



Erhvervs- og Selskabsstyrelsen

Evaluering af kvaliteten af Læs-Ind-bureauernes fakturascanning

Maj 2005

Indholdsfortegnelse

1.	Indledning og formål	1
2.	Resumé	1
3.	Metode og dataindsamling	3
4.	Kvantitativ analyse og resultater	5
4.1	Informationer, der valideres både maskinelt og kvalitativt	7
4.2	Informationer, der alene valideres maskinelt	9
5.	Kvalitativ analyse og resultater	12
6.	Konklusioner og anbefalinger	23

1. Indledning og formål

Erhvervs- og selskabsstyrelsen (EogS) ønsker en vurdering af kvaliteten af Læs-Ind-bureauernes fakturascanning og har i april 2005 indgået aftale med Rambøll Management om en sådan vurdering. I nærværende rapport beskriver Rambøll Management (RM) resultaterne af den vurdering af kvaliteten af fakturascanningen, der er gennemført.

Vurderingen af materialet, der behandles af Læs-Ind-bureauerne, er omfattet af en statistisk dokumentation af fakturamaterialet vurderet på en række parametre vedrørende kvaliteten af den dannede OIOXML-streng. Vurderingen omfatter op til 14 obligatoriske, standardiserede informationer på fakturaen, den genererede OIOXML-streng og den tilhørende tiff-fil.

Vurderingen omfatter kun *outputtet* af Læs-Ind-bureauernes scanning af fakturamaterialet og er baseret på de OIOXML-filer, der genereres af scaningsprogrammet og det billede af fakturaen, der er indlejret som en tiff-fil i OIOXML-filen. Det skal i denne forbindelse bemærkes, at der i arbejdsgangene for scanning og etablering af OIOXML-filerne foregår manuelle vurderinger, hvorfor vurderingen af de dannede OIOXML-filer omfatter vurdering af scaningsprocessen og de manuelle processer i det omfang, disse indgår.

Vurderingen er koncentreret om de seks væsentligste oplysninger, nærmere bestemt fakturatotal, SE/CVR-nummer, OCR-linje, betalingsdato, fakturanummer og EAN-lokationsnummer. Derudover er der foretaget en maskinel behandling af de øvrige obligatoriske informationer, således at de gøres til genstand for statistisk vurdering af, om de tilhørende felter i OIOXML-filen er afløftet.

Vurderingen er baseret på en stikprøvekontrol, idet kontrol af fx en fuld dagsproduktion er meget omkostningsfuld.

2. Resumé

Undersøgelsen omfatter såvel en kvalitativ som en kvantitativ undersøgelse. De seks mest betydende informationer er undersøgt visuelt og manuelt for, om de er korrekt afløftet fra fakturaen (baseret på i OIOXML filens tiff-fil) til OIOXML-filen. Alle informationer omfattet af undersøgelsen er kvantitativt undersøgt for, i hvor høj grad default-informationer anvendes.

Den kvalitative undersøgelse

Resultaterne af den manuelle og visuelle undersøgelse af, i hvilket omfang de seks mest betydende informationer er korrekt afløftet, fremgår af nedenstående skema.

Der er efterfølgende gennemført et fokusgruppeinterview med henblik på at afdække de mest generelle årsager til forkerte afløftninger af data. De kvalitative vurderinger af årsager til fejl kan findes i kapitel 5.

INFORMATION	% af fakturaer, hvor informationen er korrekt afløftet til OIOXML-filen
Fakturatotal	98,80
SE/CVR-nummer	98,45
OCR-linje ¹	
OCR Art	95,31
OCR Betalingsid	94,30
OCR PBS-Kreditornummer	95,44
OCR Kontonummer	12,17 ²
Betalingsdato	97,15
Fakturanummer	97,10
EAN-lokationsnummer	96,95

Generelt er årsager til fejlflæsninger bl.a. forårsaget af, at:

1. de oplysninger, der skal aflæses, er indlejret i tekst, fx OCR-linje og betalingsdato.
2. de oplysninger der skal aflæses, er ikke foranstillet en sigende ledetekst, fx betalingsdato og fakturanummer og er vanskelige at identificere.
3. de oplysninger, der skal aflæses, ikke findes på første side af en faktura, der fylder flere sider, fx fakturatotal og OCR-linje.
4. scanningen aflæser "snavs" på fakturaen som punktum og bindestreger, fx i fakturanummer.

Den kvantitative undersøgelse

Undersøgelsen viser som ventet, at der er en meget lille forekomst af anvendelse af default-værdier blandt de seks informationer, der er vurderet som værende mest betydende for, om en betaling kan gennemføres. Blandt de øvrige data findes en væsentlig højere forekomst af default-værdier, fordi data ikke findes på fakturaerne.

Generelle konklusioner

Kvalitativt vurderes på basis af data fra undersøgelserne og det kvalitative fokusgruppeinterview, at en meget stor andel af de informationer, der overhovedet kan afløftes fra fakturaerne, bliver afløftet, og at scanningen er stærk til at afløfte korrekte data, når den meget store mangfoldighed af informationer og fakturaer tages i betragtning.

Der er i undersøgelsen ikke forekommet fakturaer, der er helt ulæselige af det menneskelige øje på baggrund af tiff-filen. I det omfang, fakturaer ikke kan læses *fuldstændigt*, skyldes det "snavs" på fakturaen, som gør det vanskeligt at aflæses enkeltstående oplysninger.

¹ Undersøgelsen er baseret på den del af stikprøven, hvor en OCR-streng er identificeret. Det betyder, at stikprøvegrundlaget er mindre end 2000.

² Dette lave tal skyldes primært, at feltet sjældent skal udfyldes. Således skyldes 82,89 % af de ikke-korrekte forekomster af OCR, at feltet ikke skal angives, mens der kun er forekommet fejl i 4,94 % af fakturaer med OCR-linje. Fejlen ER rettet.

3. Metode og dataindsamling

Kvalitativ vurdering

For at foretage en vurdering af, om data er afløftet korrekt i scanningen, er der foretaget en manuel og visuel vurdering af de væsentligste data på tiff-filen i forhold til de udfyldte værdier i OIOXML-filen. Disse har således gennemgået en kvalitativ vurdering af korrekthed.

De data, der er manuelt og visuelt kontrolleret, er følgende seks informationer:

- Fakturatotal
- SE/CVR-nummer
- OCR-linje
- Betalingsdato
- Fakturanummer
- EAN-lokationsnummer

For den manuelle visuelle kontrol er overordnet defineret følgende kategorisering af "rigtigt afløftet" og "forkert afløftet":

Som udgangspunkt er det vurderet i hvilket omfang informationen er:

- *Korrekt afløftet*
 - Hvis feltet er tomt i OIOXML-filen (har en default-værdi), og informationen heller ikke optræder på fakturaen.
 - Hvis feltet er afløftet med en værdi i OIOXML-filen, og informationen på fakturaen har netop denne værdi.
- *Forkert afløftet*
 - Hvis feltet er tomt (har en default-værdi) i OIOXML-filen, og informationen optræder på fakturaen.
 - Hvis feltet er afløftet med en værdi i OIOXML-filen, og informationen på fakturaen ikke findes.
 - Hvis feltet er afløftet i OIOXML-filen, og informationen findes på fakturaen, men værdierne er forskellige.
- *Ubestemmeligt*
 - Det er ikke muligt at afgøre, om informationen er korrekt afløftet, fx fordi informationen er ulæselig med det menneskelige øje på tiff-filen, eller fordi informationen er tvetydig på tiff-filen, og det dermed ikke er muligt at afgøre, om den korrekte information er afløftet.

