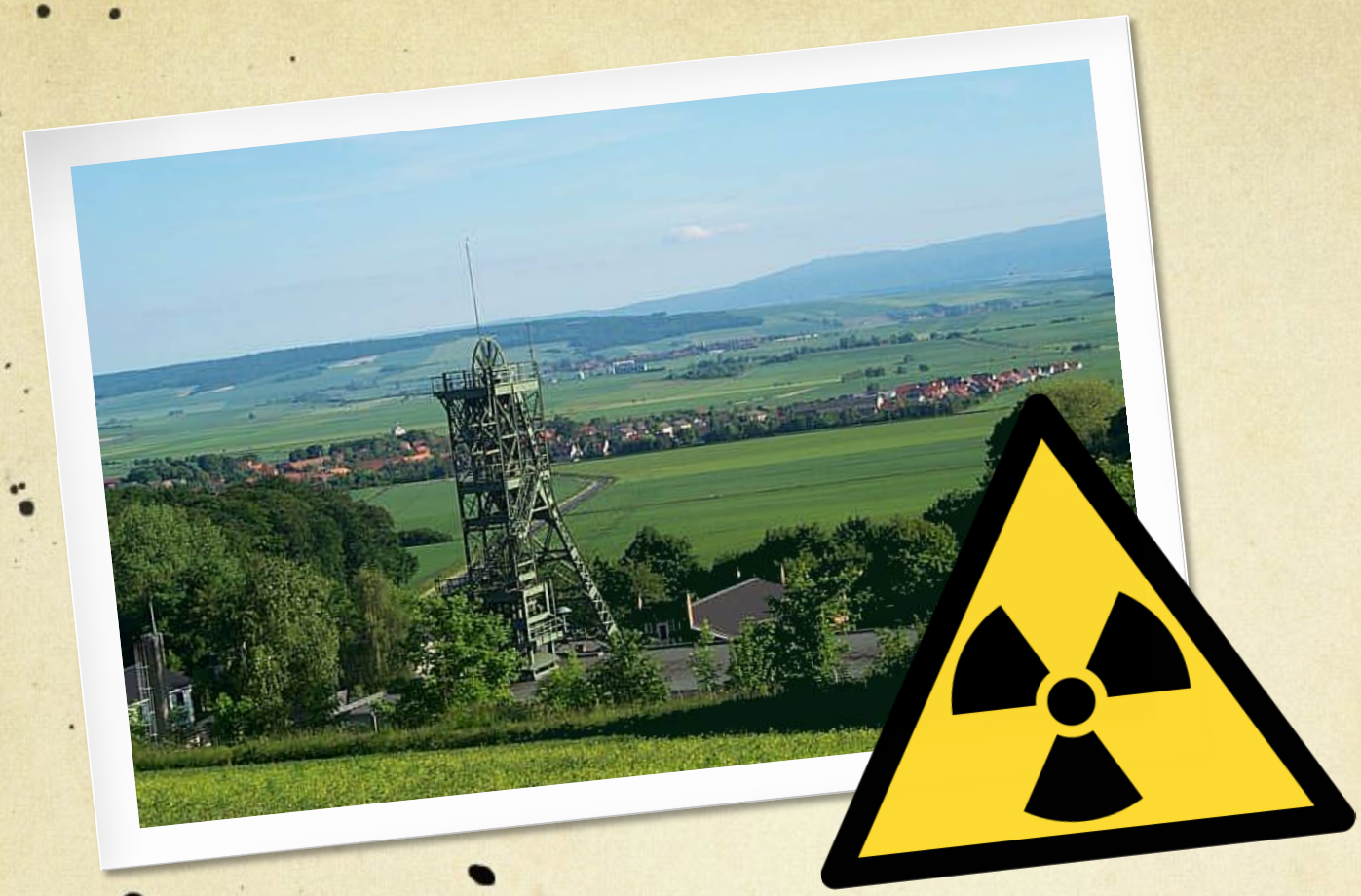
OCR of Legacy Documents

as a Building Block in Industrial Disaster Prevention

Isemann, D., Niekler, A., Preßler, B., Viereck F. and Heyer G.

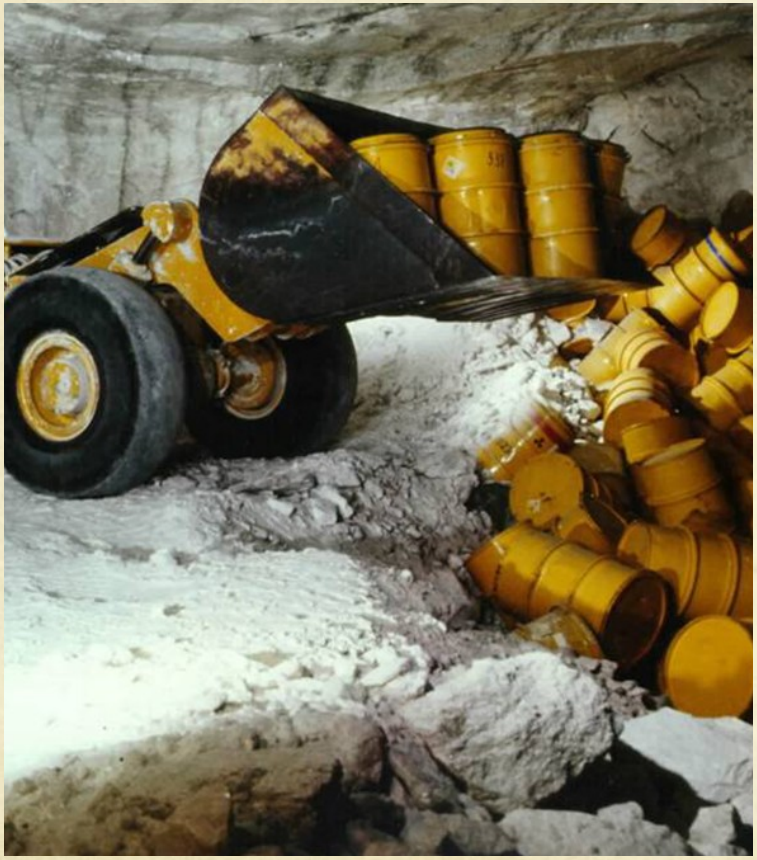


Background

Project: “Knowledge Management for Legacy Documents in Science, Administration and Industry”

The Asse Pit

Asse History

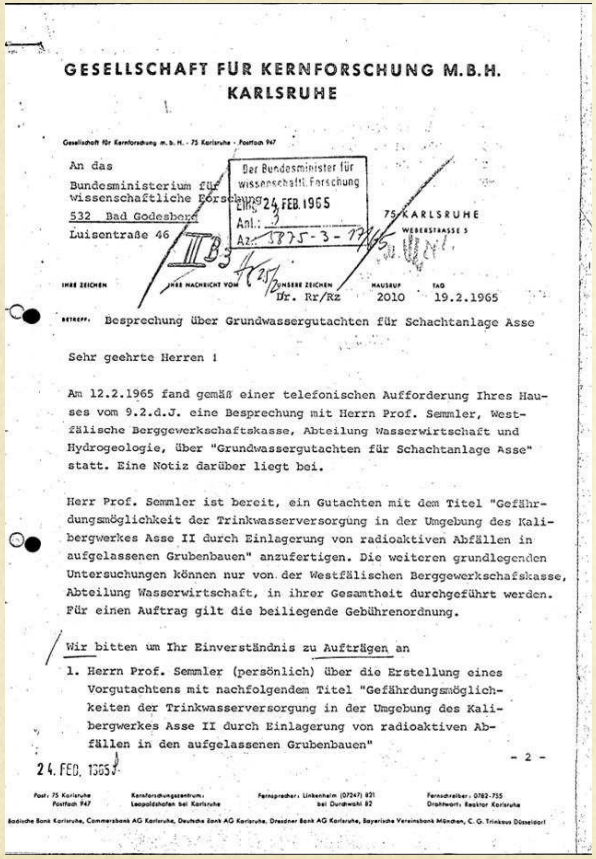


- 1909-1964: Salt mine
- 1964-1978: “Exploratory final storage” (Versuchsendlager) for radioactive waste
- 1. 1. 1979: No new radioactive material accepted
- 1979—now: Managing and maintaining the site and its contents
- 1988: Ingress of water into the pit
- 2009: Change of management, political pressures

Asse Legacy Files

- In the order of 5000 file folders documenting / including (no claim to completeness):
  - Geological and mining surveys
  - Radioactive waste management and reactive transport studies
  - Scientific and administrative correspondence
  - Telephone protocols and meeting minutes
  - Delivery notes / receipts for radioactive substances
  - Reports / assessments of extraordinary events (e.g. earthquakes, operational incidents, etc.)
  - Guidelines and written exhortations making reference to legal and regulatory frameworks governing nuclear waste disposal
- Huge source of information of political, historical and practical interest (current plan is to empty out the nuclear waste repository)

Example Documents



GESELLSCHAFT FÜR KERNFORSCHUNG M.B.H.
KARLSRUHE

Gesellschaft für Kernforschung m. b. H. · 75 Karlsruhe · Postfach 947

An das
Bundesministerium für
wissenschaftliche Forschung
532 Bad Godesberg
Luisenstraße 46

Der Bundesminister für wissenschaftl. Forschung
Eing.: 24. FEB. 1965
Anl.: 3
Az.: 1875-3-7715

75 KARLSRUHE
WEBERSTRASSE 5

IHRE ZEICHEN	IHRE NACHRICHT VOM	UNSERE ZEICHEN	HAUSRUF	TAG
		Dr. Rz/Rz	2010	19.2.1965

BETREFF: Besprechung über Grundwassergutachten für Schachtanlage Asse

Sehr geehrte Herren !

Am 12.2.1965 fand gemäß einer telefonischen Aufforderung Ihres Hauses vom 9.2.d.J. eine Besprechung mit Herrn Prof. Semmler, Westfälische Berggewerkschaftskasse, Abteilung Wasserwirtschaft und Hydrogeologie, über "Grundwassergutachten für Schachtanlage Asse" statt. Eine Notiz darüber liegt bei.

Herr Prof. Semmler ist bereit, ein Gutachten mit dem Titel "Gefährdungsmöglichkeit der Trinkwasserversorgung in der Umgebung des Kalibergwerkes Asse II durch Einlagerung von radioaktiven Abfällen in aufgelassenen Grubenbauen" anzufertigen. Die weiteren grundlegenden Untersuchungen können nur von der Westfälischen Berggewerkschafskasse, Abteilung Wasserwirtschaft, in ihrer Gesamtheit durchgeführt werden. Für einen Auftrag gilt die beiliegende Gebührenordnung.

