Svein Arne Solbakk

Wdrożenie depozytu cyfrowego w Bibliotece Narodowej Norwegii

Biblioteka 18 (27), 235-246

2014

Artykuł został opracowany do udostępnienia w internecie przez Muzeum Historii Polski w ramach prac podejmowanych na rzecz zapewnienia otwartego, powszechnego i trwałego dostępu do polskiego dorobku naukowego i kulturalnego. Artykuł jest umieszczony w kolekcji cyfrowej bazhum.muzhp.pl, gromadzącej zawartość polskich czasopism humanistycznych i społecznych.

Tekst jest udostępniony do wykorzystania w ramach dozwolonego użytku.

MUZEUM HISTORII POLSKI

NR 18 (27) BIBLIOTEKA 2014 PL ISSN 1506-3615

P R Z E K Ł A D Y

SVEIN ARNE SOLBAKK

Wdrożenie depozytu cyfrowego w Bibliotece Narodowej Norwegii[1]

Streszczenie. Artykuł omawia pokrótce strategie przyjęte w Bibliotece Narodowej (Norwegia) dotyczące gromadzenia i zdobywania publikacji tworzonych oryginalnie w wersji elektronicznej bezpośrednio od wydawców, nawet jeśli sama publikacja została wydana drukiem lub zapisana na innych nośnikach. Ponadto przedstawiono zasady wprowadzania tych strategii w bibliotece, a także doświadczenia z faktycznego procesu implementacyjnego. Proces ten obejmuje zarówno stały kontakt i dialog z wydawcą, rozwiązania techniczne stosowane w cyfrowym odwzorowaniu materiału, jak i obsługę opracowania publikacji cyfrowych w samej bibliotece. W artykule omówione zostały konkretne procesy implementacyjne dotyczące czasopism. Warto jednak odnotować, że opisane zasady działania mają charakter uniwersalny.
Zaprezentowano skrótowo, w jaki sposób najważniejsze strony WWW z domeny „.no" (Norwegia) przechwytywane są obecnie z myślą o ich indeksacji i zachowaniu w zbiorach.
Na koniec omówione zostały zasady egzemplarza obowiązkowego dotyczące materiałów cyfrowych oraz tzw. web harvesting (technika gromadzenia materiałów w internecie na potrzeby ochrony zasobów cyfrowych), przedstawione w kontekście ambitnego programu digitalizacyjnego prowadzonego w Bibliotece Narodowej.

Słowa kluczowe: cyfrowy egzemplarz obowiązkowy, czasopisma, biblioteka cyfrowa, web harvesting.

Narodowa biblioteka cyfrowa

W roku 2006 Biblioteka Narodowa Norwegii przystąpiła do ambitnego programu digitalizacyjnego. Ostatecznym celem jest stworzenie cyfrowej

[1] S.A. Solbakk, *Implementation of digital deposit at the National Library of Norway*, http://library.ifla.org/992/1/107-solbakk-en.pdf [dostęp: 15.09.2014]. Copyright © 2014 by S.A. Solbakk. This work is made available under the terms of the Creative Commons Attribution 3.0 Unported License: http://creativecommons.org/licenses/by/3.0/.

biblioteki narodowej, która nie ograniczałaby się do udostępniania wyłącznie cieszących się dużą popularnością usług w sieci. Ukazuje to całkowicie nowe podejście do efektywnego działania biblioteki w wypełnianiu swojej funkcji jako biblioteki narodowej.

Aby urzeczywistnić tak ambitne plany, należy:

- zdigitalizować istniejące zbiory wydane drukiem,
- gromadzić cyfrowe oryginały publikacji wydanych drukiem,
- gromadzić publikacje wydane wyłącznie w formacie cyfrowym,
- negocjować z właścicielami praw autorskich w sprawach związanych z możliwością udostępnienia zbiorów w formatach cyfrowych,
- udostępniać zbiory cyfrowe użytkownikom wszędzie tam, gdzie się znajdują, kiedy tylko mają na to ochotę i w sposób, jaki sobie wybiorą.