I sammenhæng med den kvalitative vurdering er der foretaget en generel vurdering af kvaliteten af informationerne på tiff-filerne gennem fokusgruppeinterview med de personer, der har gennemført valideringen. Dette giver mulighed for at foretage en subjektiv vurdering af fx årsager til tvetydighed m.m. af informationer på tiff-filerne.

Kvantitativ vurdering

Alle felter i undersøgelsen er behandlet maskinelt og optalt, således at det kan vurderes, hvor stort et omfang brugen af fx default-værdier er.

De data, der *alene* behandles maskinelt, er:

- Ordre eller rekvisitionsnummer
- Personreference eller lignende reference
- Fakturadato

- Pengeinstitutkonto eller OCR-linje ved fordringer til tredjemand (maskinel validering af, om overførslen til tredjemand vil være behæftet med usikkerhed, da dette skal afkodes fra en tekst)
- Fakturatype - kreditnota/faktura (det kan alene afgøres, hvorvidt en faktura er en kreditnota eller en nota)
- Valutakode, hvis ikke DKK.

For de seks vigtigste felter, der ligeledes er vurderet kvalitativt, og for de øvrige obligatoriske felter ovenfor er der foretaget en maskinel validering, hvor der etableres følgende statistiske informationer:

- Afløftet og forskellig fra default-værdi
- Afløftet og lig med default-værdi
- Feltet forekommer ikke i OIOXML-filen.

Stikprøveomfang

Der er udført kontrol ved en stikprøve på 1000 fakturaer pr. Læs-Ind-bureau, *det vil sige 2000 i alt*. Stikprøvens omfang er begrundet i følgende:

- Det meget store antal fakturaer, der går gennem Læs-Ind-bureauerne, gør en population på 1000 enheder troværdig.
- En kontrol af 1000 enheder er en statistisk fornuftig og troværdig løsning, der anvendes i standardmeningsmålinger og kvalitetsvurderinger.
- En forøgelse af stikprøveomfanget giver ikke undersøgelsen væsentlig mindre usikkerhed.
- En kontrol af 1000 enheder vurderes statistisk at være behæftet med højst 3 % usikkerhed.

Kontrol af stikprøvekontrol

Det er RM's erfaring fra egne kvalitetsvurderinger af store indtastninger og scanninger fra spørgeskemaundersøgelser, at der vil forekomme fejl i en begrænset del af vurderingerne, når menneskelige øjne skal vurdere et omfattende materiale. Derfor er gennemført en dobbeltkontrol, idet en del af det kvalitetsvurderede materiale er kontrolleret endnu engang med henblik på at finde den usikkerhed, der knytter sig til kvalitetsvurderingen.

Omfanget af dobbeltkontrollerede fakturaer er 10 % af den samlede stikprøve.

Processen

Processen har omfattet følgende aktiviteter:

Del 1 – Etablering af indlæsningsprogrammet og opsætning af spørgeskema til dokumentation

1. Data er fremsendt til RM. Data er sendt som OIOXML-filer og udgør en dagsproduktion. Herfra har RM udvalgt 1000 tilfældige OIOXML-filer til stikprøven.
2. Fortolkning af fejl er aftalt kort efter opstart af projektet.
3. RM har udarbejdet elektronisk spørgeskema til dokumentation.
4. Fremsendt materiale er indlæst.

Del 2 – Visuel, manuel validering og kontrol

1. RM har gennemført manuel, visuel verifikation af de seks informationer, baseret på visuel afkodning af tiff-filernes repræsentation af de indscannede fakturaer, sammenholdt med værdierne i OIOXML-strengen.
2. RM har gennemført dobbeltkontrol på 10 % af de verificerede fakturaer og OIOXML-strengene.

3. RM har gennemført fokusgruppeinterview med de personer, der har gennemført den kvalitative vurdering. Formålet har været at afdække andre kvalitetsaspekter som fx forekomst af flere ens felter på fakturaen, årsager til forkerte aflæsninger mv., som ikke har kunnet afdækkes med henblik på at indgå i en statistisk opgørelse.

Del 3 – Statistisk analyse og rapportering

1. RM har gennemført statistisk bearbejdning og analyse af resultaterne.
2. RM har gennemført fokusgruppeinterview med de personer, der har gennemført den kvalitative vurdering.
3. RM har dokumenteret undersøgelse, metode, statistisk validitet og resultater i en nærværende, kortfattede rapport til EogS.

Forudsætninger for undersøgelsen

RM's visuelle, manuelle validering af de seks vigtigste informationer og tiff-filens læsbarhed baseres alene på visuel afkodning af den tiff-fil, der er indlejret i de enkelte OIOXML-skemaer.

4. Kvantitativ analyse og resultater

For at kunne gennemføre og forstå den kvantitative analyse er det nødvendigt at gøre en række forudsætninger, som vil blive klargjort inden analysen af resultaterne. Konkret er det nødvendigt at tage stilling til, hvad det fx betyder, når en værdi er afløftet og forskellig fra default-værdien? Hvad er fastlagt som default-værdi for det pågældende felt? Er der felter, som ikke forekommer i OIOXML-filen?

Forudsætninger for den kvantitative analyse

Når en værdi er forskellig fra default-værdien er det ikke ensbetydende med, at den er korrekt afløftet, men udelukkende et udtryk for, at der er indlæst en værdi. Omvendt er en afløftet default-værdi ikke et sikkert udtryk for at værdien ikke optræder på fakturaen, da det kan skyldes fejl i aflæsning eller ubestemmelige informationer på en faktura.

Det er i kravspecifikationen til Læs-Ind-bureauer, der ønsker certificering, jf. Bekendtgørelse om "Læs-Ind"-service, § 2, stk. 2., gjort kendt angående default-værdier at:

... Kan information til de obligatoriske felter ikke lokaliseres på fakturaen, anføres default-værdien "0" i numeriske felter eller "null" i tekstfelter. Videnskabsministeriet kan anvise andre default- eller "dummy"-værdier til felter, hvor de nævnte default-værdier ikke kan anvendes.

...Læs-Ind-bureauer påfører OIOXML-Læs-Ind-meddelelser, der dannes ud fra fakturaer modtaget uden EAN-nummer, et default-EAN-nummer. Dette findes ved opslag i en database, der stilles til rådighed af Økonomistyrelsen.

Nedenstående er angivet en oversigt over de default-værdier, der er *registreret* for de respektive felter på fakturaerne i stikprøven, dvs. de default-værdier Læs-Ind-bureauerne har afløftet. Det er disse default-værdier, der ligger til grund for denne rapport's kvantitative resultater.

Tabel 4.1: Oversigt over anvendte default-værdier

Felt navn:	Bureau 1	Bureau 2
Fakturanummer	Null	Null
EAN-lokationsnummer		
Betalingsdato	0001-01-01	0001-01-01
Udsteder SE/CVR-nummer	n/a0	CVRO
Fakturatotal		
OCR: Art	Null	Null
OCR: Betalings-id	0	0
OCR: PBS	0	Null
OCR: Konto	Null	Blank
Ordrenr.	n/a	Null
Reference	n/a	Null
Fakturadato	1753-01-01	1753-01-01
Fakturanote	Blank	Blank
Bankregistreringsnummer	Null	Null
Fakturatype	PIP	PIP
Valutakode	DKK	DKK

EAN-nummer og fakturatotal har ikke nogen default-værdi, da denne fremgår af samtlige behandlede fakturaer, hvorfor default-værdien ikke kendes. Det er også værd at bemærke de to felter fakturatype og valutakode, hvor praksis er at påføre hhv. PIP eller DKK, medmindre andet er anført.

De default-værdier, der er markeret med fed, er forskellige hos de to bureauer, selv om de er tilknyttet det samme felt. Hvorvidt denne forskel skyldes, at Videnskabsministeriet som følge af bekendtgørelsen har anvist andre default-værdier, eller at Læs-Ind-bureauerne ikke følger bekendtgørelsen, vides ikke. Anvendelse af default-værdier forskellig fra bekendtgørelsen kan i værste fald have den konsekvens, at modtageren af en OIOXML-fil ikke vil være i stand til at afkode den korrekt.

