DOMSTOLSVERKET
SVERIGES DOMSTOLAR

DATUM
2015-10-26

Atom-insamling

Rättsinformationssystemets användning av Atom-flöden

Om du har några frågor med anledning av detta dokument eller angående rättsinformationsprojektet i stort, kontakta gärna Heléne Lundgren på helene.lundgren@dom.se.

Innehåll

1. Inledning

1.1. Syfte

Här beskrivs hur rättsinformationssystemet hanterar datakällor med hjälp av s.k. Atom-flöden.

1.2. Bakgrund

Informationen i rättsinformationssystemet kommer primärt från andra myndigheter. Både dessa källor och systemets sammanställning (aggregat) av dem ska representeras av Atom-flöden. Systemets aggregerade källa utgör en komplett sammanställning av alla datakällor.

1.3. Definitioner

1.3.1. Generella

Atom
: Ett standardformat i XML för att representera flöden av uppdateringar.

Atom-flöde
: En representation av posters tillväxt i en datakälla. (Eng. "feed".)

Atom-post
: En paketering av ett eller flera digitala dokument som beskriver/tillhör samma logiska resurs. (Eng. "entry".)

Arkiv
: En arkiverad representation av ett flöde med fixerade start- och slutdatum.

Uppdatering
: Ändring av informationen för en post.

Radering
: Borttagning (avpublicering) av en post (och därmed all information som den paketerar).

Kontrollsumma
: Ett digitalt "tumavtryck". En kontrollsumma beräknad på en digital (binär) representation, designad för att upptäcka fel vid lagring och överföring av information.

MD5
: En kryptografisk algoritm (öppen och fri att använda) för att beräkna kontrollsummor. Är inte garanterat unik i alla lägen, men anses tillförlitlig för att säkerställa överförd data från en pålitlig källa.

1.3.2. Specifikt för rättsinformationsdomänen

Rättsinformationssystemet

Den uppsättning datasystem som samordnar syntax, semantik, inhämtning och exponering av primär rättsinformation samt vyer av dessa.

Rättinformation

Rättsliga dokument och metadata om dessa, inkluderande digitala representationer av fysiska dokument, poster som beskriver dessa med inneboende egenskaper samt relationer sinsemellan.

2. Varför Atom

2.1. Källor

Rättsinformationssystemets datakällor utgörs av samlingar av poster som växer över tid. Posterna utgör en paketering av en "enhet" av rättsinformation (i regel ett dokument såsom en föreskrift), tillsammans med enhetens identifierare och en tidstämpel som anger när posten tillkommit eller uppdaterats i datakällan.

Eftersom myndigheterna har publiceringsansvar är systemet byggt på att dessa exponerar tillkommet material som listor av nytillkomna poster. På så vis kan både det centrala systemet och andra intressenter hämta information direkt från utgivarna allt eftersom den tillkommer. Detta sker på samma vis som elektroniska nyhetsläsare (RSS-läsare) idag hämtar ny information.

RSS (eng. "Really Simple Syndication") är ett samlingsnamn på format som uttrycker flöden (eng. "feeds") för t.ex. nyhetsuppdateringar. Ursprungligen stod RSS för "Rich Site Summary" och var ett enskilt format. Nuförtiden finns det flera versioner av detta format, och samlingsnamnet inkluderar både dem och Atom-formatet. Atom är ett specifikt, öppet, XML-baserat format som har standardiserats i form av RFC 4287 (The Atom Syndication Format).

Atom har valts för att det är öppet, standardiserat och väletablerat, samt har explicit stöd för anpassning och utökning för olika ändamål, t.ex. detaljerad hantering av paginering för t.ex. arkiverade postförteckningar, raderingar av poster med mera. Det finns bra verktygsstöd i olika programspråk/-plattformar för att hantera formatet.

2.2. Specifikationer

Följande Atom-relaterade specifikationer används för att hantera inhämtning och syndikering av innehållet i systemet:

- RFC 4287: The Atom Syndication Format
- RFC 5005: Feed Paging and Archiving
- RFC 6721: The Atom "deleted-entry" Element
- (Internet-draft) Atom Link Extensions

Övriga specifikationer beskriver de detaljer som berör hantering av stora mängder poster (m.h.a. länkade flöden), representation av raderade poster (vid felpublicering) samt checksummor av överförda dokument.