Wszystkie te działania wymagają szczegółowego skupienia się na zadaniach i środkach, które muszą pozostać aktywne przez dłuższy czas.

Oczekuje się, że proces digitalizacji zbiorów niecyfrowych w Bibliotece Narodowej zajmie 25–30 lat. W prace digitalizacyjne różnego rodzaju materiałów zaangażowanych będzie ponad 50 osób pracujących na pełnych etatach. Obecnie 80% książek, tj. 375 000 tytułów, i 20% czasopism (16 000 000 stron), jakie publikuje się w Norwegii, jest na bieżąco digitalizowanych. Dodatkowo odwzorowano cyfrowo dużą liczbę periodyków, zdjęć, rękopisów, utworów muzycznych oraz innych nagrań dźwiękowych, a także filmów i audycji radiowych.

Wiele umów z właścicielami praw autorskich zawartych w ostatnim czasie umożliwia bibliotece udostępnianie dzieł również w formie cyfrowej. Projekt „Bokhylla" z roku 2012 i towarzysząca mu umowa z wydawcami przewiduje dostęp do dzieł chronionych prawem autorskim i daje Bibliotece Narodowej prawo udostępniania wszystkich książek opublikowanych w Norwegii w roku 2000 lub w latach wcześniejszych. Dostęp mają użytkownicy korzystający z adresów IP zarejestrowanych w Norwegii. W rezultacie obecnie już 175 000 tytułów książek dostępnych jest bezpłatnie, a zakłada się, że w 2017 roku 250 000 tytułów będzie udostępnianych na tych samych zasadach.

Biblioteka Narodowa podpisała także umowy z redakcjami gazet, na mocy których upoważniona jest do udostępniania zawartości 35 tytułów gazet codziennych w formie cyfrowej w bibliotekach na terenie Norwegii. Dziewięć z tych tytułów dostępnych jest każdemu użytkownikowi. Na mocy wspomnianych umów 50% stron zdigitalizowanego lub złożonego w formie cyfrowej materiału pochodzącego z gazet codziennych dostępnych jest w bibliotekach norweskich.

Kolejnym dobrym przykładem jest umowa, jaką podpisano z głównym nadawcą radiowym kraju, w rezultacie której biblioteka może obecnie

udostępniać bezpłatnie ponad 36 000 programów radiowych, w tym programy informacyjne i wiadomości emitowane na falach radiowych od lat 30. XX wieku do wczoraj.

Zasadniczym tematem niniejszego artykułu są dwa ostatnie działania przedstawione powyżej. Dalsza część tekstu poświęcona będzie egzemplarzowi obowiązkowemu w kontekście cyfrowego dziedzictwa kulturalnego.

Cyfrowy egzemplarz obowiązkowy

Pierwszy cyfrowy egzemplarz obowiązkowy w Bibliotece Narodowej Norwegii wprowadzono pod koniec roku 2004. Od tamtego czasu audycje radiowe nadawane przez publicznego norweskiego nadawcę radiowo-telewizyjnego gromadzone są jako pliki przesyłane w otoczeniu sieciowym, a transport fizycznych taśm został definitywnie zakończony. W roku 2008 przyjęto cyfrowy egzemplarz obowiązkowy obejmujący audycje telewizyjne pierwszego kanału TV. Obecnie składanie cyfrowego depozytu drogą internetową stosowane jest już w przypadku wszystkich ogólnokrajowych audycji radiowych i telewizyjnych w Norwegii. Co więcej, depozyt cyfrowy wprowadzany jest także dla lokalnych stacji norweskiego nadawcy państwowego.

Z przejścia na cyfrowy przekaz skorzystali nadawcy, ale też Biblioteka Narodowa. Transfer jest wydajniejszy i o wiele tańszy niż dotychczasowe przenoszenie danych na nośniki fizyczne, takie jak taśmy, i transportowanie ich metodą tradycyjną.

