CCS & Goobi

Goobi Anwendertreffen - 12.5.16 - Mannheim

Claus Gravenhorst

Director Strategic Initiatives

About CCS | Some facts

- CCS - Content Conversion Specialists is a privately owned company with headquarters in Hamburg, Germany
- Technology company developing market-leading software and hardware for the creation and display of digital collections
- Founded in 1976 , 50+ employees (Germany, Romania, US)
- Participating in US research project:
  - Library of Congress (2004), NDNP specification
- Participating in European research projects:
  - METAe – The Metadata Engine (2000 – 2003)
  - ENP – Europeana Newspapers Project (2012 – 2015)
- Mass digitisation projects:
  - The British Library (books, 2007 – 2008)
  - Dutch National Library (newspapers, 2008 – 2012)

Portfolio

Consulting

Successful projects start with careful listening.

CCS has one of the best expertise in digitisation programs and project management.

We analyze your requirements and offer a customized workflow design based on your architecture and standards.

Technology

Providing cutting-edge technology to meet the requirements of the market.

nW newsWorks

dW docWorks

iT itemTracking

Content
Experience
Solutions

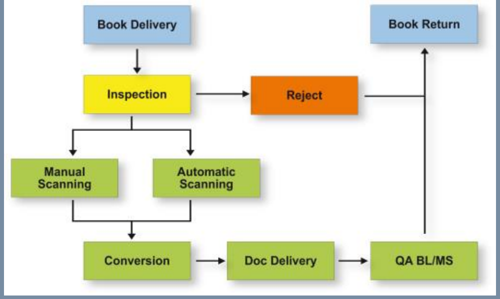
Digitisation Services

Processing more than 2 million pages every month in the lead libraries of the world.

Best expertise in applied integrated digitisation workflow. Dealing with most valuable items.

Newspapers / Books

Journals / Magazines



docWorks | Selected References

docWorks I

National Library of Norway
National Library of Finland
National Library of Luxemburg
National Library of Latvia
National Library of Estonia
National Library of Slovakia
National Library of Poland

National Library of Vietnam
National Library of Australia

University of California, Riverside
University of Texas, Austin
Library of Congress, Washington DC
Harvard University, Cambridge
Princeton University, Princeton
National Library of Medicine, Bethesda
Washington State Library, Olympia
Library of Virginia, Richmond
Queens Library, New York
Cleveland Public Library
Indiana State Library, Indianapolis
J. Paul Getty Trust, Los Angeles

National Library of Trinidad & Tobago

docWorks II

Digital Divide Data, USA
Backstage Library Works, USA
Hudson Microimaging, USA
brightsolid., UK
LETA, Latvia
CD Imaging, Singapore
Contentra Technologies, India

More than 100 million book pages
More than 20 million newspaper pages

digitizationServices

British Library, London
Royal Library of the Netherlands
National Library of Luxemburg
National Library of Finland
National Library of Norway
Royal Library of Denmark
National Library of Latvia
National Library of Austria
Wellcome Library, London

FAZ, Germany
Axel Springer Verlag, Germany

Library of Congress, Washington DC
Washington State Library, Olympia
University of California, Riverside
University of Minnesota, Minneapolis
Michigan State University, East Lansing
National Library of Australia
National Library of Singapore
National Library of New Zealand

More than 25 million book pages
More than 15 million newspaper pages

ENP – Europeana Newspapers Project

- CCS, as technical project partner, provided its expertise and docWorks technology to set up and operate a mass digitization workflow for creating high quality structured content from 2 million scanned newspaper pages provided by 5 library partners

- Page volume:

