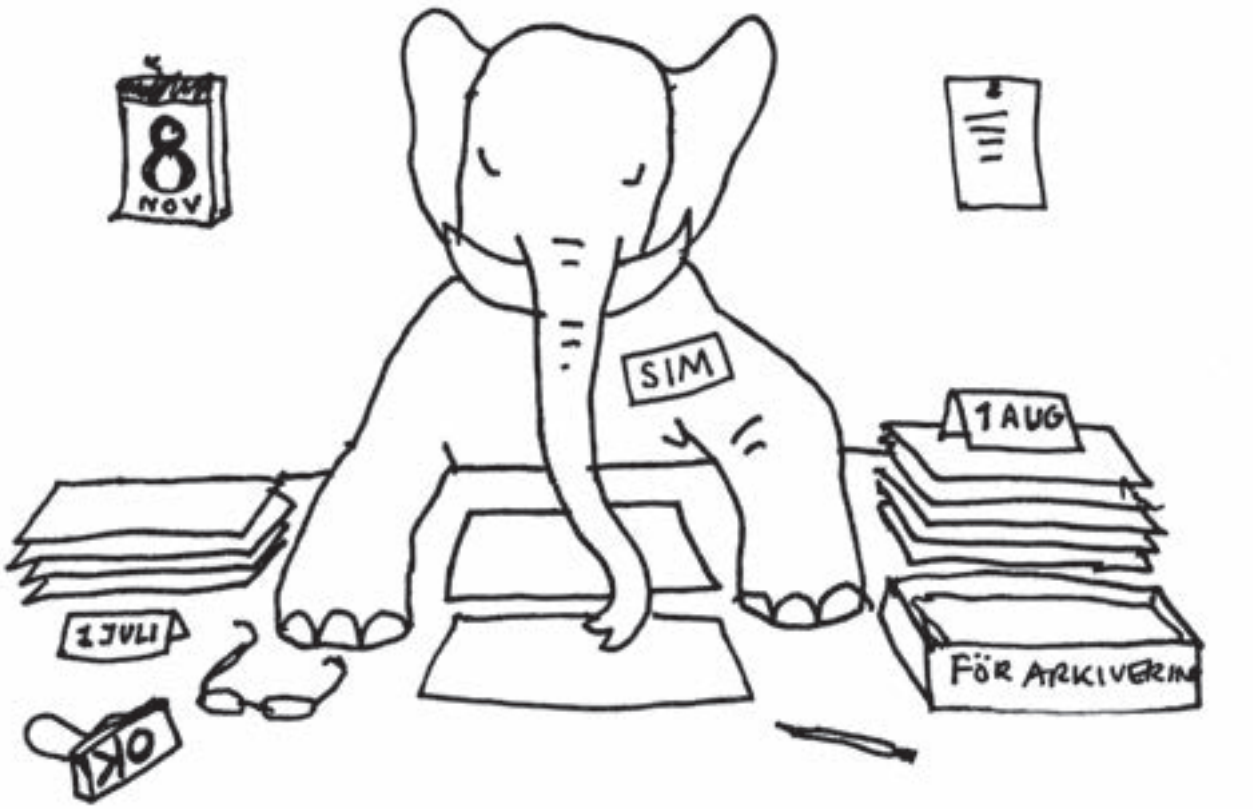
NY LAGRINGSMETOD FÖR Webbarkiv



Statisk inkrementell metod (SIM) är en ny metod för att underlätta förvaltning och hantering av inkrementella webbarkiv. Till skillnad från branschstandarden WARC så är SIM inte beroende av någon mjukvara. Matias Vangsnes, grundare av företaget Archiwwwe förklarar.

Text Mariya Lysenkova mariya.lysenkova@archiwwwe.se
Illustration Maria Lindroos

”Viktigt när vi utvecklade metoden var möjligheten att förvalta webbarkivet, det vill säga att kunna öppna filerna i arkivet utan speciell mjukvara,kunna bearbeta och uppdatera arkivet.”

WARC-FORMATET ÄR en branschstandard vid webbarkivering, ett filformat där den insamlade webbplatsens samtliga filer lagras. Formatet tillåter också inkrementell insamling (avbilder över tid). Tekniken bygger på att ”spela in” och ”spela upp” webbplatsen med en speciell mjukvara, vilket i sig är en bra teknisk lösning.

Webbplatser innehåller inte bara tillgångar för att presentera innehållet (bilder, stilmallar, skript etc.) utan ofta bifogade dokument. Dessa dokument behöver ibland bearbetas och konverteras för att kunna långtidslagras.

Och det är just förvaltning som formatet WARC lämpar sig mindre bra för. Formatet kräver en speciell mjukvara för att öppna arkivet. En annan nackdel är att formatet försvårar bearbetning av arkivet. Det finns idag inga enkla metoder för att extrahera filer ur WARC för att analysera dem eller konvertera och uppdatera arkivet.

EN BRANSCHÖVERENSKOMMELSE är att få beroenden och öppna format som skapar en bättre förutsättning för kvalitativa arkiv över tid.