W przypadku materiałów wydawanych drukiem, bez względu na fakt wprowadzenia alternatywnego depozytu cyfrowego, Biblioteka Narodowa nadal jednak domaga się otrzymania od wydawcy wersji drukowanej dzieła. Nie jest zatem do końca tak oczywiste, co w takim razie wydawca zyskuje z cyfrowego depozytu. Co więcej, jeśli chodzi o czasopisma, Biblioteka Narodowa również chce mieć dostęp w bibliotekach na terenie Norwegii do wersji elektronicznej czasopism i gazet, aby zakończyć wykonywanie ich kosztownego mikrofilmowania.

Obowiązujące w Norwegii prawo dotyczące zasad egzemplarza obowiązkowego ogranicza obowiązkowe dostawy w ramach egzemplarza obowiązkowego tylko do formatu, w jakim materiał został faktycznie wydany. Tym samym Biblioteka Narodowa nie ma prawa żądać od wydawców plików w formie elektronicznej, które używane są w produkcji publikacji drukowanej. Aby więc otrzymać takie właśnie pliki, należy podpisać odrębne umowy z wydawcą dotyczące właśnie dodatkowego depozytu cyfrowego.

Biblioteka Narodowa zaprosiła więc wydawców norweskich gazet do współpracy w ramach projektu obejmującego zarówno digitalizację, jak i egzemplarz obowiązkowy. W tego typu współpracy zadaniem Biblioteki Narodowej jest zajęcie się procesami digitalizacyjnymi i konwersją OCR, przy jednoczesnym zobowiązaniu się do długotrwałego przechowywania i zabezpieczenia cyfrowych plików głównych (danych stałych) pochodzących zarówno z digitalizacji, jak i z depozytu cyfrowego. Wydawca gazety płaci 50% kosztów związanych z digitalizacją, w zamian otrzymuje kompletny egzemplarz cyfrowy. Po nawiązaniu współpracy wprowadzony zostaje także depozyt cyfrowy danego tytułu.

Jak do tej pory umowy, które zostały podpisane z wydawcami, obejmują 35 tytułów gazet. Cały czas prowadzone są negocjacje z innymi wydawcami, tak by dodać do tej listy kolejne. Na razie skoncentrowano się na największych tytułach gazet w Norwegii, ale jednocześnie chciano by uzyskać szeroką reprezentację elektronicznie dostępnych tytułów publikowanych na terenie całego kraju. Niektóre z umów ograniczają się wyłącznie do bieżącego depozytu cyfrowego, ale całkiem spora grupa obejmuje współpracę wydawcy także przy digitalizacji numerów archiwalnych gazet.

Chociaż 35 z grupy ponad 250 obecnie wydawanych tytułów nie wydaje się szczególnie dużą liczbą, to biorąc pod uwagę fakt, że umowy dotyczą najważniejszych tytułów, można powiedzieć, że reprezentują one jednak spory procent ogólnej liczby stron gazetowych publikowanych rocznie w Norwegii. To ważny fakt, bowiem digitalizacja wydania drukowanego na miejscu w bibliotece zostaje automatycznie wstrzymana z chwilą wejścia w życie umowy na depozyt cyfrowy z wydawcą. W konsekwencji oznacza to duże oszczędności dla Biblioteki Narodowej i jednocześnie możliwość przesunięcia środków na digitalizację wydań archiwalnych.

W celu uzupełnienia cyfrowego depozytu gazet Biblioteka Narodowa rozpoczęła też w roku 2012 działanie egzemplarza obowiązkowego obejmującego wydawnictwa książek elektronicznych. Obecnie e-booki w formatach ePub oraz PDF gromadzone są i przechowywane w cyfrowym repozytorium.

