



CCS

content conversion specialists

consulting
technology
digitization services



CCS

content conversion specialists

Future Challenges for OCR Technology

1st IMPACT conference, The Hague, 6-7 April 2009

Claus Gravenhorst, Director Strategic Initiatives

- CCS stands for Content Conversion Specialists
- Founded in 1976
- Headquarters in Hamburg, Germany
- Subsidiaries and representatives in Rumania and North America
- Research Centre at Polythenica-University Bucharest
- 100+ employees
- Technology: 3 successful product lines
- Consulting & Digitisation services



- Started 1983 at CCS, “grew up” with OCR
- 1976 Raymond Kurzweil introduced the first industrial machine based on ICR technology designed to support blind people (KRM, speech output) and in 1978 the KURZWEIL Data Entry Machine (KDEM)
- Netherlands was one of the early adopters
 - Instituut voor Nederlandse Lexicologie (dicts)
 - Kluwer Rechtswetenschappen (legal database)
- Omnifont, char/word/dict based, trainable
- Machines got smaller
- Separated scanners
- OCR running on specialised co-processors
- OCR as software-only application
- Integration with 3rd party applications through SDK

- Reached a high level for recognition of characters
- Supports a wide range of languages and fonts, e.g. ABBYY OCR more than 170 languages
- With support of dictionaries OCR's decision process takes words into account
- OCR still page based
- After 33 years of R&D on ICR technologies there's still a big difference to quality of human reading

- METAe – The Metdata Engine
- EU funded FP5 research project
- Production tool for conversion of printed material to structured XML objects
- 2000 – 2003, 14 partners, UIBK as coordinator, CCS and MitCom (ABBYY) as commercial technology partners
- Product launch in 2003 under brand name docWORKS/METAe
- Output: Structured digital objects describing physical and logical document structure
- Consistent use of international XML standards like METS, DC, NISO MIX, ... and ALTO

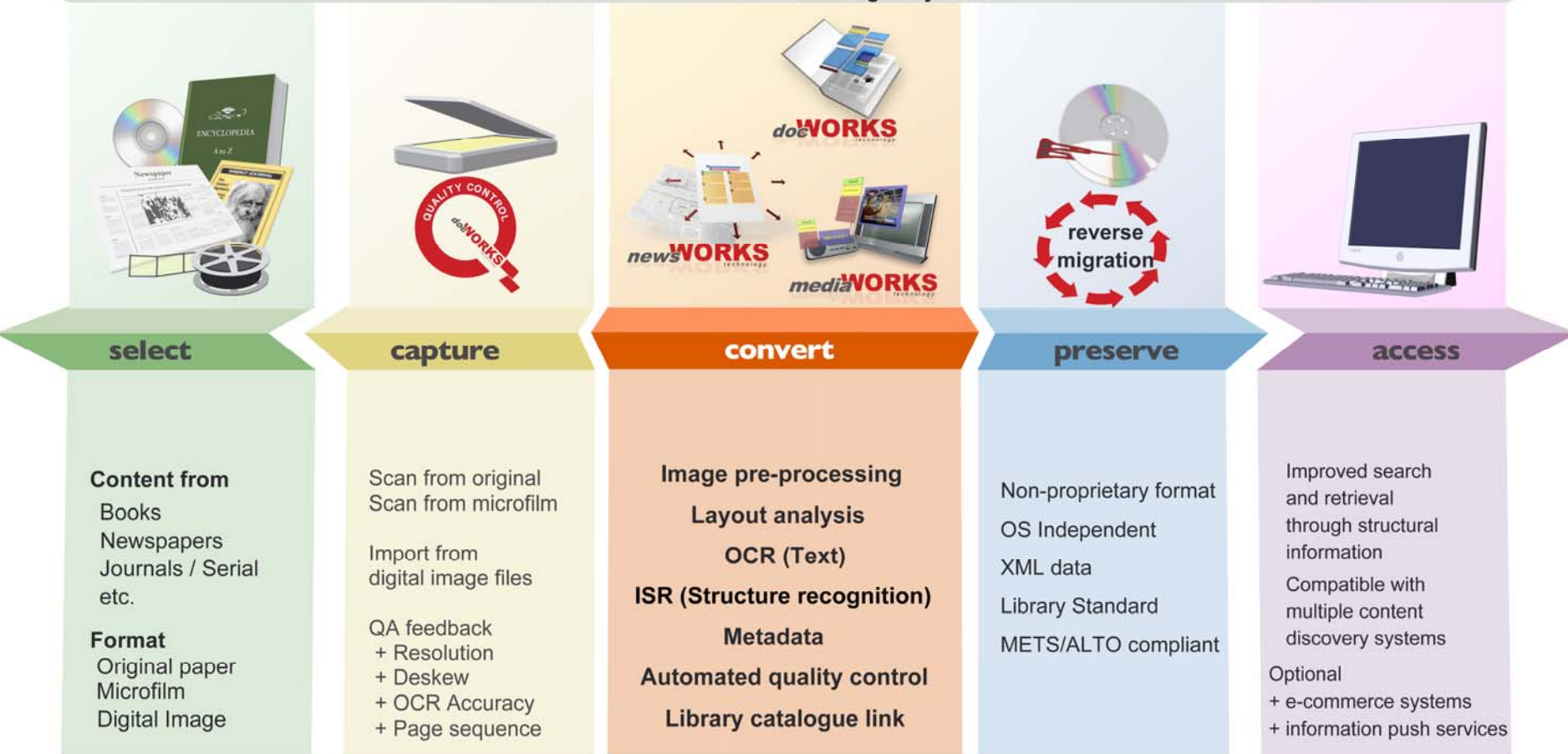
Since 2003 METAe technology has been significantly improved and is in use at ivy-league cultural and scientific institutions as well as digitisation service providers around the world, e.g.:

- National Library of Norway – in-house, books/newspapers
- National Library of Finland – in-house, books/newspapers
- Royal Library Denmark – in-house, journals
- Harvard University Library – in-house, books, journals
- Stanford University Library – in-house, books
- The British Library (mass digitisation books)
- Koninklijke Bibliotheek (mass digitisation newspapers)

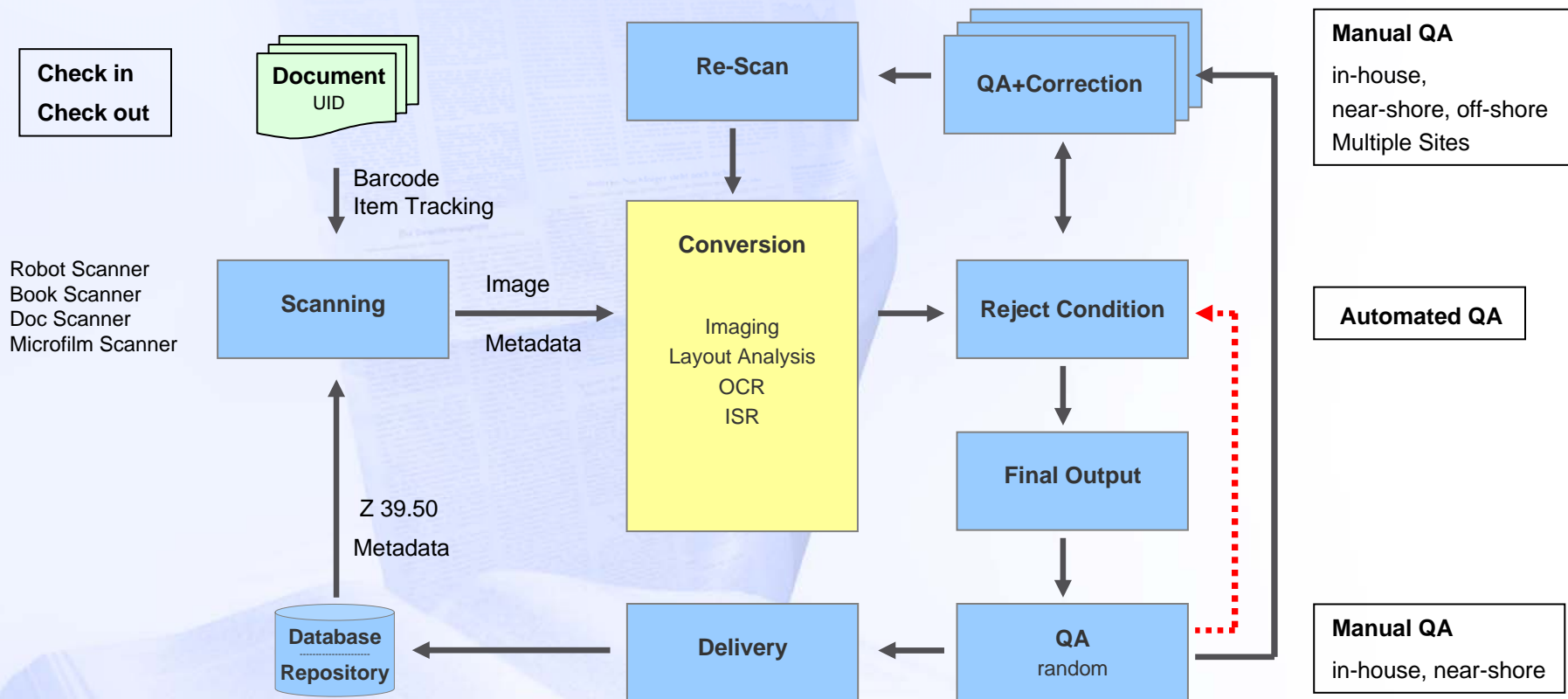
OCR – core process within the value chain

