing the scanned document (color image, 300dpi).



Binarisation and Segmentation

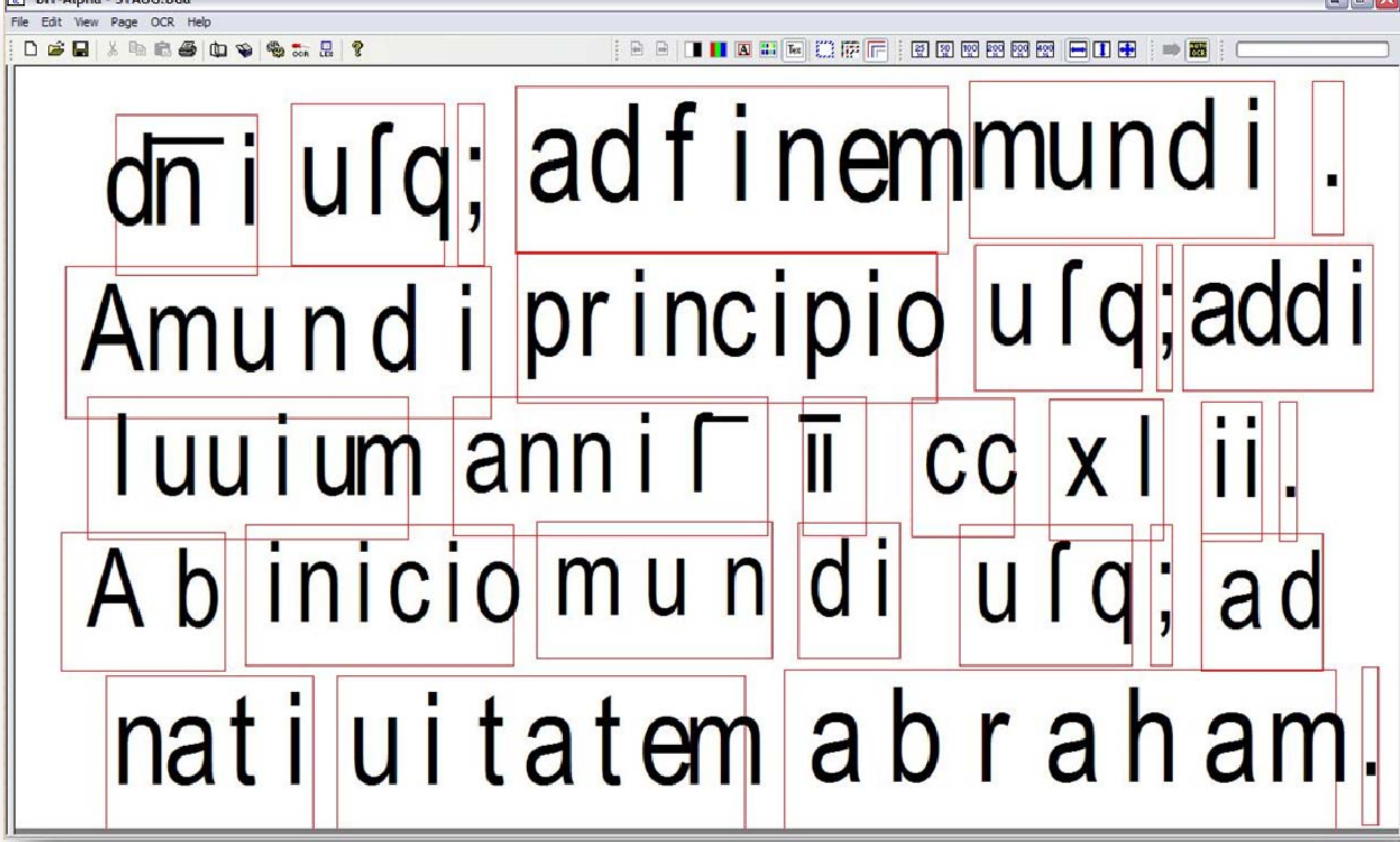
Screenshot of BIT-Alpha showing the binarized document (mod. Niblack) and the results of content capture.



dñi uſq; ad finem mundi .
A mundi principio uſq; addi
luuium anni Ī ī ccxlii.
Ab inicio mundi uſq; ad
natiuitatem abraham .

