WP5: Task 72 (Describe testing procedures for scenario 2.4)

Authors:

Eva Toller (National Archives of Sweden, RA)
……

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	P
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.1	20130411	Eva Toller	RA	First draft
0.2	20130520	Eva Toller	RA	Added four tools for local testing.
0.3	20130627	Eva Toller	RA	Added Heritrix as a tool to test.
0.4	20130819	Eva Toller	RA	Added result for user acceptance tests for archived web site.

1 TESTING PROCEDURES FOR SCENARIO 2.4

1.1 SCENARIO DESCRIPTION

“A history student interested in natural history discovers that Riksarkivet has archived the "Linnéjubilet" web site http://www.riksarkivet.se/default.aspx?id=23153 .He wonders how he can get access to it (the link www.linne2007.se obviously doesn't work anymore).“

General comment: actually, the link www.linne2007.se *does* still work, but for test purposes this is not important. Furthermore, there can be no guarantee that it will continue to be accessible.

Suggested test data: see document **DCH-RP_WP5_Scen-2-4_ID-68.pdf**

1.2 SUGGESTED TEST PROCEDURES

1. Tools (for dissemination)
    a. Investigate the existing dissemination tools and formats that may be applicable for this scenario (**WARC** should be considered as a candidate; see http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml and http://www.ltu.se/centres/Centrum-for-langsiktigt-digitalt-bevarande-LDB/Bibliotek/Webb/Om-filformatet-WARC-och-verktyg-for-formatet-1.67311).
    Use the previous results from DC-NET a basis:
    http://digital-scholarship.org/dcrg/dcrg.htm and the subsequent results from DCH-RP WP 3.1: **D3.1_DCH_RP_ver_1.0_070213.doc**
    b. When feasible, test the chosen tools locally (at RA) and determine which ones that are useable/useful and if they should be included in the SweGrid/SweStore trials.

2. SweGrid/SweStore
    a. Investigate all preparations that must be done to make usage of SweGrid and Swestore

    http://snicdocs.nsc.liu.se/wiki/Getting_started_with_SweGrid

    http://snicdocs.nsc.liu.se/wiki/Swestore

    http://snicdocs.nsc.liu.se/wiki/Accessing_SweStore_national_storage_with_the_ARC_client

    b. Arrange a meeting with SweGrid/SweStore and determine what they can do for DCH-RP.
    c. Make all the necessary preparations and agreements with SweGrid/SweStore.

d. Upload data to SweStore and test the chosen tools there.

3. Obtain permissions from RA to use and export data
    a. Obtain permission from the Electronic Archives section ("ElArk", with Christina Olsson as their representative) at RA to use the data outside the Preservation Net ("Bevarandenätet"), although still within RAs internal net – or, if needs must, on a computer not connected to the Internet.
    b. Obtain permission from the legal owner of the information ("informationsägare", with Karin Åström-Iko as their representative) to use the data outside RA – that is, to use them in the SweGrid/SweStore infrastructures.

4. Cloud archive providers apart from SweStore (theoretical exercise only)
    a. Make an inventory of the cloud archive providers (CAP) in Sweden.
    b. Compile a list of requirements appropriate for the scenario. Observe that the time horizon is 5 years according to the scenario. Examples of requirements: guarantee authenticity, migrate files to new formats when the old formats become obsolete.
    c. Construct a questionnaire and send it to the CAPs.
    d. Evaluate the answers and determine if a commercial CAP is suitable for a *small* institution like that in the scenario.

5. Access
    a. Register the web site archive in the archive information system NAD ("Nationella arkivdatabasen", the National Archive Database). This is done indirectly via registration in the internal archive information system ARKIS (if not already done).
    b. Obtain a test person (like the one in the scenario). S/he does not have to know anything about NAD or anything technical about how the web archive is stored.
    c. Let the test person find the archive, open it and evalutate how easy it was to find and use this archive.
    d. Compare how the access to the web site works when you use the original one (www.linne2007.se) and when you use the format/tools chosen in 1a. (This test must be omitted if the original site ceases to be accessible).