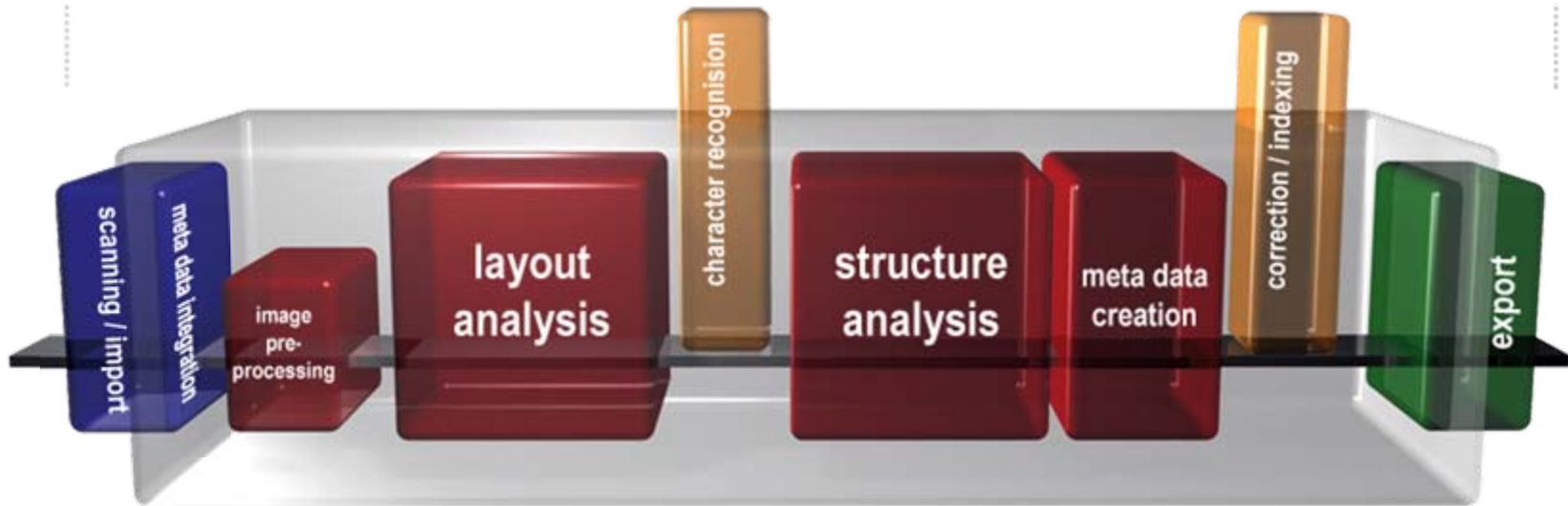
Wirtschaft

production management system
document tracking system



Typical elements of digitisation workflows





Scanning origin or microfilm, import of digital images

- Problems with image quality (due to scanning process or microfilm/print quality)
- Lighting problems
- Low contrast (foreground, background)
- Curved and wavy text lines (binding)
- Long and curved columns in newspapers
- Bleed-through, noise, broken characters