Text analysis

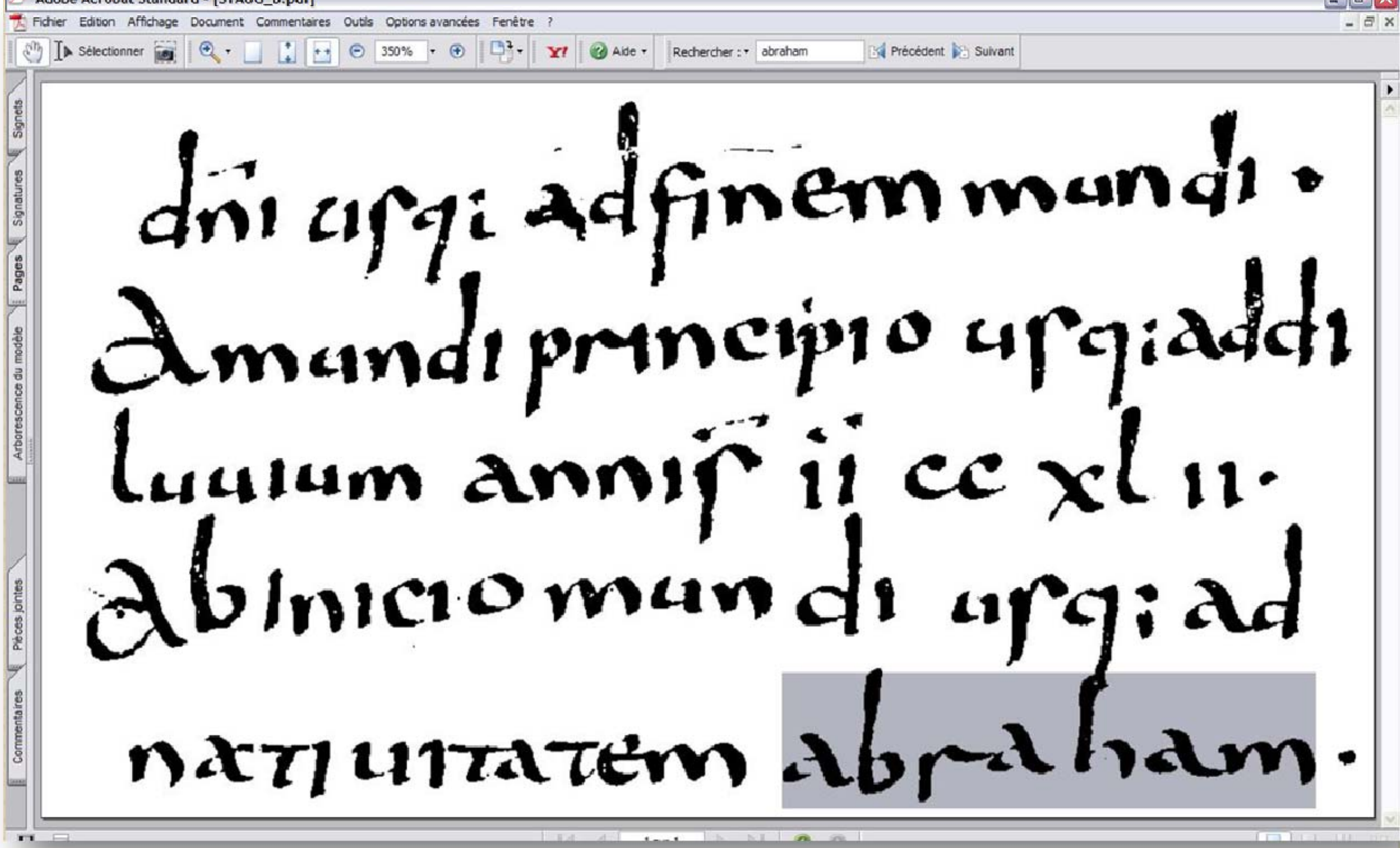
Screenshot of BIT-Alpha showing the result of text analysis (Unicode symbols are used). The reliability of the recognition is shown by colors (from blue=„perfect“ to red=„doubtful“).



dn̄i uſq; adfinemmundi .
Amundi principio uſq;addi
luuium anniſ īi cc xl ii.
Ab inicio mundi uſq; ad
natiuitatem abraham.

