FIDES – Biuletyn Bibliotek Kościelnych
nr 1-2 (22-23) / 2006, s. 47-56
ISSN 1426-3777

PETR ŽÁK[1]

WebArchiv – CZESKI PROJEKT ARCHIWIZACJI PUBLIKACJI INTERNETOWYCH[2]

Czeskie bibliotekarstwo w ostatnich dziesięciu latach szybko się rozwija i w niektórych obszarach osiąga wyniki na europejskim, a nawet światowym poziomie. Warto więc śledzić poczynania naszych południowych sąsiadów i zapoznać się bliżej z funkcjonowaniem czeskich bibliotek. Zainteresowanych tym tematem odsyłam do dwuczęściowego artykułu pt. *Sąsiedztwo warte współpracy*, opublikowanego w elektronicznym *Biuletynie EBIB*. W tym miejscu tylko wspomnę o jego zawartości, żeby zachęcić do przeczytania całego tekstu.

W pierwszej części artykułu, zatytułowanej *Czeskie biblioteki w świetle źródeł internetowych*[3], przedstawiono ważniejsze, po-

[1] Petr Žák – Absolwent Instytutu Bibliotekoznawstwa i Informacji Naukowej przy Wydziale Filozofii Uniwersytetu Karola w Pradze. Od 1979 r. pracuje w Bibliotece Narodowej w Warszawie, gdzie między innymi, przez 8 lat kierował Zakładem Uzupełniania Zbiorów. Obecnie jest bibliotekarzem modułowym – opiekuje się modułem gromadzenia systemu Innopac/Millennium i jego użytkownikami. Poza tym pomaga w BN przy gromadzeniu i opiniowaniu czeskich i słowackich publikacji.

[2] Artykuł ten, choć nie dotyczy bezpośrednio bibliotek kościelnych publikujemy w pewnym związku z artykułem księdza Krzysztofa Goneta „Wiara, biblioteki i Internet...” zamieszczonym w naszym Biuletynie w nr 1-2/1997 na s. 200-204. Projekt naszych południowych sąsiadów – WebArchiv jest bardzo interesującą próbą rozwiązania problemu sygnalizowanego przez Ks. Goneta w części IV tamtego artykułu zatytułowanej „Co pozostanie po naszej cywilizacji?”. Może Federacja FIDES powinna obecnie poważnie podjąć temat archiwizacji katolickich stron internetowych w Polsce opierając się na wzorcach z Czech?

wszechnie dostępne internetowe źródła informacji, które pozwalają bliżej zapoznać się z organizacją, funkcjonowaniem i rozwojem bibliotek w Republice Czeskiej. Podano adresy URL i krótkie charakterystyki portali i serwisów informacyjnych, przeznaczonych dla bibliotekarzy, bibliotek i ich użytkowników. Wspomniano też o elektronicznych czasopismach bibliotekarskich i innych webowych źródłach, poświęconych węższym zagadnieniom lub konkretnym przedsięwzięciom, np. dużym corocznym imprezom krajowym i międzynarodowym, organizowanym w Czechach, podczas których prezentowane są najnowsze osiągnięcia i dyskutowane bieżące problemy czeskiego bibliotekarstwa. Zwrócono uwagę, że czeskie bibliotekarskie źródła internetowe są ze sobą powiązane oraz zawierają również informacje nt. krajowej i międzynarodowej współpracy bibliotek oraz ich uczestnictwa w różnych programach i projektach.

Drugą część artykułu poświęcono „…wybranym dokumentom o charakterze strategicznym i koncepcyjnym oraz programom i projektom…”[4], które spowodowały dynamiczny rozwój czeskich bibliotek i ich wzajemnej współpracy w ostatnich latach. Zauważono przy okazji, że niektóre z tych długofalowych przedsięwzięć, zwłaszcza te związane ze stosowaniem nowoczesnych technologii, mogą stanowić interesującą przestrzeń do naśladowania lub do poszerzenia współpracy międzybibliotecznej w naszym środkowoeuropejskim regionie. Próbowano też odpowiedzieć na pytanie o przyczyny sukcesów odnoszonych przez czeskich bibliotekarzy. Jednocześnie zwrócono uwagę na mechanizmy współpracy bibliotek oraz zasady i ograniczenia finansowania programów i projektów w Czechach.