1.3 DEPENDENCIES

The following chronological dependencies should be noted:

- 1b is dependent on 1a to be at least partly finished.
- 2b is dependent on 1b and 2a to be at least partly finished.
- 2c is dependent on 2b to be finished.
- 2d is dependent on 2c to be finished.
- 3b is dependent on 3a to be finished.
- 4c is dependent on 4a and 4b to be finished.
- 4d is dependent on 4c to be finished.
- 5b is dependent on 1, 2, 3 and 5a to be finished.
- 5c is dependent on 5b to be finished.
- 5d is dependent on 1b to be finished.

[A graphical description of the dependencies may be inserted here]

1.3.1 Activities that can start immediately

- 1a (Investigate the existing dissemination tools and formats …)
- 2a (Investigate all preparations that must be done to make usage of SweGrid and Swestore…)
- 3a (Obtain permission from the Electronic Archives section…)
- 4a (Make an inventory of the cloud archive providers (CAP) in Sweden.)
- 4b (Compile a list of requirements appropriate for the scenario…)
- 5a (Register the web site archive in the archive information system…)

2 TOOLS TO BE TESTED

The choice of tools to be tested and evaluated will be guided by the following appeal from WP3:

"The developed solutions need to be tested for their simplicity of installation, management and use."

A fourth criterium will be used: generality of solution.

See **DCH-RP_WP5_Scen-2-4_ID-ToolTests.pdf** for the results of the tool tests. There, these four criterias will be graded on a scale from 1 (very bad) to 5 (very good).

For each tool, there is a (postulated) reason for why that tool has been chosen for tests.

2.1 WARC Tools

For download and information, see: https://code.google.com/p/warc-tools/, http://www.ltu.se/centres/Centrum-for-langsiktigt-digitalt-bevarande-LDB/Bibliotek/Webb/Om-filformatet-WARC-och-verktyg-for-formatet-1.67311 (in Swedish)

“The main goal of WARC Tools is to facilitate and promote the adoption of the WARC file format for storing web archives by the mainstream web development community by providing an open source software library, a set of command line tools, web server plug-ins and technical documentation for manipulation and management of web archive files, or WARC files. WARC files are produced by web archiving crawlers, such as **Heritrix**, the open-source, extensible, Web-scale, archiving quality Web crawler developed by the Internet Archive with the Nordic National Libraries, and Hanzo's own commercial crawlers.”

Postulated reason: WARC is an ISO standard and widely used as a dissemination format.

2.2 Web Curator Tool

For download and information, see: http://webcurator.sourceforge.net/

“The Web Curator Tool (WCT) is an open-source workflow management application for selective web archiving. It is designed for use in libraries and other collecting organisations, and supports collection by non-technical users while still allowing complete control of the web harvesting process. It is integrated with the **Heritrix** web crawler and supports key processes such as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata.”

Postulated reason: oriented towards DCH sector and non-technical users. Integrated with Heritrix. Compared to simpler tools (like HTTrack) it has features like quality review and collection of descriptive metadata.

2.3 SWAT

For download and information, see: http://sourceforge.net/projects/swat-archiving/

"SWAT (Snappy Web Archiving Tool) is a tool designed for archiving web sites and displaying the archive in a simple way. Besides harvesting all files from the web site, SWAT generates snapshots of each page to TIFF files and describes the entire archive in a METS-file."

Postulated reason: a tool that has not been extensively tested but that may be good for dissemination purposes (and uses the metadata standard METS).

2.4 HTTRACK WEBSITE COPIER

For download and information, see: http://www.httrack.com/

"HTTrack is a free (GPL, libre/free software) and easy-to-use offline browser utility. It allows you to download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer."

Postulated reason: if it is real easy to use it may be worthwhile, but the target format may not be ideal for dissemination purposes.

2.5 Heritrix

For download and information, see: https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

"Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project."

Postulated reason: both WARC Tools and Web Curator Tool refer to Heritrix, and Web Curator Tool is even integrated with it.