2.3. Paket respektive innehåll

Systemet använder Atom enbart för att beskriva tillkomst av poster av information i systemet. Dessa kopplas med sin identifierare till den faktiska resurs (återigen, primärt rättsinformationsdokument) som ingår i rättsinformationen. Data (i form av digitala dokument) som beskriver och representerar dessa resurser "paketeras" av posten (d.v.s. länkas till).

På detta vis hålls den tekniska hantering av tillkomst och den information som tillhör denna process åtskiljd från rättsinformationsdomänens specifika information.

2.4. Systemet som primär, auktoritativ datakälla

Systemets kontrollerade sammanställning (aggregat) av källorna exponeras på precis samma vis som källorna till systemet. Detta sammanställda flöde representerar tillkomsten av kontrollerade poster som hämtats från källorna. Denna källa utgör den primära källan för data ur rättsinformationssystemet, och kan konsumeras av allehanda rättsinformationsintressenter.

3. Representationer i Atom

Atom-flöden uttrycks som postförteckningar i ett eller flera XML-dokument. Respektive post är försedd med en unik identifierare i form av en URI, samt en tidstämpel som anger uppdatering.

Minst ett förteckningsdokument ska finnas för en källa till systemet. Ett sådant dokument kallas prenumerationsdokument ("subscription feed", se RFC 5005: Feed Paging and Archiving för detaljer) och är det dokument som ska länkas till för att komma åt datakällan som flödet representerar.

Exempel på strukturen i ett feed-dokument:

```
<?xml version="1.0"?>
<feed xmlns="http://www.w3.org/2005/Atom"
      xml:lang="sv">

  <id>tag:riksarkivet.se,2009:rinfo:feed</id>

  <title>Riksarkivets författningssamling</title>

  <updated>2009-01-23T12:42:32Z</updated>

  <author>
    <name>Riksarkivet</name>
    <uri>http://www.statensarkiv.se/</uri>
    <email>rinfo@riksarkivet.ra.se</email>
```

```
  </author>

  <link rel="self"
        href="https://www.statensarkiv.se/rafs/feed/index.atom"/>
  <link rel="prev-archive"
        href=
      "https://www.statensarkiv.se/rafs/feed/archive/2008/index.atom"
      />

  <entry>
    <id>http://rinfo.lagrummet.se/publ/ra-fs/2006:6</id>
    <updated>2007-02-09T00:00:00.000Z</updated>
    <published>2007-02-09T00:00:00.000Z</published>
    <title>Föreskrifter om ändring av Riksarkivets föreskrifter och
        allmänna råd (RA-FS 2004:2) om gallring och återlämnande av
        handlingar vid upphandling;</title>
    <summary>Ändringen innebär att undantaget för myndigheter under
        Försvarsdepartementet har tagits bort och att de allmänna
        försäkringskassorna har tagits bort ur uppräkningen.</summary
>
    <content type="application/pdf"
        src=
      "https://www.statensarkiv.se/Sve/RAFS/Filer/ra-fs-2006-06.pdf"
        hash="md5:3c5fc4bdc3306ae6541e97b89dbf4d16"/>
    <link rel="alternate"
        type="application/rdf+xml"
        href=
"https://www.statensarkiv.se/Sve/RAFS/showrdf?doc=2006-6"
        length="2317" hash="md5:e4e242a358b72405fdbf5ba8bc044b1a"/>
  </entry>

  <entry>
    <id>http://rinfo.lagrummet.se/publ/ra-fs/2004:2</id>
    <updated>2004-09-27T00:00:00.000Z</updated>
    <published>2004-09-27T00:00:00.000Z</published>
    <title>Riksarkivets föreskrifter och allmänna råd om gallring och
        återlämnande av handlingar vid upphandling;</title>
    <summary></summary>
    <content type="application/pdf"
          src=

"https://www.statensarkiv.se/Sve/RAFS/Filer/ra-fs-2004-02.pdf"
          hash="md5:ca68b77f41ad2231586cf3e4d7970629"/>
    <link rel="alternate"
        type="application/rdf+xml"
        href=
"https://www.statensarkiv.se/Sve/RAFS/showrdf?doc=2004-2"
        length="2414" hash="md5:8051c695ef087fcc1183e51694c9c8c1"/>
  </entry>

</feed>
```

3.1. Förteckningsdokumentets beståndsdelar

3.1.1. Flödets ID

Nödvändig information!

