NUUG
CSS Mailboks 70
Middelthunsgate 25B
0368 OSLO

Tlf: +47 85234290
Kontonr: 7050.20.09921
Orgnr: NO 979 629 079
E-post: sekretariat@nuug.no
Hjemmeside: http://www.nuug.no/

Oslo, 2017-04-28

Riksarkivaren
Pb. 4013 Ullevål Stadion
0806 OSLO
post@arkivverket.no

Deres ref.: 2016/9840 HELHJO Vår ref.: 2017-NOARK-PR

Høringsuttalelse til Riksarkivarens forskrift

Vi viser til høring sendt ut 2017-02-17, og tillater oss å sende inn noen innspill til Riksarkivarens arbeid med ny forskrift om utfyllende tekniske og arkivfaglige bestemmelser om behandling av offentlige arkiver (riksarkivarens forskrift)[1]. Innspillene følger samme rekkefølge som forskriften, så langt det lot seg gjøre.

Kommaseparert format i § 5-12 bør klargjøres

Forskriftens § 5-12 (Tekstfilformater i arkivuttrekk) punkt b) omtaler tegnseparert format (kommaseparert format) uten å henvise til en fri og åpen standard som beskriver hva slags format en mener i detalj. Kommaseparerte filer er beskrevet av IETF RFC 4180[2]. For å sikre en entydig beskrivelse av formatet foreslår vi at forskriften henviser til denne formatbeskrivelsen. Slik forskriften står i dag mangler for eksempel forskriften krav om beskrivelse av hvordan poster som inneholder feltskilletegn, linjeskift, anførselstegn og så videre skal håndteres, mens dette er klart beskrevet i RFC 4180.

Oversikten over godkjente dokumentformater ved innlevering bør klargjøres

Punkt § 5-16 1a) nevner formatet «TXT» uten nærmere beskrivelse, uten å henvise til § 5-11 som lister opp godkjente tegnsett, og uten å forklare hvordan en skal vite hvilket tegnsett som er brukt i «TXT»-formatet. Hvis poenget er at §§ 5-12 og 5-12 er ment å beskrive «TXT»-formatet nærmere bør det legges inn en kryssreferanse til §§ 5-11 og 5-12 i punkt 1a. Det er ikke klart fra forskriften konkret hva slag format TXT er. Beskrivelsen «ren tekst» kan være så mangt, og både HTML og XML består jo av en «ren» tekstlig beskrivelse. Er det ment å være tilsvarende text/plain definert i IETF RFC 2046[3]? Der beskrives den som

> «"text/plain", which is a generic subtype for plain text. Plain text does not provide for or allow formatting commands, font attribute specifications, processing instructions, interpretation directives, or content markup. Plain text is seen simply as a linear sequence of characters, possibly interrupted by line breaks or page breaks. Plain text may allow the stacking of several characters in the same position in the text. Plain text in scripts like Arabic and Hebrew may also include facilitites that allow the arbitrary mixing of text segments with opposite writing directions.».

Vi tror det er lurt at «TXT» beskrives nærmere, og at det henvises til klargjørende spesifikasjoner som for eksempel text/plain i IETF RFC 2046 for å forklare hva som menes.

Punkt § 5-16 1b) omtaler TIFF som «tekst med objekter». Bildeformatet TIFF anses normalt ikke som et

[1] http://lovdata.no/dokument/SF/forskrift/1999-12-01-1566

[2] https://tools.ietf.org/html/rfc4180

[3] https://tools.ietf.org/html/rfc2046

tekstformat. Vi antar en her snakker om bilder av papir med tekst. Men TIFF er bilde av tekst, ikke ren tekst og det virker misvisende å likestille den med TXT, XML og PDF/A. PDF/A kan inneholder innskannede dokumenter i bildeformat, men bør vel i slike tilfeller anses å være bilder, ikke tekst. Bilde av tekst, uavhengig av om det er pakket inn som TIFF eller PDF, bør heller legges under punkt 1c.

Punkt § 5-16 1f) nevner «PCM-basert wave» uten å beskrive nærmere hva som menes. Det er dermed uklart hvilken av lydformatene med wave i navnet det refereres til. Hvis det er Waveform Audio File Format definert i Multimedia Programming Interface and Data Specification 1.0 av IBM Corporation og Microsoft Corporation i august 1991 det gjelder, så er det beste å referere til en klar og offentlig spesifikasjon, helst en fri og åpen standard, som beskriver formatet.