3 ACCEPTANCE TEST FOR DOWNLOADED LINNÉ WEB SITE

This was a test for end-user acceptance. Originally, the test was formulated like this:

1. Let the test person find the archive Linnéjubileet, open it and evalutate how easy it was to find and use this archive
2. Compare how the access to Linnéjubileet works when you use the original one (www.linne2007.se) and when you use the format/tools chosen for scenario 2.4. (This test must be omitted if the original site ceases to be accessible)

For test #1, it was intended that the archived Linnéjubileet web site would have been registered in NAD, Nationella Arkivdatabasen (the Swedish National Archive Database), and the test person would evaluate how easy it was to find the Linnéjubileet web archive there.

There is several ways to find information about the site. One of them (currently, often the most easy way to find Riksarkivet's digital archives) is a temporary file on Riksarkivet's web site:

Figure 1: Linnéjubileet (1)

However, the most natural way to find it would be to look for it in NAD. Unfortunately, we thought the site hadn't been registered, so the evaluation was rather of how easy it was to ascertain that the web archive could *not* be found in NAD. After digging a little deeper, I found this page where the web site is one item out of many (B8):

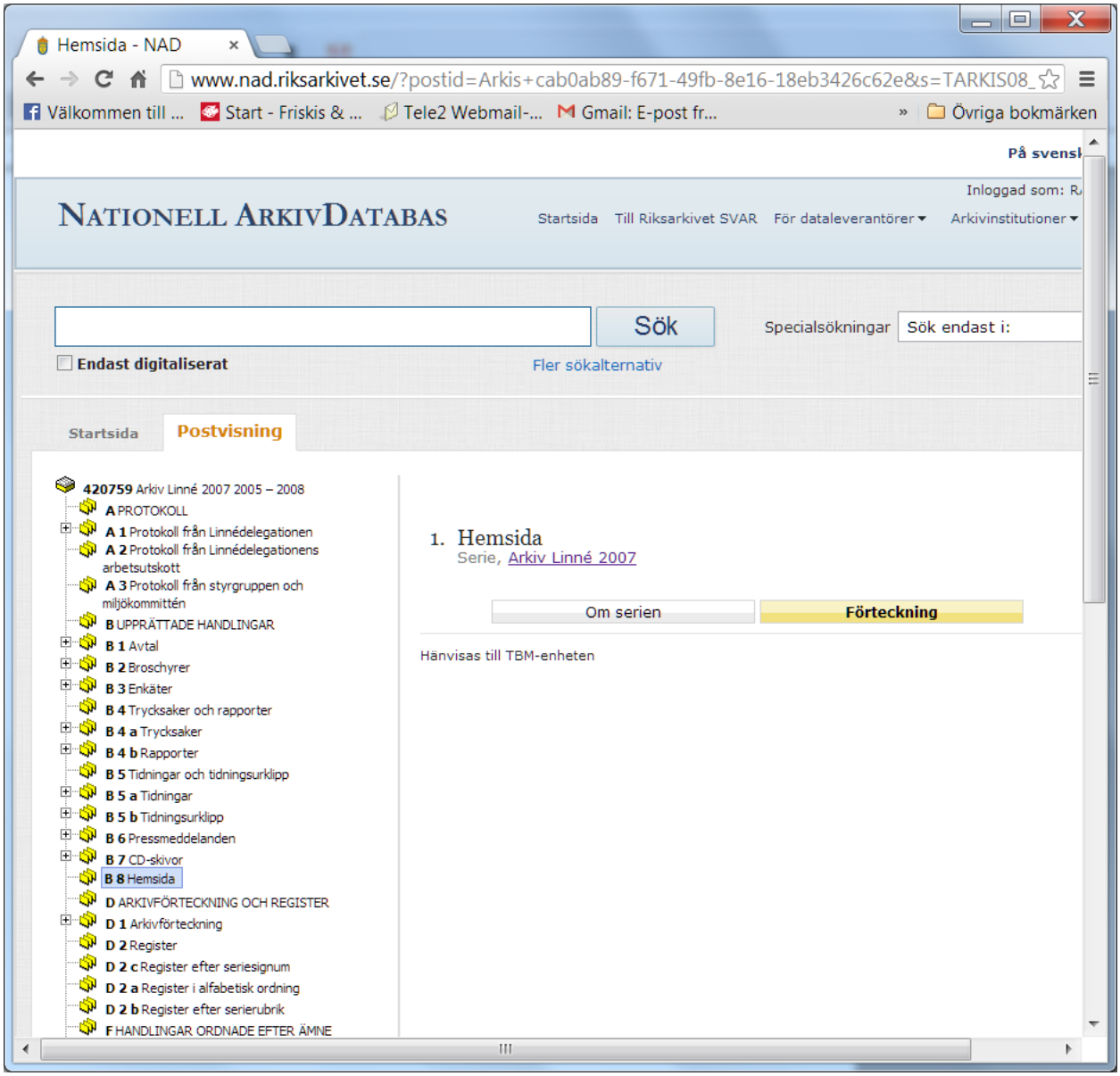

Figure 2: Linnéjubileet (2)

