

Archiwizacja stron internetowych szkół wyższych, instytucji naukowych i kulturalnych.

Rozwój technologii informacyjnych opartych na nowoczesnych rozwiązaniach informatycznych, takich jak rozwój globalnej sieci internetowej stawia przed archiwistyką nowe wyzwania. Gwałtowny rozwój stron internetowych powoduje, że stają się one istotnym elementem życia społeczeństwa, przez co ich wartość źródłowa ciągle wzrasta. Stwarza to konieczność zajęcia się problemem archiwizacji stron WWW. Problem ten dotyczy również uczelni wyższych i instytucji kulturalnych, których życie coraz częściej i coraz dokładniej prezentowane jest na stronach internetowych. Celem niniejszego referatu będzie próba przeglądu problemów związanych z zagadnieniem archiwizacji stron internetowych przez uczelnie wyższe oraz instytucje kulturalne. Archiwizacja stron internetowych jest ważkim problemem dla wspomnianych instytucji zwłaszcza w perspektywie digitalizacji zasobów i co z tym związane tworzenia tzw. bibliotek i archiwów cyfrowych.

Z zagadnieniem archiwizacji stron internetowych wiąże się wiele problemów wynikających z charakteru stron WWW, jak i lawinowego przyrostu liczby i rozmiarów witryn internetowych. Jednym z podstawowych problemów z jakim styka się użytkownik Internetu jest tymczasowy charakter zarówno całych stron internetowych, jak i ich zawartości. Wystarczy nadmienić, że średnia długość życia strony WWW to zaledwie 44 dni. W obliczu sytuacji, w której coraz więcej osób powołuje się w swoich publikacjach na treści zawarte w Internecie coraz ważniejszym problemem staje się możliwość zweryfikowania informacji nawet po dłuższym czasie. Konieczna więc staje się archiwizacja przynajmniej niektórych stron WWW. Z punktu widzenia archiwisty tymczasowość to jeden z podstawowych problemów. Wiąże się z nim bowiem wiele istotnych dla archiwizacji zagadnień. Tymczasowość stron, jak i ciągła zmiana ich treści wpływa bowiem na metody archiwizacji. Ponieważ treść witryn internetowych zmienia się często wielokrotnie w ciągu dnia, pojawiają się problemy natury technicznej.

Trudno sobie wyobrazić szanującą się uczelnię wyższą lub instytucję kulturalną nie posiadającą własnej mniej lub bardziej rozbudowanej strony WWW. Każda witryna internetowa składa się z wielu elementów, które są zróżnicowane zarówno pod względem formy jak i treści. Cechą charakterystyczną stron WWW jest hipertekstowość i często multimedialność. Część stron internetowych może zawierać w sobie elementy prezentacji stworzonych za pomocą dodatkowych technologii (np. Flash), część tworzona jest dynamicznie w momencie wysłania przez przeglądarkę zapytania do serwera. Hipertekstowość to organizacja danych w postaci niezależnych leksji połączonych hiperłączami. Hipertekst cechuje nielinearność i niestukturalność układu leksji. Oznacza to, że nie ma z góry zdefiniowanej kolejności czytania leksji, a nawigacja między nimi zależy wyłącznie od użytkownika. Podsumowując obecnie prawie każda bardziej rozbudowana strona internetowa składa się zarówno z elementów tekstowych, grafiki jak i filmów, muzyki czy też animacji.

Strony internetowe omawianych instytucji są zróżnicowane również pod względem treści. Zawierają one zarówno ogólne informacje dotyczące funkcjonowania uczelni lub instytucji, jej historii, struktury władz itd., jak też ogłoszenia dotyczące bieżącego funkcjonowania uczelni (np.: rozporządzenia władz uczelni). W ostatnich latach coraz popularniejsze staje się przeprowadzanie rekrutacji za pośrednictwem Internetu, bardzo ważną rolę odgrywa w tym procesie strona internetowa.