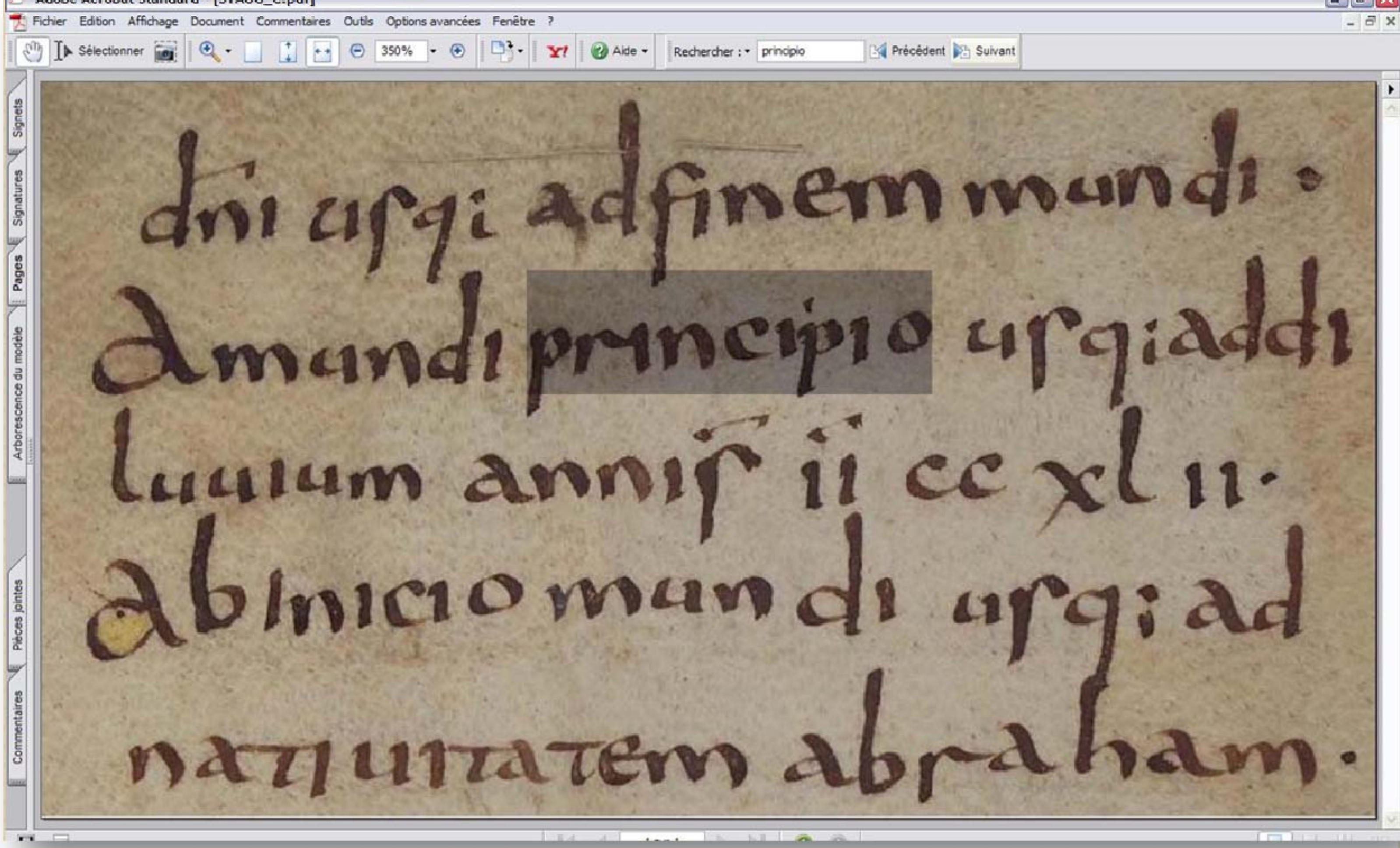
Text transcription

Screenshot of BIT-Alpha showing the transcribed text.