Wir bitten um Ihr Einverständnis zu Aufträgen an

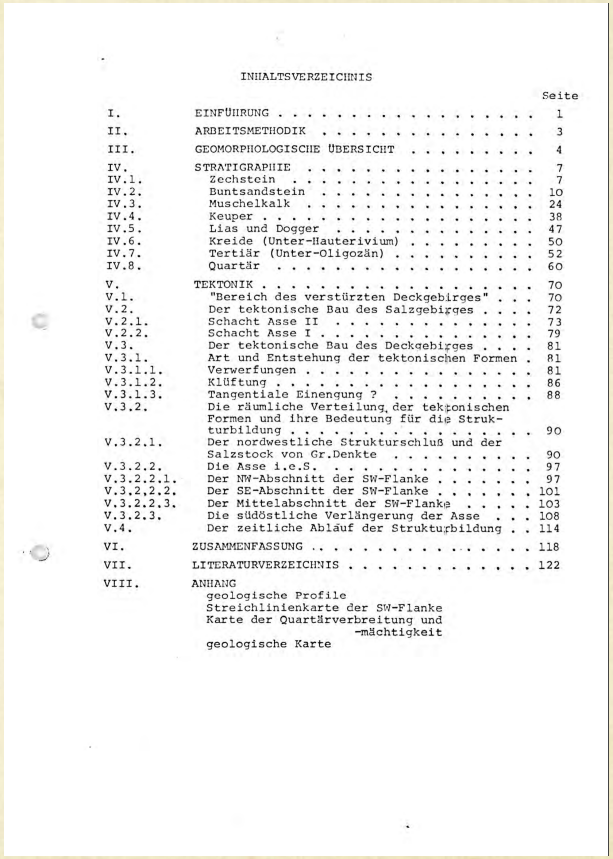
1. Herrn Prof. Semmler (persönlich) über die Erstellung eines Vorgutachtens mit nachfolgendem Titel "Gefährdungsmöglichkeiten der Trinkwasserversorgung in der Umgebung des Kalibergwerkes Asse II durch Einlagerung von radioaktiven Abfällen in den aufgelassenen Grubenbauen"

- 2 -

24. FEB. 1965

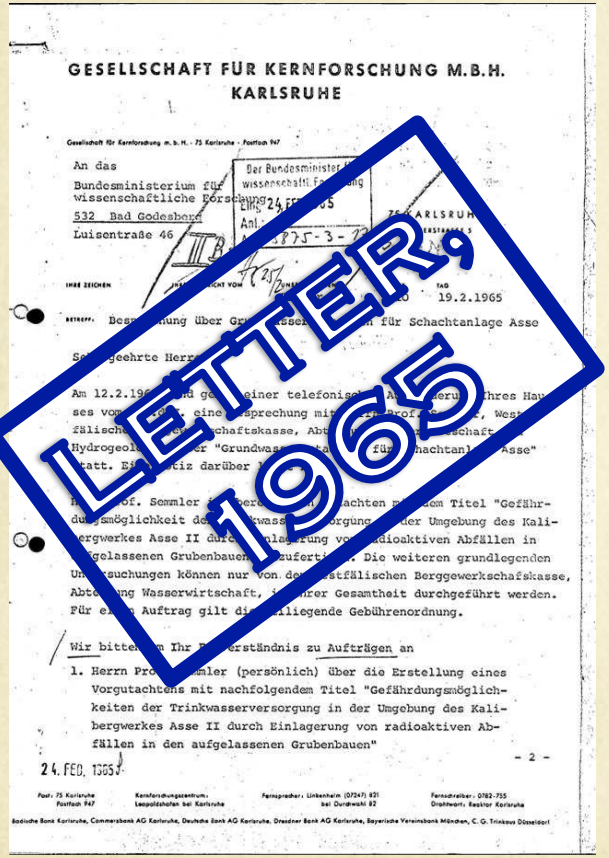
Post: 75 Karlsruhe Postfach 947 · Kernforschungszentrum: Leopoldshafen bei Karlsruhe · Fernsprecher: Linkenheim (07247) 821 bei Durchwahl 82 · Fernschreiber: 0782-755 Drahtwort: Reaktor Karlsruhe

Badische Bank Karlsruhe, Commerzbank AG Karlsruhe, Deutsche Bank AG Karlsruhe, Dresdner Bank AG Karlsruhe, Bayerische Vereinsbank München, C. G. Trinkaus Düsseldorf



INHALTSVERZEICHNIS

Example Documents



GESELLSCHAFT FÜR KERNFORSCHUNG M.B.H.
KARLSRUHE

Gesellschaft für Kernforschung m. b. H. · 75 Karlsruhe · Postfach 947

An das
Bundesministerium für
wissenschaftliche Forschung
532 Bad Godesberg
Luisenstraße 46

Der Bundesminister für wissenschaftliche Forschung
Eing. 24. FEB. 1965
Anl.

IHRE ZEICHEN

TAG 19.2.1965

BETREFF: Besprechung über Gr... für Schachtanlage Asse

Sehr geehrte Herr...

Am 12.2.196... einer telefonisch... Ihres Hau... ses vom ... eine Besprechung mit ... Prof. ..., Wes... fälische ... schaftskasse, Abt... schaft ... Hydrogeol... r "Grundwas... ... für ... achtanl... Asse" ... statt. Ei... tiz darüber ...

... of. Semmler ... achten ... em Titel "Gefähr... dungsmöglichkeit de... kwas... sorgung ... der Umgebung des Kali... bergwerkes Asse II durc... lagerung vo... dioaktiven Abfällen in ... gelassenen Grubenbauen ... zufert... . Die weiteren grundlegenden Un...suchungen können nur von de... stfälischen Berggewerkschafskasse, Abt... ung Wasserwirtschaft, ... rer Gesamtheit durchgeführt werden. Für e... n Auftrag gilt di... liegende Gebührenordnung.

Wir bitten ... Ihr ... erständnis zu Aufträgen an

1. Herrn Pro... mler (persönlich) über die Erstellung eines Vorgutachtens mit nachfolgendem Titel "Gefährdungsmöglichkeiten der Trinkwasserversorgung in der Umgebung des Kalibergwerkes Asse II durch Einlagerung von radioaktiven Abfällen in den aufgelassenen Grubenbauen"

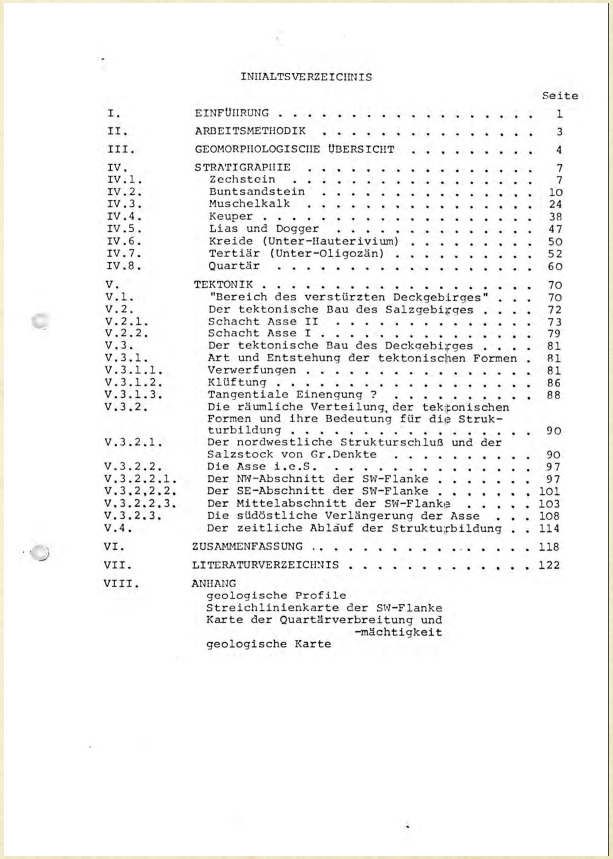
- 2 -

24. FEB. 1965

Post: 75 Karlsruhe Postfach 947 · Kernforschungszentrum: Leopoldshafen bei Karlsruhe · Fernsprecher: Linkenheim (07247) 821 bei Durchwahl 82 · Fernschreiber: 0782-755 Drahtwort: Reaktor Karlsruhe

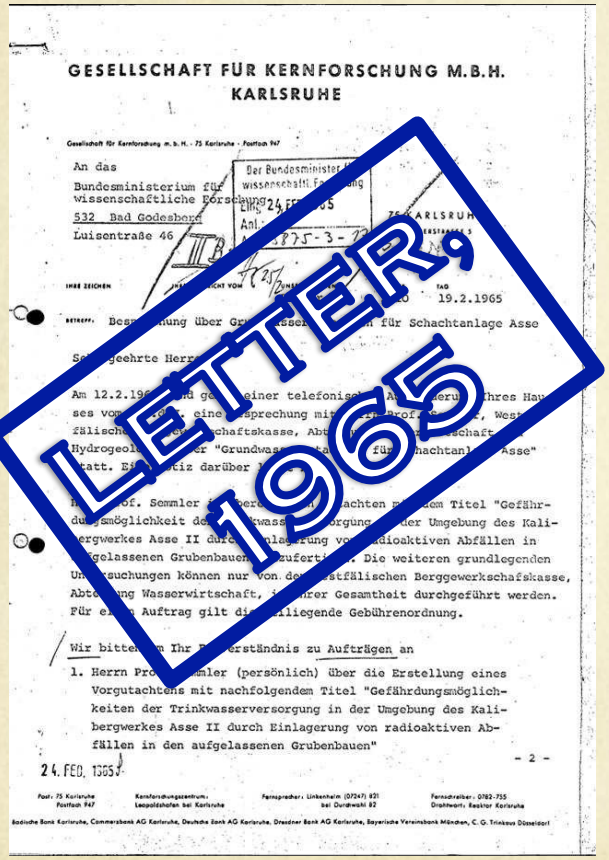
Badische Bank Karlsruhe, Commerzbank AG Karlsruhe, Deutsche Bank AG Karlsruhe, Dresdner Bank AG Karlsruhe, Bayerische Vereinsbank München, C. G. Trinkaus Düsseldorf

LETTER, 1965



INHALTSVERZEICHNIS

Example Documents



GESELLSCHAFT FÜR KERNFORSCHUNG M.B.H.
KARLSRUHE

An das
Bundesministerium für
wissenschaftliche Forschung
532 Bad Godesberg
Luisenstraße 46

19.2.1965

Betreff: Begutachtung über Gefährdungsmöglichkeiten für Schachtanlage Asse

Sehr geehrte Herren,

Am 12.2.1965 wurde gelegentlich einer telefonischen Aussprache Ihres Hauses von Ihnen eine Besprechung mit Prof. [illegible], Westfälische Berggewerkschaftskasse, Abteilung Wasserwirtschaft, Hydrogeologe, über "Grundwasserbegutachtung für Schachtanlage Asse" statt. Ein Protokoll darüber ist Ihnen zugegangen.