Punkt § 5-16 1h) nevner HTML, men nevner ikke hvordan eksterne referanser i HTML (for eksempel skrifttyper, JavaScript-kode, bilder og video) skal håndteres for å kunne gjenskape en web-side. § 5-19 sier derimot at formatet ikke skal brukes. Kanskje HTML ikke burde være på listen over godkjente dokumentformater, eller i det minste ha en henvisning til § 5-19?

Internett-e-post (IETF RFC 5322) bør inn som godkjent dokumentformat

Forskriftens § 5-16 med oversikt over godkjente dokumentformater ved innlevering bør inneholde Internett-e-post. Det meste av forvaltningens korrespondanse gjennomføres i dag på epost, og det bør være mulig å lagre korrespondansen i originalformatet i stedet for å konvertere til for eksempel PDF/A. Det gjør det også mulig å direkte svare på epost i arkivet med et epostprogram, ved for eksempel å hente ut eposten fra arkivet og gjøre den tilgjengelig for epostprogrammet.

Vi foreslår at det legges inn et nytt punkt k) under punktet om websider i § 5-16. Det nye punktet kan for eksempel lyde slik:

> k) For Internett-e-post aksepteres følgende:
> RFC822 som spesifisert i IETF RFC 5322. Eventuelle vedlegg lagres i tillegg som separate vedleggsdokumenter i aksepterte dokumentformater der dette er mulig.

Vi foreslår å bruke «RFC822» som formatnavn for slik e-post, da IETF RFC 822[4] (1982-08-13) er opprinnelig beskrivelse av Internett-e-post, siden etterfulgt av IETF RFC 2822[5] (2001-04) og IETF RFC 5322[6] (2008-10), med oppdatert beskrivelse av formatet.

E-post er omtalt i Noark 5 versjon 3.1[7] på sidene 193 til 207. Side 193 dokumenterer det som for oss ser ut til å være en misforståelse:

> «Selv om RFC2822 definerer syntaksen for e-posttransaksjoner, er den ingen standard som definerer dataformatet som skal bli brukt når e-posttransaksjonen er fanget som dokumenter.»

En kan jo «fange en e-post som et dokument» for lagring ved å ta vare på teksten som utgjør e-posten, dvs. hode og kropp. Det er standard måte å ta vare på e-post, for eksempel i mbox, maildir, mh og en rekke andre kjente måter å ta vare på e-post, og håndteres av ethvert e-postprogram. En e-post lagres tradisjonelt ved å lagre tekstlinjene som utgjør e-posten. Formatet er enkelt, der e-postens hode består av tekstlinjer med feltnavn, kolon og feltverdi, så en blank linje og deretter tekstlinjer som utgjør e-postens kropp. Det er beskrevet i e-postens hode hvordan tekstlinjene i kroppen skal tolkes. Detaljene er beskrevet i detalj i IETF RFC 5322.

---

[4] `https://tools.ietf.org/html/rfc822`

[5] `https://tools.ietf.org/html/rfc2822`

[6] `https://tools.ietf.org/html/rfc5322`

[7] `http://www.arkivverket.no/arkivverket/Offentleg-forvalting/Noark/Noark-5`

Krav 8.1.8 i Noark 5 versjon 3.11 (side 195) lyder som følger:

> Ved arkivering av e-post i Noark 5-løsningen, skal e-posten med eventuelle vedlegg automatisk arkiveres i et enhetlig, samlet format som gjengir både e-posthode og e-postmelding.

Men det står ingenting om hvordan dette skal gjøres, og når epost ikke er på listen over godkjente formater blir det utfordrende.

Vi savner dermed informasjon i forskriften om hvordan Noark 5-løsninger skal ta vare på e-post uten informasjonstap. Det bør være mulig å gjenskape hele den originale e-posten slik den så ut ved mottak hvis en skal kunne sjekke kryptosignaturer (for eksempel der S/MIME (IETF RFC 5751[8]) eller OpenPGP (IETF RFC 4880[9]) er brukt). Det vil også gjøre det mulig å sjekke hvor e-posten kommer fra ved hjelp av felter brukt av Domain Keys Identified Message (DKIM) i e-posten. DKIM er definert i IETF RFC 6376[10].

