Aan de slag met webarchivering

Een checklist voor het starten met webarchivering

Deze checklist is door de NCDD (Nationale Coalitie Digitale Duurzaamheid) gemaakt ten behoeve van de workshop “aan de slag met webarchivering” als onderdeel van de studiedag “een web van webarchieven” in 2016. De checklist kan worden gebruikt door organisaties die van plan zijn websites (of sociale media) te gaan archiveren. Aan de hand van een aantal vragen in vier categorieën wordt je geleid door een voorbereidingsproces. De vragen helpen u bij het maken van beleid en het inrichten van processen. De vier stappen zijn:

A. Het maken van beleid ten aanzien van webarchivering
B. Het inrichten van processen
C. Het uitvoeren van processen
D. Het analyseren van het te archiveren materiaal

Door het beantwoorden van de vragen in deze vier stappen kan de aanpak voor webarchivering nader worden bepaald en ingericht. Deze checklist is een eerste versie en is zeker niet compleet. Aanvullingen en opmerkingen zijn dan ook zeer welkom. De expertgroep webarchivering zal deze gebruiken om een volgende versie van de lijst te maken. Discussie over deze checklist is ook welkom en kan via het kennisplatform webarchivering gevoerd worden.

Voor meer informatie:
Nationale Coalitie Digitale Duurzaamheid
Postbus 90407
2509 LK Den Haag
info@ncdd.nl

2017, Marcel Ras. Nationale Coalitie Digitale Duurzaamheid.

A. Beleid maken

1. Waarom webarchivering?

Beschrijf de redenen voor webarchivering: cultuurhistorisch belang, wettelijke bepalingen, wetenschappelijk gebruik. (vallen websites onder de archiefwet?)

2. webarchivering en collectiebeleid/selectiebeleid

Is er een koppeling met het collectie- of selectiebeleid van mijn organisatie? Kijk daarbij oa naar de te bewaren websites (thema's, alleen de website van uw eigen organisatie, regionaal, etc)

3. Wie zijn de beoogde gebruikers?

5. Met wie kan ik samenwerken?

Zijn er voorbeelden van gelijksoortige initiatieven waarvan ik kan leren of een samenwerking mee kan aangaan?

B. Inrichten van het proces

1. vaststellen om welke *webuitingen* het gaat

Wat gaan we archiveren? Websites, social media, … Om welke hoeveelheden gaat het?

2. Wie gaat dit doen?

Gaan we dit zélf doen binnen de organisatie of uitbesteden?

3. Wat heb ik daarvoor nodig?

Welke kennis heb ik nodig, welke personele inzet, en welke financiële middelen en technische infrastructuur?

4. in geval van zelf doen: welke tools heb ik nodig?

Tools voor beheer, binnenhalen, kwaliteitscontrole, toegang, duurzame opslag

5. welke rechten rusten er op de te selecteren websites?

6. wat is de frequentie van binnenhalen?

Hoe vaak worden de geselecteerde websites “geharvest”?

7. Lange termijn toegang

Hoe lang dienen de gearchiveerde websites bewaard te worden en hoe wordt duurzame toegang geregeld? Is dit gebaseerd op het duurzaamheidsbeleid van uw organisatie (preservation policy)

C. Uitvoeren van het proces

Hieronder zijn de volgende stappen te onderscheiden waarover vooraf over nagedacht moet worden en daar waar mogelijk ingeregeld.

1. Selecteren
    - Welke websites of andere webuitingen ga ik bewaren?
    - Moet ik de gehele website/alle informatie wel bewaren of zijn slechts bepaalde delen van belang ihkv bewijsvoering en verantwoording?
    - Waarop baseert u de selectie?
    - Voldoet dit aan he collectie- selectiebeleid?
    - Is dit gekoppeld aan selectielijsten?
2. Technische aspecten van het te archiveren materiaal inventariseren (*)
    - Inventarisatievragen zie onder inventariseren
3. Rechten en notificeren
    - hoe zit het met auteursrechten?
    - Hoe zit het met content die mogelijk inbreuk maakt op de persoonlijke levenssfeer?
    - Hoe ga ik de eigenaren van de geselecteerde websites benaderen?
    - Heb ik een “opt-out” message?
4. Binnenhalen
    - Ga ik zelf harvesten of ga ik dit uitbesteden?
    - Welke tools heb ik hiervoor tot mijn beschikking?
    - Wat zijn de kosten daarvan?
    - Wat zijn de technische en juridische beperkingen?
    - Met welke frequentie worden websites binnengehaald?
    - Wordt daarin gedifferentieerd?
5. Controleren van de kwaliteit
    - Hoe ga ik de kwaliteit van de binnengehaalde websites controleren?
    - Wat doe ik met geharveste sites die niet voldoen aan mijn kwaliteitseisen?
6. Opslaan
    - Hoe ga ik de binnengehaalde websites opslaan?
    - Hoeveel opslagruimte is daarvoor ter beschikking?
7. Duurzaam beheren
    - En hoe ga ik deze duurzaam beheren?
    - Voldoe ik daarvoor aan standaarden?
    - Is dit gekoppeld aan de preservation policy van mijn organisatie?
    - Moet het mogelijk zijn om reeds gearchiveerde sites te verwijderen uit het archief en zijn daar procedures voor?
8. Toegang regelen
    - Is het mogelijk om toegang tot de webcollectie te bieden?
    - Zo ja, hoe en wat zijn de beperkingen?
9. Overzicht van gearchiveerde websites
    - Is het mogelijk om een overzicht te geven van de gearchiveerde websites, bij voorkeur online?
    - Draagt deze bij aan de nationale catalogus van gearchiveerde websites in Nederland?

D. Inventarisatievragen

Deze bepalen mede de te kiezen aanpak en de wijze van het gebruik van de beschikbare tooling.

Inventariseren van de soorten content/informatieobjecten op de website en de opbouw van de website.

- Bevat de website interactieve JAVA-achtige onderdelen als scrolldown menu's, aanvink vakjes, dynamische infographics
- Bevat de website AV-content of niet. Zo ja staat er een filmpje van YouTube op de website? De harvester neem alleen gekoppelde bestanden op wanneer ze op dezelfde webserver staan als de hele website. Bovendien is er geen garantie dat YouTube eeuwig blijft bestaan, en zou een externe koppeling naar YouTube dus niet duurzaam zijn. Marktpartijen lijken hier wel aanvullende techniek voor te hebben ontwikkeld.
- Bevat de website zoekfunctionaliteit en zo ja biedt de website de optie om ook door alle beschikbare zoekresultaten te scrollen omdat deze op één webpagina gegroepeerd staan en benaderbaar zijn via een linkje? Of is het zoekfunctionaliteit van een website die op de server wordt uitgevoerd, bijvoorbeeld zoeken in een database? Die werkt in de meeste gevallen niet omdat de zoekmachine niet vastgelegd kan worden door de huidige crawltechniek;
- Bevat de website er social mediafeeds op de website of newsfeeds?
- Bevat de website een forum of reactie mogelijkheid op de website. Dit komt allemaal niet mee in een harvest en men moet zich daar vooraf bewust van zijn
- Bevat de website content achter wachtwoorden of persoonlijke accounts? Ook die content komt niet mee in een harvest
- Bevat de website links naar externe sites? Dan moet de harvesting tool daarop ingesteld zijn dat die externe links wel of niet gevolgd worden en hoe ver er links in die externe website gevolgd moeten worden
- Bevat de website een kalenderfunctie: kan een zogenaamde 'crawl trap' zijn in de zin van dat een harvesting tool dan oneindig content blijft binnen halen