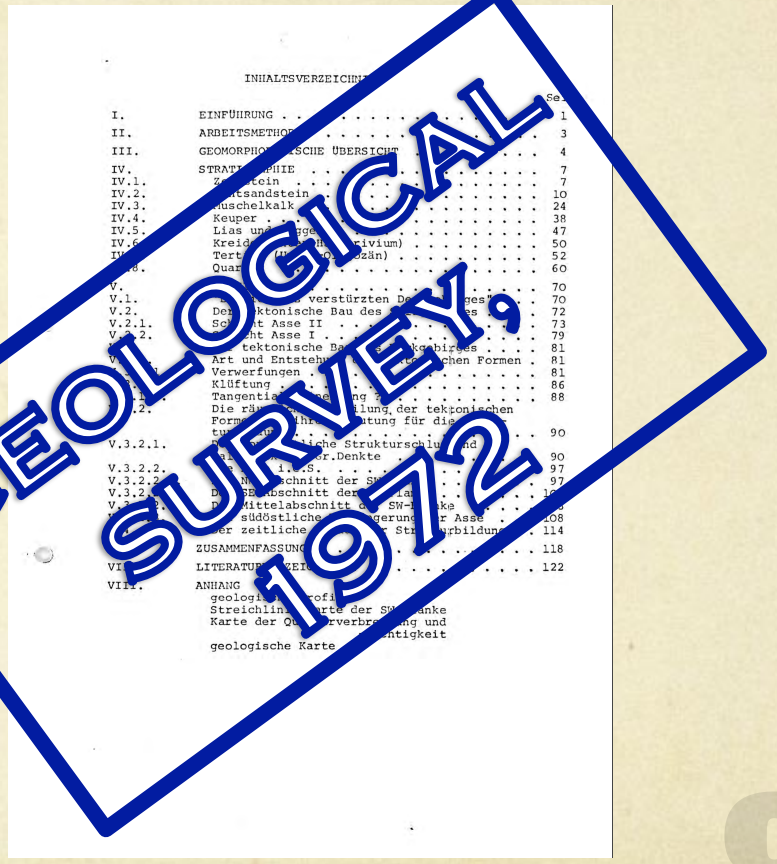
Prof. Semmler hat sich bereit erklärt, ein Gutachten mit dem Titel "Gefährdungsmöglichkeit der Trinkwasserversorgung in der Umgebung des Kalibergwerkes Asse II durch Einlagerung von radioaktiven Abfällen in den aufgelassenen Grubenbauen" zu erstatten. Die weiteren grundlegenden Untersuchungen können nur von der Westfälischen Berggewerkschaftskasse, Abteilung Wasserwirtschaft, in ihrer Gesamtheit durchgeführt werden. Für einen Auftrag gilt die beiliegende Gebührenordnung.

Wir bitten um Ihr Einverständnis zu Aufträgen an

1. Herrn Prof. Semmler (persönlich) über die Erstellung eines Vorgutachtens mit nachfolgendem Titel "Gefährdungsmöglichkeiten der Trinkwasserversorgung in der Umgebung des Kalibergwerkes Asse II durch Einlagerung von radioaktiven Abfällen in den aufgelassenen Grubenbauen"