Rozwiązania strategiczne dla cyfrowego depozytu gazet

Wyznaczonym celem jest ustanowienie cyfrowego egzemplarza obowiązkowego dla większości aktywnych tytułów gazet ukazujących się w Norwegii. Zakładana rama czasowa na wykonanie prac to pięć

lat. Oznacza to, że biblioteka będzie w stanie monitorować i kontrolować cyfrowy przepływ pracy regularnie i codziennie dla około 250 tytułów gazet.

Należy mieć jednak świadomość, że wydawcy wykorzystują różne cyfrowe systemy produkcyjne i że zastosowali u siebie znacznie zróżnicowaną automatyzację procesów i organizację pracy. Aby uniknąć konieczności wprowadzenia aż 250 różnych rozwiązań obejmujących zasady depozytu cyfrowego, Biblioteka Narodowa opracowała standardy obowiązujące w cyfrowym depozycie gazet. Określają one:

– sposób oznaczania (nazywania) plików przeznaczonych do depozytu,
– format pliku,
– rozdzielczość obrazu,
– sposób przesyłu.

Każda storna gazety powinna jednocześnie być oddzielnym plikiem PDF, a ilustracje (obrazy) powinny zachować taką samą jakość jak te wykorzystane w oryginalnych plikach używanych do przygotowania druku wersji papierowej. Wszystkie pliki składające się na kompletne wydanie numeru powinny być zawarte w formacie tar (program komputerowy służący do archiwizowania plików, powszechny dla systemu Unix) albo w pliku skompresowanym (zip file). Spakowany plik powinien zostać umieszczony na serwerze FTP wydawcy i udostępniony Bibliotece Narodowej. W spakowanym pliku muszą znaleźć się wszystkie strony wydania gazety w jej wersji drukowanej.

Ścieżka UNC (ścieżka do folderu kopii dystrybucyjnej) dla pojedynczej strony będzie więc miała następującą postać:

```
<newspaper-name_subname_zone_date_volume_number_issue>-<sequencenumber_pagenumber_appendixname>.pdf
```

Przykład – pierwsza strona dodatku z wiadomościami do porannego wydania gazety codziennej „Aftenposten" z 7 czerwca 2006 roku, do wydania krajowego, strefa 1:

```
aftenposten_morgen_1_20060607_147_253_1-1_001_nyheter.pdf
```

Zatem konwencja nazewnicza dla spakowanego pliku wygląda następująco:

```
<newspaper-name_subname_zone_date_volume_number_issue>.zip/.tar
```

Wszystkie znaki pisane są od małej litery, nie powinno też być żadnych podkatalogów. Nazwy plików zawierają wystarczające informacje do odtworzenia drukowanej wersji gazety w środowisku cyfrowym.

W celu zachowania wymaganej jakości zdefiniowano także zalecaną opcję do wykonania md5 checksum (sprawdzenia sumy kontrolnej pojedynczego pliku) spakowanego wydania numeru oraz dołączenia go w oddzielnym pliku wraz z listą wszystkich nazw plików zawartych w spakowanym wydaniu numeru. A zatem ścieżka UNC dla tego pliku wyglądałaby następująco:

`<newspaper-name_subname_zone_date_volume_number_issue>.md5`

Jeśli któryś ze zdefiniowanych elementów wzorca nie dotyczy danego czasopisma, wstawia się symbol „null". W definiowaniu konwencji nazewniczej dla plików gazet jako studium przypadku użyto tytułu gazety o najbardziej rozbudowanej i złożonej strukturze. Większość gazet nie ma podtytułów lub nie ma wydań regionalnych.

Z reguły transfer każdego tytułu odbywa się codziennie z serwera FTP wydawcy. Jednakże wymagane jest, aby wydawca posiadał wystarczającą ilość miejsca na swoim serwerze FTP, tak aby mogły się tam zmieścić wydania numerów z ostatnich 14 dni. W ten sposób zapewniony zostaje niewielki bufor, jeśli dzieje się coś nieoczekiwanego.