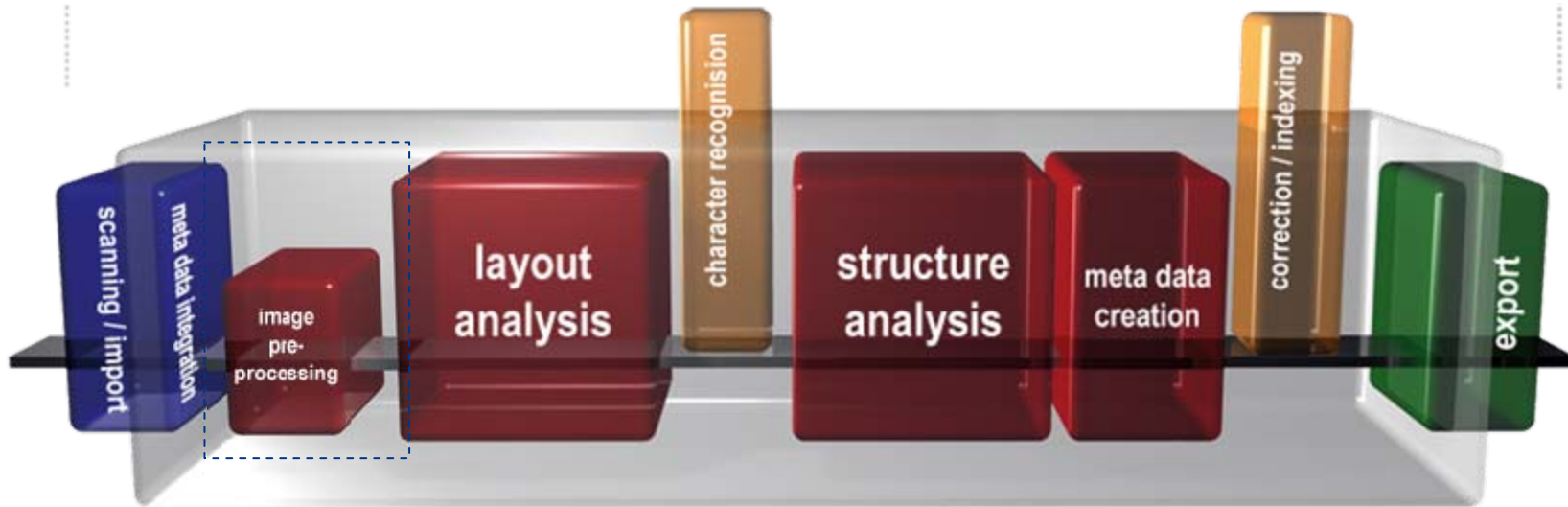
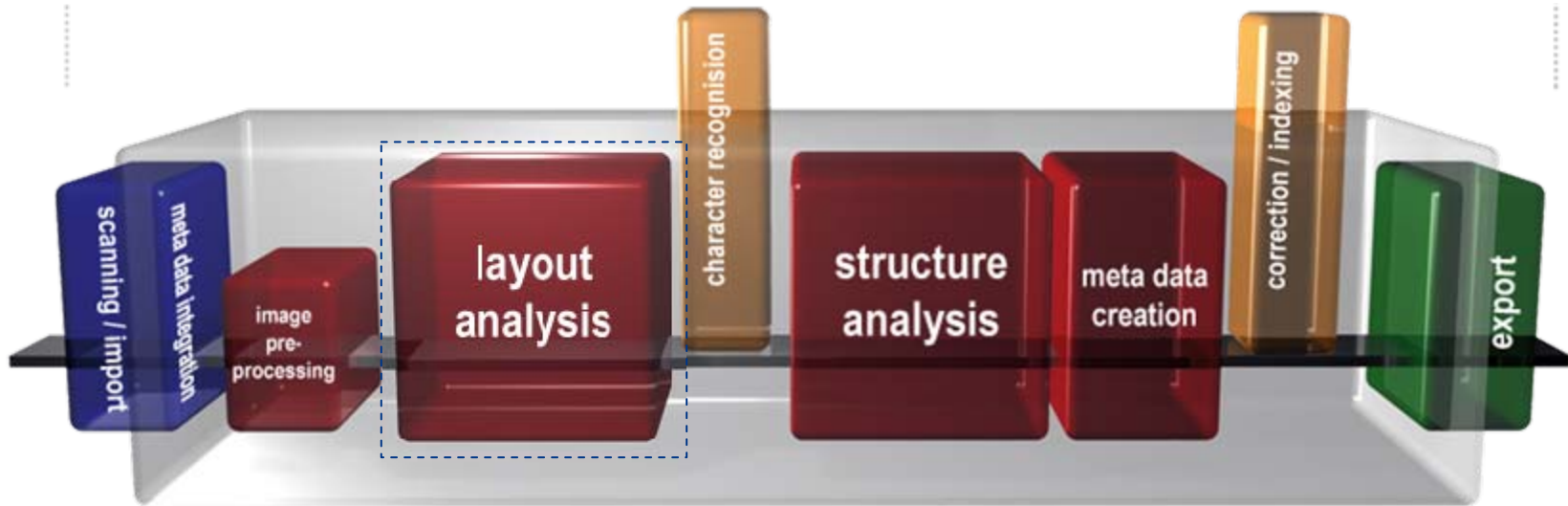