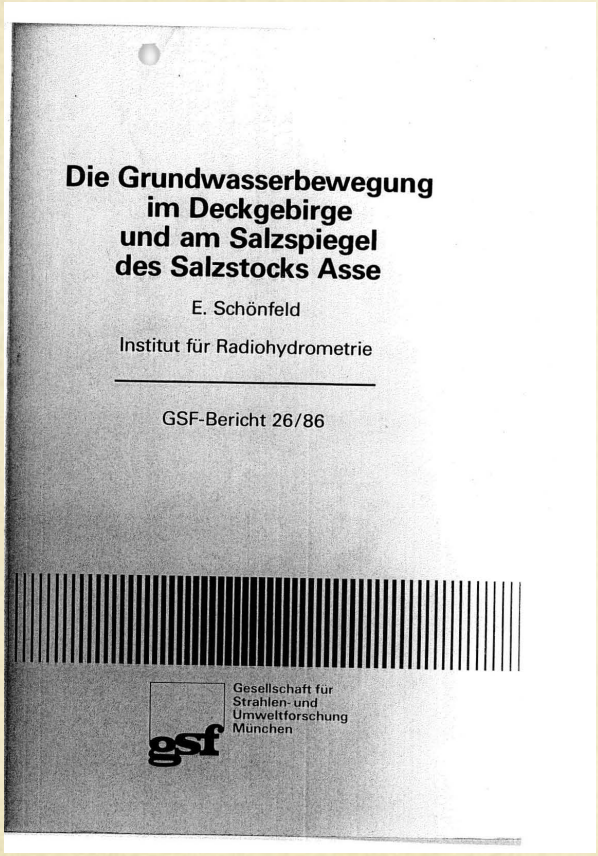
- 2 -

24. FEB. 1965

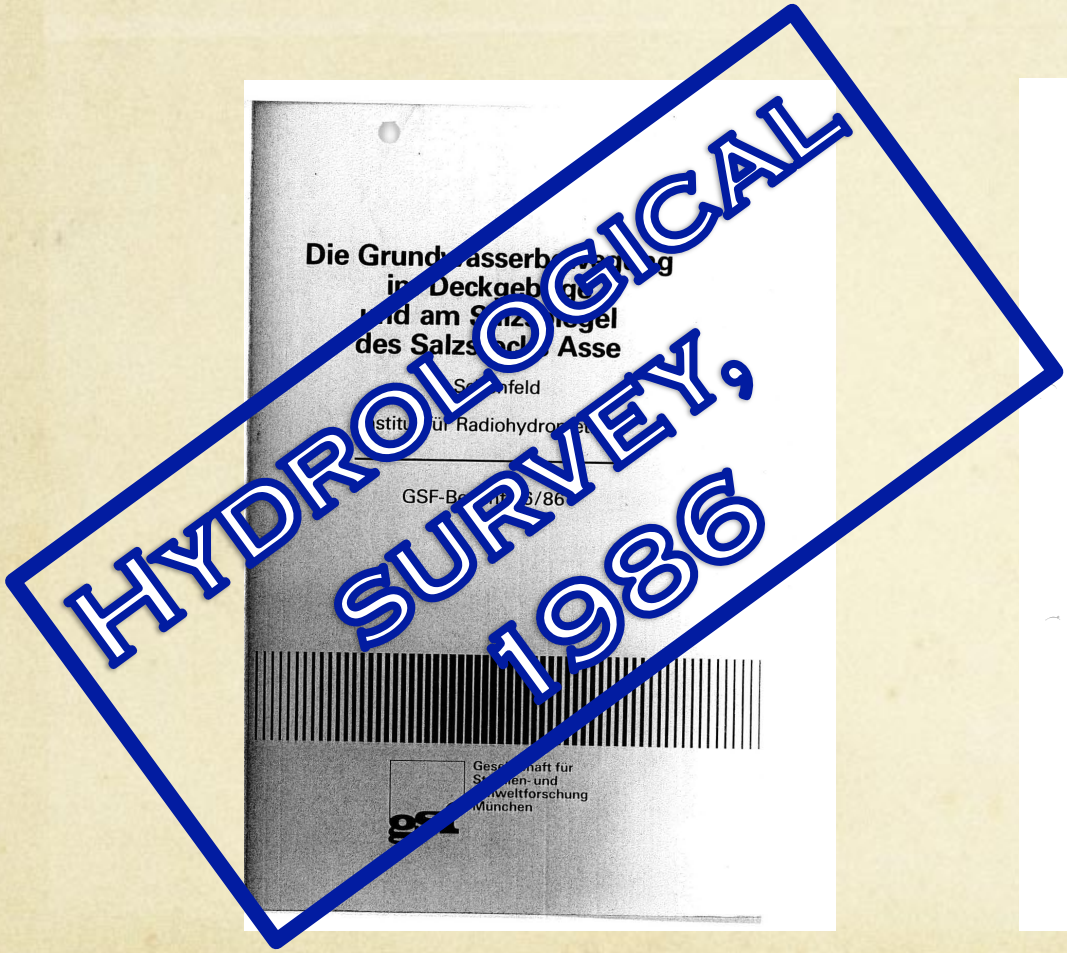


INHALTSVERZEICHNIS

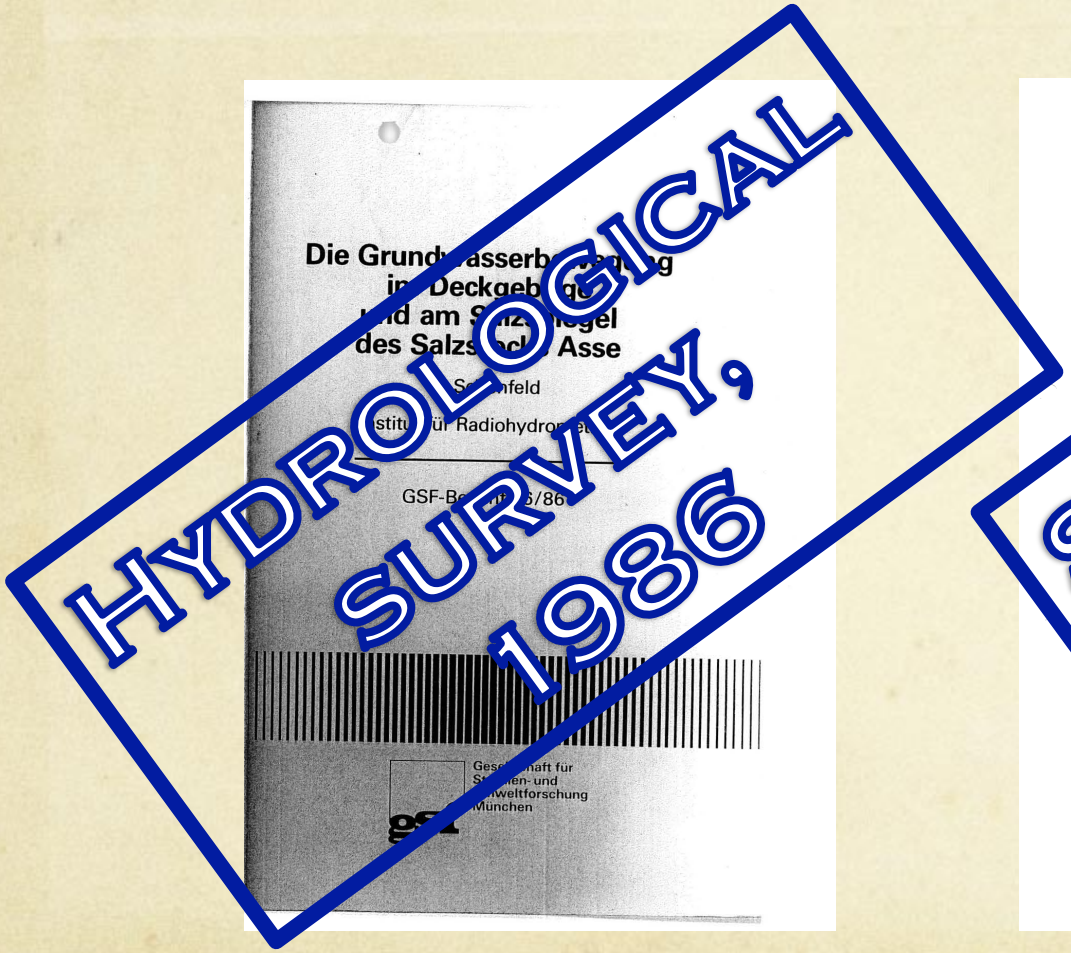
Example Documents

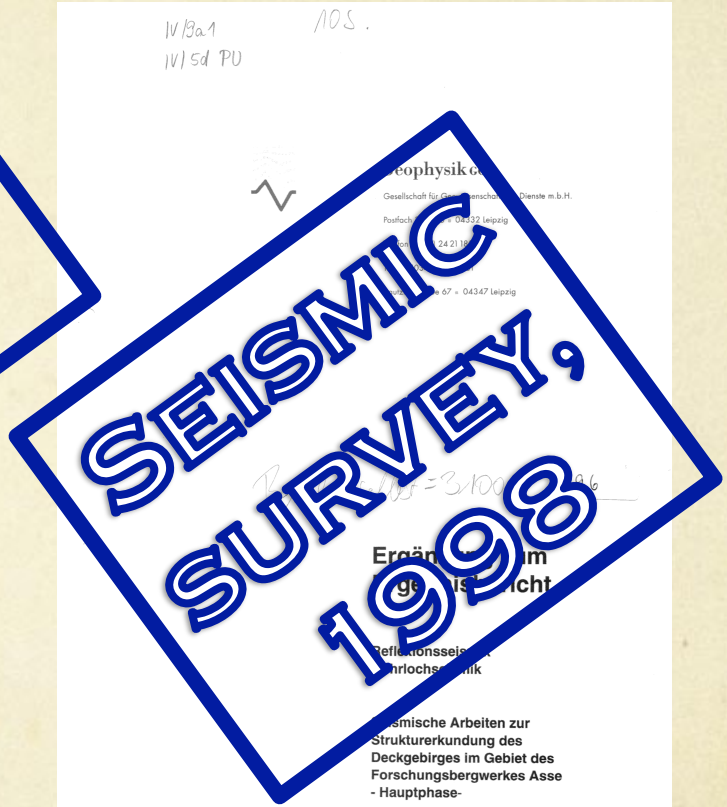


Example Documents

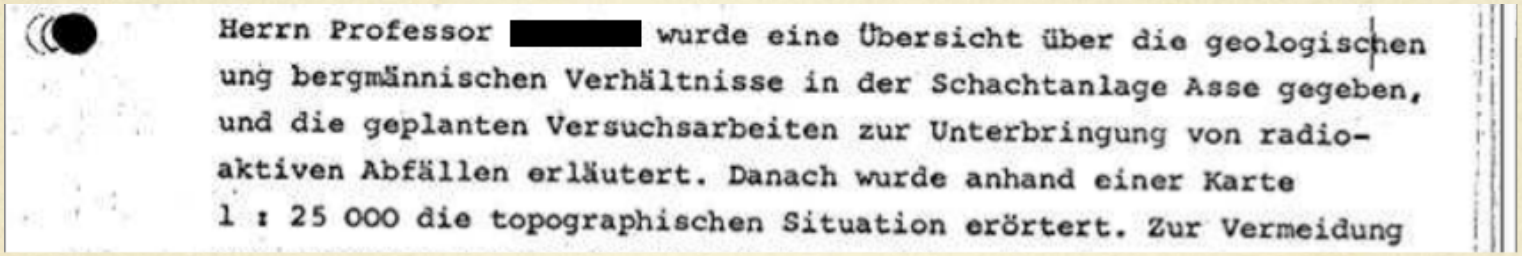


Example Documents





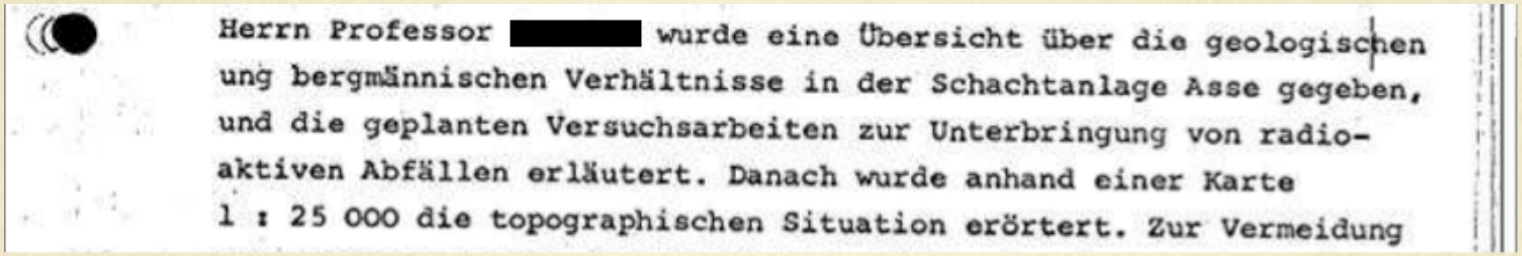
Challenge for OCR


Herrn Professor [redacted] wurde eine Übersicht über die geologischen
ung bergmännischen Verhältnisse in der Schachtanlage Asse gegeben,
und die geplanten Versuchsarbeiten zur Unterbringung von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Google Drive OCR

. | Herrn Professor XXXXX wurde eine: Übar:=icht über? du: genlug:.scfuan
um; hergmännischen Uerhältnisse in der: Schachtanlage ,um:
und die qaplanten versuchsarbaitan zu: Unterbringung man radia
MEäILEn erläutert. Danach wurde anhand einer Harta
1 : 25 üüü dia top-ugraphischen Eituatinn arä:terr:.. Zur vermeiflung

Challenge for OCR

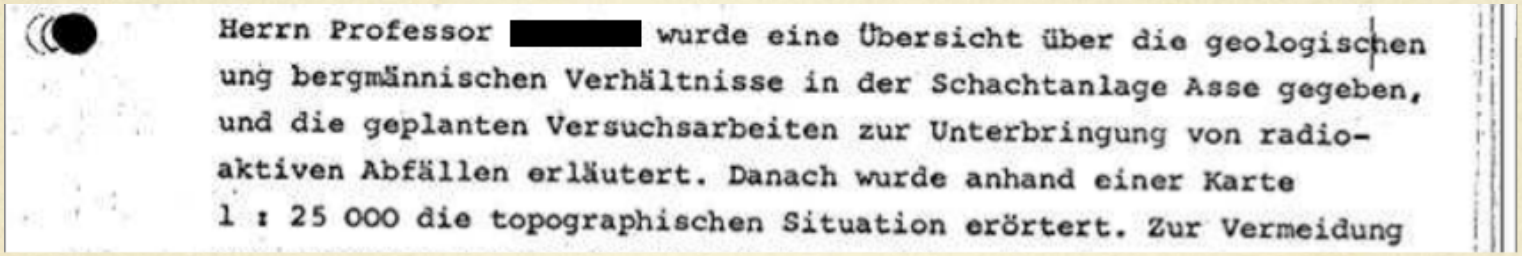


Herrn Professor [redacted] wurde eine Übersicht über die geologischen
ung bergmännischen Verhältnisse in der Schachtanlage Asse gegeben,
und die geplanten Versuchsarbeiten zur Unterbringung von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Google Drive OCR (correctly identified tokens)

. | Herrn Professor XXXXX wurde eine: Übar:=icht über? du: genlug:.scfuan
um; hergmännischen Uerhältnisse in der: Schachtanlage ,um:
und die qaplanten versuchsarbaitan zu: Unterbringung man radia
MEäILEn erläutert. Danach wurde anhand einer Harta
1 : 25 üüü dia top-ugraphischen Eituatinn arä:terr:.. Zur vermeiflung

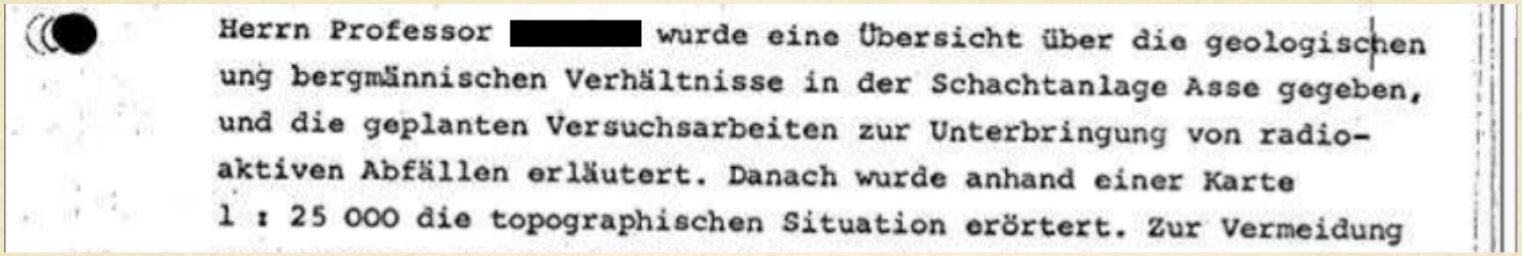
Challenge for OCR


Herrn Professor [redacted] wurde eine Übersicht über die geologischen
ung bergmännischen Verhältnisse in der Schachtanlage Asse gegeben,
und die geplanten Versuchsarbeiten zur Unterbringung von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Google Drive OCR (correctly identified tokens)

Herrn Professor wurde
in
und die
erläutert. Danach wurde anhand einer
1 : 25 Zur

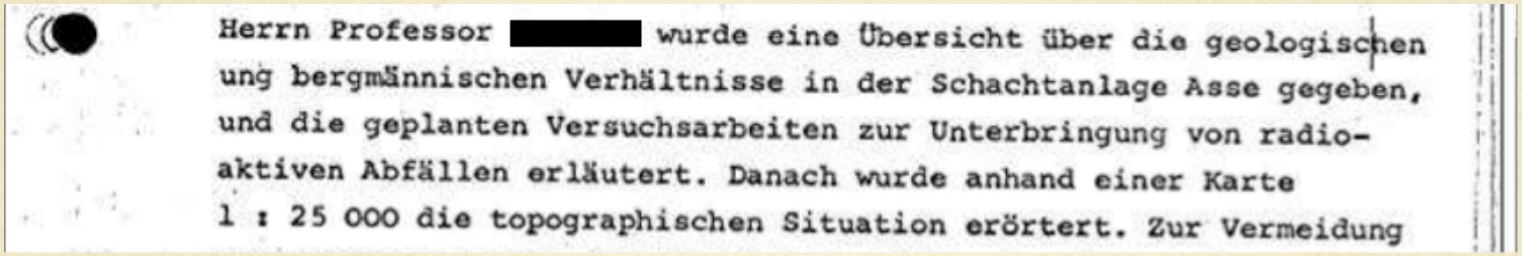
Challenge for OCR


Herrn Professor [redacted] wurde eine Übersicht über die geologischen
ung bergmännischen Verhältnisse in der Schachtanlage Asse gegeben,
und die geplanten Versuchsarbeiten zur Unterbringung von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Google Drive OCR (correctly identified tokens)

Herrn Professor wurde
in
und die
erläutert. Danach wurde anhand einer
1 : 25 Zur

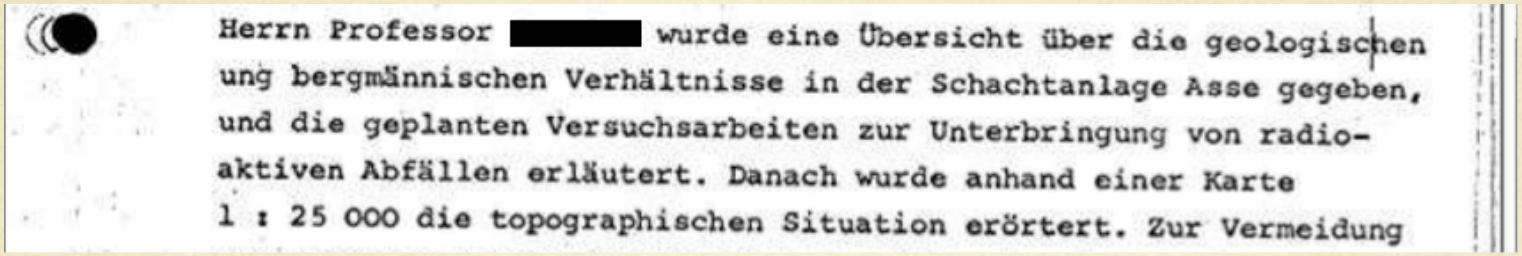
Challenge for OCR


Herrn Professor [redacted] wurde eine Übersicht über die geologischen
ung bergmännischen Verhältnisse in der Schachtanlage Asse gegeben,
und die geplanten Versuchsarbeiten zur Unterbringung von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Online OCR service