Elementet `id` anger en *persistent* identifierare (URI) för datakällan. Exempel:

```
<feed ...>
```

```
<id>tag:riksarkivet.se,2009:rinfo:feed</id>
```

Rent tekniskt spelar det ingen roll vad för slags URI som används så länge denna inte ändras. Men i detta exempel används en TAG URI. Sådana är praktiska i feed-sammanhang, eftersom de inte är bundna till http-platsen för dokument. Eftersom identifieraren ska vara persistent över tid kan en HTTP-uri som inte längre leder någonstans vara vilseledande.

3.1.2. Flödets titel

Exempel:

```
<title>Riksarkivets författningssamling</title>
```

Anger någon slags titel på detta feed. Används inte till någonting av Rättsinformationssystemet. Men den kan var användbar för någon som tar del av denna källa via andra verktyg eller i andra sammanhang.

3.1.3. Flödets tidstämpel

Exempel:

```
<updated>2009-01-23T12:42:32Z</updated>
```

Ska ange tidstämpel då något i detta feed-dokument ändrats. Anges som W3C Date Time Format (DTF).

Observera att i exempet används "Zulu Time". Det är inte samma sak som svensk tid. En del programmeringsspråk har inbyggt stöd för att generera dessa, i andra fall finns det i regel tredjepartsbibliotek för DTF.

(Anledningen till att vi väljer Zulu Time här är för att dessa är oberoende av tidzoner, *inklusive* sommartidavvikelser. Tekniskt är det okej att istället för "Z" ange t.ex. "+01:00", för svensk vintertid, "+02:00" för sommartid osv. Korrekt parsning av DTF ska resultera i samma tidpunkt oavsett.)

3.1.4. Metadata för flödets utgivare

Nödvändig information!

Exempel:

```
<author>
  <name>Riksarkivet</name>
  <uri>http://www.statensarkiv.se/</uri>
  <email>rinfo@riksarkivet.ra.se</email>
</author>
```

Anger kontaktinformation till ansvarig för publiceringen av denna datakälla.

Planen är att Rättsinformationssystemet ska använda denna information för att rapportera om eventuella felaktigheter har uppstått vid inhämtning av informationen.

Därför är det lämpligt att e-postadressen går till en funktionsbrevlåda istället för en enskild handläggare.

Namn och URL används däremot inte av Rättsinformationssystemet, men kan vara nyttig i andra sammanhang.

3.1.5. Addressen till flödets representation

Exempel:

```
<link rel="self"
    href="https://www.statensarkiv.se/rafs/feed/index.atom"/>
```

Anger den adress där detta feed-dokument ligger. Är inte strikt tekniskt nödvändig, men underlättar för någon som t.ex. sparar ner dokumentet lokalt utan att notera ursprungsplats.

3.1.6. Länk till arkiv för äldre poster

Viktig!

Exempel:

```
<link rel="prev-archive"
      href=
    "https://www.statensarkiv.se/rafs/feed/archive/2008/index.atom"
    />
```

Anger plats för ett tidigare dokument i den sammanhängande kedjan av publicerade feed-dokument över tid.

Se RFC 5005: Feed Paging and Archiving. Rättsinformationssystemet kräver att dessa är konsistenta, samt beständiga över tid så pass att det är garanterat att systemet har hunnit hämta in informationen. (Exakta krav är inte fastställda, men livslängder på minst ett år är rimliga.)

Observera att arkiverade feed-dokument inte är nödvändiga. Om man har en lågt antal rättsinformationsdokument kan samtliga förekomma i samma feed-dokument under en överskådlig tid. Syftet med feed archiving är att hantera storleken på feed-dokument över tid, så att dessa inte växer till ohanterliga storlekar. Se asvnittet om Arkiv nedan för detaljer.

3.2. Tidstämplar i rättskällor respektive det centrala systemet

Atom-flöden representerar en växande tidlinje av poster över tid. Varje post har en tidstämpel i sitt `updated`-element som anger när den uppstod, eller när information i posten senast förändrats.

Det är viktigt att särskilja denna maskinella tidsangivelse från de formella datum och tider som gäller för det faktiska rättsinformationsdokumentet (såsom utgivningsdatum och liknande).