Z lektury ww. artykułu wynika, że szczególnie warte uwagi są te rozwojowe projekty i przedsięwzięcia, które są związane ze stosowaniem nowoczesnych technologii, takie jak:

- retrokonwersja katalogów bibliotecznych z zastosowaniem oryginalnej technologii *RETROKON*, pozwalającej również, dzięki dodatkowym narzędziom RIS i NRIS, racjonalnie sterować całym proce-

[3] Žàk, Petr [Dok. elektr.] (2006). Sąsiedztwo warte współpracy. Cz. 1. Czeskie biblioteki w świetle źródeł internetowych. *Biuletyn EBIB* nr 8(78). Tryb dostępu: http://www.ebib.info/2006/78/a.php?zak [odczyt: 26.06.2007].

[4] Žàk, Petr [Dok. elektr.] (2007). Sąsiedztwo warte współpracy. Cz. 2. *Biuletyn EBIB* nr 5(86). Tryb dostępu: http://www.ebib.info/2007/86/a.php?zak [odczyt: 6.08.2007].

sem w jednej bibliotece, jak i całym projektem ogólnokrajowym (bliżej zob. http://retrokon.nkp.cz i http://visk.nkp.cz/VISK5.htm)

- digitalizacja dokumentów i tworzenie bibliotek cyfrowych (zwłaszcza znany już w świecie system *Manusriptorium*, który przekształca się w „wirtualne środowisko badawcze" i projekt międzynarodowy UE – *ENRICH* (bliżej zob. http://www.manuscriptorium.com ; http://digit.nkp.cz/projekty/VZ2004_2010/2006/ZPRAVA2006.pdf)
- archiwizowanie i udostępnianie krajowych źródeł elektronicznych (internetowych) – *WebArchiv* (bliżej zob. WebArchiv http://www.webarchiv.cz).

Szczególnie interesujący – przynajmniej z polskiego punktu widzenia – wydaje się ostatni z tych projektów, *WebArchiv*, gdyż temat selekcji i archiwizacji publikacji elektronicznych, dostępnych online, nie został dotychczas w Polsce (Bibliotece Narodowej czy innej książnicy gromadzącej egzemplarz obowiązkowy) podjęty, nawet na poziomie analitycznym i koncepcyjnym. Jest już jednak dostrzegany i to również w kontekście gromadzenia oraz przechowywania egzemplarza obowiązkowego[5].

Czeski projekt *WebArchiv* stanowi część szerszego planu działań, obejmujących archiwizowanie różnego rodzaju dokumentów, w tym elektronicznych. Aczkolwiek został zainicjowany już w 2000 r., wpisuje się odpowiednio do nieco później przyjętej *Koncepcji stałego przechowywania zbiorów tradycyjnych i elektronicznych dokumentów w bibliotekach RCz do roku 2010*, która z kolei powstała w wyniku realizacji jednego z zadań sformułowanych w *Koncepcji rozwoju bibliotek w Republice Czeskiej na lata 2004-2010*. Warto przypomnieć, że w ramach *Koncepcji stałego przechowywania zbiorów tradycyjnych i elektronicznych dokumentów w bibliotekach RCz do roku 2010* podzielono dokumenty na cztery grupy: dokumenty tradycyjne, digitalizowane dokumenty historyczne, digitalizowane dokumenty nowsze oraz publikowane dokumenty elektroniczne. Ten