  BNF=1.000 k, NLE=500 k , SUB HH=480 k, NLF=90 k, SBB=10 k

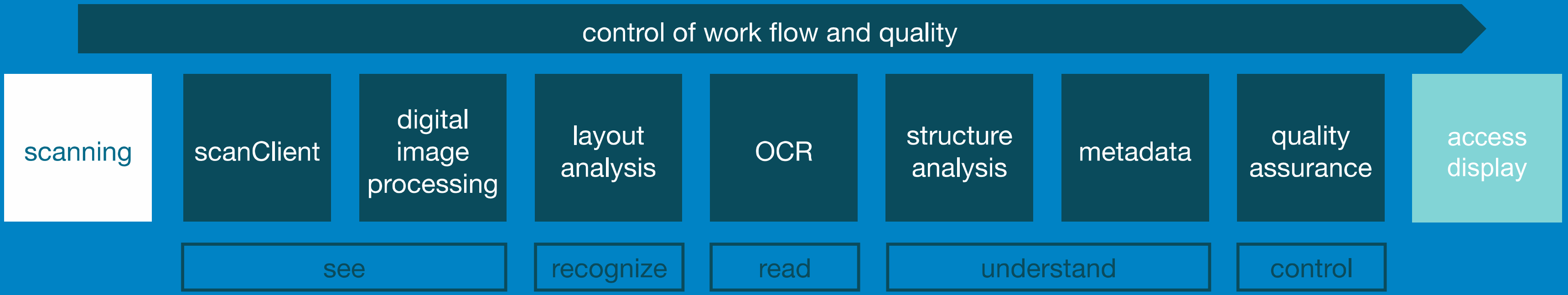
- The distributed OLR workflow enabled the contribution of project partners (content providers) to the integrated quality assurance process