Pierwszym krokiem po transferze pliku jest walidacja spakowanych numerów gazet. Jeżeli jest dostępny opcjonalny plik md5, to md5 checksum dla otrzymanego spakowanego pliku porównywane jest z sumą kontrolną w pliku md5 file, a lista zawartych plików – z plikami otrzymanego pakietu danych. Kolejnym działaniem jest weryfikacja poprawności formatu PDF za pomocą JHOVE (2014). Jeśli wypada ona pomyślnie dla kompletnego pakietu, numer wydania gazety zostaje usunięty z serwera FTP wydawcy. Jeśli są podejrzenia, że coś jest nie tak, opracowanie numeru zostaje wstrzymane i następuje kontakt z wydawcą w celu uzyskania poprawnych plików.

Następny krok to uruchomienie każdego z plików numeru gazety za pomocą docWorks (2014), tak aby wydobyć tekst ze stron i sformatować go do formatu XML/ALTO (2014). Większość stron będzie miała tekst ukryty użyty bezpośrednio, ale często niektóre części zawierają również elementy graficzne. W takich przypadkach elementy te przepuszczane są przez program dokonujący optycznego rozpoznawania znaków (OCR), aby wydobyć z nich tekst. Na ogół dotyczy to ogłoszeń i reklam. Pliki XML/ALTO używane są zarówno do indeksacji tekstu z gazet, która później pozwala na przeszukiwania za pomocą usług oferowanych przez bibliotekę cyfrową, jak i do umożliwienia podświetlenia (wytłuszczenia) poszukiwanych słów lub fraz w tekście.

Po obróbce w docWorks generowane są tzw. access quality files JPEG2000 (2007) dla każdej strony w celu uzyskania możliwości obróbki

zdigitalizowanego materiału i gazet zdeponowanych cyfrowo w serwisie biblioteki cyfrowej.

Następnie wszystkie pliki związane z danym wydaniem numeru gazety zawijane są w METS container (2014), a metadane zostają wyodrębniane i umieszczone w bazie Mavis (2014), tj. bazie danych wykorzystywanej przez Bibliotekę Narodową Norwegii dla całej gamy różnego typu obiektów cyfrowych. Kompletny METS container jest inkorporowany do cyfrowej długoterminowej infrastruktury przechowywania danych, pliki dostępowe są transferowane do serwera obrazowego (image server), a pliki METS i XML/ALTO są wykorzystywane do umożliwienia pełnej przeszukiwalności danego numeru gazety oraz do udostępnienia w serwisach biblioteki cyfrowej.

Jeżeli dana gazeta nie posiada już chroniących ją praw autorskich lub podpisano umowę zezwalającą na dodatkowe usługi, tekst staje się widoczny i udostępniony także przez zewnętrzne wyszukiwarki, a usługa oferowana przez serwis obejmuje również możliwość transferu danych w wersji PDF wraz z tekstem ukrytym. Format PDF generowany jest „w locie" z plików JPEG2000 i odpowiadających im plików XML/ALTO. Do takiej sytuacji dochodzi jednak rzadko i rzadko jest to opcja dla cyfrowo deponowanych gazet, choć zdarza się o wiele częściej w przypadku odwzorowanych cyfrowo gazet historycznych.

Dotychczas cyfrowy depozyt wprowadzono dla 15 tytułów gazet. Przyjęcie cyfrowych zalgorytmizowanych procedur wydawcy obejmujących zespół czynności koniecznych do dostarczenia egzemplarza do Biblioteki Narodowej, tak by spełniał wszystkie zdefiniowane standardy, zajmuje jednak sporo czasu. Ponadto wszelkie zmiany w zautomatyzowanych procedurach wydawcy często przynoszą nieoczekiwane wyzwania dla biblioteki.