[5] Chociażby na spotkaniu przedstawicieli bibliotek, otrzymujących egzemplarz obowiązkowy, które miało miejsce w Bibliotece Narodowej w Warszawie na początku czerwca b.r. Mówiono tam m.in. o potrzebie stworzenia centralnego repozytorium gazet i czasopism elektronicznych. Notatkę ze spotkania można znaleźć w witrynie elektronicznej BN – w rubryce Aktualności na stronie http://www.bn.org.pl/index.php?id=1&archiwum

podział odzwierciedla specyfikę poszczególnych grup dokumentów, które, według twórców *Koncepcji stałego przechowywania zbiorów* należy w przyszłości integrować do jednego, przyjaznego dla użytkownika środowiska, które będzie miało większy zasięg i asortyment usług od tych, jakie proponują dzisiejsze biblioteki, portale, wyszukiwarki etc. (przykładem tworzenia takiego środowiska „wyższego rzędu" jest system Manuscriptorium, rozwijany w NK ČR). W *Koncepcji stałego przechowywania zbiorów* starano się przede wszystkim sformułować prawne, organizacyjne oraz techniczne warunki dla gromadzenia, archiwizowania oraz udostępniania publikowanych elektronicznych i zdigitalizowanych dokumentów w najbliższych latach.[6]

Kwestia egzemplarza obowiązkowego publikacji elektronicznych (internetowych) nie została w Czechach uregulowana prawnie, nie rozwiązano też odpowiednio problemu udostępniania tych wydawnictw w świetle prawa autorskiego. Nie przeszkodziło to Czechom podjąć na przełomie 20 i 21 w. tematu „narodowego web-archiwu" i potraktować go jako integralną część Narodowej Biblioteki Cyfrowej, w skład której wchodzą również inne (zdigitalizowane) dokumenty elektroniczne, objęte odrębnymi projektami jak *Kramerius* i *Manuscriptorium*.[7] Zdawano sobie od samego początku sprawę ze złożoności i „płynności" materii – z zagranicznych szacunków wynikało, że ponad 90% ogólnej liczby publikacji dostępnych w Internecie stanowią dokumenty istniejące tylko w postaci elektronicznej (tzw. digital born), przy czym ok. 40% publikacji w ciągu roku znika z sieci, a kolejnych 40% – zmienia formę. Tylko 20% sieciowych dokumentów elektronicznych jest po roku dostępnych w pierwotnym kształcie. Żeby ściągnąć i zapisać

[6] Wyraźnie ale stwierdzono – i to na samym początku opracowania - że zasadniczą sprawą jest zabezpieczenie środków (min. 214 mln koron na pięć lat). To mocne akcentowanie materialnego zabezpieczenia zamierzonych działań wynikało z nienajlepszych doświadczeń czeskich bibliotekarzy w uczestnictwie w długookresowych programach i projektach. Bliżej zob. Žák Petr [Dok. elektr.] (2007).

[7] Warto zauważyć, że problem gromadzenia, przechowywania i udostępniania źródeł elektronicznych publikowanych w Internecie, traktowanych jako część dziedzictwa narodowego czy narodowego zasobu informacji, jest w świecie podejmowany dopiero od połowy lat 90-tych – informację o „Web archiving" w poszczególnych krajach zob. http://www.nla.gov.au/padi/topics/92.html).

wszystkie (a nawet tylko wybrane) czeskie źródła internetowe, należało być przygotowanym do obsługi i udostępniania zasobu o wielkości setek GB, przy czym najpierw należało je zidentyfikować, opisać, zindeksować itd.[8] Trzeba było podjąć szereg strategicznych decyzji w sprawach identyfikacji i doboru źródeł, skanowania (penetracji) Internetu i ściągania zasobów sieciowych, ich archiwizowania, aktualizacji itd.[9] Do tego wszystkiego potrzebne było odpowiednie, wciąż doskonalone oprogramowanie jak też inne narzędzia, pozwalające projekt przetestować, wdrożyć i dalej prowadzić.

To duże wyzwanie zostało w Czechach podjęte w postaci projektu *WebArchiv*, w którym wydawnictwa elektroniczne dostępne online (internetowe) potraktowano jako część narodowego zasobu informacyjnego (czy też dziedzictwa kulturowego) oraz bibliografii narodowej. Projekt powstał w 2000 r. w ramach programu badań i wdrożeń (R&D) Ministerstwa Kultury *Rejestracja, ochrona i udostępnianie krajowych źródeł elektronicznych w sieci Internet*. Od początku jest realizowany przez NK ČR w Pradze przy współpracy z Morawską Biblioteką Ziemską i Instytutem Techniki Komputerowej Uniwersytetu im. Masaryka w Brnie. Instytucje te razem, powoli budują archiwum czeskich źródeł informacji internetowej oraz starają się zapewnić również ich udostępnianie na zasadach uwzględniających realia krajowe.