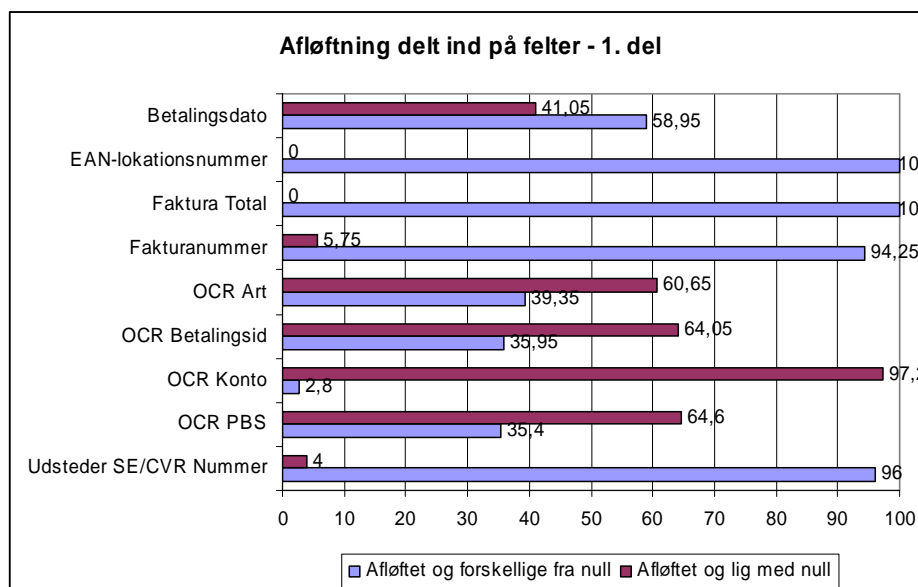
Alle de undersøgte felter forekommer i stikprøven – der er således INGEN af de undersøgte felter, der IKKE afløftes.

Analyse og resultater af den kvantitative analyse

I nedenstående figurer er angivet resultaterne af den kvantitative analyse, det vil sige opgørelsen af, hvorvidt informationerne optræder og er afløftet med default eller en anden værdi. Figurerne er inddelt på Læs-Ind-bureau, og det er valgt at dele felterne fra hvert Læs-Ind-bureau op i to figurer, for at gøre det mere læsevenligt. Opdelingen er foretaget, så de felter, der efterfølgende foretages kvalitativ validering på, kommer først.

4.1 Informationer, der valideres både maskinelt og kvalitativt

Frekvensen af forekomster af default-værdier i de informationer, der *også valideres kvalitativt*, er vist i figuren nedenfor. Disse værdier er vurderet at være det væsentligste grundlag for, at en betaling kan gennemføres og valideres senere kvalitativt. Det er på forhånd forventeligt, at disse data som oftest er afløftet og forskellig fra default.



Generelt er der gennemsnitlig en meget høj afløftningsprocent af værdier forskellig fra default i denne gruppe af felter. En default-værdi kan være afløftet, enten fordi scanningen ikke har fundet den, eller fordi værdien i fakturatotalen ikke kunne bestemmes. For OCR-feltet er sammenhængen mere kompleks og er beskrevet senere i afsnittet.

EAN-nummer

EAN-nummeret er altid afløftet med en værdi forskellig fra default, da en faktura ikke kan sendes fra scanningsbureauerne uden EAN-nummer, jf. kravspecifikationen til Læs-Ind-bureauer, hvori der skrives:

...Læs-Ind-bureauer påfører OIOXML-Læs-Ind-meddelelser, der dannes ud fra fakturaer modtaget uden EAN-nummer, et default-EAN-nummer. Dette findes ved opslag i en database, der stilles til rådighed af Økonomistyrelsen...

Fakturatotal

Fakturatotal er altid afløftet, her er forekomster af default-værdien 0 et udtryk for, at scanningen enten ikke fandt en fakturatotal, eller at fakturatotalen har værdien "0".

Udsteder SE/CVR-nummer og Fakturanummer

Felterne Udsteder SE/CVR-nummer og Fakturanummer har en høj afløftningsprocent på 96 % afløftede værdier forskellig fra default. Dette synes logisk, da alle disse felter kan betragtes som basale for en faktura, hvorfor de i langt de fleste tilfælde vil optræde på fakturaen og derfor alene er et spørgsmål om succesfuld afløftning.

Betalingsdato

Betalingsdato har en afløftning af værdier forskellige fra default på 59 % og med default-værdier i 41 % af tilfældene. Følgende er fra kravspecifikationen til Læs-Ind-bureauerne vedr. betalingsdato:

...Når en betalingsdato ikke er direkte anført på en regning, men derimod blot angivet ved en regel (eksempelvis løbende måned + 30 dage el.), skal følgende betalingsdato anføres i Payment Means-klassen:
<com:PaymentDueDate>1753-01-01</com:PaymentDueDate>³

Værdien 1753-01-01 er, som beskrevet under forudsætninger til dette afsnit, default-værdien for feltet, hvilket betyder, at alle datoer *fastsat som en regel* vil optræde som en default-værdi, hvilket sandsynligvis vil udgøre en større andel af de 41 % afløftede default-værdier. Dette behandles yderligere i den kvalitative analyse.

OCR-felter

I forhold til alle OCR-felterne skal det bemærkes, at disse er gensidige afhængige, forstået på den måde, at de alle er en del af det, der kendes som en OCR-linje. En OCR-linje indledes af en OCR-art, som fortæller, hvilken type betaling der er tale om. Ved betalinger af typen 01 og 04 skal der. jf. vejledningen⁴, tilknyttes et OCR-betalings-id og en OCR-konto, hvorimod der ved betalinger af typen 71, 73 og 75 skal tilknyttes OCR-betalings-id og OCR-PBS-kreditornummer. Vejledningen kan på dette punkt opfattes tvetydigt. Ved kortart 01 og 73 optræder OCR-betalings-id normalt ikke.

Da forekomsten af OCR-felter både er betinget af, at scanningen har scannet en OCR-linje, og at scanningsprogrammet efterfølgende har fortolket afhængigheder mellem felterne, er det vanskeligt at konkludere på tallene fra den kvantitative analyse.

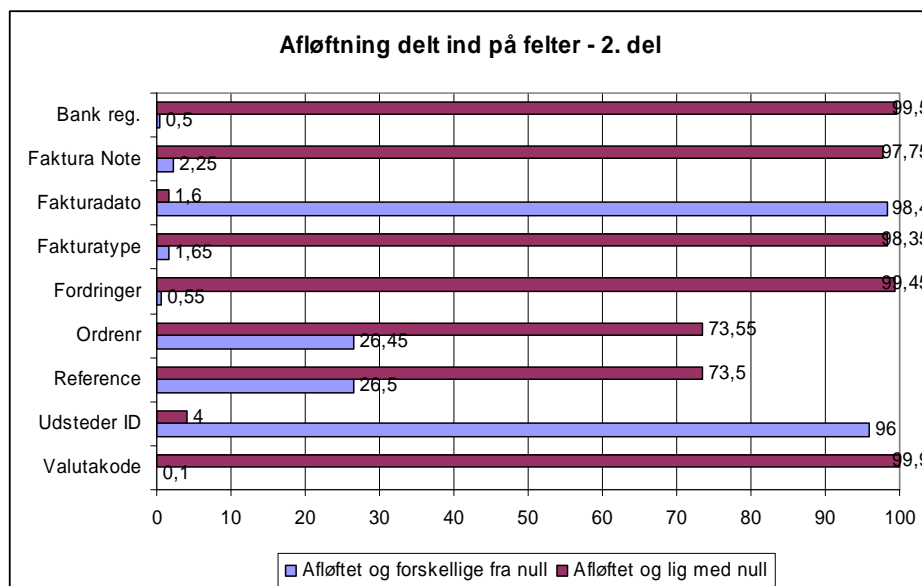
³ Fakturadato er default "1753-01-01" i populationen, og dette antages at være korrekt..

