To Harmonize Quality and Quantity

Heidi Rosen; the National Library of Sweden. Torsten Johansson; the National Library of Sweden.

Abstract

An ever-growing Internet places an increasing demand on cultural institutions ability to deliver digital material. Not too long ago internal production lines managed the demand, but today the demand of making digitised material available is so large, that many institutions no longer cope with digitising within their own ranks. In recent years, many institutions got into so-called mass digitising projects to be able to live up to the high demands from the public, but also in the hope that in such a production environment finds cost-effective methods that enable the digitising of the wealth of material that they have to administer.

These large development projects are often financially driven by different funds and contributions in order to build a production line. But what happens when these projects are subjected to reality, that is, moving from a development environment to a production environment where a lot of money can be involved? What's the expectation of the end product; will it differ from the expectations and requirements one has on products from internal production lines?

The digitising dilemma

Clearly, one can digitise very large volumes in a short time, but why do it if the quality is poor or barely acceptable? And who decides what is acceptable quality, the customer, the supplier or the end user? As a client you often hear that it's impossible, that you cannot expect a perfect result just because it is about large volume digitising. One has to accept what is offered, otherwise one is perceived as unrealistic. For an example, would a customer who buys a new TV set accept stripes across the screen, even if the TV set was made on an assembly line?

As a national library we do have high quality standards, even in the digitising that happens in-house. Since the digitised material often will be used instead of the original material / object, original similarity is therefore of great importance. Therefore, we mean that it is impossible to accept the argument that high quantity must automatically mean lower quality. For us, there is no natural connection between the concepts. However, it is important to us that quality control is done with fully automated processes in order to keep production costs down.

Digitising of newspapers often involves large volumes and thus inevitably high costs. Moreover, the earlier newspaper material is often brittle and it is doubtful whether it is capable of further digitising occasions.

In our case, we may never get the opportunity to digitise “Svenska Dagbladet” from the 27th of October, 1884 again as the paper can only handle one pass through the scanner, so it must be right straight away.

In order to determine how much a particular increase in quality is worth, one must first have a basic price list to start with. We have carefully calculated prices for all the different newspaper types. When we work out a new method to secure quality, the price might change. In the end it is KB, in its role as purchaser and National Library, who must make the final decision if the increased quality is related to the price increase.

Work on quality in project Digidaily

The concept of quality is subjective and each organization must carefully consider which quality standards are desirable and where the lower quality limit lies. The nature of the originals, the intended use, data storage facilities, technical equipment for digitising and last but not least, financing, all are aspects that will affect the quality level you choose.

At the National Library of Sweden, we use three quality levels, each with Metamorfoze as a basis.

1. Items with high demands on colour accuracy
2. Items with standard requirements for colour accuracy
3. Items that are digitised in gray scale

As an example we have in our ongoing newspaper project "Digidaily", which we presented at last year's conference, chosen level 2. We have had the opportunity to be able to change and improve the specifications as we went along, which has been a great benefit. For an example, in the beginning of the project we were sure that gray scale images saved in TIFF was right for us. It turned out that the files, despite the gray scale, would have been far too large, which was not realistic and financially feasible when you have in mind that we wanted to begin digitising 1/3 of our collection of 122 million pages.

We therefore worked on and came to the conclusion that if we saved the files in jpg2000 instead, we could afford to reproduce the newspapers pages in colour, which would be perceived as an added value to the user.

Another important aspect of quality that we have improved during the project is segmentation of the pages. For a long time we had the idea that segmentation at page level would be sufficient for the user's needs, but after listening to arguments from Simon Tanner [1] and Edwin Klijn [2] we thought some more and then re-wrote the requirements, including segmentation to article level.

The system we use for segmentation and OCR-interpretation enables us to automatically segment pages at article level, and the results so far are acceptable, we may end up at an accuracy of around 80%, but for us it's *good enough*, gives no extra cost and increase the added value for the end users.

Since the beginning, the idea behind our project has been to automate as many processes as possible. Our thesis has been to minimize human labour wherever possible, to keep the final price at a level as low as possible.

Digitising costs

The high cost in terms of newspaper digitising is the manual preparation of the pages before digitising and the image capture.

© Copyright 2013; Society for Imaging Science and Technology

Other parts of the digitising chain, such as OCR, automated QC etc. are more or less machine hours and these can be pressed but that will not impact in the final price too much.

Below is an example of the different newspaper categories and the production price. Please note that the price in the second table is in Swedish Krona.