Herrn Professor XXXXXX wurde ein* übercicht über dio geologischen
14.11/4.9111111
une bergmännischen Verhältnisse in der Schachtanlage Asse 9.agebenf
und die geplanten Verauchsarbeiten zur Unterbrin9ung von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Challenge for OCR

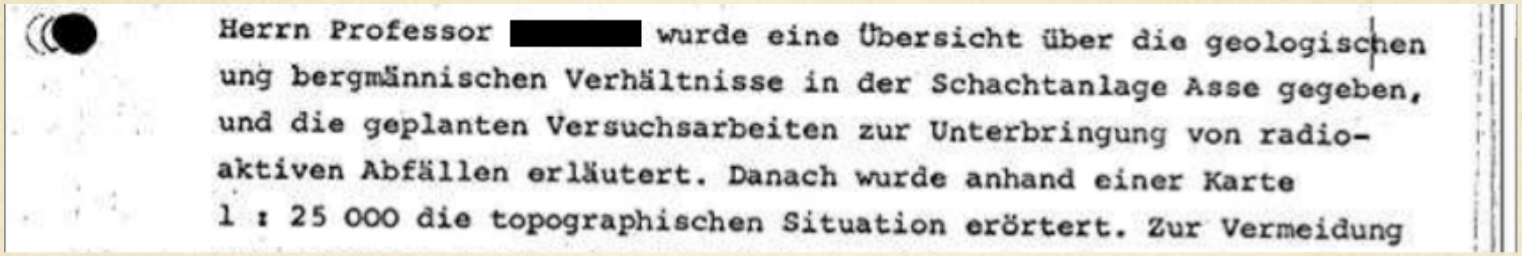

Herrn Professor [redacted] wurde eine Übersicht über die geologischen
ung bergmännischen Verhältnisse in der Schachtanlage Asse gegeben,
und die geplanten Versuchsarbeiten zur Unterbringung von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Online OCR service (correctly identified tokens)

Herrn Professor wurde über geologischen

bergmännischen Verhältnisse in der Schachtanlage Asse
und die geplanten zur von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Challenge for OCR


Herrn Professor [redacted] wurde eine Übersicht über die geologischen
ung bergmännischen Verhältnisse in der Schachtanlage Asse gegeben,
und die geplanten Versuchsarbeiten zur Unterbringung von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Better, but keywords are missing

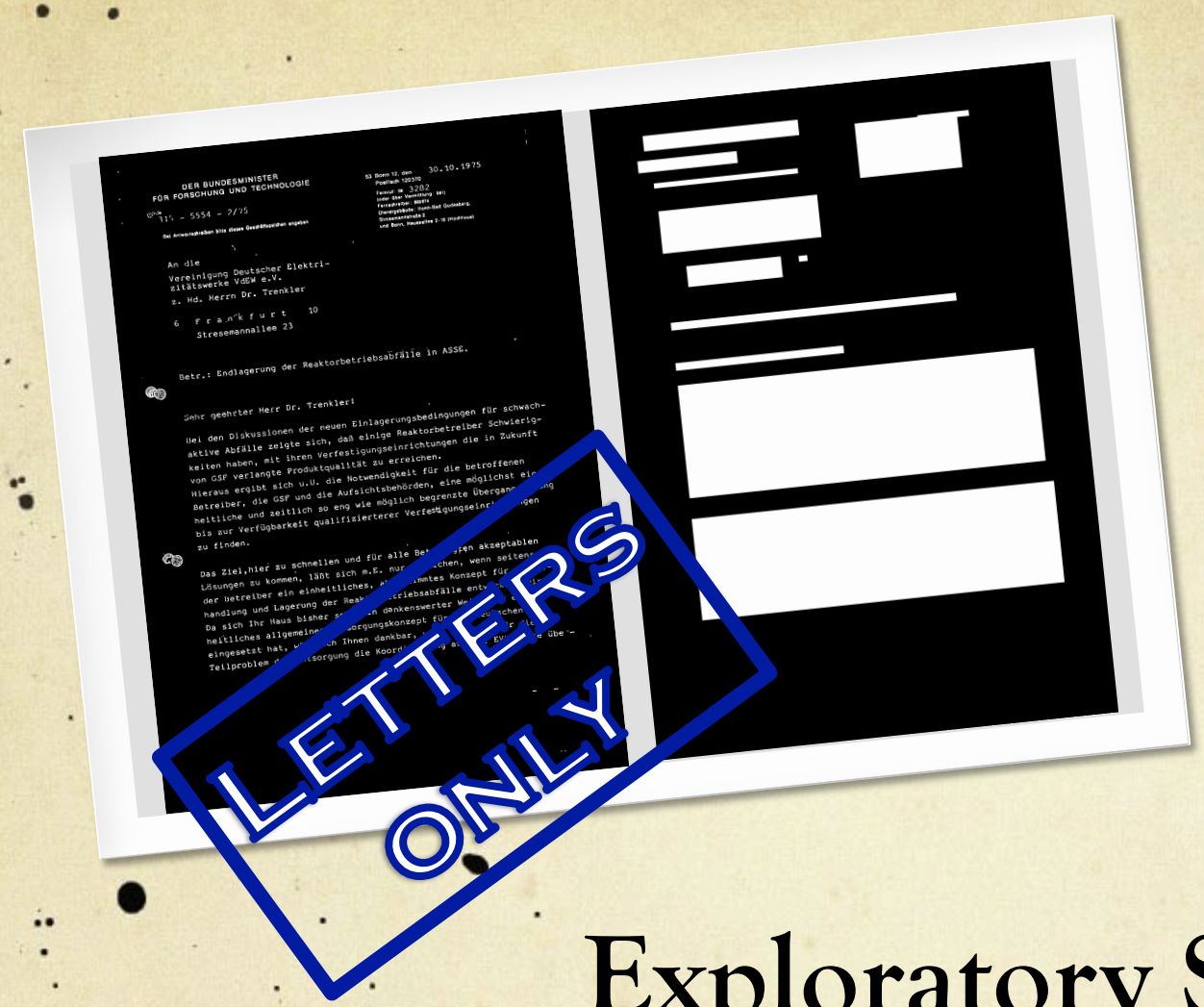
Online OCR service (correctly identified tokens)

Herrn Professor wurde über geologischen

bergmännischen Verhältnisse in der Schachtanlage Asse
und die geplanten zu von radio-
aktiven Abfällen erläutert. Danach wurde anhand einer Karte
1 : 25 000 die topographischen Situation erörtert. Zur Vermeidung

Vision for the Project

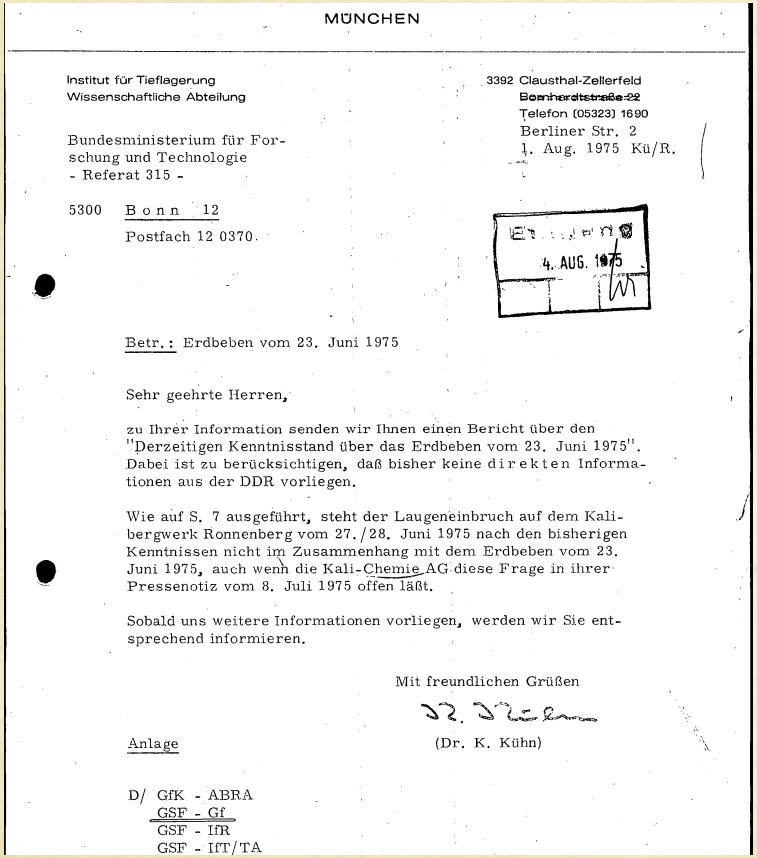
- Archiving: Searchable access to legacy paper documents
- Repurposing: Extracting metadata, meaningful connections, ontology learning
- Integrated processing chain from OCR to semantics
- First exploratory study: Segmentation of letterheads



Exploratory Study

Letterhead segmentation for legacy correspondence

Data



MÜNCHEN

Institut für Tieflagerung
Wissenschaftliche Abteilung

3392 Clausthal-Zellerfeld
~~Bernhardtstraße 22~~
Telefon (05323) 1690
Berliner Str. 2
1. Aug. 1975 Kü/R.

Bundesministerium für Forschung und Technologie
- Referat 315 -

5300 Bonn 12
Postfach 12 0370

4. AUG. 1975

Betr.: Erdbeben vom 23. Juni 1975

Sehr geehrte Herren,

zu Ihrer Information senden wir Ihnen einen Bericht über den "Derzeitigen Kenntnisstand über das Erdbeben vom 23. Juni 1975". Dabei ist zu berücksichtigen, daß bisher keine direkten Informationen aus der DDR vorliegen.

Wie auf S. 7 ausgeführt, steht der Laugeneinbruch auf dem Kalibergwerk Ronnenberg vom 27./28. Juni 1975 nach den bisherigen Kenntnissen nicht im Zusammenhang mit dem Erdbeben vom 23. Juni 1975, auch wenn die Kali-Chemie AG diese Frage in ihrer Pressenotiz vom 8. Juli 1975 offen läßt.

Sobald uns weitere Informationen vorliegen, werden wir Sie entsprechend informieren.