---

[8] M.in. zob. Celbová, Lidmila; Simonová, Markéta, Žabička, Petr (2003). WebArchiv – od výzkumu k (tvrdé) realitě. *Knihovny Současnosti,* s. 70-81.

[9] Biblioteki podchodzą do tych spraw różnie. I tak Australijska BN archiwizuje tylko te webowe źródła, które wcześniej oceni i zakwalifikuje bibliotekarz. Dzięki takiemu podejściu australijski web-archiv liczy po prawie 10-ciu latach tylko ok. 8,500 adresów. Wersją takiego podejścia może być tworzenie tematycznych zbiorów webowych źródeł, np. dokumentów opublikowanych w Internecie, w czasie wyborów prezydenckich w USA. Takie jakościowe podejście wymaga jednak dużo ludzkiej pracy, dlatego większość bibliotek budujących narodowe web-archivy stosuje zautomatyzowane, całościowe skanowanie i archiwizowanie wszystkich dokumentów spełniających te kryteria, które można stosować automatycznie. Można też przyjąć podejście mieszane, korzystające z obu możliwości.
Bliżej zob. Celbová, Ludmila [Dok. elektr.] (2005). Archivace a zpřístupnění elektronických online zdrojů v evropském kontextu [Referat na konferencję CASLIN 2005]. Tryb dostępu: http://www.webarchiv.cz/files/dokumenty/seminar/celbova.doc [odczyt: 26.06.2006].

Jak już wspomniano, czeskie wydawnictwa elektroniczne dostępne online (internetowe) potraktowano jako część narodowego zasobu informacyjnego i bibliografii narodowej, a także narodowej biblioteki cyfrowej. W związku z tym przyjęto, że należy opisać i zachować dla następnych generacji również i tego rodzaju nietradycyjne dokumenty (zwłaszcza te wartościowe pod względem kulturowym, artystycznym i historycznym) i to jak najszybciej, gdyż lawinowo wzrasta liczba źródeł opublikowanych tylko w Internecie i jednocześnie część z nich z tego środowiska bezpowrotnie znika lub zmienia swoją postać. Podstawowym celem jest więc zidentyfikować i archiwizować wszystko to, co było opublikowane (udostępnione) w ramach czeskiego webu.

Naturalnie zdawano sobie sprawę z tego, że z czysto technicznych powodów nie da się tego celu w pełni osiągnąć, ale i z tego, że tak naprawdę nie ma potrzeby opisywać i archiwizować wszystkich „opublikowanych" w Internecie źródeł, chociażby reklam. Zastosowano więc kombinację dwóch metod: zautomatyzowanego „zbierania" całej powierzchni narodowego webu (large-scale automated harvesting, obejmujący również metadane) oraz selektywnego archiwizowania (na podstawie URL najbardziej interesujących źródeł wybranych według przyjętych kryteriów). Zdecydowano się również na tworzenie „tematycznych kolekcji" odzwierciedlających ważne sprawy bieżące, np. wybory, powódź a nawet projektowanie nowego gmachu Národní Knihovny České Republiky (BN RCz) w Pradze. Metody te są cały czas testowane i doskonalone. Pozyskiwanie dokumentów i danych z webu jest więc na ogół – od strony technicznej i ilościowej – zautomatyzowanym procesem, podczas którego są w oparciu o zadane parametry i za pomocą specjalnego oprogramowania ściągane zbiory i metadane, które są następnie indeksowane i układane do cyfrowego archiwum.[10] Obecnie są w tym celu stosowane ogólnie dostępne narzędzia SW z otwartym kodem źródłowym (Heritix, rozwijany przez IIPC – The International Internet Preservation Consortium) na serwerze przeznaczonym do archiwizacji.

[10] Harvester służy do tworzenia archiwum dokumentów elektronicznych (kopiując, np. raz na pół roku, obraz całej sieci lub wybierając z niej zakwalifikowane dokumenty).