Category	**Price**	**Comment**
1. Bound, torn, fragile paper, the biggest format size	1.03 USD	About 4 % of the collection
2. Bound, most of it can be taken apart, only a few are saved still bounded, fair paper quality.	0.51 USD	Main part of the collection, 49 %
3. Tabloids stapled but not bound.	0.34 USD	About 47 % of the collection
4. National copies – not to be taken apart. Or very poor paper quality	2.8 USD	Less than 1 % of the collection

Category 2	**Pages/shift**	**Price/page (SEK)**	**% off the total cost**
Logistics	50 000	0,08 SEK	3%
Preparation - go through	7 676	0,54 SEK	17%
Preparation - take apart	6 550	0,63 SEK	20%
Image capture	2 946	1,40 SEK	43%
QC	124 459	0,03 SEK	1%
Segmentation & export	96 322	0,04 SEK	1%
OCR-click		0,50 SEK	15%
Total price		3,23 SEK	100%

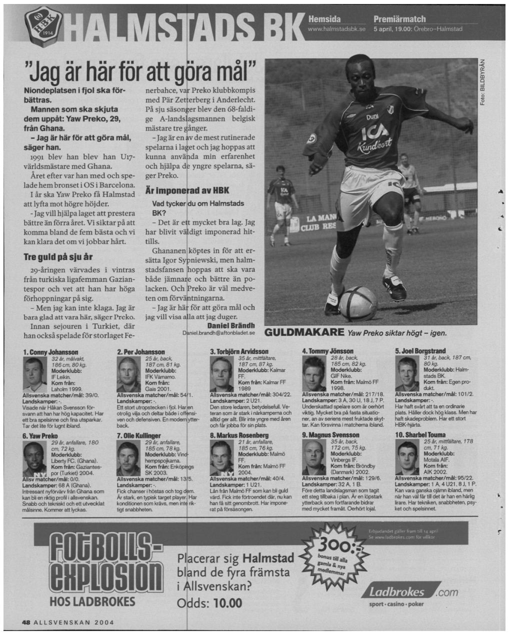
We also try to automate the quality control. All image capture equipment are checked and calibrated with test targets at specific intervals. The evaluation of the test targets is automated and the results are stored in the production system so that it is possible to see changes over time. The test results are also saved in a performance file in XML, which will be stored in the in SIP file along with the newspaper images.

So far the QC has been more or less manually done. However, this is completely unrealistic in a process that will handle millions of images every year. Therefore we are now working on developing a statistical sampling method that will minimize the material to be inspected manually.

Practical problem solving

We are always actively looking for new rational ways to scan newspaper pages. But with new scanner models there also arise unexpected quality problems. We became unpleasantly surprised when we noticed that our new Supag MediaScan all too often gave vertical stripes across the newspaper pages. And the reason for that was dust. Instead of a "dust spot" like on a traditional scanner, dust on these types of scanners instead give coloured vertical stripes. So the effect was that the much cheaper and more efficient scanner gave rise to an entirely new problem. *Good enough*? No, not for us, stripes are not OK.

After many internal discussions at KB we have decided to accept artefacts in the sense coloured lines if they do not exceed two lines per digitised page and that the lines are a maximum of three pixels wide. Thicker lines than three pixels are not accepted at all.



HALMSTADS BK

Hemsida
www.halmstadsbk.se

Premiärmatch
5 april, 19.00: Örebro–Halmstad

"Jag är här för att göra mål"

Niondeplatsen i fjol ska förbättras.

Mannen som ska skjuta dem uppåt: Yaw Preko, 29, från Ghana.

– Jag är här för att göra mål, säger han.

1991 blev han blev han U17-världsmästare med Ghana.

Året efter var han med och spelade hem bronset i OS i Barcelona.

I år ska Yaw Preko få Halmstad att lyfta mot högre höjder.

- Jag vill hjälpa laget att prestera bättre än förra året. Vi siktar på att komma bland de fem bästa och vi kan klara det om vi jobbar hårt.

Tre guld på sju år

29-åringen värvades i vintras från turkiska ligafemman Gaziantespor och vet att han har höga förhoppningar på sig.

– Men jag kan inte klaga. Jag är bara glad att vara här, säger Preko.

Innan sejouren i Turkiet, där han också spelade för storlaget Fenerbahce, var Preko klubbkompis med Pär Zetterberg i Anderlecht. På sju säsonger blev den 68-faldige A-landslagsmannen belgisk mästare tre gånger.

– Jag är en av de mest rutinerade spelarna i laget och jag hoppas att kunna använda min erfarenhet och hjälpa de yngre spelarna, säger Preko.

Är imponerad av HBK

Vad tycker du om Halmstads BK?