Mit freundlichen Grüßen

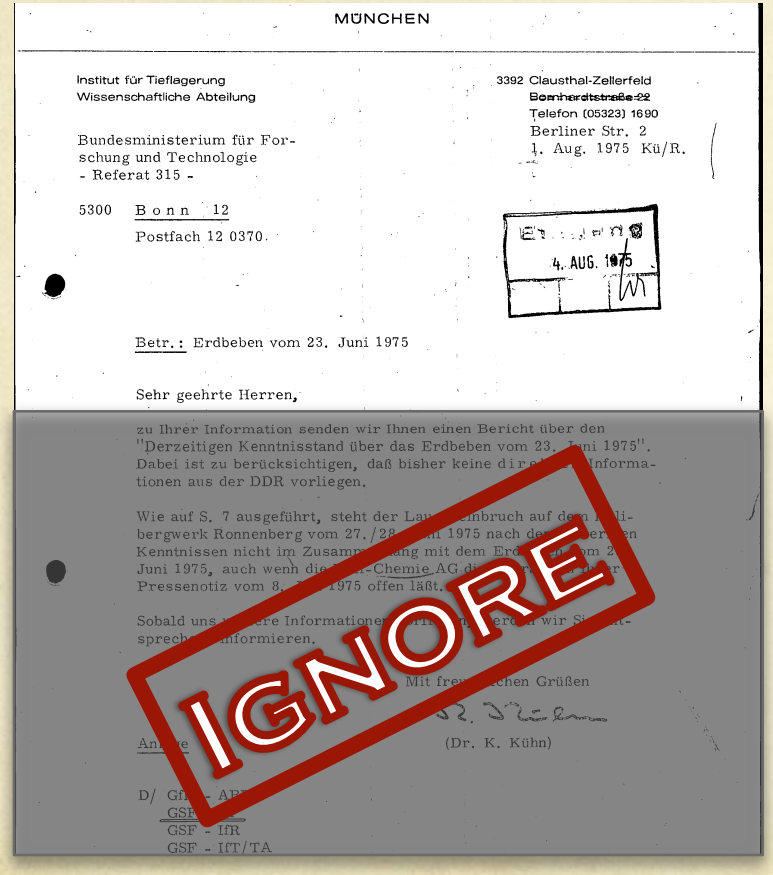
(Dr. K. Kühn)

Anlage

D/ GfK - ABRA
GSF - Gf
GSF - IfR
GSF - IfT/TA

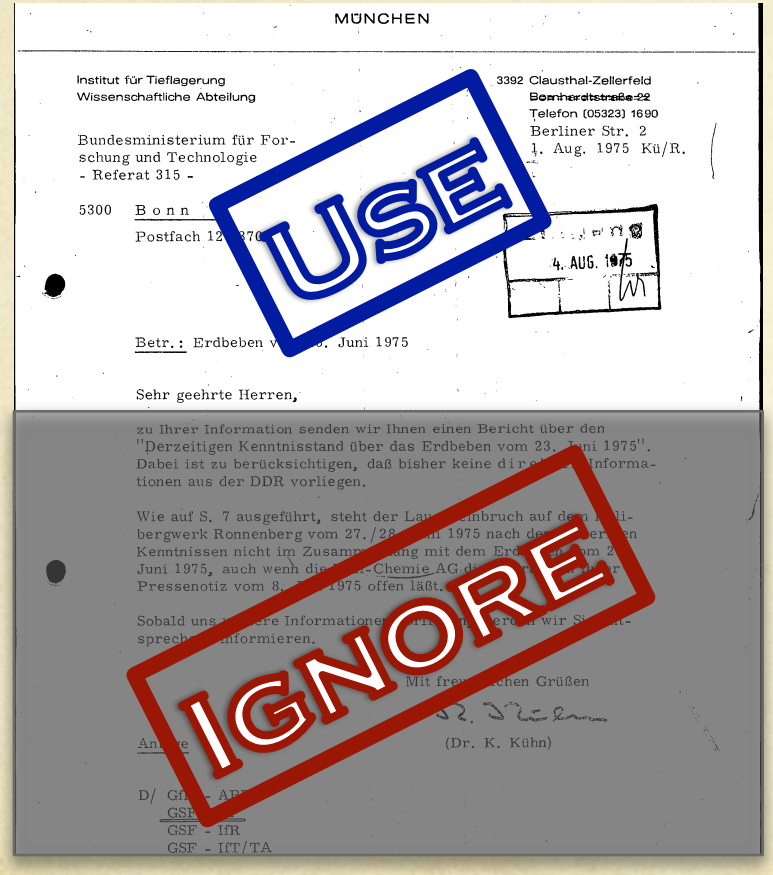
- Want to identify letterhead elements that form a unit
- 24 letters from an initial batch of scanned documents (approx. 80 documents)
- OCR optimised by external project partner
- Text body stripped away, just letterheads

Data



- Want to identify letterhead elements that form a unit
- 24 letters from an initial batch of scanned documents (approx. 80 documents)
- OCR optimised by external project partner
- Text body stripped away, just letterheads

Data



- Want to identify letterhead elements that form a unit
- 24 letters from an initial batch of scanned documents (approx. 80 documents)
- OCR optimised by external project partner
- Text body stripped away, just letterheads

Method

```
<span class='ocr line'
    title='bbox
  261 2648 1064 2719'
      style=
  'position:absolute;
left:93.2142857143px;
top:313.928571429px'>
    Forschung und</
      span><br />
```

Fig.: Snapshot from hOCR output

- Unsupervised clustering approach
- Use only positional information (for now)
- Agglomerative, bottom-up clustering
  - Complete-linkage
  - Euclidian distance

Method

```
<span class='ocr line'
     title='bbox
  261 2648 1064 2719'
       style=
  'position:absolute;
left:93.2142857143px;
top:313.928571429px'>
    Forschung und</
      span><br />
```

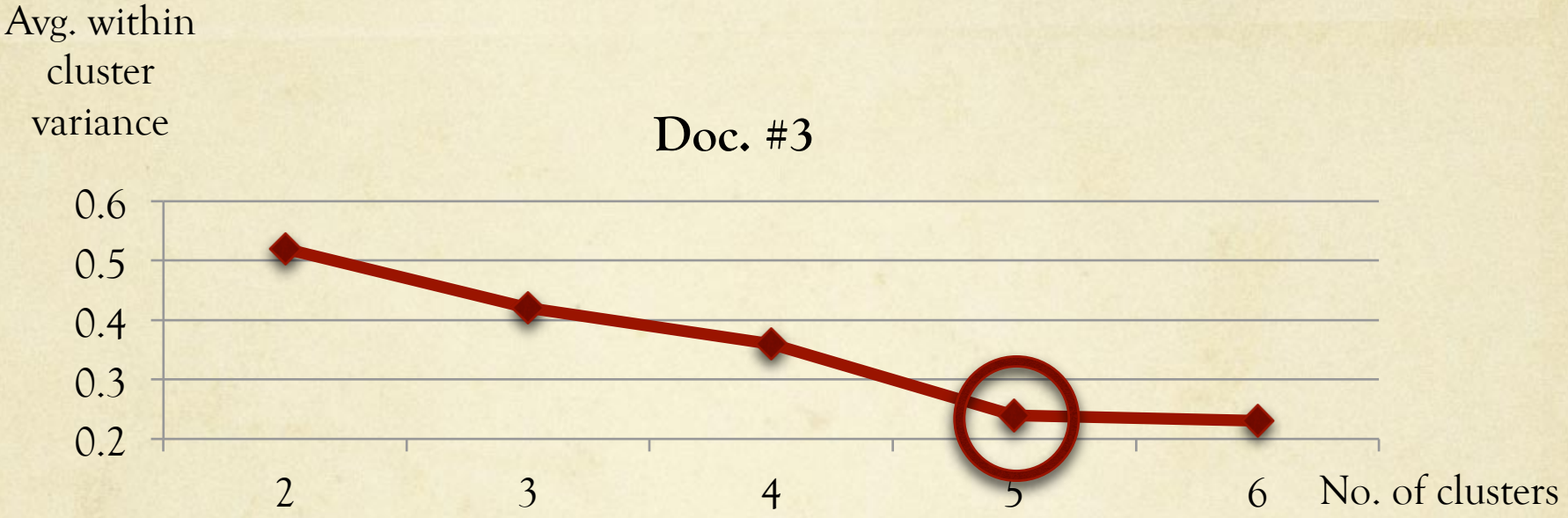
Fig.: Snapshot from hOCR output

- Unsupervised clustering approach
- Use only positional information (for now)
- Agglomerative, bottom-up clustering
  - Complete-linkage
  - Euclidian distance

Method

- "Flatten" clustering by identifying significant increase in average within cluster variance ("spot the elbow joint")
- This is likely to indicate that two clusters far away from each other have erroneously been merged

Method



- "Flatten" clustering by identifying significant increase in average within cluster variance ("spot the elbow joint")
- This is likely to indicate that two clusters far away from each other have erroneously been merged

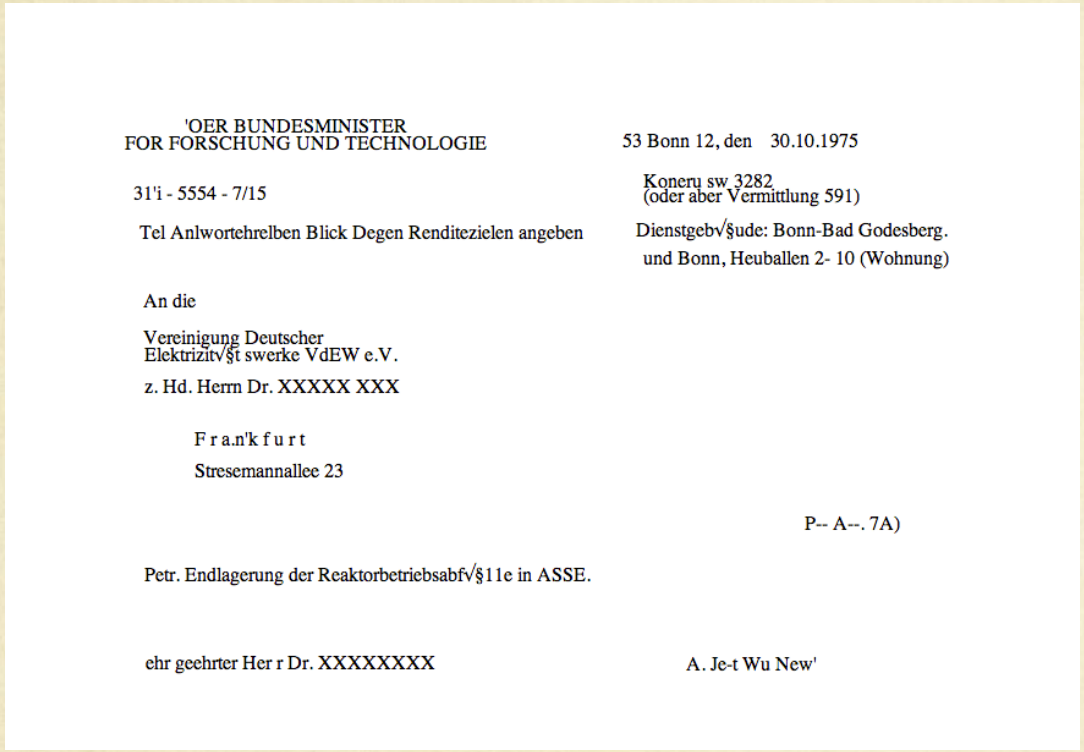
Method

- "Flatten" clustering by identifying significant increase in average within cluster variance ("spot the elbow joint")
- This is likely to indicate that two clusters far away from each other have erroneously been merged