Ściągnięte dane (webowe źródła i metadane) są układane i przechowywane na specjalnym serwerze, podłączonym do krajowej sieci akademickiej CESNET i zarządzanym przez NK ČR przy współpracy z Instytutem Techniki Komputerowej Uniwersytetu im. Masaryka w Brnie. Naturalnie dane są wprowadzane w odpowiednich formatach wspieranych przez IIPC, które powinny zapewnić w trakcie rozwoju technologii informacyjnych bezkolizyjne przenoszenie danych i ich stałą dostępność. Obok serwera archiwizacji służy macierz dyskowa (redundant disk array – RAID). W bliskiej przyszłości przewiduje się transfer danych do nowego urządzenia z bardzo dużą pamięcią, które ma być zainstalowane w NK ČR. Jest to niezbędne posunięcie, ponieważ już w maju 2006 r. zarchiwizowano ok. 26 mln plików o łącznej pojemności 2 TB. Dane są udostępniane on-line na drugim serwerze, który ledwo wystarcza do eksperymentalnego udostępniania niewielkiego zbioru. W celach pełnotekstowej indeksacji stosuje się otwarty system Nutch, dostępny poprzez narzędzia Nutchwax i WERA. Opisy dokumentów wybieranych dla czeskiej bibliografii narodowej oraz narodowy zasób archiwalny są wprowadzane w systemie Aleph, który jest wspierany protokołem Z39.50 na poziomie klienta i serwera oraz OAI-PMH na poziomach repository i harvesting z profilem dla MARC21 i kwalifikowanego Dublin Core. We wszystkich obszarach są konsekwentnie przestrzegane międzynarodowe standardy: MARC21, Dublin Core, XML dla opisu dokumentów; ISSN i URN dla identyfikacji źródeł; format ARC dla archiwizacji.

Pierwszy, eksperymentalny i testujący okres realizacji projektu był trudny. Po trzech latach uruchomiono pierwszy projekt pilotażowy, dzięki czemu udało się rozwiązać wiele technicznych i organizacyjnych problemów w zakresie identyfikacji źródeł („skanowania" Internetu), pozyskiwania oraz archiwizowania krajowych elektronicznych dokumentów sieciowych. Ich udostępnianie z WebArchiv okazało się problematyczne ze względu na niedostosowanie prawa autorskiego oraz braki (lub problemy z wykładnią przepisów) dwóch ustaw o egzemplarzu obowiązkowym, zwłaszcza ustawy dotyczącej drukowanych wydawnictw ciągłych. Zmusiło to organizatorów *WebArchiv* do zawierania umów z poszczególnymi wydawcami dokumentów internetowych. Opierają się one na zapisach Międzynarodowej Deklaracji dot. Przekazywania Dokumentów Elektronicznych

do Narodowego Zasobu Bibliotecznego, opracowanej przez CENL oraz FEP w 2000 r. Zawarto próbnie 12 umów umożliwiających NK ČR w Pradze wyszukiwanie, ściąganie, kopiowanie, archiwizowanie i udostępnianie elektronicznych dokumentów, początkowo jedynie użytkownikom biblioteki (tylko do oglądania i tylko na wybranych terminalach). Wydawca podpisujący umowę zgadza się jednocześnie na włączenie opisów tych dokumentów do czeskiej bibliografii narodowej oraz zobowiązuje się do tworzenia lub zamieszczania danych o dokumencie elektronicznym w standardzie Dublin Core. Wydawca ubiegający się o przydzielenie numeru ISSN powinien wypełnić formularz, zamieszczony na stronie internetowej Czeskiego Ośrodka Narodowego ISSN oraz podać w nim dane o wydawanym przez siebie dokumencie elektronicznym. Obecnie zawarto umowy z kilkuset wydawcami, których wykaz zamieszczono na stronie http://www.webarchiv.cz/partneri, gdzie można też zobaczyć formularz i inne informacje przeznaczone dla wydawcy zainteresowanego współpracą. Nawiązano też współpracę z Ministerstwem Informatyki w celu uzyskania zezwolenia na ściąganie również dokumentów administracji publicznej, które nie są już publicznie dostępne wskutek utraty ważności. Ministerstwo było zainteresowane oprogramowaniem stosowanym w WebArchiv (wówczas Nedlib harvester) oraz generatorem metadanych. Chciało je wykorzystać do tworzenia elektronicznego katalogu dokumentów administracji publicznej.