Det är *av yttersta vikt* att tidstämplar respekteras med avseende på:

- Att förändringar *måste* tidstämplas kronologiskt.
- Att inga tidstämplar för exponerat material modifieras i efterhand.

Eftersom systemet inhämtar resurser från rättskällorna måste det lita på att kommunicerade tidstämplar gäller, så att allting före en påträffad tidpunkt kan garanteras gälla information som redan beaktats.

Eftersom systemet exponerar all inhämtad (och egen) information på samma vis som källorna — d.v.s. som Atom-flöden — kommer poster som lästs in att få *nya* tidstämplar som återger inhämtningstillfället.

Originaltid (och URL:er) sparas ner för respektive post för att kunna återge detaljer kring den inlästa posten.

3.3. Posternas innehåll

Varje post representeras av ett s.k. Atom Entry. Här anges URI:n för dokumentet (i elementet `id`), primär representation (i elementet `content`) samt referenser till alternativa representationer.

```
  <entry>
    <id>http://rinfo.lagrummet.se/publ/ra-fs/2004:2</id>
    <updated>2004-09-27T00:00:00.000Z</updated>
    <published>2004-09-27T00:00:00.000Z</published>
    <title>Riksarkivets föreskrifter och allmänna råd om gallring och
        återlämnande av handlingar vid upphandling;</title>
    <summary></summary>
    <content type="application/pdf"
          src=

"https://www.statensarkiv.se/Sve/RAFS/Filer/ra-fs-2004-02.pdf"
          hash="md5:ca68b77f41ad2231586cf3e4d7970629"/>
    <link rel="alternate"
          type="application/rdf+xml"
          href=
            "https://www.statensarkiv.se/Sve/RAFS/showrdf?doc=2004-2"
          length="2414" hash="md5:8051c695ef087fcc1183e51694c9c8c1"/>
  </entry>
```

3.3.1. Postens ID

Nödvändig information!

För att koppla en post till det rättsinformationsdokument den paketerar används den officiella URI:n för dokumentet som värde i Atom-postens ID.

```
<id>http://rinfo.lagrummet.se/publ/ra-fs/2004:2</id>
```

Rättsinformationssystemet matchar id:t i entry för ett rättsinformationsdokument mot den officiella URI:n (den elektroniska identifieraren) för detta dokument. Den kommer att användas av alla system som hanterar poster i rättsinformationssystemet.

För att rättsinformationsdokument-URI:er ska bli korrekta har rättsinformationssystemet en algoritm för att konstruera dessa. Se dokumenetet URI-principer för hur dessa ska konstrueras.

3.3.2. Postens tidstämplar

Nödvändig information!

```
<updated>2007-02-09T00:00:00.000Z</updated>
```

Uppdateringstiden i `updated` anger när något i posten modifierats i källan. Rättsinformationssystemet använder denna för att avgöra om posten för ett rättsinformationsdokument har uppdaterats. Det är därför kritiskt att denna är korrekt angiven.

```
<published>2007-02-09T00:00:00.000Z</published>
```

Publiceringstiden i `published` anger när posten officiellt publiceras i källan. För merparten av dokument kommer denna att vara densamma som updated (se ovan), då förändringar av rättsinformationsdokument i huvudsak görs i form av nya rättsinformationsdokument (såsom rättelseblad och ändringsförfattningar). Enbart i undantagsfall ska poster uppdateras (eller tas bort) elektroniskt.

3.3.3. Titel och beskrivning

```
<title>Riksarkivets föreskrifter och allmänna råd om gallring och
    återlämnande av handlingar vid upphandling;</title>
<summary>...</summary>
```

Dessa element måste enligt Atom-specifikationen finnas i en post. Det är dock inte tekniskt nödvändigt att de innehåller något. Innehåll i dessa element används heller inte av Rättsinformationssystemet. Vi rekommenderar ändå att något läsbart värde används för att andra verktyg ska kunna presentera läsbar information av detta feed. Utan värde kommer t.ex. många vanliga feed-läsare inte att kunna skapa en användbar presentation av Atom-flödet.

3.3.4. Huvudsakligt innehåll

Nödvändig information!

```
    <content type="application/pdf"
          src=

"https://www.statensarkiv.se/Sve/RAFS/Filer/ra-fs-2004-02.pdf"
          hash="md5:ca68b77f41ad2231586cf3e4d7970629"/>
```

En primär representation ska representera det faktiska rättsinformationsdokumentet på ett av Rättsinformationssystemet godkänt format.