Method

- "Flatten" clustering by identifying significant increase in average within cluster variance ("spot the elbow joint")
- This is likely to indicate that two clusters far away from each other have erroneously been merged

Example Segmentation



'OER BUNDESMINISTER
FOR FORSCHUNG UND TECHNOLOGIE

53 Bonn 12, den 30.10.1975

31'i - 5554 - 7/15

Koneru sw 3282
(oder aber Vermittlung 591)

Tel Anlwortehrelben Blick Degen Renditezielen angeben

Dienstgebv§ude: Bonn-Bad Godesberg.
und Bonn, Heuballen 2- 10 (Wohnung)

An die

Vereinigung Deutscher
Elektrizitv§t swerke VdEW e.V.

z. Hd. Herrn Dr. XXXXX XXX

F r a.n'k f u r t

Stresemannallee 23

P-- A--. 7A)

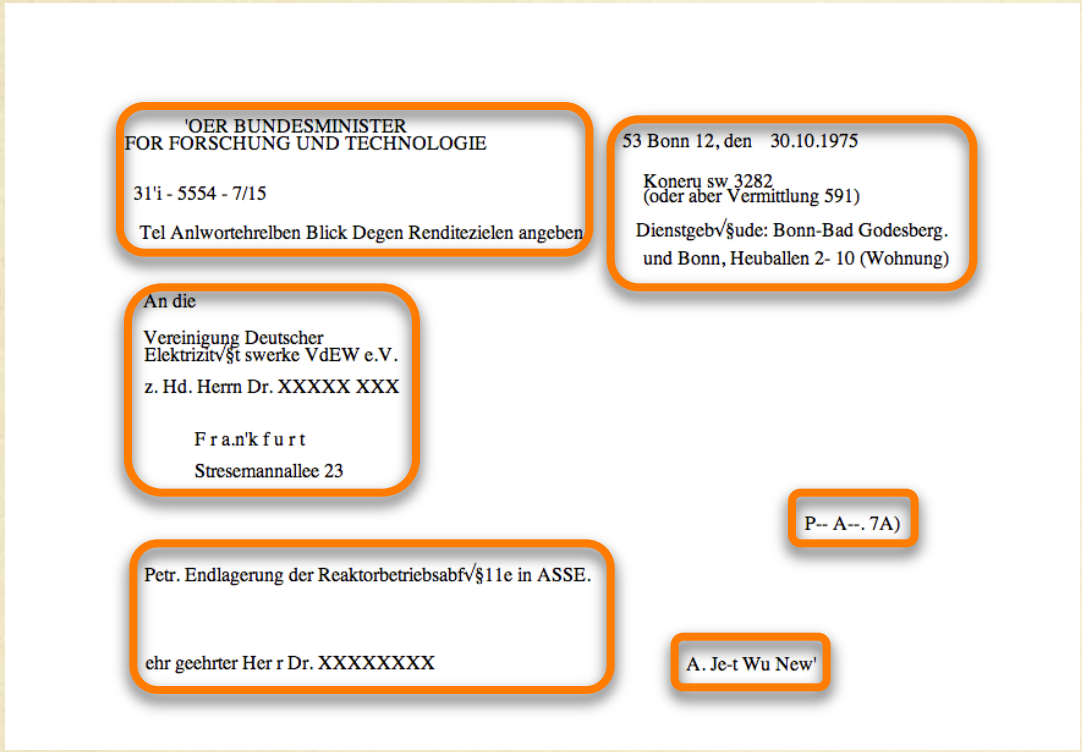
Petr. Endlagerung der Reaktorbetriebsabfv§11e in ASSE.

ehr geehrter Her r Dr. XXXXXXXX

A. Je-t Wu New'

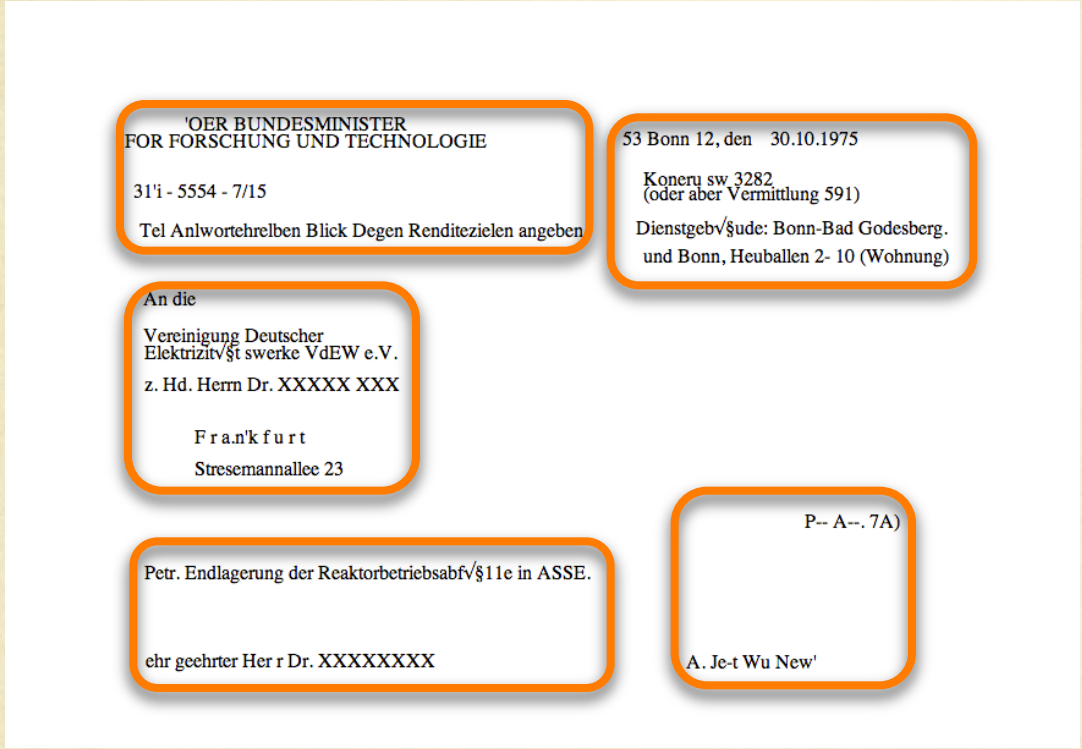
No. of clusters	Within cluster variance

Example Segmentation



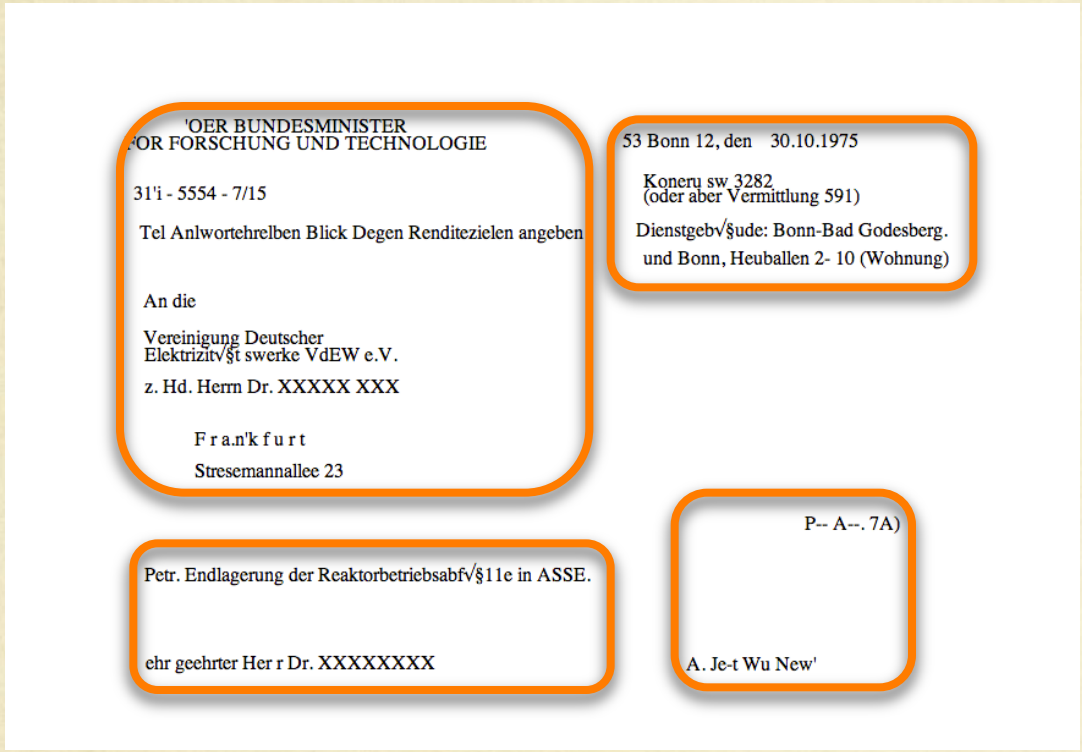
No. of clusters	Within cluster variance
6	0.23

Example Segmentation



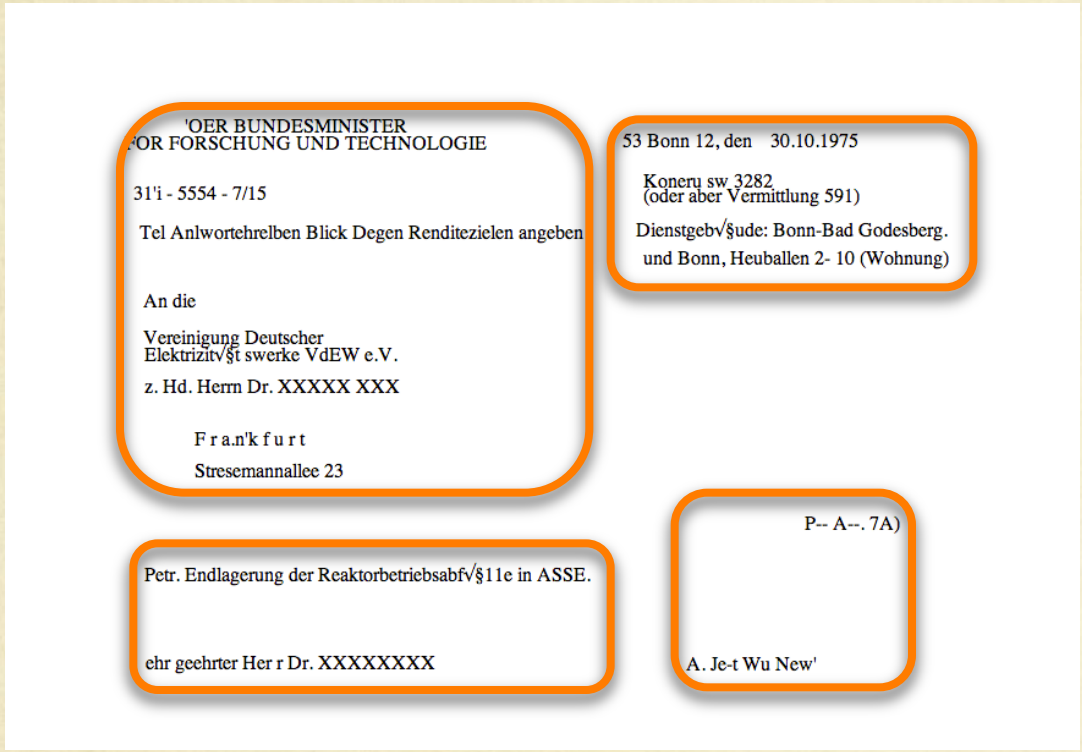
No. of clusters	Within cluster variance
6	0.23
5	0.24