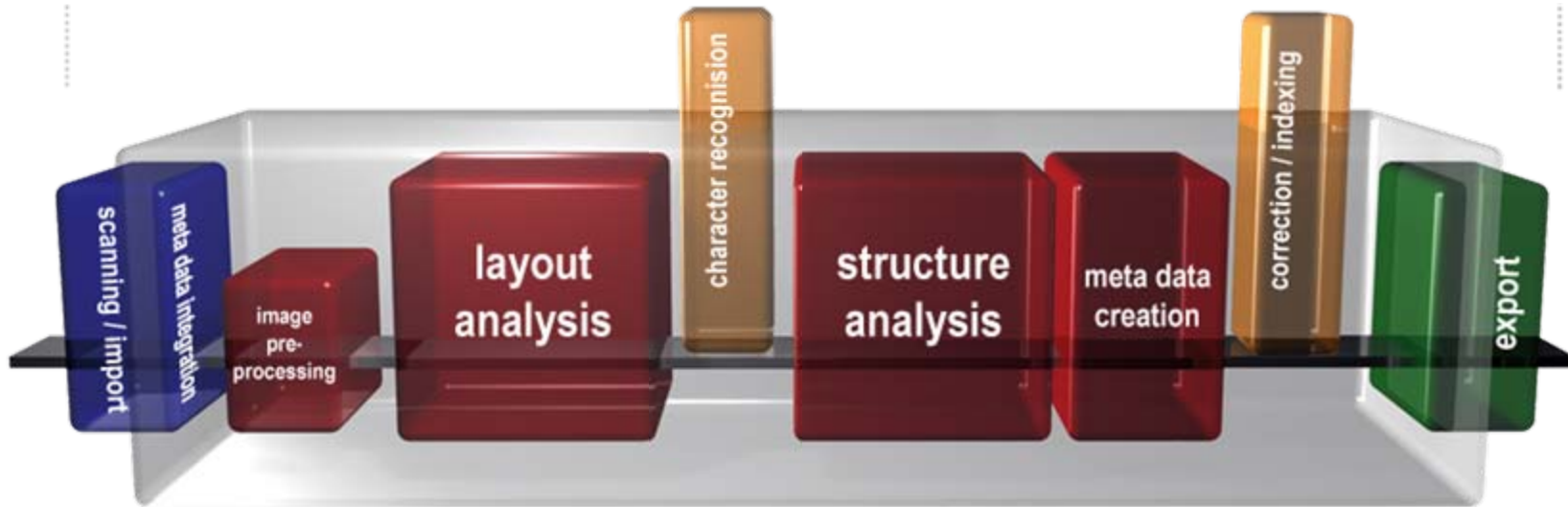
Image enhancement to support layout analysis and OCR

- Splitting of double page images
- Despeckling, noise reduction
- Deskewing
- Various image treatments to create an optimized image for OCR
- Problem to deal with uneven lighting on page or page sequence
- Problem to deal with bleed through