Złożoność budowy i treści zmusza do zastanowienia się nad zakresem archiwizacji stron internetowych. Główny problem stanowi to czy archiwizować całą stronę czy tylko wybrane jej elementy. Wydaje się być konieczna klasyfikacja wewnętrzna treści i elementów składowych stron internetowych jak i określenie częstotliwości archiwizacji. Dynamika stron internetowych to jeden z podstawowych problemów, który należy rozwiązać. Myśląc o archiwizacji stron WWW należy ciągle pamiętać o tym, że nie są one tradycyjnym źródłem. Od wydawnictw czy też tradycyjnej dokumentacji papierowej różni ją możliwość ciągłej zmiany treści jak i formy. Treść raz zapisana na tradycyjnym nośniku, jakim jest papier, w praktyce nie może być usunięta bez wyraźnej ingerencji w nośnik. Dynamika stron WWW sprawia, że jeżeli przed wprowadzeniem na nich jakichkolwiek zmian nie zarchiwizujemy usuwanych treści lub elementów, tracimy możliwość dostępu do nich. Problem ten wywołał gorącą dyskusję wśród archiwistów. Kazimierz Szmidt zaproponował, aby archiwizować tylko i wyłącznie zmieniane elementy stron WWW. W takim wypadku zarchiwizowana została by cała struktura i układ strony internetowej oraz wszelkie rodzaje plików na nią się składające. W momencie wprowadzenia zmiany na stronie archiwizowany byłby tylko

zmieniany element, a nie ponownie zachowywana cała strona. Na przykład, jeśli archiwizujemy stronę internetową, do której dodawane jest nowe zdjęcie to archiwizacji podlega tylko i wyłącznie nowo dodane zdjęcie wraz z metadanymi określającymi umiejscowienie tego zdjęcia w strukturze strony WWW. Takie rozwiązanie przynosi wiele korzyści. Główną z nich jest oczywiście zaoszczędzenie miejsca na nośnikach do archiwizacji. Nie rozwiązuje jednak problemu bardzo częstych zmian wprowadzanych na stronach, które ciągle musiałyby być monitorowane i sprawdzane pod kątem pojawiających się nowych elementów. Alternatywą dla powyższej metody jest archiwizowanie całych stron internetowych, napotyka to jednak na wiele przeszkód natury technicznej. Metoda ta wymaga znacznie większej pamięci przeznaczonej do przechowywania archiwizowanych stron.

Potocznie za archiwizację w świecie informatyki uznaje się popularne backup-y, czyli kopie zapasowe. Na rynku oprogramowania można znaleźć szereg programów służących do tworzenia kopii zapasowych. Programy te w czasowych odstępach zapisują na dowolnym nośniku dowolny rodzaj danych, również strony WWW. Problem jednak polega na tym, że operacja ta bardziej przypomina aktualizację i zabezpieczenie aktualnej wersji, niż rzeczywistą archiwizację. Backup-y nie dają możliwości cofnięcia się do poprzednich wersji strony. Za archiwizację można uznać zapisywanie kolejnych wersji witryn internetowych z możliwością dostępu do każdej z nich.

Jedną z podstawowych metod porządkowania i gromadzenia danych na temat witryn WWW są popularne w sieci katalogi, takie jak np.: www.google.pl , www.katalog.gemsnet.pl, www.katalog.info , www.quickdir.info itd. Katalogi te również nie stanowią formy archiwizacji, gdyż zawierają one jedynie linki oraz krótkie opisy. W przypadku wygaśnięcia witryny lub zmiany jej lokalizacji dotarcie do strony jest praktycznie niemożliwe. Poza tym katalogi dają jedynie ścieżkę dostępu do istniejących aktualnie stron, a nie do ich wersji archiwalnych. W sieci można jednak już znaleźć strony, które za główny cel swej działalności stawiają sobie pewną formę internetowej „archiwizacji”. Jednym z nich jest „Internet Archive” (www.archive.org). Internetowe archiwum zostało powołane w 1996 roku w USA. Jego zadaniem jest udostępnianie badaczom, historykom, studentom i innym osobom archiwalnych wersji stron internetowych. Omawiana witryna nie przechowuje jednak archiwalnych wersji stron WWW na własnych nośnikach, prowadzi jedynie swoisty katalog linków do zarchiwizowanych wersji stron zgromadzonych przez właścicieli tych witryn. W odróżnieniu od zwykłych katalogów stron internetowych daje ona możliwość przeglądania archiwalnych wersji danej strony. Jednak częstotliwość archiwizowania jest różna i zależy od administratora danej strony. W „Internet Archive” znajdują się również odnośniki do

archiwów polskich stron, w tym stron uczelni wyższych. Ponadto w roku 1997 zaczęło ukazywać się w Stanach Zjednoczonych czasopismo dotyczące archiwizacji zasobów sieciowych: „Journal of Internet Cataloging. The International Quarterly of Digital Organization, Classification and Access” (Przegląd katalogowania Internetu. Międzynarodowy Kwartalnik Organizowania, Klasyfikacji i Dostępu do Zasobów Cyfrowych).