Zalgorytmizowany odpowiedni zespół czynności w Bibliotece Narodowej został już w sporym zakresie zautomatyzowany. A jednak pewne krytyczne elementy procesu nadal wymagają kontroli ludzi. Po pierwsze, w liczbie 250 aktywnych tytułów znajdujemy dość szeroki zakres częstotliwości publikacyjnej. Niektóre tytuły mają kilka wydań dziennie, inne ukazują się tylko raz w tygodniu. Do tej pory monitorowano to ręcznie w zależności od tego, czy wydanie elektroniczne danego tytułu gazety lub czasopisma otrzymywano według oczekiwanego wzoru ukazywania się tytułu. Jeśli oczekiwany numer gazety nie pojawiał się na czas, kontaktowano się z wydawcą, aby wyjaśnić problem. Taki tryb postępowania jest możliwy przy 15 tytułach, ale byłby niezwykle czasochłonny przy niespełna 250 tytułach.

Co więcej, pewne odchylenia w przyjętych procedurach, które mogą doprowadzić do zatrzymania numeru gazety, jak do tej pory nie zostały

zgłoszone przez system. Trzeba było jednak sporo wysiłku, aby konsekwentnie i regularnie kontrolować automatyzację procesów w celu natychmiastowego wykrycia jakichkolwiek odchyleń od normy i rozwiązania problemów.

Aby przeskalować depozyt cyfrowy tak, by obejmował wszystkie aktywne tytuły czasopism, opracowano nowy system, którego zadaniem jest wspomaganie depozytu cyfrowego czasopism. Obecnie system jest testowany, a jego wdrożenie planowane jest na wrzesień 2014 roku.

Oczekiwane częstotliwości publikacyjne dla każdego z tytułów wprowadzane są do systemu komputerowego, a jego zadaniem jest monitorowanie poprawności deponowania każdego tytułu pod względem przyjętych oczekiwań. Jeśli zdarzają się odchylenia, system automatycznie wysyła wiadomość mailową do wydawcy, zamawiając dostarczenie oczekiwanego numeru.

Ponadto system kontroluje poszczególne fazy zespołu cyfrowych procedur i, jeśli tylko natrafi na jakieś odchylenia od normy, uruchamia alarm. Działanie systemu można śledzić na konsoli, która umożliwia podgląd wszystkich operacji systemu, tzn. jeśli jakakolwiek procedura deponowania tytułów została naruszona, informacja o tym pojawia się natychmiast na ekranie. Co więcej, istnieje możliwość monitorowania aktualnego statusu każdego tytułu. W czasie rozwiązywania danego problemu czy sytuacji odbiegającej od normy operator może ustawić ich nowy status, po to aby poinformować system, że dany błąd właśnie jest naprawiany.

Doświadczenia z nowym systemem są bardzo obiecujące i z niecierpliwością czekamy na jego wdrożenie jesienią.

Plany na przyszłość

Biblioteka Narodowa negocjuje obecnie z grupą wydawców skonsolidowanych tytułów gazet (około 80 tytułów). Oczekujemy, że dojdzie do podpisania umowy dotyczącej depozytu cyfrowego w roku 2014. Następnie depozyt cyfrowy obejmujący 80 nowych tytułów zostanie wprowadzony w ciągu kolejnych dwóch do trzech lat. Będzie to stanowiło kamień milowy w naszym programie cyfrowego depozytu czasopism i pozwoli na znaczną relokację środków na digitalizację czasopism historycznych.

Jednocześnie dokonywany jest przegląd i korekta ustaleń prawnych dotyczących zasad norweskiego egzemplarza obowiązkowego. Oczekuje się, że nowa interpretacja prawna włączy do egzemplarza obowiązkowego oryginalną wersję cyfrową dzieła wydawanego następnie drukiem. Wprowadzenie stosownych zapisów umożliwi Bibliotece Narodowej

zwiększenie depozytu cyfrowego w szybszym tempie. Bez względu na wynik przyszłych ustaleń Biblioteka Narodowa nadal będzie koncentrowała się na doprowadzaniu do sporządzania nowych umów z wydawcami dotyczących wprowadzenia cyfrowego depozytu na jak największą liczbę aktywnych tytułów czasopism ukazujących się w Norwegii.