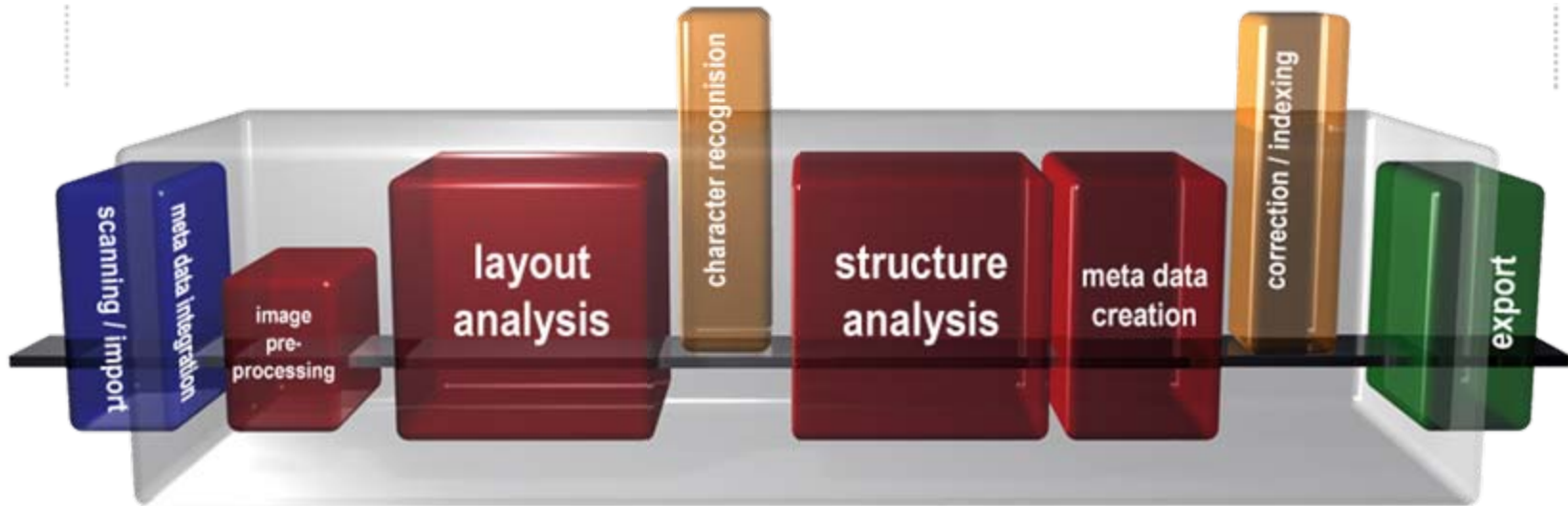
Collecting zone information through separating and tagging page elements

- Purely based on image analysis methods



OCR (1)

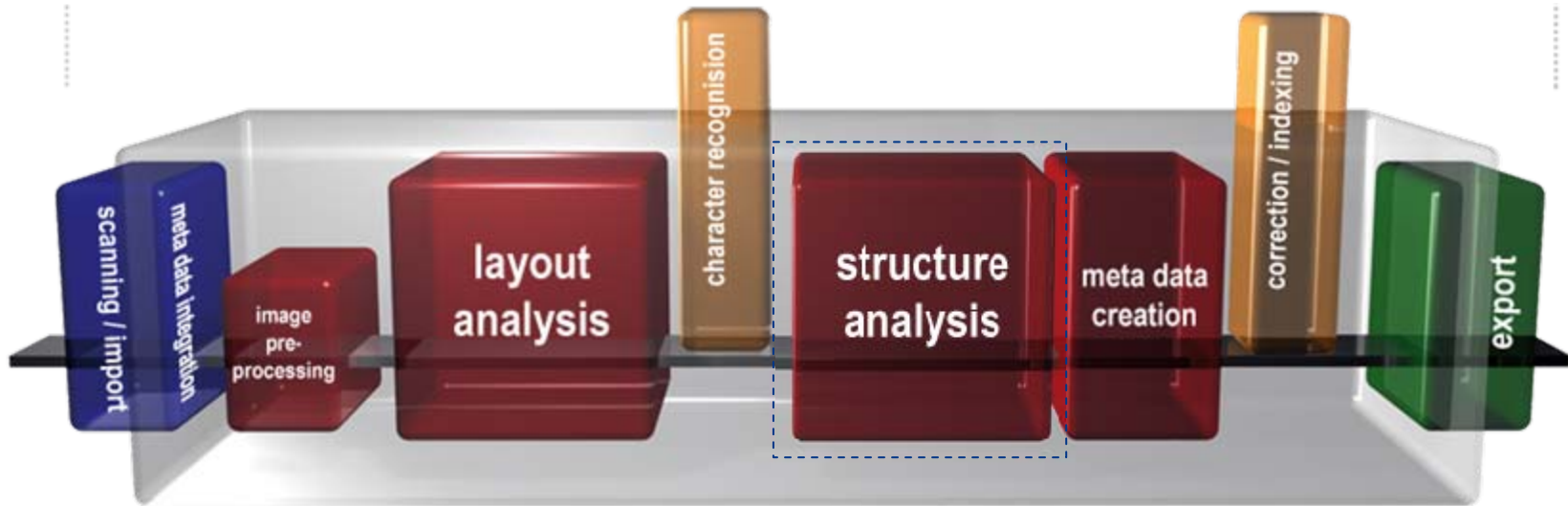
- On whole text corpus to create fulltext (running text)
- On page numbers to build page sequence
- On chapter titles, article headings, captions, marginalia, footnotes, etc.
- Recognized text as additional input for structure analysis (e.g. page classification)



