Experimental workflow design and development in digitisation

Zeki Mustafa Dogan (1), Clemens Neudecker (2),
Sven Schlarb (3) and Gerd Zechmeister (4)

(1) Goettingen State and University Library, Platz der Goettinger Sieben 1, 37073
Goettingen, Germany; Tel.: +49 551 39 52 12; dogan@sub.unigoettingen.de
(2) Koninklijke Bibliotheek; Prins Willem-Alexanderhof 5; 2595 BE Den Haag;
The Netherlands; Tel. +31 70 3140781; Email: clemens.neudecker@kb.nl
(3) Austrian National Library; Josefsplatz 1, A-1015 Vienna, Austria; Tel.:
+43-1-53410-491; Email: sven.schlarb@onb.ac.at
(4) Austrian National Library; Josefsplatz 1, A-1015 Vienna, Austria; Tel.:
+43-1-53410-506; Email: gerd.zechmeister@onb.ac.at

Library collections are composed of non-digitally born assets representing
text, such as books, newspapers or manuscripts, where no digital version -
at least with satisfying quality of searchable text - has been created so far.
Setting up a workflow for the digitisation of such collections is a complex
exercise that includes several processes and requires considerable hardand
software resources as well as skilled personnel.
In this abstract we present the idea of an extensible platform for
digitisation of historic printed material with a focus on Optical Character
Recognition (OCR). Our primary assumption is that before setting up a
productive digitisation workflow, the software based processes and
available components should be evaluated beforehand in order to assess
the general feasibility and to identify additional resources that might be
required.
The technical framework for the platform is built within the IMPACT
project which is funded by the European Commission and aims to
significantly improve access to text-based European cultural heritage. In
order to establish interoperability between the very specific tools which are
being developed in the project, IMPACT creates a set of services and
workflow modules, comprising pre- and post-processing steps from image
enhancement down to language correction tasks. From these fragments,
more complex workflows will be developed and demonstrated by the
content-holding institutions as part of the project. Albeit these workflows
refer especially to the tools developed and used by IMPACT, they can also
be seen as a general guideline for creating OCR related digitisation
workflows.
The experimental environment used in IMPACT allows the setup and
evaluation of such workflows using a wide variety of components that are
made available in the framework. It provides the means for collaborative
management and assessment of distributed software components with
regard to their usability, quality, effectiveness, and scalability especially in
terms of their aptitude within a mass digitisation project.

Keywords: Digitisation, Digital Library, OCR, Text Recognition,
Workflow

http://www.impact-project.eu