Planuje się rozszerzenie funkcjonalności nowego systemu wspomagającego, tak aby obejmował obsługę innych typów publikacji cyfrowych. Już podjęto prace nad przystosowaniem systemu do obsługi cyfrowego depozytu materiału radiowego i telewizyjnego. Liczba ogólnokrajowych programów radiowych i telewizyjnych znacząco wzrosła w ciągu ostatnich lat i stąd potrzeba rozszerzenia systemu wspomagającego.

Następnym krokiem może okazać się konieczność adaptacji systemu do obsługi depozytu czasopism. Udało się już podpisać umowy na digitalizację numerów archiwalnych z kilkoma norweskimi wydawcami czasopism. Ich cyfrowy depozyt wydaje się oczywisty.

Gromadzenie i przechowywanie materiałów dostępnych online

Zachowanie i ochrona narodowego dziedzictwa kulturowego Norwegii i dorobku kulturowego dostępnego w sieci wymaga wprowadzenia całkowicie nowego zestawu narzędziowego. Już od 15 lat Biblioteka Narodowa jest aktywnie zaangażowana w prace nad rozwiązaniami dotyczącymi technik gromadzenia materiału internetowego i jego późniejszego przechowywania. Pierwsze pełne przeszukanie domeny „.no" (Norwegia) przeprowadzono w roku 2001. Do 2003 roku skandynawskie biblioteki narodowe współpracowały nad tym zagadnieniem, prowadząc projekt Nordic Web Archive. W roku 2003 kraje skandynawskie wspólnie przystąpiły do międzynarodowej inicjatywy tworzącej konsorcjum International Internet Preservation Consortium (IIPC) przy współudziale innych bibliotek narodowych i Internet Archive. Z biegiem czasu IIPC (2014) stał się międzynarodową siłą napędową prac związanych z archiwizowaniem stron internetowych i dzisiaj liczy już 49 instytucji członkowskich z 25 krajów, w tym biblioteki narodowe, uniwersyteckie, regionalne, a także archiwa i niektórych dostawców usług.

Wiele z odpowiednich narzędzi zostało opracowanych przez członków IIPC i większość z nich opisana jest na stronie domowej IIPC, która dostępna jest w internecie. Sporą część narzędzi udostępniono na zasadach open source. Najważniejszym narzędziem do przeszukiwania i archiwizowania stron internetowych jest oprogramowanie Heritrix, a do

umożliwiania dostępu do archiwów internetowych OpenWayback. Oba zostały opracowane przez Internet Archive, przy współpracy innych członków IIPC. Obecnie IIPC kontynuuje swoje działania i zamierza przyjąć na siebie większą odpowiedzialność za przyszły rozwój dalszych istotnych narzędzi do archiwizowania stron internetowych.

W chwili obecnej norweskie archiwum stron internetowych obejmuje ponad 8 miliardów plików. Nie wykonaliśmy jeszcze pełnego przeszukania domeny „.no" od roku 2008, a to za sprawą toczącej się i niezakończonej debaty politycznej dotyczącej zasad prywatności w sieci. Obecnie około 1000 stron harwestowanych jest bardzo często. Strony domowe przeszukiwane są co godzinę, podczas gdy reszta stron w regularnych odstępach czasu. Oczywiście, aby tego dokonywać, musimy informować właściciela strony internetowej o naszym zamiarze, zanim jeszcze przystąpimy do harwestowania strony.

Oczekuje się, że wspomniane przewidywane uzupełnienia dotyczące ustaleń prawnych egzemplarza obowiązkowego w Norwegii powinny wyjaśnić i sprecyzować, w jaki sposób, przy instytucjonalnym archiwizowaniu stron internetowych, należy stworzyć procedury, które brałyby pod uwagę problem prywatności ich zawartości. Mając to na względzie, Biblioteka Narodowa przygotowuje się właśnie do przeskalowania swoich prac, tak by obejmowały zbieranie informacji o stronach WWW dla wszystkich stron internetowych z domeny „.no". Ponadto Biblioteka Narodowa podpisała już umowy z kilkoma właścicielami stron dotyczące zarówno archiwizowania tych stron, jak i udostępniania tych, które wcześniej zostały już zarchiwizowane. Tym samym Biblioteka Narodowa umożliwi, ograniczony jeszcze, otwarty dostęp do archiwów sieciowych tych stron, a także do zarchiwizowanych stron internetowych agencji rządowych.