⁴

<http://www.oio.dk/XML/standardisering/eHandel/materialer/OIOXMLFaktura/laesind>, oplyst af Mikkel Hippe Bruun, IT og Telestyrelsen.

4.2 Informationer, der alene valideres maskinelt

Nedenfor er illustreret afløftning af default-værdier for den gruppe informationer, der *alene* er valideret maskinelt. Disse vurderes at være mindre væsentlige for, om en betaling kan gennemføres, og der forventes derfor på forhånd en lav grad af afløftning.



Her en meget lav afløftning af værdier forskellige fra default i denne gruppe af felter. Forklaringen kendes ikke, men baseret på vurderinger fra fokus-gruppeinterviewet i forbindelse med den kvalitative undersøgelse skyldes det som ventet, at informationerne sjældent findes på fakturaerne.

Valutakode

Valutakode er speciel ved, at DKK både udgør den hyppigste forekomst af valutakode og samtidig er default-værdi, fordi denne valutakode vælges, hvis ingen anden forekommer. Denne dualitet gør det umuligt ud fra den maskinelle validering at afgøre, om DKK optræder, fordi den entydigt er identificeret, eller fordi valutakoden ikke fremgår af fakturaen.

Da valutakode forskellige fra DKK desuden optræder meget sjældent i stikprøven, vurderes det, at det er vanskeligt at konkludere på afløftning af valutakode.

Fakturatype

Den samme fortolkning af default-værdi ved valutakode kan anvendes ved fakturatype, hvor default-værdien er PIP samtidigt med, at PIP er den hyppigst forekomne fakturatype. PIP er et udtryk for en almindelig faktura, mens PCP, som er den anden værdi, fakturatype kan antage, er at sidestille med en kreditnota. Dualiteten gør det det umuligt at afgøre ud fra den maskinelle validering, om fakturatypen PIP optræder, fordi den entydigt er identificeret, eller fordi typen er vanskelig at afgøre ud fra fakturaen.

Overdragelser til tredjemand

Felterne fakturanote, bankregistreringsnummer og fordringer har indbyrdes afhængigheder, hvor feltet "fordringer" er en afledt variabel, konstrueret til denne analyse og dermed ikke et felt, der optræder i OIOXML-filen. Fordringer er afledt af feltet *fakturanote* ved en søgning på teksten, som bør optræde ved fordringer overdraget til tredjemand, jf. kravspecifikationen til Læs-Ind-bureauerne:

...Følgende tekst skal anføres i Faktura-Note-feltet:

"Fordringen er overdraget til tredjemand. Betaling med frigørende virkning af fakturabeløbet kan alene ske til fordringshaver."

Bankregistreringsnummeret er en indikation af, om der er tale om en fordring overdraget til tredjemand, idet informationen kun optræder ved sådanne, såfremt der er tale om betalingsformen kontooverførsel. Dermed er felterne fordring og bankreg. begge indikatorer for en fordring overdraget til tredjemand.

Med udgangspunkt i disse afhængigheder og relationer mellem de tre felter fakturanote, bankreg. og fordringer, er der en afløftning på 2,25 % af fakturanotefeltet, hvoraf 0,55 % er registreret som værende fordringer til tredjemand. Ved visuel validering af de data er konstateret, at i de forekomster af fakturaer, hvor der er tale om fordringer til tredjemand, er bankreg. afløftet med værdi forskellig fra default, når der er tale om en kontooverførsel. Dette styrker sandsynligheden for, at fordringerne er korrekt afløftet.

Forskellen mellem de 2,25 % ved fakturanote og 0,55 % ved fordringer er et udtryk for, at fakturanotefeltet bliver brugt til andet, bl.a. til beskeder om, hvorvidt der er anvendt valid eller invalid digital signatur ved e-mail.

Reference og ordrenummer

De to felter reference og ordrenummer har meget sammenfaldne karakteristika ved, at de begge er afløftet med omkring 26,5 %. Afløftning er et udtryk for, at scanningen har identificeret værdier, der er godtaget som værende korrekte. Dette skal i forhold til disse to felter tages med et vist forbehold pga. den meget store diversifikation, der kan være i angivelsen af fx ordrenummer, hvilket gør dem svære at identificere. Ved at se i data for de to felter ved begge bureauer, kan det også konstateres, at hovedparten af afløftninger synes korrekt.

Fakturadato

Fakturadato har som det eneste felt i denne del en høj afløftning på 98,4 %. På grund af de meget ensformige måder, som fakturadato fremstilles på fakturaerne, synes det sandsynligt, at der reelt er afløftet en dato i samtlige tilfælde. Det bekræftes ved at se i data, at der i samtlige tilfælde er tale om en afløftet dato, men i kraft af, at der på en faktura kan optræde en række forskellige datoer, er der mulighed for, at scanningen fx fejlagtigt har taget betalingsdato som værende fakturadato.

Konklusioner

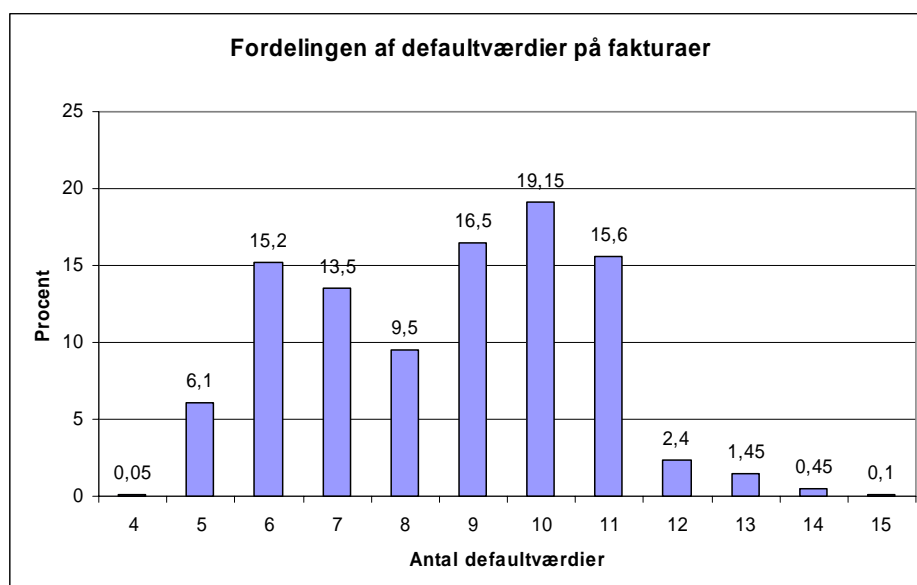
Antagelsen forud for den kvantitative analyse af data har været, at de væsentligste informationer – svarende til informationerne:

- Fakturatotal
- SE/CVR-nummer
- OCR-linje
- Betalingsdato
- Fakturanummer
- EAN-lokationsnummer

som oftest var afløftet, mens de øvrige mindre vigtige obligatoriske informationer ofte er afløftet med default. Denne antagelse er bekræftet i den kvantitative undersøgelse.

I forhold til konklusionerne på den kvantitative analyse er det vigtigt at pointere, at den udelukkende kan registrere mulige problemstillinger og ikke entydigt finde fejlkilden.