Formatet for Internett-e-post er som nevnt spesifisert i IETF RFC 5322, og består i grove trekk av et sett med tekstlinjer som e-posthode, en blank linje som skilletegn og så et sett med tekstlinjer som består av e-postens kropp. E-postens kropp kan være strukturert i underdeler i tråd med MIME-formattering og inneholde for eksempel både ren tekst, HTML-sider og vedlegg med formater uten en klar og offentlig spesifikasjon (for eksempel proprietære binærformater). E-posten kan også henvise til eksterne filer (for eksempel bilder i HTML-e-post) som kun er tilgjengelig i en tidsbegrenset periode. Arkivering av e-post på formatet beskrevet i IETF RFC 5322 bør derfor kanskje ha en del begrensninger når det gjelder aksepterte vedlegg, for å sikre at e-post og tilhørende vedlegg kan forstås også i fremtiden, selv når det ufrie programmet som kan tolke slike proprietære binærformater er gått tapt eller ikke lenger lar seg bruke (noe som kan skje hvis programmet er avhengig av nett-tjenester som ikke lenger eksisterer for å fungere), eller nett-tjenesten der eksternt henviste filer befant seg er borte, for eksempel hvis det henvises til en Youtube- eller Facebook-side.

En naturlig begrensning for slike e-postarkiver kunne være å kreve at e-postvedlegg skal være et av de godkjente dokumentformatene omtalt i forskriften. Et spørsmål som må besvares i den forbindelse er hva som skal gjøres med mottatt arkivverdig e-post som inneholder vedlegg på formater som ikke er godkjent for arkivering. En konvertering til andre formater før arkivering vil ødelegge for kontroll av eventuelle kryptosignaturer i e-posten. Det grunnleggende spørsmålet er om arkivet skal ta vare på informasjon det er uklart hvordan skal tolkes eller ikke. Det tryggeste er kanskje å både ta vare på originalinformasjonen og i tillegg kreve at der det er mulig blir det lagret omformede utgaver der det er klart hvordan vedleggene skal tolkes.

Det bør også avklares hvor metadata fra en epost lagres i Noark 5-strukturen, men det hører kanskje ikke hjemme i forskriften. For å automatisk kunne finne hvilke e-poster som «henger sammen» når nye e-poster skal arkiveres er det hensiktsmessig å raskt kunne søke opp e-post ved hjelp av e-postens Message-ID. En e-post inneholder referanser til hvilken e-post den er svar på i e-posthodefeltene In-Reply-To og References. Disse feltene inneholder referanser til tidligere e-post ved hjelp av verdier hentet fra e-posthodefeltet Message-ID. Disse feltverdiene kan dermed brukes til å finne aktuelle mapper å foreslå for å plassere en ny e-post som skal arkiveres. Det vil være hensiktsmessig å standardisere hvilket felt i Noark 5-strukturen dette skal lagres i for å sikre gjenfinning i deponerte arkiver. En god kandidat for Message-ID kan være feltet «filnavn» fra spesifikasjonen for NOARK 5 Tjenestekatalog side 104.

Klargjøre hvordan SMS/MMS skal arkiveres

Det står ingenting i forslaget til forskrift hvordan øyeblikksmeldinger som SMS/MMS skal arkiveres.

---

[8] `https://tools.ietf.org/html/rfc5751`

[9] `https://tools.ietf.org/html/rfc4880`

[10] `https://tools.ietf.org/html/rfc6376`

Lagring av SMS/MMS er en utfordring flere etater har i dag, og det er kanskje naturlig å gi instrukser eller anbefalinger til forvaltningen om hvordan slike meldinger bør langtidslagres. Kan det være en ide å bestemme et eget XML-basert format for lagring av SMS? Eller kanskje det er mer fornuftig å arkivere originalmeldingen slik den ble oversendt til mobilen? SMS-formatet er beskrevet og vedlikeholdes av 3GPP som TS 23.041[11]. MMS-formatet er beskrevet og vedlikeholdes av Open Mobile Alliance som OMA MMS[12]. Kanskje XML-format i MMS-spesifikasjonen kan brukes?

Uansett bør det kanskje nevnes hva slags informasjon som bør registreres for hver SMS? Som et minimum bør nok selve teksten, sendertelefonnummer, mottaker(e), sendetidspunkt og mottakertidspunkt tas vare på.

For foreningen NUUG

Hans-Petter Fjeld
Leder i NUUG

---

[11] `http://www.3gpp.org/DynaReport/23041.htm`

[12] `http://www.openmobilealliance.org/`