For test#2, it was intended that the test person would compare the experiences between the original site (http://www.linne2007.se/) and the archived version (archiving being made by **HTTrack**). However, since the original site is unreachable intermittently, and the test time was one of those times, this couldn't be done. Instead, the test person evaluated how believable it was that this could indeed be the real, original Linné site.

When the web archive itself would be opened, the test person was confronted with the folder where HTTrack had stored it:

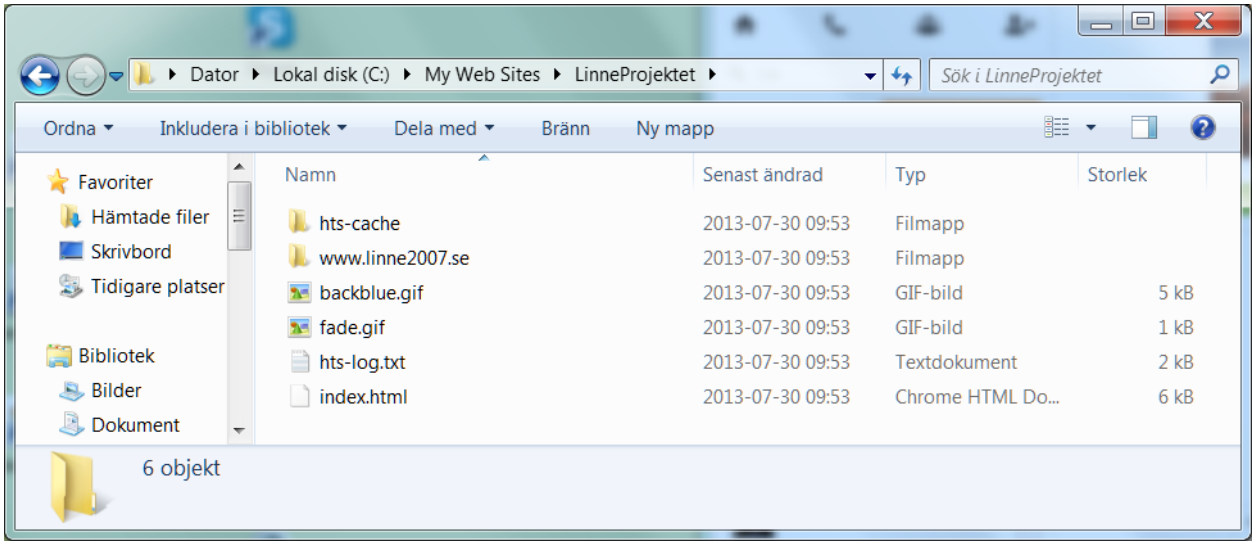

Figure 3: Download folder of the Linnéjubileet web site

Although the test person was not familiar with how to open a web site in this way (for example, that you can almost always open a site with the **index.html** file), by experimenting a bit she could open it quickly.

The test person thought that it was very believable that this could be the real site. However, HTTRack didn't manage to archive everything; for example, the English version of the site was missing (I checked later that it really was present in the original site). Some links were dead, and nothing happened when you pressed “Kontakt” (Contact), although it was later discovered that this was also the case in the original site. (Indeed, for finished projects, there is no need/possibility to have up-to-date contact information).