

WARC – ett nytt ISO-filformat för lagring av flera miljarder Internetuppgifter

En webbsida som finns i dag kanske inte finns i morgon. Nu finns en ny ISO-standard som garanterar att den omfattande och ofta mycket värdefulla information som läggs upp på Internet inte försvinner när en sida ändras eller tas bort. Standarden finns tillgänglig hos SIS, Swedish Standards Institute.

Nu finns en ISO-standard som gör det möjligt att arkivera webbsidor. Med WARC-formatet är det möjligt att länka samman många dataobjekt till en lång fil. Formatet kan användas för att skapa applikationer för insamling, hantering, åtkomst och utbyte av innehåll på webbsidor.

– Med den nya standarden blir det enklare att hantera, strukturera och lagra flera miljarder resurser som har samlats in från Internet och andra platser, säger Jörgen Wyke som är projektledare på SIS, Swedish Standards Institute.

Webbsidor har sparats även tidigare. Antikvarier och arkivarier som arbetar med digitala medier har haft en utmanande uppgift i att hålla reda på det svindlande antalet webbplatser och webbsidor. Kungliga Biblioteket i Stockholm har ca 1,7 miljarder objekt från ca 3,2 miljoner webbservrar i sitt webbarkiv.

– Vi har sparat webbsidor sedan 1997. Eftersom det inte fanns någon standard att utgå ifrån utvecklade vi ett eget arbetssätt som också delvis ligger till grund för den nya ISO-standard. Idag får vi inte publicera vårt material externt men om man besöker biblioteket är det fullt möjligt att surfa på webben så som den har sett ut från 1997 och framåt, säger Allan Arvidson, IT-specialist på Kungliga Biblioteket som har varit med och utvecklat standarden för webbarkivering, ISO 28500.

För ytterligare information:

Jörgen Wyke, projektledare SIS, 08-555 520 24, jorgen.wyke@sis.se

Allan Arvidson, IT-specialist på Kungliga Biblioteket, 08-463 40 55, allan.arvidson@kb.se

Erika Messing, pressansvarig SIS, 08-555 520 97, 070-948 06 74, erika.messing@sis.se

Fakta:

- Standarden heter ISO 28500:2009, Information and documentation – WARC file format
- WARC-formatet är en utvidgning av ARC-filformatet, som sedan 1996 har använts för att lagra "spindlar" – som representerar utdrag av hela webbsidor och länkarna på dessa.
- Motivationen till att utvidga ARC är ett resultat av diskussioner och erfarenheter inom organisationerna i International Internet Preservation Consortium (IIPC) – vars främsta uppgift är att inhämta, bevara och tillgängliggöra kunskaper och information från Internet för framtida generationer.
- WARC-formatet skiljer sig från ARC genom att det ger fler möjligheter, framför allt för registrering av huvuden vid HTTP-anrop och godtyckliga metadata, tilldelning av en identifierare för varje fil, hantering av dubletter och flyttade poster samt uppdelning av poster.
- WARC-filer är avsedda för att lagra alla typer av digitalt innehåll, oavsett om det har hämtats med HTTP eller med något annat protokoll.

SIS är en del av det europeiska och globala nätverk som utarbetar internationella standarder. Genom att delta i standardiseringsarbetet kan svenska företag och organisationer påverka utformningen av standarder inom sin marknad. Standarder befrämjar handel över gränser och bidrar till att processer och produkter blir säkrare.

SIS är en fristående ideell förening med medlemmar från både privat och offentlig sektor. SIS omsatte 2008 MSEK 216 och har 170 anställda. SIS arbetar inom och är medlem i de internationella standardiseringsorganisationerna CEN (europeisk) och ISO (global).