OCR (2)

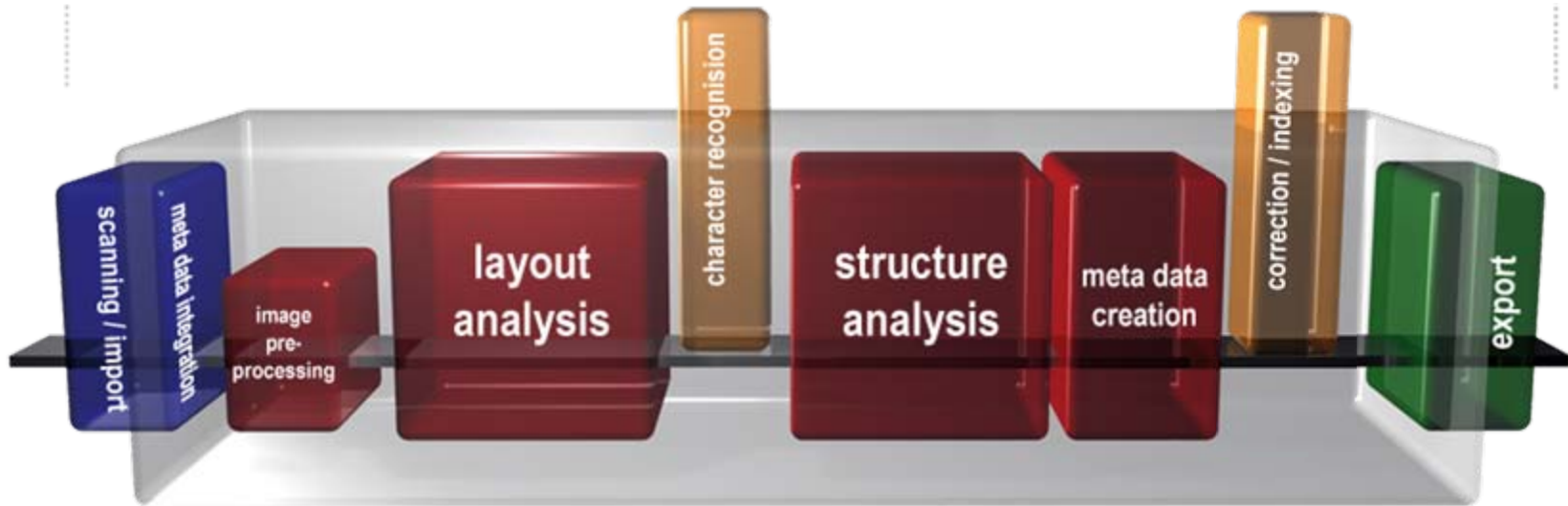
- missing dictionaries for old languages
- missing support for historic fonts
- problems with curved columns and text lines
- bad recognition of outline fonts (headlines)
- bad results on titles, headlines with specific fonts
- insufficient recognition of formulas
- no support for handwritten

Extol
Das Streiflicht
Frankfurter



structure analysis automatically recognizes logical entities like chapters, contributions, articles, ...