– Det är ett mycket bra lag. Jag har blivit väldigt imponerad hittills.

Ghananen köptes in för att ersätta Igor Sypniewski, men halmstadsfansen hoppas att ska vara både jämnare och bättre än polacken. Och Preko är väl medveten om förväntningarna.

– Jag är här för att göra mål och jag vill visa alla att jag duger.

Daniel Brändh
Daniel.brandh@aftonbladet.se

Foto: BILDBYRÅN

GULDMAKARE *Yaw Preko siktar högt – igen.*

1. Conny Johansson
32 år, målvakt, 186 cm, 80 kg. **Moderklubb:** IF Leikin. **Kom från:** Laholm 1999.
Allsvenska matcher/mål: 39/0.
Landskamper: -.
Visade när Håkan Svensson försvann att han har hög kapacitet. Har ett bra spelsinne och fina utsparkar. Tar det lite för lugnt ibland.

2. Per Johansson
25 år, back, 187 cm, 81 kg. **Moderklubb:** IFK Värnamo. **Kom från:** Gais 2001.
Allsvenska matcher/mål: 54/1.
Landskamper: -.
Ett stort utropstecken i fjol. Har en otrolig vilja och deltar både i offensiven och defensiven. En modern ytterback.

3. Torbjörn Arvidsson
35 år, mittfältare, 187 cm, 87 kg. **Moderklubb:** Kalmar FF. **Kom från:** Kalmar FF 1989
Allsvenska matcher/mål: 304/22.
Landskamper: 2 U21.
Den store ledaren, betydelsefull. Veteran som är stark i närkamperna och alltid ger allt. Blir inte yngre med åren och får jobba för sin plats.

4. Tommy Jönsson
28 år, back, 185 cm, 82 kg. **Moderklubb:** GIF Nike. **Kom från:** Malmö FF 1998.
Allsvenska matcher/mål: 217/18.
Landskamper: 3 A, 30 U, 18 J, 7 P.
Underskattad spelare som är oerhört viktig. Mycket bra på fasta situationer, an av seriens mest fruktade skyttar. Kan försvinna i matcherna ibland.

5. Joel Borgstrand
31 år, back, 187 cm, 80 kg. **Moderklubb:** Halmstads BK. **Kom från:** Egen produkt.
Allsvenska matcher/mål: 101/2.
Landskamper: -.
Har haft svårt att ta en ordinarie plats. Håller dock hög klass. Men har haft skadeproblem. Har ett stort HBK-hjärta.

6. Yaw Preko
29 år, anfallare, 180 cm, 72 kg. **Moderklubb:** Liberty FC. (Ghana). **Kom från:** Gaziantespor (Turkiet) 2004.
Allsv matcher/mål: 0/0.
Landskamper: 68 A (Ghana).
Intressant nyförvärv från Ghana som kan bli en riktig profil i allsvenskan. Snabb och tekniskt och ett utvecklat målsinne. Kommer att lyckas.

7. Olle Kullinger
29 år, anfallare, 185 cm, 78 kg. **Moderklubb:** Vindhempspojkarna. **Kom från:** Enköpings SK 2003.
Allsvenska matcher/mål: 13/5.
Landskamper: -.
Fick chanser i höstas och tog dem. Är stark, en typisk target player. Har konditionen som krävs, men inte riktigt snabbheten.

8. Markus Rosenberg
21 år, anfallare, 185 cm, 76 kg. **Moderklubb:** Malmö FF. **Kom från:** Malmö FF 2004.
Allsvenska matcher/mål: 40/4.
Landskamper: 1 U21.
Lån från Malmö FF som kan bli guld värd. Fick inte förtroendet där, nu kan han få sitt genombrott. Har imponerat på försäsongen.

9. Magnus Svensson
35 år, back, 172 cm, 75 kg. **Moderklubb:** Vinbergs IF. **Kom från:** Bröndby (Danmark) 2002.
Allsvenska matcher/mål: 129/6.
Landskamper: 32 A, 1 B.
Före detta landslagsman som tagit ett steg tillbaka i plan. Är en löpstark ytterback som fortfarande bidrar med mycket framåt. Oerhört lojal.

10. Sharbel Touma
25 år, mittfältare, 178 cm, 71 kg. **Moderklubb:** Motala AIF. **Kom från:** AIK 2002.
Allsvenska matcher/mål: 95/22.
Landskamper: 1 A, 4 U21, 8 J, 1 P.
Kan vara ganska ojämn ibland, men när han väl får till det är han en härlig lirare. Har tekniken, snabbheten, psyket och spelsinnet.