Jednym z elementów ograniczających możliwości archiwizacji stron internetowych są przepisy dotyczące własności intelektualnej. Każda strona internetowa jest własnością osoby bądź instytucji, która ją stworzyła lub nabyła praw do niej. W związku z tym nie można bez zgody tej osoby gromadzić i udostępniać zarówno struktury jak i treści strony internetowej. W większości krajów brak jest szczegółowych regulacji prawnych dotyczących praw do archiwizacji stron WWW. Jeden z amerykańskich projektów (Przechowywanie Informacji Cyfrowej) zakłada, aby początkowo to właściciel praw autorskich zobowiązany był do archiwizowania, ale miałby on możliwość przekazania tych praw jakiejś powołanej do tego instytucji. Rozważając zagadnienia prawne związane z Internetem należy mieć na uwadze istnienie dwóch warstw Internetu, istnieje bowiem Internet „płytki” i „głęboki”. Internet „płytki”, jest to ogólnie-dostępna część sieci. Nie może ona funkcjonować bez „głębokiego Internetu”, który jest strukturą informacji, która wraz z oprogramowaniem umożliwia korzystanie z sieci. Szacuje się, że „głęboki Internet” jest ok. 500 razy większy niż „płytki” dodatkowo ze względu na szereg zabezpieczeń jego archiwizacja jest niemożliwa.

Coraz mocniej problem archiwizacji stron internetowych jest dostrzegany w krajach Europy zachodniej i USA gdzie rozwój sieci jest o wiele większy niż w Polsce. Jedną z najważniejszych inicjatyw koordynujących gromadzenie i przechowywanie zasobów cyfrowych jest program NDIIPP (National Digital Information Infrastructure and Preservation Program). Program ten został przyjęty przez Kongres USA w grudniu 2000 roku i wzywa on Bibliotekę do objęcia przywództwa nad ogólnokrajowymi działaniami w zakresie planowania długotrwałego przechowywania treści cyfrowych. Biblioteka Kongresu zinterpretowała to zadanie możliwie szeroko włączając w to konsultacje z innymi bibliotekami narodowymi, które rozpoczęły działania w podobnym zakresie. Biblioteki narodowe uznały archiwizację zasobów WWW za ważną funkcję, czy to poprzez wybrane zestawienia konkretnych stron WWW czy też przechwytywanie całych domen dla konkretnych terytoriów poszczególnych krajów. W ramach tej współpracy zaleca się pilną obserwację prac prowadzonych w Wielkiej Brytanii, Francji, Australii i Skandynawii, oraz badanie możliwości wspólnego opracowywania narzędzi i procedur postępowania. W ramach programu NDIIPP zaleca się

wykorzystanie standardów OAIS (Open Archival Information System). Jest to pierwszy międzynarodowy standard w obszarze przechowywania cyfrowego. Oferuje on szerokie możliwości praktycznego wykorzystania i został powszechnie przyjęty w środowisku bibliotekarskim.

We Francji pod kierownictwem Biblioteki Narodowej prowadzone są badania dotyczące gromadzenia i zachowywania stron internetowych opartych na bazach danych. W grudniu 2000 roku zrealizowano szereg eksperymentów dotyczących archiwizacji krajowej domeny WWW. Celem tych eksperymentów jest oszacowanie kosztów oraz określenie procedur selekcji, transferu i przechowywania witryn internetowych. Celem tych badań było również stworzenie odpowiedniego narzędzia programowego do archiwizacji zasobów sieciowych. Natomiast w Wielkiej Brytanii istnieje zróżnicowana grupa instytucji kulturalnych realizujących inicjatywy w zakresie przechowywania cyfrowego. Inicjatywy te cieszą się szerokim uznaniem na całym świecie. Na uwagę zasługuje program Joint Information System Committee (Komisja ds. Wspólnych Systemów Informacyjnych Komitetów ds. Wyższego i Dalszego Kształcenia). Instytucja ta jest mocno zaangażowana w działania na rzecz przechowywania cyfrowego i opracowuje nowy program, który obejmować będzie dalsze prace badawcze poświęcone systemom cyfrowym oraz nowe usługi w zakresie przechowywania cyfrowego dla sektora szkolnictwa wyższego. Głównym polem działalności jest archiwizacja dużej liczby publikacji elektronicznych wykorzystywanych w brytyjskim sektorze naukowo-badawczym i edukacji. Częścią tego projektu będzie również archiwizacja stron WWW. Podobną inicjatywę można spotkać również w Holandii. Biblioteka Narodowa Holandii zajmuje się opracowaniem systemu Digital Archive Store Project (Projekt Cyfrowego Archiwum). Wraz z IBM prowadzi badania dotyczące długoterminowego przechowywania materiałów cyfrowych. Zaznaczyć trzeba jednak, że na razie Biblioteka Narodowa Holandii nie gromadzi żadnych materiałów audiowizualnych, filmowych czy radiowo-telewizyjnych, oprogramowania, gier lub baz danych. Pomimo tego w kręgu jej zainteresowań jest przyszła archiwizacja holenderskiej domeny internetowej.