Przełomowym okazał się rok 2005, kiedy po pięciu latach zmagań osiągnięto pierwsze praktyczne wyniki, a mianowicie udostępniono część cyfrowego archiwum w trybie on-line. Chodzi właśnie o te źródła elektroniczne, publikowane przez wydawców, z którymi NK ČR podpisała ww. umowy. Ten sukces został należycie doceniony podczas INFORUM 2006, gdzie projekt WebArchiv otrzymał nagrodę jako jeden z najważniejszych i najlepszych produktów, usług lub przedsięwzięć, związanych z elektronicznymi źródłami informacji, zaistniałych w roku 2005.

Kolejnym, bardzo istotnym wydarzeniem było przyjęcie organizatorów WebArchivu do międzynarodowego konsorcjum IIPC (International Internet Preservation Consortium). Misją IIPC jest gromadzenie i przechowywanie wiedzy i informacji „opublikowanych" w Internecie tak, żeby były dostępne dla przyszłych generacji, a także wspieranie ogólnoświatowej wymiany tych informacji oraz do-

świadczeń. Dzięki członkostwu w IIPC projekt przekroczył granice Czech i uzyskał międzynarodowe wsparcie.[11]

****

O dalszych losach WebArchivu zadecydują dwa czynniki: legislacyjny i finansowy. W pierwszym przypadku poczyniono krok do przodu, a mianowicie od połowy 2006 r. weszło w Czechach w życie znowelizowane – zgodnie z Directive 2001/29/EC – prawo autorskie. Oznacza to, że można już udostępniać cały WebArchiv dla celów naukowych jak też na użytek prywatny. Mimo to wciąż zawierane są umowy z wydawcami ważniejszych źródeł elektronicznych, uprawniające bibliotekę do ich udostępniania w trybie on-line. Przygotowywane są też nowe rozwiązania prawne, dostosowujące przepisy o egzemplarzu obowiązkowym do nowej „elektronicznej" rzeczywistości. W drugim przypadku chodzi o to, żeby zapewniono stabilne finansowanie WebArchivu i to na niezbędnym poziomie. Do tej pory projekt finansowano prawie w stu procentach z corocznych grantów, co było „od biedy" do przyjęcia na etapie konceptualizacji, prób i testowania. Realizatorzy projektu musieli się zadowolić przeciętnie z 600 tys. koron na rok (na HW i usługi Instytutu Techniki Komputerowej Uniwersytetu im. Masaryka) oraz dwoma pracownikami w NK ČR, zatrudnionymi w komórce zajmującej się elektronicznymi źródłami on-line. Dla potrzeb pełnego wdrożenia i dalszej eksploatacji WebArchivu taki sposób zapewniania środków nie może być wystarczający.

Tempo tworzenia, opracowania, archiwizowania i udostępniania narodowego zasobu sieciowych wydawnictw elektronicznych w Czechach nie było w pierwszych latach imponujące, nie pokonano wszystkich barier prawnych oraz nie zapewniono stabilnego finansowania projektu. Najważniejsze jest jednak to, że projekt *WebArchiv* został podjęty, przemyślany i wdrożony (łącznie z udostępnianiem części źródeł), że jest kontynuowany i doskonalony oraz że

[11] Bliżej zob. Bliżej zob. Celbová, Ludmila; Coufal, Libor [Dok. elektr.] (2007). WebArchiv se účastní mezinárodní spolupráce při archivaci webu. *Ikaros* R. 11, č. 5. Tryb dostępu: http://www.ikaros.cz/node/4085 [odczyt: 1.08.2007].

uczestniczy w międzynarodowej współpracy nad archiwizacją webu. Czeskie doświadczenia uzyskane podczas jego tworzenia oraz wdrażania mogą być niezwykle cenne i przydatne dla tych, którzy podejmą się podobnego przedsięwzięcia w Polsce.