PDF (binary)

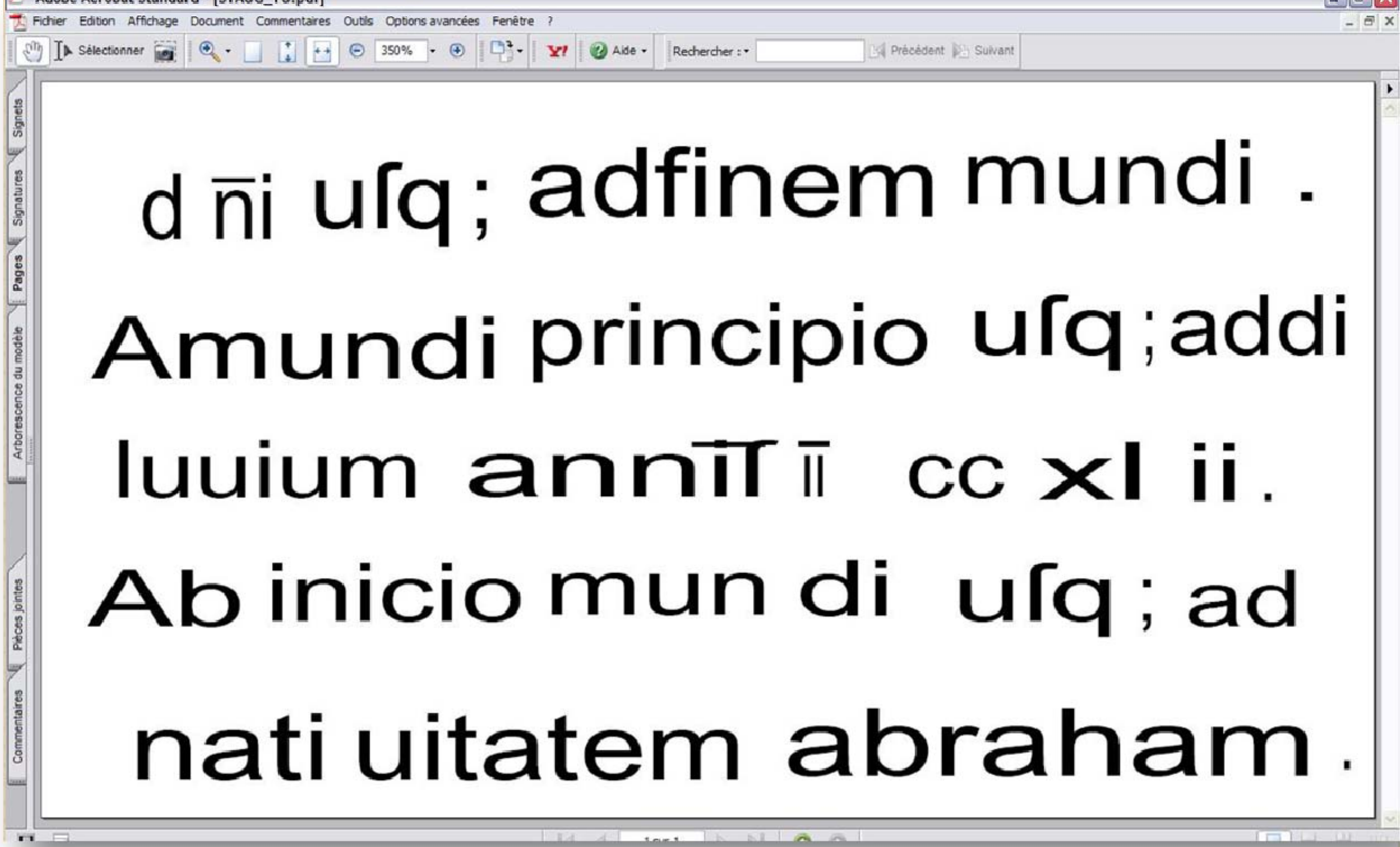
Screenshot of a PDF file with binary image which has been exported by BIT-Alpha. The word „abraham“ was searched and has been found.