– Vi har sedan starten provat och utmanat de rådande standarder och metoder som finns för att kunna leverera en så bra lösning som möjligt. Vi upptäckte tidigt att branschen hittills lagt mest resurser på tekniken för själva insamlandet av webbplatser och inte så mycket på att underlätta förvaltningen, säger Matias Vangsnes.

För honom blev det extra tydligt att det fanns behov av en ny metod när hans företag Archiwwwe skulle hjälpa sina offentliga kunder att leverera inkrementella webbarkiv till Riksarkivet. Riksarkivet accepterar nämligen inte WARC som leveransformat.

– Viktigt när vi utvecklade metoden var möjligheten att förvalta webbarkivet, det vill säga att kunna öppna filerna i arkivet utan speciell mjukvara, kunna bearbeta och uppdatera arkivet. Därför utgick vi från att filerna i arkivet bör lagras i en filstruktur och länka till varandra relativt. Sedan tittade vi på hur vi skulle lösa problematiken med just inkrementella webbarkiv för att undvika dubbletter, det vill säga de filer i arkivet som inte ändrats över tid, säger han. ✖

Så här fungerar SIM

I följande simplifierade exempel utgår vi från att samla in en webbplats (www.example.com) månatligen och lagra den med SIM. Exempelwebbplatsen består av en HTML-sida som länkar till en bild. Så här ser källan för index.html ut:

```
<html>
  <body>
    <img src=”images/img.jpg”>
  </body>
</html>
```

Låt oss titta på hur resultat blir efter att vi insamlat webbplatsen och lagrat den med metoden SIM.

Efter första insamlingen, 1 augusti 2017, klockan 10.00:

```
.
└── www.example.com
    ├── 1501581600
    |   ├── images
    |   |   └── img.jpg
    |   └── index.html
    └── .sim
```

Observera att filerna från den första insamlingen är lagrat i en folder namngiven med en unik identifierare (i detta fallet en tidsstämpel i UNIX-format när insamlingen ägde rum) samt en dold folder (.sim) för att spara metadata fortsättningsvis. Tittar vi på källan för index.html så ser den likadan ut som orignalkällan:

```
<html>
  <body>
    <img src=”images/img.jpg”>
  </body>
</html>
```

Låt oss titta på vad som händer efter att vi har insamlat webbplatsen en månad senare. Webbplatsen har under tiden uppdaterats med en länk till ett PDF-dokument. Låt oss titta på hur detta återspeglas när vi lagrar filerna med metoden.

Efter andra insamlingen, 1 september 2017, klockan 10.00:

```
.
└── www.example.com
    ├── 1501581600
    |   ├── images
    |   |   └── img.jpg
    |   └── index.html
    ├── 1504260000
    |   ├── files
    |   |   └── info.pdf
    |   └── index.html
    └── .sim
```

Arkivet innehåller nu:
• 1501581600/index.html
- oförändrad
• 1501581600/images/img.jpg
- oförändrad
• 1504260000 – en adderad folder
• 1504260000/index.html – en uppdaterad version av index.html
• 1504260000/files/info.pdf – en adderad fil (PDF)
• .sim – oförändrad (tom folder)

Notera att 1504260000/index.html länkar till den nya filen (info.pdf) och bilden (img.jpg) som är oförändrad och länkar till den första avbilden (1 augusti 2017, klockan 10.00). På detta sätt sparas statiska foldrar och länkas mellan varandra. Se exempel:

```
<html>
  <body>
    <img src=”../1501581600/images/img.
jpg”>
    <a href=“files/info.pdf”>Download
info</a>
  </body>
</html>
```

Nu har vi gått igenom hur avbilder skapas när information uppdateras och adderas, men hur hanterar SIM avbilder när något tas bort? Då inga filer tas bort från arkivet använder vi oss av metadata för att spåra förändringar.

Om vi genomför en tredje insamling där bilden och PDF-dokumentet är borttagna och länkarna i HTML-filen inte längre finns, lagrar vi filer på följande sätt.
Efter tredje insamlingen, 1 oktober 2017, klockan 10.00:

```
.
└── www.example.com
    ├── 1501581600
    |   ├── images
    |   |   └── img.jpg
    |   └── index.html
    ├── 1504260000
    |   ├── files
    |   |   └── info.pdf
    |   └── index.html
    ├── 1506852000
    |   └── index.html
    └── .sim
        └── 1506852000-deletions.txt
```

Vi har nu adderat 1506852000/index.html tillsammans med en metadatafil .sim/1506852000-deletions.txt som innehåller en lista med filer som tagits bort i avbilden.
Innehåll i .sim/1506852000-deletions.txt:

```
images/img.jpg
files/info.pdf
```

SIM är alltså en metod för att spara och namnge filer, inte en produkt eller ett program. Med metoden kan digitala kanaler insamlas inkrementellt och samtidigt vara tillgängligt för bearbetning (exempelvis konvertering) över tid. SIM är gratis, vem som helst som har tillgång till en dator kan börja använda SIM idag.

– Vi har med gott resultat provlevererat webbarkiv med SIM till Riksarkivet, säger Matias Vangsnes.

För mer information

Läs mer detaljerat om SIM på https://archiwwwe.com/om-webbarkivering/sim