(Obs! Det är tänkt att systemet (åtminstone initialt) ska tolerera avsaknad av en digital dokumentrepresentation, så länge RDF-data om dokumentet finns tillgänglig. Denna tolerans är enbart till för de fall då det inte *går* att få fram en digital representation, inte för att minska kraven på vad rättskällorna ska publicera.)

3.3.5. Alternativa format

Viktig!

De alternativa formaten som beskriver dokumentet länkas till som följer.

```
    <link rel="alternate"
          href=
            "https://www.statensarkiv.se/Sve/RAFS/showrdf?doc=2004-2"
          type="application/rdf+xml"
          length="2414" hash="md5:8051c695ef087fcc1183e51694c9c8c1"/>
```

Då dokumenten ska beskrivas med RDF-data *måste* minst ett sådant dokument på ett av rättsinformationssystemet godkänt RDF-format finnas länkat till.

Se dokumentet Introduktion till rättsinformationssystemet - För implementatörer och tekniskt ansvariga för detaljer kring krav på inlämningsformat och innehåll.

3.3.6. Checksummor på länkade dokument

Som angivet i exemplen ovan ska respektive länkad resurs vara annoterad med storlek, mediatyp och en kontrollsumma i MD5. Se Atom Link Extensions, sektion 3.2 för specifikationen av länk-attributet för kontrollsummor.

Denna specifikation förlitar sig på Atom Link Extensions för att ange checksummor. Denna är ännu en I-D, vilket inte garanterar en stabil specifikation. Sedan systemet specades har denna förändrats. Den tidigare

formen på checksummor, specificerad i den nu osboleta revision 2 av Atom Link Extensions, såg ut likt följande:

```
  <feed ...
        xmlns:le="http://purl.org/atompub/link-extensions/1.0">
    ...
      <content type="..." src="..."
               le:md5="d41d8cd98f00b204e9800998ecf8427e"/>
```

Denna är nu ändrad till:

```
      <content type="..." src="..."
               hash="md5:d41d8cd98f00b204e9800998ecf8427e"/>
```

För att vara bakåtkompatibel stödjer inhämtningen båda dessa former. Även om denna I-D ännu inte har blivit en RFC förväntas den inte ändras substantiellt längre. Därför rekommenderas denna nya form.

3.3.7. Bilagor

Enclosures representerar någon slags extern bilaga till den primära resursen. Detta omfattar alltså bara de fall då bilagorna är separata filer, och inte t.ex. PDF:er som inkluderar både huvuddokumentet och tillhörande bilagor.

```
<feed ...
      xmlns:dct="http://purl.org/dc/terms/">
  ...
  <entry>
    ...

    <link rel="enclosure"
          href=
    "http://www.slv.se/upload/lagstiftning/2005-2006/2006_22bil1.pdf"
          type="application/pdf"
          dct:isFormatOf=

"http://rinfo.lagrummet.se/publ/livsfs/2006:22#bilaga_1"
          length="1168037"
          hash="md5:1115651b5e8380aff8d"/>
```

Attributet `@dct:isFormatOf` skapar en koppling mellan identifieraren för bilagan och själva filen som innehåller bilagan.

Externa bilagor hämtas in och sparas på en ny URL baserad på att antingen:

- lösa värdet i `@href` mot postens primära innehållsreferens (`content/@src`), basera basen på postens formella URI (angivet i `id`)
- använda `@dct:isFormatOf` plus filändelse för angiven mediatyp.

4. Förändringar av existerande poster

4.1. Uppdateringar

Uppdateringar är tillåtet för *viss form* av nytillkommen post-data och -representationer. Eftersom rättsinformation är statisk i betydelsen att den när den väl är utgiven inte förändras, bör i teorin inga digitala förändringar som leder till uppdaterade dokument ske. I praktiken kan dock vissa korrigeringar vara nödvändiga, så länge de representerar maskinellt relaterade justeringar, t.ex. format- och syntax, samt metadata.

Den information som den juridiska texten utgör förändras inte med mindre än att nya rättsinformationsdokument tillkommer. Detta gäller även konsoliderade versioner av författningar. Varje konsolidering som görs till följd av att en ändringsförfattning antagits, betraktas inom systemet som en ny post.