dni usqi adfinem mundi ·
amundi principio usqiaddi
luuium annis ii cc xl ii·
abinicio mundi usqi ad
natiuitatem abraham·

PDF (color)

Screenshot of a PDF file with color image which has been exported by BIT-Alpha. The word „principio“ was searched and has been found.



d n̄i uſq; adfinem mundi .
Amundi principio uſq;addi
luuium anniī ī cc xl ii .
Ab inicio mun di uſq ; ad
nati uitatem abraham .

PDF (text only)

Screenshot of a PDF file with only text which has been exported by BIT-Alpha.

Scribe identification

- In paleography the identity of the person that has written a text (scribe) is of great interest.
- Due to the detailed geometric properties that are delivered BIT-Alpha can be a valuable tool for scribe identification.
- The information provided by BIT-Alpha can be grouped into textline specific and character specific data.
- Such data typically consists of expectation values and standard deviations.

Scribe identification

- Textline specific data may include (without going into details):
  - Height and thickness of textlines
  - Distance between words
  - Length of words
  - Curvature of textlines
  - Slope of textlines
  - …

Scribe identification

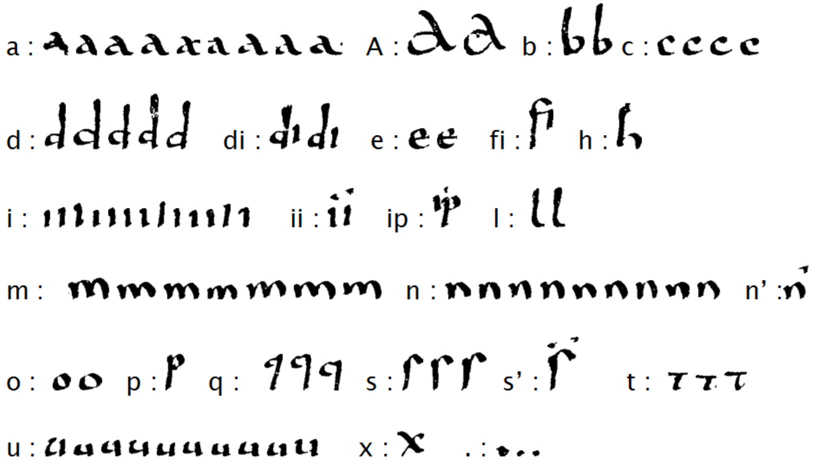
- Character specific data may include (without going into details):
  - Size of characters
  - Size of upper/lower-case characters
  - Difference of size between the largest and smallest character within a line
  - Size of accents and i-points
  - Vertical position of accents and i-points
  - …

Scribe identification

- The pattern recognition capabilities of BIT-Alpha can be used to measure the distance (similarity) between characters written by different scribes.
- Combined with the aforementioned textline and character specific data this suggests the definition of a distance measure for handwritings, which however can only be conceived to be done under the guidance of paleographers.

Scribe identification

- □ The patterns of characters can be exported as bitmaps by BIT-Alpha and will automatically be grouped into subdirectories according to the symbol to which they correspond.



Conclusion

- The transcription of handwritten documents requires a combination of pattern analysis and linguistic considerations.
- Our approach allows to conceive of a tool that can help with the transcription of texts and scribe identification. The contribution of specialists would be essential, especially to the definition of a distance measure for handwritings.