FOTBOLLS-EXPLOSION
HOS LADBROKES
Placerar sig Halmstad bland de fyra främsta i Allsvenskan?
Odds: 10.00
300:- bonus till alla gamla & nya medlemmar
Erbjudandet gäller fram till 14 april
Se www.ladbrokes.com för villkor
Ladbrokes.com
sport · casino · poker

48 ALLSVENSKAN 2004

Figure 1. Vertical line across the newspaper pages

Newspapers are dusty by nature, particularly old newspapers, and this type of scanner uses a conveyor belt that grips the paper and drags the page into the scanner. Dust and small pieces of paper can then whirl around and lie down on the glass above the scanner sensor.

To scan dusty old newspapers pages on this type of scanner therefore requires optimal cleanliness, i.e. continuous cleaning. Therefore, we now have a first measure improved the cleaning procedures. In order to bind the dust, we have also tested a humidifier, but it gave no appreciable improvement.

But how do you solve it practical and cost effective? The idea to manually go through all the digitised images was never an option; it must be solved with an automated method. So we found out a method to let software look for longitudinal lines of a certain pixel width and alert if any are detected. Only then will the image will be sent to manual control and then eventual go back for re-scan. The cost for this function is minimal when a computer performs it. The only extra cost is if there will be a re-scan.

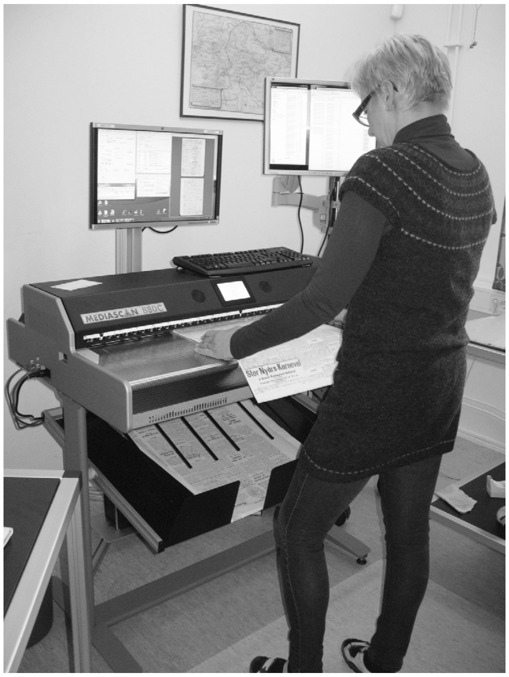
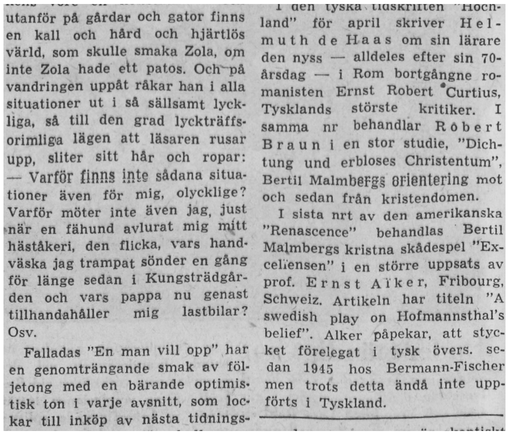


Figure 2. *Supag MediaScan*

Other problems we experienced with MediaScan scanner are stitch-problems and wavy text. If the newspaper page is not completely flat, the scanner software has difficulties in putting together the different images as one, with the result that some images may have wave-like text, or text with a misfit.


utanför på gårdar och gator finns en kall och hård och hjärtlös värld, som skulle smaka Zola, om inte Zola hade ett patos. Och på vandringen uppåt råkar han i alla situationer ut i så sällsamt lyckliga, så till den grad lyckträffsorimliga lägen att läsaren rusar upp, sliter sitt hår och ropar: — Varför finns inte sådana situationer även för mig, olycklige? Varför möter inte även jag, just när en fähund avlurat mig mitt häståkeri, den flicka, vars handväska jag trampat sönder en gång för länge sedan i Kungsträdgården och vars pappa nu genast tillhandahåller mig lastbilar? Osv.

Falladas "En man vill opp" har en genomträngande smak av följetong med en bärande optimistisk ton i varje avsnitt, som lockar till inköp av nästa tidnings-

I den tyska tidskriften "Hochland" för april skriver Helmuth de Haas om sin lärare den nyss — alldeles efter sin 70-årsdag — i Rom bortgångne romanisten Ernst Robert Curtius, Tysklands störste kritiker. I samma nr behandlar Róbert Braun i en stor studie, "Dichtung und erbloses Christentum", Bertil Malmbergs orientering mot och sedan från kristendomen.