En uppdaterad post *ska* ha ett senare datum än publiceringsdatumet. Detta för att undvika att en ny post av misstag fått en redan befintlig identifierare.

Atom-syntaxen kräver inte närvaron av *published* för att ange den ursprungliga tidstämpel då posten publicerades. Rättsinformationssystemet kräver dock närvaron av detta. Detta tydliggör de fall då en post uppdaterats sedan ursprunglig publicering (och likaledes då den är oförändrad, genom att både publiceringstid och uppdateringstid är identiska).

Således måste både ursprungstid och eventuell uppdateringstid sparas av det publicerande systemet.

4.2. Raderade poster

För att hantera *felaktigt inkomna* poster exponeras även *raderade* resurser.

(Detaljer kring hur dessa förekommer i flödesdokumenten beskrivs i kommande avsnitt, specifikt för Arkiv-flöden. I Kompletta flöden behövs inte explicita raderingar.)

Raderingar representeras av ett `deleted-entry`-element, se The Atom "deleted-entry" Element. Dessa element innehåller ID:t för posten som raderats samt raderingsdatum.

Exempel på radering:

```
<feed xmlns="http://www.w3.org/2005/Atom"
      xmlns:at="http://purl.org/atompub/tombstones/1.0">
  ...

  <at:deleted-entry
      ref="http://rinfo.lagrummet.se/publ/ra-fs/3010:0"
      when="2010-03-08T15:55:34+0100"/>
```

(Notera att Atom-syntaxen förväntar sig att utökningar såsom dessa för raderade poster ska placeras före `entry`-elementen. Tidsordningen garanteras enbart via tidstämplarna.)

Obs! Detta stödjs enbart för att ta bort tekniskt *felaktiga* poster, t.ex. då misstag vid publikation skett, såsom fel eller ännu ej färdigställt dokument har publicerats, eller formatmässigt trasiga filer lagts upp. Raderingar har ingenting med upphävning av dokuments legala giltighet eller liknande att göra.

Om möjligt ska *uppdateringar* av sådana felaktiga publiceringar göras i första hand. Bara om systemet exponerat dokument som ännu inte är giltiga, eller poster med icke-giltiga URI:er i postens `id` är det lämpligt att uttrycka en radering i en postförteckning.

4.2.1. Beständiga raderingar

Notera också att raderingar *måste* kommas ihåg av den som publicerar sådana. Detta eftersom konsumenter av postförteckningen måste underrättas om detta oavsett när de sist läste av denna.

Det enda fall då en raderad post kan anses irrelevant i framtiden är om en korrekt publicerad post med samma URI i `id`:t för posten görs.

5. Kompletta flöden

Feeds markerade som kompletta (se RFC 5005, sektion 2) är lämpliga för att publicera en lagom mängd dokument i en samling som inte väntas växa avsevärt över tid.

Ett flödesdokument markeras som komplett med:

```
<feed xmlns="http://www.w3.org/2005/Atom"
      xmlns:fh="http://purl.org/syndication/history/1.0">
  ...

  <fh:complete/>
```

Dessa flöden representeras av ett sammanhållet Atom-dokument. Mängden poster i sådana bör vara "rimligt liten", förslagsvis under tusentalet som mjukt gränsvärde. En hantering av sådana skulle kräva att alla poster från en given källa ska kunna sökas ut av systemet, och vid en inhämtning skulle alla de som inte finns listade i det kompletta flödet betraktas som raderade och tas bort ur det lokala lagret.

6. Arkiv

Större samlingar (eventuellt med många uppdateringar och/eller raderingar) kan och bör delas upp i flera feed-dokument som länkar till varandra, s.k. paginering av feeds. Det går inte att kombinera kompletta flöden enligt sektion 5 med sådana arkiv.

Paginering möjliggör långa tidsorterade listor av poster. För en stabil inhämtning av sådana listor har systemet valt att både samla in och exponera Atom-arkiv (se RFC 5005, sektion 4 för specikationen av sådana).

Tekniskt kan rimligen upp till något tusental entry-element i samma dokument hanteras utan större problem. Men om man från början uppskattar att tillväxttakten kommer att göra arkiverade feeds nödvändiga är det möjligt att max några hundra poster per dokument är en rimligare maxgräns. Detta är beroende på filstorlek och överföringskapacitet, d.v.s webbinfrastruktur i stort. Beakta att insamlande system kan komma att hämta hem åtminstone prenumerationsdokumentet många gånger per dag, eller t.o.m. per timme.