Podsumowując należy stwierdzić, że obecnie nie ma jednego programu międzynarodowego koordynującego wszystkie działania zmierzające zarówno do digitalizacji zasobów czy to archiwalnych, czy też bibliotecznych jak i archiwizacji zasobów internetowych w tym oczywiście stron internetowych. Istnieje wiele inicjatyw międzynarodowych zajmujących się tymi sprawami np.: Sieć Przechowywania i Udostępniania Zasobów Elektronicznych – EPRANET, Projekt Sieciowej Europejskiej Biblioteki Depozytowej (NEDLIB) czy też OAIS. W poszczególnych krajach również

istnieje wiele inicjatyw zajmujących się archiwizowaniem szerokopojętych zasobów sieciowych, brak natomiast jest jednego spójnego programu. Taki stan rzeczy spowodowany jest oczywiście świeżością zagadnienia, gdyż archiwizowanie i przechowywanie cyfrowe nadal jest nowym elementem w działalności bibliotek i archiwów.

W Polsce brak jest szerszych inicjatyw dotyczących gromadzenia i przechowywania zasobów sieciowych. Istnieje i rozrasta się natomiast Polska Biblioteka Internetowa, która zajmuje się digitalizacją wydawnictw. W Poznaniu z inicjatywy poznańskich środowisk naukowych i bibliotek uczelnianych została powołana Wielkopolska Biblioteka Cyfrowa, która gromadzi, przechowuje i udostępnia w Internecie swe zasoby. Wracając do zagadnienia gromadzenia i przechowywania zasobów sieciowych - stworzono warunki prawne umożliwiające archiwizację zawartości Internetu. Z ustawy o informatyzacji działalności podmiotów realizujących zadania publiczne z dnia 17 lutego 2005 roku wynika, że zarówno dokument elektroniczny jak i zawartość Internetu jeżeli ma znaczenie „jako źródło informacji o wartości historycznej, działalności państwa Polskiego, [...] organizacji i rozwoju nauki, kultury i sztuki [...]” wchodzi w skład narodowego zasobu archiwalnego.

Na podstawie ustawy o dostępie do informacji publicznej z dnia 6 września 2001 roku określone instytucje zostały zobowiązane do prowadzenia w Internecie podmiotowej strony Biuletynu Informacji Publicznej. Wśród tych podmiotów są również szkoły wyższe. Strony BIP-u można uznać za część strony danej uczelni, a na pewno jako źródło informacji. Strony zawierają szereg istotnych informacji statystycznych oraz ogłoszeń publicznych dotyczących min. przetargów, konkursów itd. Rozporządzenie MSWiA z dnia 17.05.2002 r. wprowadziło obowiązek systematycznego kopiowania informacji zgromadzonych zarówno na głównej stronie BIP-u jak i na stronach podmiotowych na odrębne elektroniczne nośniki informacji nie rzadziej niż raz na dobę. Taki stan rzecz możemy uznać za archiwizowanie, gdyż z praktycznych względów zapisywane są kolejne wersje witryn. Poza codobowym archiwizowaniem, również wprowadzenie jakiejkolwiek zmiany wymusza ponowne zapisanie strony. Archiwizacja podmiotowych stron BIP-u powinna odbywać się na nośnikach zewnętrznych czyli nie zintegrowanych bezpośrednio ze strukturą dysku na którym posadowiona została podmiotowa strona BIP. Dopuszcza się rozwiązania archiwizacyjne oparte o ZIP-ym JAZZY-y, nośniki MO (magneto-optyczne), tasiemki streamerów DAT, oraz CD-RW, dyski kieszonowe oraz pamięci masowe.