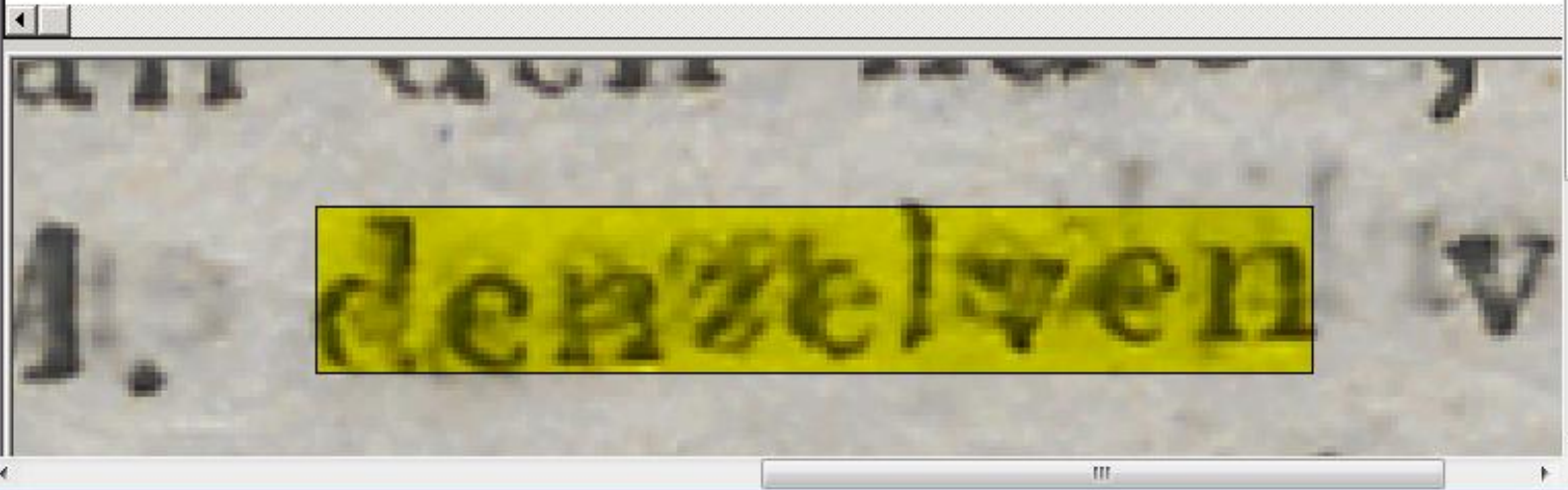
- major capabilities of OCR play a minor role
- important are font attributes like size and style (bold, italic, spaced)
- in certain cases also font types will be utilized



Interactive QA

- Correction of logical structure
- No correction of running text in mass digitisation projects
- Correction of specific text elements like chapter titles, headlines, captions of illustrations, footnotes, page numbers,

20. Aai. d«.i huis-, hof- en staats kanselier Prins van Mn'.ernwh , waar-B
 bij Z. M. deß*fc)wi verzekering geeft van de erkenning zijner ■
 sten jegens -zijn verbeven vader, zijn.bons en den staat,, hém (08 v.-:Al
 zetting zijner riiersAn uiüioodigt en hem telnat alle zpn<? <-.*,J,i*ho;>;
 binnen- en buitahslahda van de bevestiging in hunne .p<s*9A kennis tel
 geven. I
 3". AindengfasfYan. H.trdegg , generaal der kavellerle en president»
 vaa don raad van oorlog , waarbij dezelve wordt aangeschreven „ omviel



- avoid decision for special characters within words being part of running text
- Support of a wider range of historic fonts beyond black letter fonts (Fraktur)
- Increase OCR-speed to lower hardware cost

- Digitisation and conversion of historic printed items reached a good quality level, but still reveals problems in scanning, imaging, OCR and correction
- Beside improving scanning, imaging and interactive correction technologies the goal should be getting OCR to the next level to reduce the gap between machine and human reading
- Linguistic, semantic and statistical methods seem to have the potential to reach that goal
- Mass digitisation need “next level OCR”

Thank you!

Claus Gravenhorst
Director Strategic Initiatives

CCS Content Conversion Specialists GmbH

information:accessible

Weidestr. 134, D-22083 Hamburg, Germany

+49 (0) 402 271316 phone

+49 (0) 402 2713011 fax

+49 (0) 163 271316 mobile

claus.gravenhorst@content-conversion.com

Internet: www.content-conversion.com