I sista nrt av den amerikanska "Renascence" behandlas Bertil Malmbergs kristna skådespel "Excellensen" i en större uppsats av prof. Ernst Aïker, Fribourg, Schweiz. Artikeln har titeln "A swedish play on Hofmannsthal's belief". Alker påpekar, att stycket förelegat i tysk övers. sedan 1945 hos Bermann-Fischer men trots detta ändå inte uppförts i Tyskland.

Figure 3. *Wave-like text with bad stitching*

We have not yet found automated methods for finding these problems. To get around it all, we are now testing how pressing with a flatiron or flattening with a hydraulic press may reduce these errors. But this includes manual handling, which could result in a cost increase of 0,08-0,12 USD/page so the question is whether this increase in quality is worth it. The solution with a hydraulic press is cheaper and we think that the price increase will be about 0-0,08 USD/page.

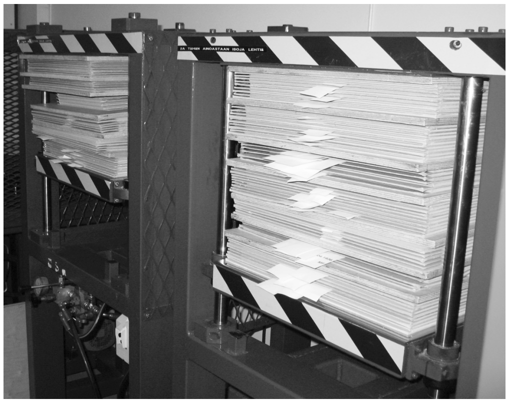


Figure 4. *Pressing pages at Centre for Preservation and Digitisation, Mikkeli, Finland*

What we have seen so far is that the wavy text will not affect the OCR interpretation so the quality increase would only be to get a more aesthetically pleasing image. We have not yet taken a decision how to act in this matter, and discussions are still ongoing.

Having to choose

A third scan we have conducted tests on is a document scanner. Our initial tests indicate a production level that far exceeds the other scanners we use today. However, the limitations are tabloid format and that the material must not be brittle or damaged.

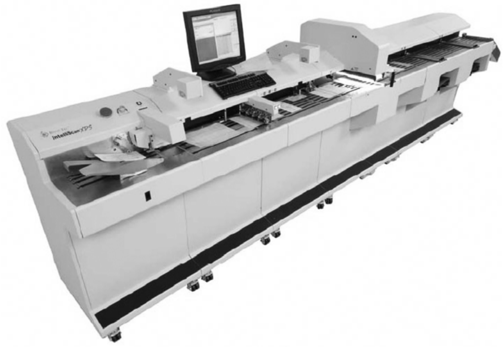


Figure 5. *High-speed document scanner*

As a comparison:

- Zeutschel OS 14000 scanner - 800 pages/shift
- Supag MediaScan - 4000 pages/shift
- Document scanner - 10000 pages/shift (or more)

The problem is that the document scanner might give an image quality lower than the requirements specified in KB´s specifications. It is particularly noise in the dark areas that might

© Copyright 2013; Society for Imaging Science and Technology

be below target, where the standard deviation maximum is set to be STD ≤ 4.

We do not believe that the higher level of noise will negatively affect the OCR result. The discussion “good enough” is here very relevant; a much higher production rate or a noise level that will meet the Metamorfoze standard?

Our methods to combine quality with quantity are based on a chain as automated as possible. It is machines, not people that will find quality deficiencies. The technical equipment is tested and the results are saved as an XML file and added to the metadata for the sake of traceability, all to ensure that the technical equipment emits such a good result as possible. Quality control also needs to be as automated as possible, and a file only goes to manual control in cases where the automated checker finds quality problems.

So our simple thesis and conclusion is to automate as much as possible, for it is only then one can combine quality and quantity.

Quality ♥ Quantity = True

References

[1] Simon Tanner, Director of Digital Consultancy, Dept of Digital Humanities at King´s Collage London, England

[2] Edwin Klijn, The NIOD Institute for War, Holocaust and Genocide Studies. Former project manager at Koninklijke Bibliotheek, Holland

Author Biography

Heidi Rosen has studied Graphic Arts Technology at KTH, Royal Institute of Technology in Stockholm. She is working as a project manager at the Newspaper Division at the National Library of Sweden, where she currently is responsible for the library' s part of Project Digidaily.