Nedenfor er vist det samlede antal af default-værdier pr. faktura sat ind i en fordeling for at give overblik over brugen af default-værdier:



Det er, som omtalt i analysen ovenfor, meget vanskeligt at drage nogle sikre konklusioner på basis af default-værdierne, men de er en indikation for, hvor meget eller lidt der afløftes for hver enkelt faktura.

Der er ingen med under 3 afløftede default-værdier, hvilket skal ses i lyset af, at der ved valutakode og fakturatype i langt de fleste tilfælde (formentlig korrekt) anvendes en default-værdi, fordi disse to default-værdier også repræsenterer information om fakturaen i modsætning til de andre default-værdier. Der er meget få fakturaer med 12 eller flere default-værdier. Dette fortæller, at der næsten altid afløftes 5 værdier og i hvert fald altid 3 værdier. Af de tre værdier der altid registreres, er fakturatotal og EAN-lokationsnummer – som vist i analysen – altid to sikre afløftninger af data forskellig fra default.

5. Kvalitativ analyse og resultater

Den kvalitative analyse bygger på en manuel og visuel verifikation af de seks informationer og foretaget ved, at en gruppe personer har sammenholdt de indscannede fakturaer med værdierne i OIOXML-strengen. En sådan manuel verifikation kan og vil imidlertid være behæftet med fejl, hvorfor der er foretaget en selvkontrol af processen.

Forudsætninger for den kvalitative analyse

Basisstrukturen for tolkningen af den kvalitative analyse kan noget firkantet fremstilles som nedenfor. Den helt præcise fortolkning af korrekt, forkert og ubestemmeligt er dog afhængig af feltet, der analyseres, idet der kan være særlige forhold som gør tolkningen anderledes. De specifikke forhold vil blive taget op løbende, bl.a. i forhold til OCR-linjen.

Korrekt	Værdi forekommer og er afløftet rigtigt / Værdi optræder ikke og er derfor korrekt afløftet med default-værdi
Forkert	Værdi forekommer og er afløftet, men er scannet forkert / Værdi forekommer, men er ikke afløftet
Ubestemmeligt	Værdi forekommer og er forsøgt afløftet, men det er ikke muligt at vurdere, om værdien er rigtig eller forkert / Værdi forekommer og er sat til default-værdi, fordi faktura er ubestemmelig

Ud over en gennemgang og fortolkning af resultaterne for det respektive felt, vil der blive suppleret med kommentarer fra fokusgruppeinterviewet, hvor disse kan bidrage eller uddybe til forståelsen.

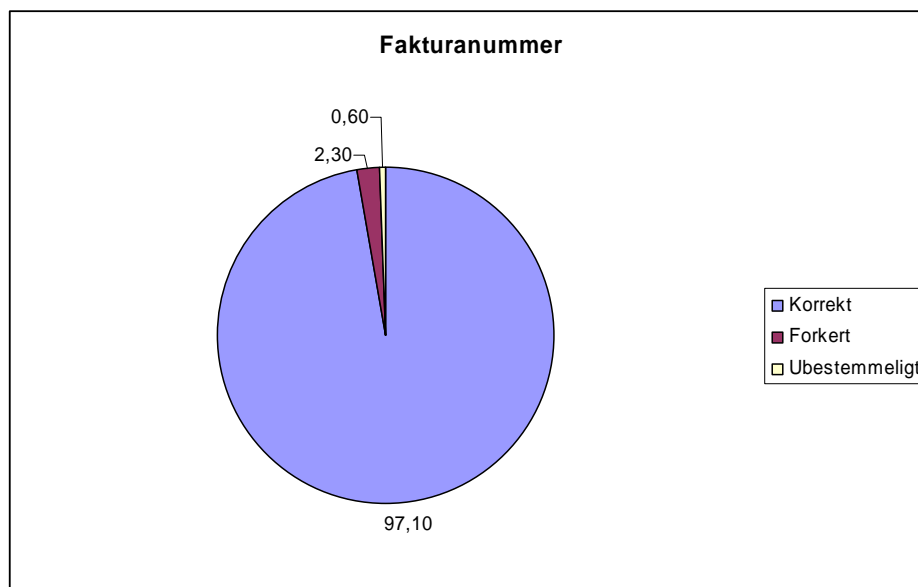
Resultat af egenkontrol

Der blev foretaget egenkontrol på i alt 200 af de gennemførte manuelle og visuelle valideringer, svarende til 10 % af den samlede stikprøve, fordelt med 100 på hvert bureau. Denne kontrol viste, at processen generelt har været af meget høj kvalitet og med få fejl. De konstaterede fejl er næsten alene fundet i forbindelse med håndteringen af OCR-linjen. Dette skyldes, at der i begyndelsen af valideringen var tvivl om betydningen af vejledningen om afløftning af felter.

Analyse og resultater af den kvalitative analyse

I dette afsnit vurderes afløftning og kvalitet i de enkelte informationer.

Fakturanummer

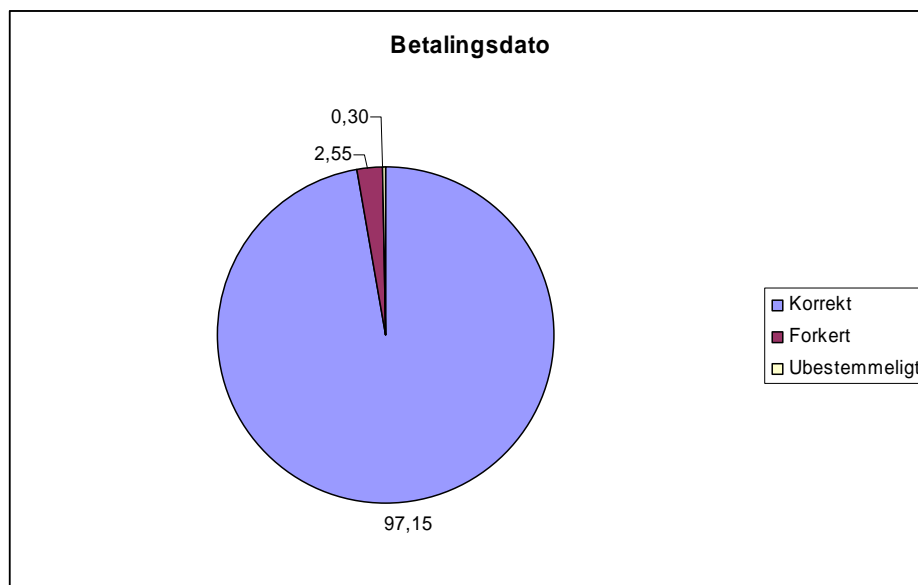


- Generelt er aflæsningen af fakturanummer af høj kvalitet, idet scanningen korrekt har kunnet aflæse feltet i mindst 97 % af tilfældene. Kun i få procent af tilfældene har den kvalitative vurdering været, at feltet på fakturaen er ubestemmeligt, dvs. tvetydigt eller ikke læseligt med øjet. Som oftest skyldes det, at et felt er ubestemmeligt på fakturaen, at det ikke entydigt fremgår, at der er tale om et fakturanummer, og dette kan forveksles med andre referencer på fakturaen.
- De tilfælde, hvor scanningen *ikke* mener, der findes et fakturanummer på tiff-filen, og hvor der dermed er afløftet en default-værdi, er tilsvarende af høj kvalitet. Kun i omkring 2,3 % af tilfældene er afløftningen forkert, dvs. indsat en default-værdi, hvor det menneskelige øje entydigt har kunnet identificere et fakturanummer på tiff-filen eller indsat en default-værdi, hvor en værdi fremgår.

Kommentar fra fokusgruppe:

- Fakturanummeret er generelt svært at finde på fakturaen. Det er ofte ikke angivet efter en ledetekst, der entydigt angiver, at der er tale om et fakturanummer.
- I tilfælde, hvor der findes et fakturanummer på tiff-filen, og hvor scanningen afløfter værdien forkert, skyldes det ofte, at scanningen aflæser tegn, der ikke findes i fakturanummer, fx punktum eller bindestreg.

