National Library of Sweden

Kravspecifikation DD v.1.2

Dokumenttyp	Status	Version	Versionsdatum
Kravspecifikation	Godkänd	1.2	2011-04-29

Ändringshistorik

Datum	Ändring	Ändrad av
2011-03-22	Gråskalebild blir färgbild, Visningsfil tas bort, PDF/A tillkommer, kvalitetsdata tillkommer	Heidi Rosen

Inledning

Detta dokument, *Kravspecifikation DD v1.2*, föreslås för godkännande av styrgruppen på KB. Då Digidaily är ett utvecklingsprojekt, kommer kravspecifikationen att vara ett levande dokument som vid behov, under projektets gång, kan komma att uppdateras.

OCR tolkningen och sökbarheten i tidningsmaterialet har varit det primära vid framtagandet av denna kravspecifikation.

Färg:

Tidningarna sparas i färg (8 bitar/kanal).

Filformat och filer:

- **Arkivfil** i JPEG2000 som komprimeras efter KB:s önskemål enligt bilaga *Preliminär specifikation för JPEG2000 (arkivfil).* [1]

  Filen får max vara 300ppi om det inte verkar negativt för läsbarheten.

  Arkivbilden ska vara obearbetad.

  För bildkvalité, se kvalitetsnivå 2 för färgbilder i bilagan *Kvalitetsnivåer2010_nyversion.* Anvisningar för mätningar av bildkvalité finns i dokumentet *Riktlinjer_för_kvalitetskontroll.doc*

  Arkivfilen kommer även att användas som visningsfil. Detta betyder att endast en bildfil per tidningssida ska levereras till KB.[2]

---

[1] Kontakter med deltagare i IMPACT-projektet för OCR-tolkning har påvisat att försiktig förstörande komprimering ej påverkar OCR-resultatet.

[2] En separat visningsfil ger möjlighet att optimera filen för visning på skärm genom skärpning och lokal kontrasthöjning. Det finns minst en bildserver på marknaden (JHelioviewer) som kan skärpa bilden i samband med leverans till ett visningsgränssnitt varför det är möjligt att vissa egenskaper hos en separat visningsfil även kan erhållas då endast en arkivfil lagras. Relevanta detaljer kravställs av projektet när visningsgränssnittet planeras.

National Library of Sweden

- **PDF-fil** som innehåller tidningsnumrets samtliga sidor. Filen ska följa standarden PDF/A, önskvärt är PDF/A-1a, men allra lägst PDF/A-1b . Bildernas färgdjup ska företrädesvis reduceras till 1-bit (dvs. tvåton), men även gråskala accepteras. Bilderna ska ha samma upplösning och storlek som arkivfilen. Den teckentolkade texten ska lagras i ett lager bakom bilden som föreställer tidningssidan. Bilderna ska optimeras för läsning på bildskärm.

- **Kvalitetsdata** med information om den senaste kvalitetsmätningen lagras i XML. Se dokumentet *imageQualityTemplate.xml* för en specifikation av filens innehåll. Under taggen <patchData> sparas färginformation för målets samtliga patchar. Under taggen <measurements> sparas medelvärden beräknade för samtliga patchar. Endast de taggar som fylls med data under <gainModulation> behöver inkluderas i filen.

Beskärning:
Arkivfil ska beskäras så att andelen bakgrund minimeras.

Upprätning:
Kan förlagan inte skannas rakt får upprätning ske. Upprätning får även utföras om förbättrat OCR resultat kan uppnås. MKC genomför tester för att undersöka detta.

Upplösning:
Upplösningen baseras på kvalitén på OCR-läsningen. Upplösningen kommer att vara beroende av tidningens format och fontens grad (storlek). Projektet på MKC jobbar fram rutiner för detta.

OCR, Antikva:
Om möjligt en ordsannolikhet på 80 % . Går detta inte att uppnå ska målet vara att uppnå en teckenriktighet på 95 %. Dessa siffror avser faktiska sannolikheter och inte de siffror som härrör från program för OCR-tolkning. För att uppnå målet kan ordlistor användas.

OCR, Fraktur:
Ingen OCR-tolkning utförs på tidningar som huvudsakligen är tryckta i fraktur. Arkivfilen sparas i originalupplösning för framtida OCR-hantering. OCR-kravspecifikation levereras vid senare tillfälle.
Kostnader för hantering av tidningar med fraktur tas fram av projektet för framtida bruk.

Segmentering:
Segmentering sker på sidnivå. Artiklar och bilder segmenteras ej.

Format på tidningar:
Se bilaga *Fysisk beskrivning*.

Teknisk metadata:
Information om teknisk metadata, filnamngivning samt hur paketen ska byggas redovisas i separat metadataspecifikation.
Alla textfiler sparas i UTF-8.