Automatic Identification and Preservation of National Parts of the Internet Outside a Country's Top Level Domain

Eld Zierau
Department of Digital Preservation
The Royal Library of Denmark
Søren Kierkegaards Plads 1
DK-1016 København K
ph. +45 91324690
elzi@kb.dk

ABSTRACT

Preservation of our cultural heritage on the Internet is increasingly in danger of getting lost due to the challenges faced when collecting it. An increasing amount of national webpages are moving to generic Top Level Domains like .com or .org. The movement is so fast that we are at risk of losing it, since we do not get in time to identify the change before it has disappeared again. Therefore this question becomes increasingly crucial for organizations covering digital national heritage including web archives for a specific country.

This poster presents the results from a research project that evaluated two different automated approaches to recognise webpages outside a country's Top Level Domain which are part the country's cultural heritage. One suggested approach has been to base extraction of national material on a snapshot of the entire Internet in form of a worldwide crawl. Another suggested approach is more silo oriented, based on harvests of web pages referred to by webpages within a National Top Level Domain.

More specifically the research project aimed to identify automatic procedures for evaluating the two suggested approaches, and for identifying Danish web content on websites outside the national Top Level Domain ".dk". The datasets used were links from a 30TB Danish 2012 bulk harvest and the 360 TB Internet Archive *wide-0005 crawl,* since these two harvests are comparable in time frame.

The poster will present

- The two methods and the difference in their results
- Indications that the two approaches find very different material
- The general method used to evaluate the nationality of web material over time

The general method mentioned here is important, since the very basis for any harvesting approach is defining a collection scope by deciding what is seen as national webpage. Automation of such definitions is far more difficult than originally anticipated. The automation here is based on a wide range of general criteria that are implemented (e.g. language recognition, national terms like 'je suis Charlie' or phone number patterns). An additional outcome of the project has been a generally applicable list of collection criteria, which is based on a cooperative effort between representatives within the fields of scholarship, the Danish web archive, and computer science.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

Preservation, web archive, collection strategies

Note

It should be noted that the last mentioned criteria method part has been presented at the RESAW 2015 conference, but in a closed forum, - and the first part with the results have been presented at the IIPC GA in a presentation, but not as a poster which opens a better possibility to discuss and understand in depth, as well as exchange ideas based on the results.