Betalingsdato

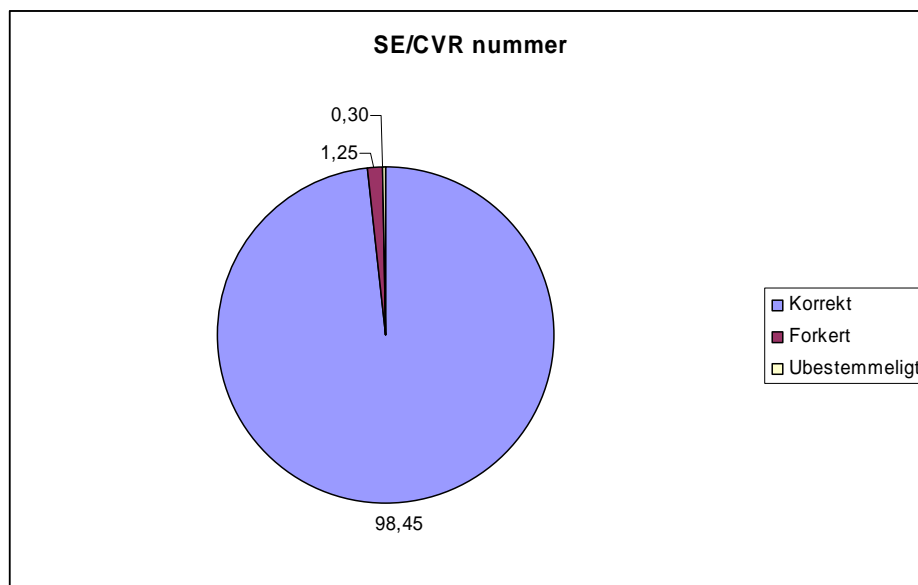


- Afløftningen af betalingsdato er af høj kvalitet med over 97 % korrekte afløftninger af betalingsdatoen. Bag disse tal viser det sig, at betalingsdatoen ikke optræder på fakturaen i omkring 40 % af tilfældene, hvorfor den korrekte afløftning i næsten halvdelen af tilfældene skyldes, at datoen ikke er påført fakturaen eller påført ved hjælp af en beregningsregel. Kun i under 3 % af tilfældene er datoen ikke afløftet, eller en forkert dato er afløftet.

Kommentar fra fokusgruppe:

- Generelt meget imponerende aflæsning, når mangfoldigheden af dato-fremstillinger tages i betragtning. Disse kan have mange formater og være indlejret i tekst.
- Scanningen kan have svært ved at aflæse dato, når den er skrevet ind i tekst.
- Kun meget sjældent tager scanningen fejl og aflæser fakturadato.

Udsteder CVR/SE

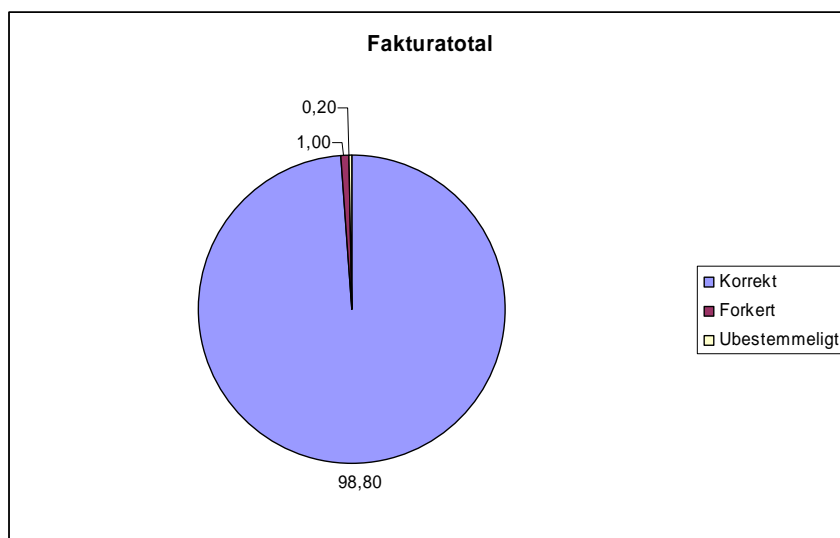


- Udsteder CVR afløftes på over 98 % af fakturaerne. Disse afløftninger er tæt på 100 % korrekte.
- Ved 1,25 % af fakturaerne afløftes CVR med en forkert værdi. Årsagen kendes ikke, men kan have årsag i fakturaernes mangfoldighed. Selv en meget lav fejlprocent vil give anledning til et større administrativt arbejde og efterfølgende krediteringer for de involverede virksomheder, når fakturaen sendes til en forkert modtager. Det er især Læs-Indbureauerne, der belastes af administrativt arbejde med efterfølgende krediteringer og fornyet fremsendelse af fakturaer.

Kommentar fra fokusgruppe:

- Hvis der fremgår flere CVR/SE-numre på fakturaer er de oftest identiske.
- Hvis såvel myndigheds som fakturaudsteders CVR/SE-nummer optræder på fakturaen, tager scanningen aldrig fejl. Scanningen afløfter *altid udsteder CVR/SE-nummer*.
- Hvis CVR/SE-nummer ikke optræder, er der oftest tale om simple "bagerkvitteringer" eller lignende.

Fakturatotal

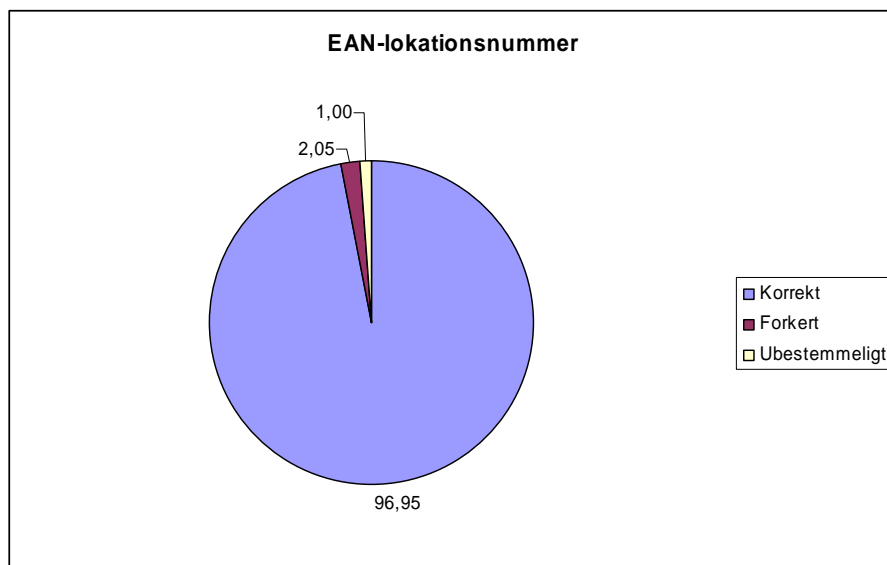


- Fakturatotal afløftes altid og med høj kvalitet, idet 98,80 % af tilfældene er afløftet korrekt, svarende til 988 ud af 1000 fakturaer.

Kommentar fra fokusgruppe:

- Når scanningen aflæser en forkert værdi, er det oftest ved komplekse beregninger af fakturatotal, der strækker sig over mere end én side.
- Scanningen synes at afløfte det højeste tal på fakturaen, hvilket betyder, at der ved mellemregninger (krediteringer) kan afløftes en højere værdi end fakturatotal. Her optræder en subtotal, fordi et udestående ikke modregnes.
- Hvis det entydigt er angivet på en faktura, at der er tale om en kreditnota, er dette vurderet at være scannet korrekt. Der kan være angivet både "-" og "+" som fortegn på en kreditnota.