- CCS has also contributed to the specification of the ENMAP* metadata model

* ENMAP = Europeana Newspapers Mets Alto Profile

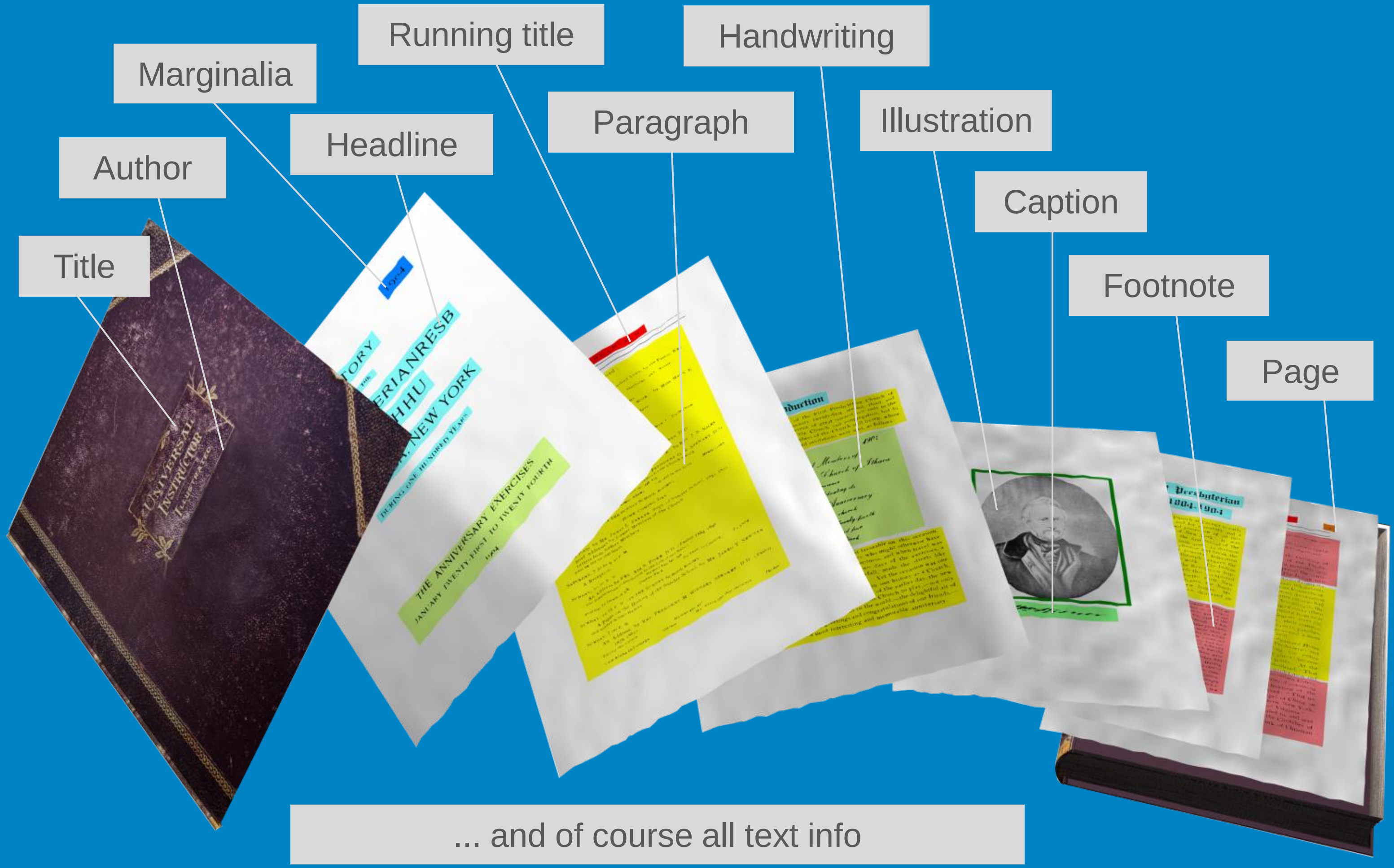
What is docWorks

- docWorks is a continuous digitisation workflow software including automated and interactive quality assurance options for every processing step.
- It’s a highly scalable solution serving small, medium and large scale digitisation projects.



- According to the required output features processing steps can be activated and deactivated.

Structure Analysis | Monograph



Title

Author

Marginalia

Headline

Running title

Paragraph

Handwriting

Illustration

Caption

Footnote

Page

... and of course all text info

Structure Analysis | Newspaper

- General rule system enables recognition of words, text lines, text blocks, columns and classification of text blocks, illustrations, advertisements, tables and the following page types:

  - title page (the title page of an issue)
  - content page (a page that consists of content/text only)
  - illustration page (a page that has at least one illustration)
  - advertisement page (a page that contains adverts only)

- Structure analysis through classification of headlines and grouping of zones into articles
  (incl. article continuation)

L'ÉCHO DE PARIS
RÉFLEXIONS sur la guerre d'Espagne
L'Allemagne et l'Italie se retirent définitivement du système de contrôle
LES CHANCES DU CABINET CHAUTEMPS

L'ÉCHO DE PARIS
RÉFLEXIONS sur la guerre d'Espagne
L'Allemagne et l'Italie se retirent définitivement du système de contrôle
LES CHANCES DU CABINET CHAUTEMPS

CCS & Goobi – Overview

- Goal: enhance quality for Goobi users by integrating docWorks features “Image pre-processing”, “Layout analysis”, “OCR” and “Structure Analysis” into the Goobi workflow
- 1-year development project with Saxon State Library Dresden
- Beside newspapers also other document types like magazines and books will be supported
- Automated server-based processing
- Decision on re-use scenario by the end of 2016

docWorks & Goobi – Roadmap

- Q1/2016
  - kick-off, start of specification
- Q2/2016
  - finalize specification
  - set-up docWorks test system on SLUB server
- Q3/2016
  - implement data exchange, develop missing components,
  - proof of concept with monograph processing
- Q4/2016
  - integrated OCR and layout/structure recognition of docWorks proves to be robust and allows high volume processing
  - support for magazines and newspapers
  - beta version and first official release

Thank you!

Claus Gravenhorst

Director Strategic Initiatives

CCS Content Conversion Specialists GmbH
Weidestr. 134
22083 Hamburg
Germany
T +49 40 227130-16
F +49 40 227130-11
M +49 176 12713016
c.gravenhorst@content-conversion.com
www.content-conversion.com