Example Segmentation



No. of clusters	Within cluster variance
6	0.23
5	0.24
4	0.36

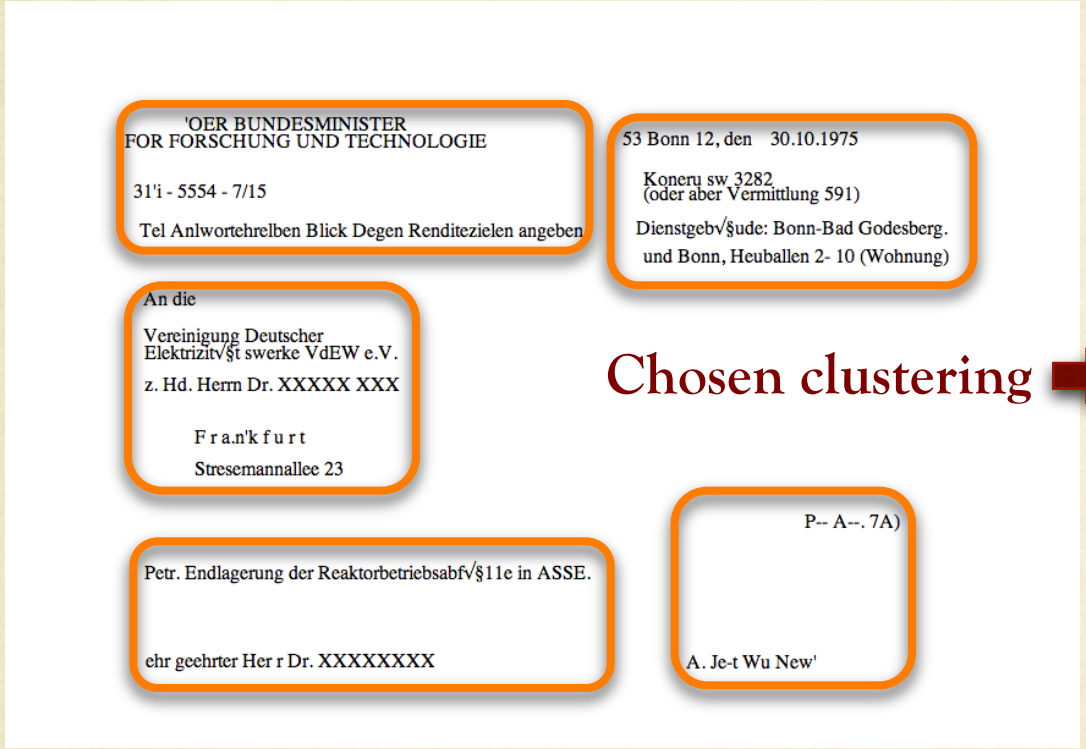
Example Segmentation



No. of clusters	Within cluster variance
6	0.23
5	0.24
4	0.36

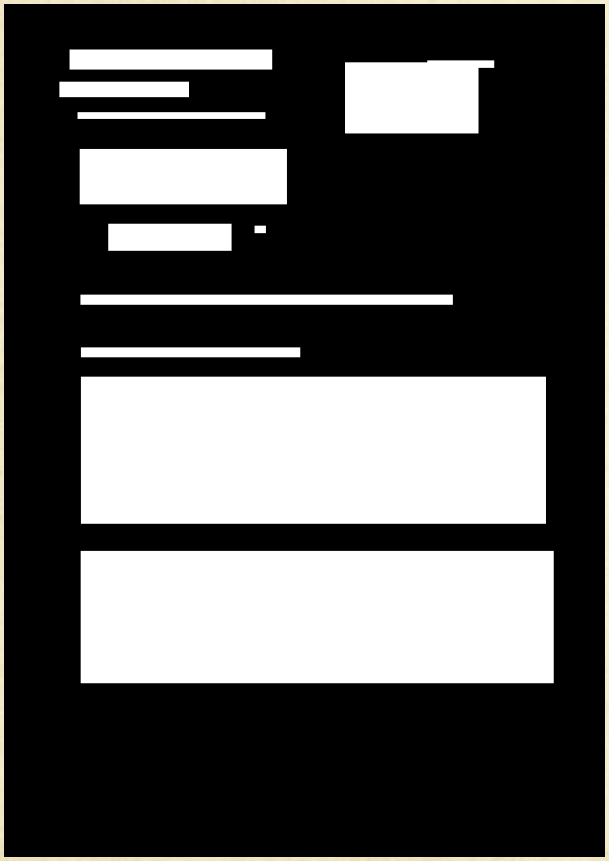
Significant increase

Example Segmentation



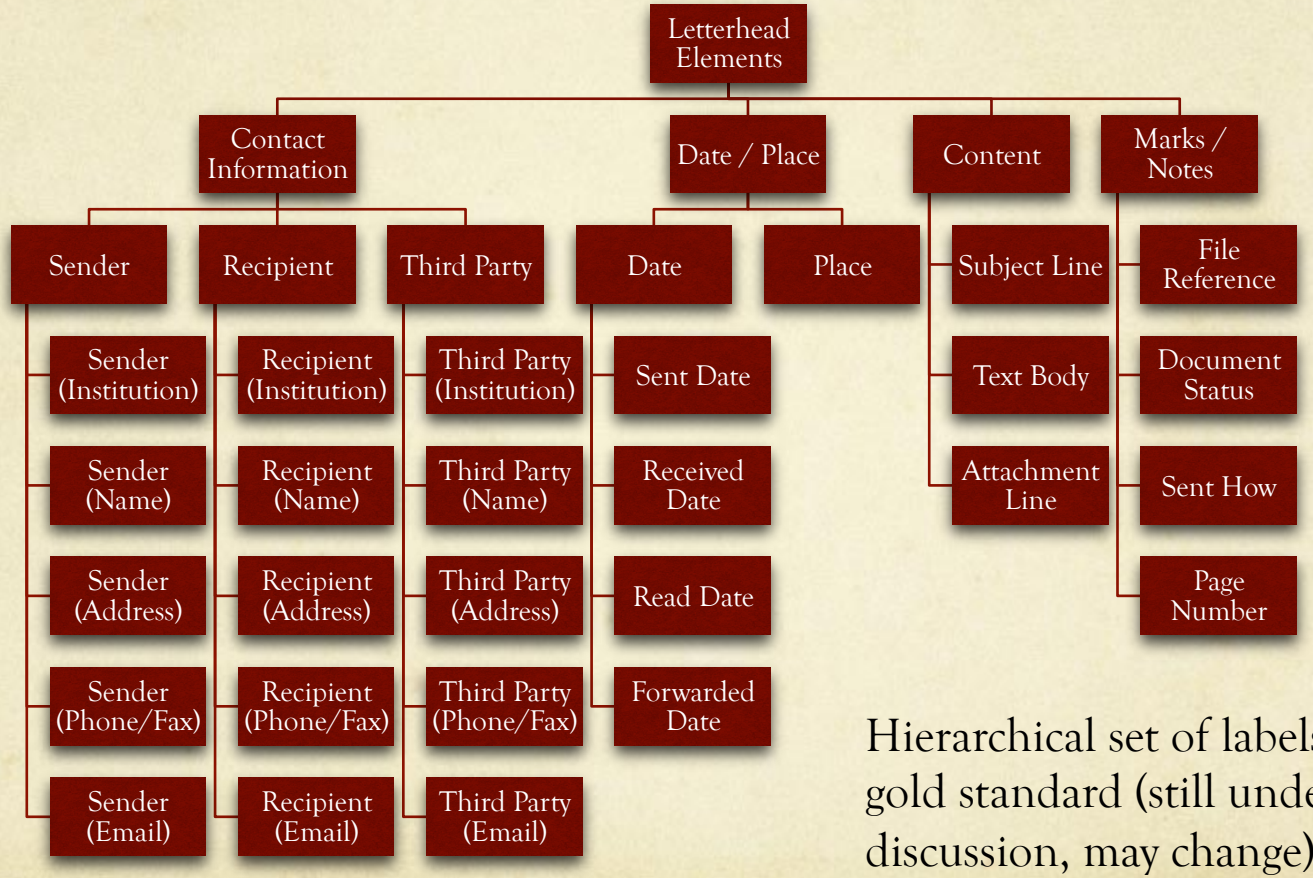
No. of clusters	Within cluster variance
6	0.23
5	0.24
4	0.36

Future Work



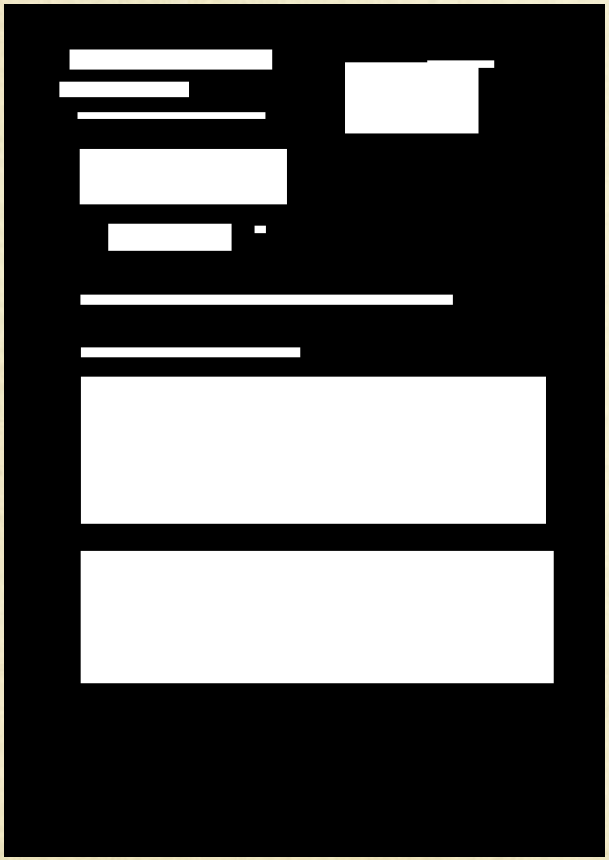
- Currently: Classify letterhead segments according to their "semantics" (address, date etc.) using Naïve Bayes and similar
- Layout and content analysis (possibly) not separate
- Next step: Create a gold standard for a small set of letters (ca. 60) and letterhead elements (ca. 1000-1500)
- Evaluate against gold standard

Future Work



Hierarchical set of labels for gold standard (still under discussion, may change)

Future Work



- Currently: Classify letterhead segments according to their "semantics" (address, date etc.) using Naïve Bayes and similar
- Layout and content analysis (possibly) not separate
- Next step: Create a gold standard for a small set of letters (ca. 60) and letterhead elements (ca. 1000-1500)
- Evaluate against gold standard

Thank you for your attention

Questions? Comments?

Back-Up / Alternative Slides

Method

- "Flatten" clustering by identifying significant increase in average within cluster variance ("spot the elbow joint")
- This indicates that two clusters far away from each other have (erroneously?) been merged