EAN-nummer



- EAN-nummeret afløftes altid og med høj kvalitet, idet 96,95 % afløftes korrekt.
- De fakturaer, hvor EAN nummeret er vurderet 'Forkert' eller 'Ubestemmeligt' er ofte fakturaer, hvor EAN nummeret på *fakturaen* er omsat til et EAN nummer, der er slået op i en speciel default database, fx fordi EAN nummeret på fakturaen ikke er i brug. Disse forekomster er altså ikke nødvendigvis forkerte i forhold til, om modtageren kan modtage fakturaen – men EAN numrene på faktura og i OIOXML fil stemmer ikke overens. Erfaringsmæssigt skal der for ca. 2 % af fakturaerne i populationen hentes et EAN nummer i default databasen, hvilket stemmer godt overens med de ca. 2 % 'Forkerte' EAN numre.

Kommentar fra fokusgruppe:

- Placering og fremstilling af EAN-nummer er meget forskellig.
- Der er ved 20-30 % af fakturaerne påført EAN-nummer med håndskrift. Det afløftede EAN-nummer og det håndskrevne er altid identiske, hvilket kan være et udtryk for en manuel arbejdsgang med påføring af EAN-nummer hos bureauerne.
- Når et forkert EAN-nummer er afløftet, er der sjældent kun ét ciffer til forskel på den påtrykte og afløftede værdi. Som oftest er 3-5 cifre forskellige. Årsagen kendes ikke.

OCR-linjen

Der er kun beregnet tal for den mængde af fakturaer, hvor der er konstateret en OCR-linje. Populationen er derfor mindre end 2000.

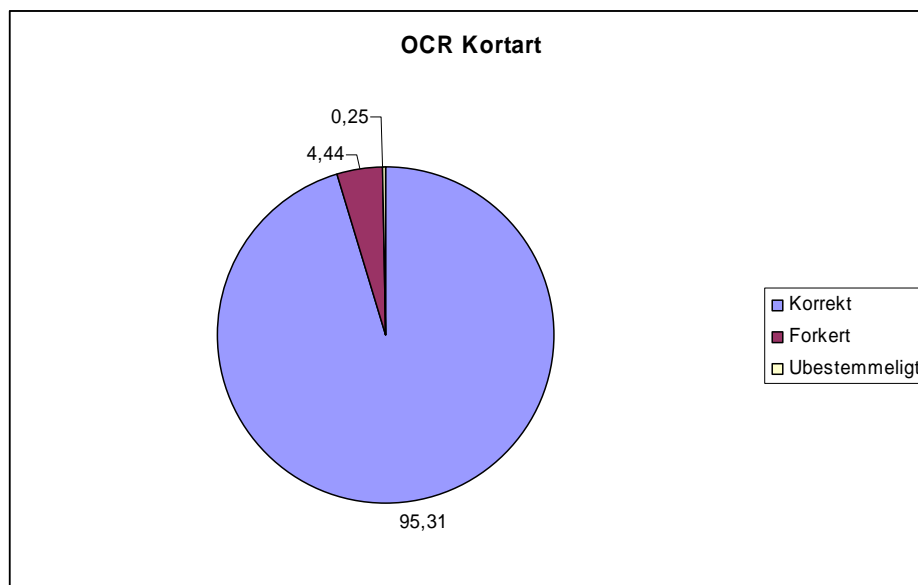
OCR-art er blevet brugt som indikator for, om der fremgår en OCR-linje eller ej på fakturaen.

I de tilfælde, hvor et af felterne, OCR-konto eller PBS-kreditnummer, i OCR-linjen ikke skal anvendes, jf. vejledningen, er det pågældende felt af tekniske grunde registreret som *ubestemmeligt* under den visuelle vurdering, dvs. her fortolket som "uinteressant". I forhold til tolkningen af disse to felter betyder det, at der som følge vil være en stor forekomst af *ubestemmeligt*, som i stedet skal tolkes som, at bureauet korrekt ikke har angivet en værdi.

Forudsætninger

- 9 ud af 66, hvor det er angivet som korrekt i stedet for ubestemmeligt ved OCR-linje.
- Under den kvalitative vurdering fremgik det, at scanningen som oftest var i stand til at identificere indholdet af en OCR-linje, selv om data ikke var angivet som en standard OCR-linje. Fx kunne data være indlejret i tekst midt på siden i form af betalingsoplysninger. Derfor er også forekomster af OCR-oplysninger, der *ikke* er angivet som en scanningsbar standard OCR-linje medtaget som værende forekommende på fakturaen. Det betyder, at i de (få) tilfælde, hvor scanneren alligevel har overset OCR-oplysninger, men hvor den visuelle validering entydigt har identificeret oplysninger, svarende til indholdet af en OCR-linje, er dette registreret som "forkert".

OCR – Art

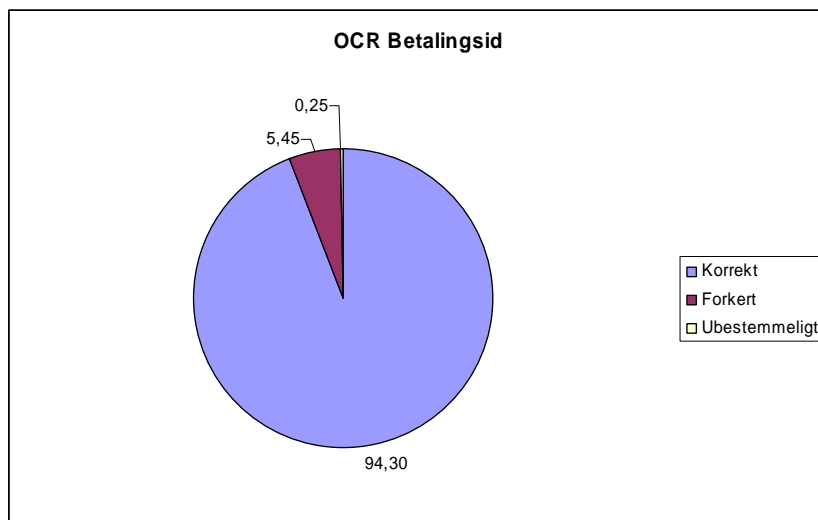


- OCR-art optræder ofte og er afløftet korrekt i 95,31 % af tilfældene. I 4,4 % af tilfældene afløftes OCR-art forkert, selv om informationerne visuelt kan identificeres på fakturaen. Dette skyldes forhold som nævnt ovenfor – det vil sige, at oplysningerne forekommer, men er ikke standardiseret (og kan da næppe tolkes som værende fejl) – eller kan skyldes forhold beskrevet i kommentarerne fra fokusgruppen herunder.

Kommentar fra fokusgruppe:

- Hvis OCR findes på den sidste side af flere sider med fx et tomt girokort fornedet, afløftes den oftest ikke.
- Scanningen kan have overset OCR-linje, hvis den først optræder på side 2 eller senere og resten af data i øvrigt har fremgået af side 1.
- Scanner overser kun OCR-linje, hvis den er blandet ind i tekst.