Podsumowanie

Biblioteka Narodowa angażuje sporą część zasobów ludzkich do obsługi egzemplarza obowiązkowego obejmującego publikacje wydane drukiem. Niestety, przy wprowadzeniu cyfrowego deponowania materiałów nadal istnieje potrzeba obsługi publikacji drukowanych. W konsekwencji brakuje pracowników, którzy potrzebni są do nowych zadań związanych z obsługą cyfrowego egzemplarza obowiązkowego. Choć wiele z działań zostało zautomatyzowanych, trzeba przyznać, że depozyt cyfrowy również wymaga alokacji zasobów ludzkich do jego obsługi.

Wprowadzając nowe narzędzia obsługujące depozyt cyfrowy gazet, dokonano więc dużego kroku do przodu w przystosowaniu biblioteki

do przeskalowania w przyszłości szerzej rozumianego depozytu cyfrowego. A przecież nadal będzie istniała potrzeba ścisłego monitorowania depozytu cyfrowego, z punktu widzenia zarówno perspektywy rozwoju i utrzymania systemów ICT, jak i zapewnienia jakości obsługi przyszłego egzemplarza obowiązkowego w bibliotece.

Do czasu kiedy ustalenia prawne na temat egzemplarza obowiązkowego nie zostaną uzupełnione o odpowiednie punkty dotyczące oryginałów cyfrowych publikacji drukowanych, istnieje także potrzeba skoncentrowania się na negocjacjach i ustaleniach zawartych w osobnych umowach z wydawcami, tak by być przygotowanym na otrzymywanie tych wersji cyfrowych.

Przeskalowanie depozytu cyfrowego znacząco ułatwia realokację naszych środków przeznaczonych na bieżącą digitalizację na dalsze prace związane z archiwizacją materiałów historycznych. W ten sposób przybliżamy się do osiągnięcia naszego celu, jakim jest stworzenie narodowej biblioteki cyfrowej!

Ponadto nadal istotną sprawą pozostaje ochrona i zabezpieczenie przynajmniej najważniejszych treści narodowej domeny internetowej. Krótka historia światowego zasobu WWW pokazuje nam, że ta część spuścizny kulturowej narodów, która jest udostępniana przez internet, zwykle jest bardzo ulotna i ma krótki żywot. Dlatego też Narodowa Biblioteka Norwegii będzie nadal odgrywać bardzo aktywną rolę w IIPC, po to aby wspierać międzynarodowe zainteresowanie zbieraniem informacji o stronach WWW, ich indeksacją i archiwizowaniem narodowego dziedzictwa kulturowego umieszczonego na stronach internetowych, jak również udostępnianie archiwów sieciowych dla celów badawczych i dokumentacyjnych.

Przeł. Tomasz Olszewski

SVEIN ARNE SOLBAKK

Implementation of digital deposit at the National Library of Norway

Abstract. The paper will outline the strategies at the National Library of Norway for capturing digital born publications directly from the publishers, even though the publication itself is on paper or other physical media. Also the actual implementation

of these strategies and lessons learned in the process will be presented. This includes both the dialog with the publishers, the technical solutions for digital capture, and the processes for handling digital publications within the library. The paper will focus on a concrete implementation for newspapers. However, the principles are general. The paper will also briefly cover how the most important parts of the Norwegian Internet are currently captured for preservation. Finally, the digital deposit and the web harvesting will be placed into the context of the ambitious digitization program at the National Library of Norway.

Key words: digital deposit, newspapers, digital library, web harvesting.