Här spelar också HTTP-mekanismer för tidstämplar och ETags en stor roll, vilka rekommenderas för att undvika överflödig inhämtning av oförändrade dokument. Se detaljer kring "conditional GET" i specifikationen för HTTP/1.1 för detaljer.

6.1. Tillväxtproblem

En risk med detta är om en mängd uppdateringar sker från någon källa, t.ex. en massiv korrigering av tusentals publicerade poster. I det fallet kommer arkiv-listorna bli mer och mer ohanterbara för initial inhämtning, d.v.s. då en konsument vill "grundladda" från systemet.

6.1.1. Arkivhushållning

En möjlig lösning på detta är att *"kompaktera"* arkiven. Med detta menas att hushålla i de äldre flödesdokumenten genom att städa i gamla arkiv-feeds då poster uppdateras, och då helt plocka bort inaktuella exponerade poster och/eller raderingar. Detta är tekniskt tillåtet enligt RFC 5005, även om det inte uttryckligen förespråkas. Genom ett sådant förfarande förloras formell historik, men samtidigt uppnås en förbättring för konsumtionen av flödet eftersom behovet av att hoppa över inaktuellt material kan elimineras.

En liknande metod är att då och då skapa helt nya arkivdokument genom att omindexera alla ingående poster och då inte exponera uppdaterade historiska poster, eller raderade som sedemera blivit återskapade. För att uppfylla kraven från RFC 5005 är det tillrådigt att låta de nya arkiven få nya URL:er, eftersom de representerar nyskapade arkivlistor (om än med gammalt material). En sådan omskrivning av arkiv kommer att störa för insamlande system, men kan anses acceptabel om tiden för störningen kan hållas till ett minimum (genom

att t.ex. inte länka om prenumerationsdokumentets första arkiv förrän de nya arkiven är klara, och först efter det plocka bort de gamla arkivdokumenten.)

7. Krav på publiceringen

7.1. Checklista

I korthet ska följande beaktas vid skapandet av Atom-flöden som ska utgöra källor till rättsinformationssystemet.

- Flödesdokument ska dela samma, för källan beständiga och unika feed-id.
- Posternas entry-id:n ska motsvara URI:n för resursen (rättsinformationsdokumentet) som posten levererar. (Se URI-principer för hur dessa ska konstrueras.)
- Den primära digitala representationen för dokumentet ska länkas till i postents `content`-element.
- RDF-data som beskriver postens dokument ska länkas till i `content` (om det primära innehållet även innehåller metadata), eller i en `alternate`-link.
- Alla länkade dokument ska ha en korrekt angiven MD5-kontrollsumma.
- Datum i respektive flödesdokument ska följa kronologisk ordning (per feed-dokument; inbördes entry-ordning på XML-nivå är inte nödvändig då `updated`-värdet räcker för att ange detta).
- Fullständighet måste iakttagas, d.v.s. arkiven ska representera alla aktuella uppdateringar och raderingar som gäller för rättskällan.

8. Insamlingsmetod

Denna sektion beskriver hur rättsinformationssystemet centralt hanterar insamling. Denna beskrivning är enbart för informativa syften.

8.1. Generellt

Per sida:

1. Följ flödet bakåt i tid:
    - notera aktiva poster, senaste (yngsta) tidstämpel per entry-id som antingen är:
        1. uppdaterad
        2. raderad
    - följ `prev-archive` (bara för arkiv; se nedan)
2. Samla in en aktiv post respektive radering i taget, i ordningen äldst till yngst.

(Detta säkerställer att den yngsta insamlade posten markerar vad det insamlande systemet har lyckats samla in. På så vis garanteras att påföjande insamlingar inte missar gammalt material som av någon anledning inte kunnat hämtas in.)

Viktigt! Notera att insamlande system måste hålla koll på en posts ursprungstidstämpel för updated och samtidigt exponera en ny tidstämpel för updated om systemet aggrererar flera källor och inte fullständigt kan lita på och samla in dessa källors poster i kronologisk ordning ("breadth-first").