OCR – Betalings-id

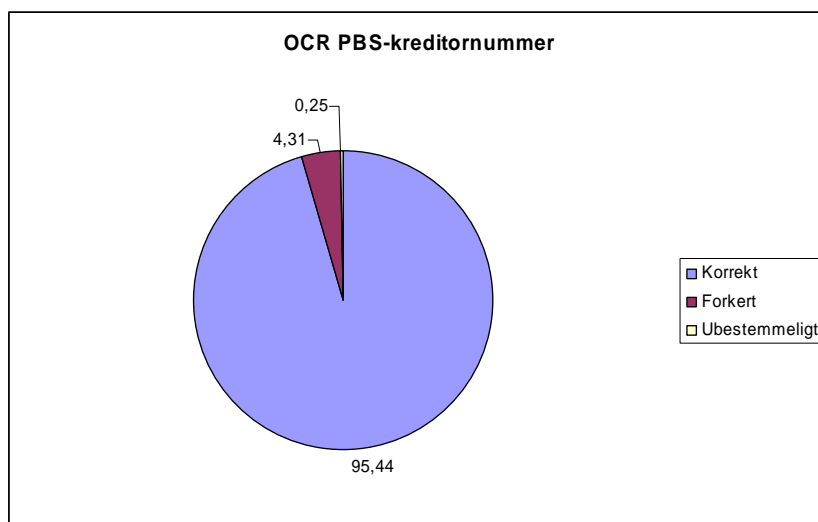


- Betalings-id afløftes korrekt i 94,3 % af forekomsterne. Den høje korrekte afløftning skyldes antageligt, at den, jf. vejledningen, ikke *skal* afløftes for visse kortarter.
- Af de afløftede værdier er 5,45 % forkert, hvilket som regel skyldes fejl i ét ciffer.
- Når der er afløftet forkert værdi, skyldes det oftest, at OCR strengen er indlejret i tekst. Ofte afløftes betalings-id ikke, selv om den findes. Vejledningen er tvetydig for så vidt angår krav om afløftning af betalings-id ved kortart forskellig fra 71,73 og 75 (FIK).

Kommentar fra fokusgruppe:

- Der er store variationer i, hvornår en betalings-id afløftes. Der er ikke noget entydigt billede, når kortart er forskellig fra 71, 73 og 75 (FIK).
- Betalings-id afløftes for kortart 71, 73 og 75 (FIK).

OCR – PBS-kreditornummer

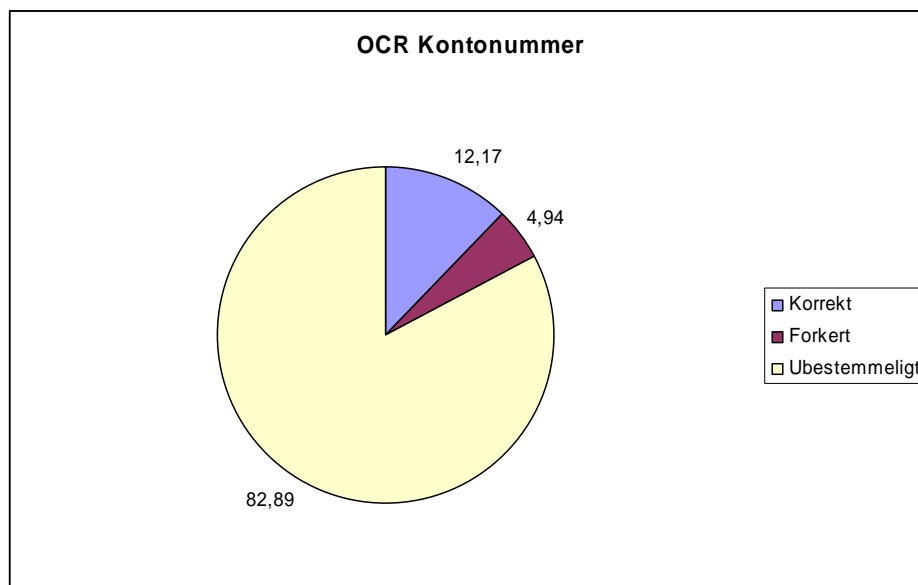


- Kvaliteten i afløftningen af PBS-kreditornummer minder meget om betalings-id, og det samme gælder de iagttagelser, der kan gøres på baggrund heraf.
- Årsagerne vurderes kvalitativt at være de samme som ved betalings-id.

Kommentar fra fokusgruppe:

- Kendetegn for afløftning af PBS-kreditornummer er de samme som for betalings-id.

OCR – Konto



- OCR-konto afløftes korrekt i 12,17 % af tilfældene, mens det i 4,94 % af tilfældene er forkert.
- At 83 % af forekomsterne er ubestemmelige skyldes alene, at udfyldelse af feltet ikke er påkrævet, jf. vejledningen (kortart 71,73, 75).

Kommentar fra fokusgruppe:

- OCR-konto indsættes i PBS-kreditorfelt i stedet for ved 04-betalingstyper i nogle tilfælde. Dette er vurderet som forkert, og det udgør en stor del af de forkerte afløftninger⁵.

⁵ Denne fejl ER nu rettet

6. Konklusioner og anbefalinger

Kvalitativt vurderes på basis af data fra undersøgelsen og det kvalitative fokusgruppeinterview, at en meget stor andel af de informationer, der overhovedet kan afløftes fra fakturaerne, bliver afløftet, og at scanningen er stærk til at afløfte korrekte data, når den meget store mangfoldighed af informationer og fakturaer tages i betragtning.

Når data ikke afløftes, og der anvendes default-værdier, skyldes det oftest, at informationerne ikke findes på fakturaerne. Anvendelse af default-værdier på flere specifikke felter blandt de kvantitativt undersøgte er markant.

Generelt er fakturamaterialet meget forskelligt – fra håndskrevne "bagerkvitteringer" til velstrukturerede entydige fakturaer. De helt eller delvist håndskrevne fakturaer udgør ikke et særligt problem i forhold til kvaliteten af afløftningerne, hvilket antageligt skyldes at de gennemgår manuel afkodning og De håndskrevne fakturer vurderes at udgøre ca. 5 % af fakturaerne.

Der er i undersøgelsen ikke forekommet fakturaer, der er helt ulæselige af det menneskelige øje på baggrund af tiff-filen. I det omfang, fakturaer ikke kan læses *fuldstændigt* skyldes det "snavs", som gør det vanskeligt at aflæses enkeltstående oplysninger, eller vanskeligheder med at skelne tal (fx 3 fra 8) i talkombinationer, specielt i OCR strengen. Det skal dog her bemærkes, at opløsning og kvalitet af grafik på den pc, der anvendes, kan være betydende for kvaliteten af læsbarheden.

Følgende er områder, hvor nærmere analyse kan give anledning til identifikation af indsatser, som kan højne kvaliteten endnu mere:

1. Vejledning og specifikation af afløftning af data vedr. OCR-linje kan opfattes tvetydigt, og dette vil muligvis kunne give anledning til, at modtager har vanskeligt ved at afkode data. Det gælder alle data i OCR-linjer for andre korttyper end FIK. Det er fx vanskeligt at afgøre præcist, hvornår meddelelses-id, PBS-kreditnummer og kontonummer skal afløftes.
2. Afløftning af OCR-linjers oplysninger om kreditornummer eller kontonummer er behæftet med fejl i visse situation. Algoritme for afkodning af disse informationer kan gennemgås mht. fejlretning.
3. Scanning af OCR-linje/-oplysninger, når en faktura strækker sig over flere sider, synes at være forbundet med vanskeligheder og forårsager fejl i få tilfælde. Algoritme for scanningen af OCR-linje kan evt. gennemgås med henblik på at fortsætte søgning ud over første side.
4. Scanning af fakturatotal i tilfælde af komplekse beregninger over flere sider kan være årsager til afløftning af en forkert fakturaværdi i få tilfælde. Algoritme for scanning af fakturatotal kan gennemgås med henblik på at sikre, at det ikke nødvendigvis er det største tal, der afløftes – fx registreres mellemregning i form af en kreditering eller modpostering i enkelte tilfælde.
5. For en række felter anvendes forskellige default-værdier, hvilket kan give anledning til, at modtager kan have svært ved at afkode data.