W dobie przyspieszonej i coraz większej informatyzacji społeczeństwa wzrasta znaczenie stron internetowych jako źródła historycznego. Olbrzymia część aktywności społecznej przenosi się w cyberświat. Poprzez Internet możemy załatwiać szereg spraw

urzędowych i osobistych. Internet staje się źródłem bardzo ważnych i coraz bardziej rzetelnych informacji. Sieć stanowi istotny element życia współczesnego człowieka i w przyszłości bez dostępu do zarchiwizowanych zasobów sieciowych historycy nie będą w stanie przedstawić prawdziwego obrazu dzisiejszego świata. Wartość stron internetowych jako źródła historycznego została już zauważona przez wspomnianą już ustawę o informatyzacji. Jednocześnie każda osoba korzystająca kiedykolwiek z zasobów Internetu spotkała się ze stronami, których zawartość nie odpowiadała jej oczekiwaniom. Przed archiwistyką stoi zadanie odpowiednie klasyfikacji stron internetowych. Nie jesteśmy bowiem w stanie przechowywać wszystkich stron, podobnie jak dokumentacji tradycyjnej. Podział stron internetowych na poszczególne kategorie wydaje się być niezbędny. Ewa Perłakowska z Archiwum Państwowego w Krakowie zaproponowała podział na kategorie stron internetowych tworzonych przez podmioty. Zgodnie z tą propozycją zasady tworzenia i obsługi stron WWW oraz własna strona powinny mieć w jednostce macierzystej kategorię A, natomiast materiały źródłowe do tej strony kategorii B5. Reasumując oprócz przepisów dotyczących BIP-u nie istnieją inne przepisy zobowiązujące uczelnie wyższe i instytucje kulturalne do archiwizacji własnych stron internetowych, nie wspominając o innych stronach. Nie dotarliśmy również do innych przepisów określających metody archiwizacji zasobów sieciowych.

Latem tego roku przeprowadziliśmy za pomocą poczty elektronicznej ankietę wśród około 100 wyższych uczelni i instytucji kulturalnych na temat archiwizacji stron Internetowych. Odzew był znikomy. Z przeprowadzonych badań wynika że nie prowadzi się żadnej archiwizacji na większą skalę. Najczęściej strona WWW jest archiwizowana na własne potrzeby przez pracowników odpowiedzialnych za zarządzanie stroną. W większości przypadków wykorzystuje się do tego płyty CD oraz twarde dyski komputerów.

Podsumowując należy stwierdzić, że problem archiwizacji stron internetowych jest w naszym kraju dopiero dostrzegany. Powinien być on jednak ciągle monitorowany, gdyż nie możemy dopuścić do sytuacji aby w przyszłości historycy badający nasze czasy nie mieli możliwości wykorzystania tak istotnego źródła. W związku z tym widzimy konieczność poczynienia pewnych kroków w celu szerszego zajęcia się tym zagadnieniem. Przede wszystkim należy sklasyfikować strony internetowe tzn. określić które instytucje państwowe, organizacje społeczne, uczelnie itd. powinny archiwizować własne strony. Należy bowiem podkreślić, że archiwizacją stron internetowych powinien zajmować się ich twórca, czyli każda ważna instytucja powinna mieć obowiązek archiwizacji swojej strony gdyż mało prawdopodobne jest aby ktoś (np. Archiwa Państwowe) miały za zadanie to robić. Dla

archiwizacji stron WWW istotne jest również poszerzenie programu kształcenia archiwistów o przedmioty informatyczne, a także zatrudnianie informatyków w archiwach instytucji zobowiązanych do archiwizacji własnych stron WWW. Przechowywanie stron internetowych z pewnością jest jednym z głównych wyzwań współczesnej archiwistyki.

Bibliografia:

1. Justyna Adamus, *Narzędzia klasyfikacji dokumentu elektronicznego*
2. Neil Beagrie, *Narodowe inicjatywy w zakresie przechowywania cyfrowego. Przegląd przedsięwzięć w Australii, Francji, Holandii w Wielkiej Brytanii* [w:] Archeion, Tom CVII, Warszawa 2004.
3. *Jurnal of Internet Cataloging, Witryna Elektroniczna*; [w:] *Biuletyn Informacyjny Biblioteki Narodowej* 4/1555/2000.
4. Henryk Niestrój, *Dokument elektroniczny – zarys problematyki*;
5. Ewa Perłakowska, *Propozycje klas (Archiwizacja stron WWW)* [w:] <http://adacta.archiwa.net>
6. Halina Robótka, *Metody i zasady selekcji dokumentacji tradycyjnej i elektronicznej a potrzeby badań obecnych i przyszłych*;
7. Wanda Krystyna Roman: *O potrzebie i problemach archiwizowania stron WWW*; [maszynopis]

Strony WWW:

1. <http://www.bip.gov.pl>
2. <http://www.ifar.lublin.ap.gov.pl>
3. <http://www.archiwa.net>
4. <http://adacta.archiwa.net>