8.2. Insamling av kompletta flöden

Kompletta flöden är enkla att samla in, förutsatt att det insamlande systemet sparar ner en lista på alla poster för respektive källas feed-id. Gör följande för att samla in poster i sådana:

1. hämta det kompletta flödet,
2. använd feed-id:t för att läsa upp lista på alla poster i systemet som kommer från denna källa,
3. radera alla poster i listan från systemet som inte finns med i det nya inlästa flödets förteckning,
4. sortera posterna på "updated" i stigande ordning, d.v.s. äldst (minsta tidstämpeln) först,
5. för varje post, samla in ny data om dess updated är nyare än exiserande insamlad posts ursprungstidstämpel.

8.3. Insamling av arkiv-flöden

En metod för att samla in poster från sådana arkiv är att:

1. följ "subscription"-flödet och "prev-archive",
2. för varje sida: sortera poster (och *raderingar*) på "updated" i fallande ordning, d.v.s. yngst (största tidstämpeln) först,
3. spara entry-id + updated i mapp om updated inte finns för id:t eller är större än (yngst) tidigare sparat updated-värde,
4. stoppa när id+updated-par stöts på som finns lokalt inhämtad,
5. klättra framåt i tiden längs inhämtade feeds och spara ner de poster som finns i mappen med samma värden på id+updated.

Detta säkerställer följande:

- En markerad början görs (den senaste posten i "current" vid starten för inhämtning). Detta säkerställer att inhämtningen inte fortgår på obestämd tid i det (förvisso osannolika) fall då källan uppdateras snabbare än inhämtningen.

- Inhämtad data sparas ner i kronologisk ordning, så att ingen slags fel leder till "luckor" i tidsordningen (detta så att återupptagen inhämtning kan lita på att lokalt inhämtat entry inte "skuggar" äldre, ännu ej sedda poster).
- Inga onödiga inhämtningar av redan uppdaterade poster sker. (Detta är viktigt eftersom gamla poster inte behöver länka till historiska dokument, och därför kan ange felaktiga (d.v.s. gamla) kontrollsummor och filstorlekar, alternativt peka på dokument som tagits bort eller sakna länkar till nytillagda sådana.)

(Obs! Naturligtvis måste alla feed-sidor som klättras vara försedda med korrekt angivna updated-datum, så att inga yngre poster finns längre bak i arkiven.)

9. Notifieringar om förändringar

Rättsinformationssystemet innehållen en lista på de rättskällor som information ska samlas in ifrån. Systemet är konfiguererat så att det då och då läser av alla källor, en i taget, för att samla in poster som är nya eller förändrats/raderats sedan förra insamlingsrundan utförts.

Exakt hur ofta systemet gör denna insamling är inte fastställt idag. Det är en relativt billig operation som kan förväntas göras många gånger per dag, eventuellt oftare än en gång i timmen.

Idag kan rättsinformationssystemet även "pingas" om en källa uppdateras och önskas inhämtas omedelbart. Detta sker genom att URL:en till en datakällas prenumerationsdokument ("subscription feed") POST:as till systemet. Om så sker, och källan är med i listan över godkända källor, kommer systemet att påbörja en insamling direkt.

Kontakta rättsinformationsprojektet för detaljer kring hur denna notifieringsmekanism ska anropas.

10. Konsekvenser

Alla poster hämtas in då och då över tid. Dessa sparas i det centrala systemet och får där nya datumstämplar som motsvarar inhämtningstillfället.

Med denna insamlingsmetodik skapas en arkivlogg som utgör en komplett lista på poster (inklusive uppdateringar och raderingar). Denna exponeras av rättsinformationssystemet och utgör den primära postförteckningen för systemet, och är således den primära datakällan för konsumenter intresserade av den kompletta, tekniskt validerade rättsinformationen.

Stora, sammanställda arkivloggar kan bli svåranvänd p.g.a. många korrigeringar i källorna. En lösning på detta är att, som nämns i avsnitt 6 ovan, "hushålla" arkiv. Detta kan åstadkommas genom att så fort en post

uppdateras eller raderas plocka bort den tidigare förekomsten (som kan ligga i något äldre arkiv-dokument bak i tiden). I nuläget gör systemet inte detta, men en lösning är planerad och kan komma att bli aktuell beroende på hur många korrigeringar som i praktiken visar sig göras i de respektive rättskällorna.