Maria Nowina Konopka

Problematyka zarządzania informacją w procesie archiwizacji zasobów Internetu

Media – Kultura – Komunikacja Społeczna 10/1, 28-43

2014

Maria Nowina Konopka

Problematyka zarządzania informacją w procesie archiwizacji zasobów Internetu

Słowa kluczowe: archiwizacja, zarządzanie informacją, Internet
Key words: web archiving, information management, Internet

The Internet is the media of our time.
Participation has never been so widespread
in the elaboration of any media before.
Size and variety of content exposed has never been so large.
For this reason, the Internet is becoming the most important source of information
on our society and will be a key resource with which to think about it,
now and in the future.
Internet Memory Foundation

Wstęp

Jedną z głównych cech charakteryzujących internautów ze względu na sposób zaspokajania przez nich swoich potrzeb informacyjnych jest traktowanie globalnej Sieci jako niemal nieograniczonego zbioru informacji. Użytkownicy Sieci pragną jednak być nie tylko biernymi odbiorcami dostępnych treści, lecz także, a obecnie może przede wszystkim, chcą mieć aktywny wpływ na zamieszczane w niej informacje. Stąd też w treściowych zasobach tego medium dostępne są gigadane konieczne do realizacji wspomnianych potrzeb oraz pozwalające na poszerzenie zakresu wiedzy w każdym niemal obszarze. Dane te, w sposób często niezauważalny, stanowią również wierny zapis naszego życia. Nic więc dziwnego, że coraz częściej podnoszone są głosy wzywające do ochrony tego dziedzictwa kulturowego, do archiwizowania świadectw życia społeczno-politycznego. Uchwycenie i zatrzymanie na nośnikach danych ważkich dziś informacji pozwoli w przyszłości na zdystansowaną ocenę zachodzących zjawisk oraz szerokie badania retrospektywne. Będą one jednak możliwe jedynie w sytuacji dokonania pełnego zapisu, bez uszczerbku dla całości mających obecnie miejsce wydarzeń, zjawisk i procesów. Tymczasem z badań prowadzonych przez Junghoo Cho i Hectora Garcia-Molinę wynika, że przeciętnie „życie" strony internetowej trwa około 50 dni[1], a według szacunków

[1] J. Cho, H. Garcia-Molina, *The Evolution of the Web and Implications for an Incremental Crawler*, s. 4, [online] <http://ilpubs.stanford.edu:8090/376/1/1999-22.pdf>, dostęp: 10.12.2013.

British Library – 75 dni[2]. Maria A. Jankowska uznaje, że 44% witryn internetowych znika w ciągu pierwszego roku funkcjonowania, co w pewnym tylko sensie potwierdza zespół badawczy Daniela Gomesa, twierdząc, że odsetek ten sięga nawet 80%[3].

Przyczyn krótkotrwałości wirtualnego kontentu wymienić można wiele, wśród najistotniejszych z nich wskazuje się na zanik zainteresowania samego twórcy stworzonym dziełem, następstwa ataków skutkujących przeciążeniem serwera, brak zainteresowania odbiorców lub, paradoksalnie, nagły wzrost ich zainteresowania, powodujący szybkie wyczerpanie się transferu dostępnego na wykupionym przez właściciela koncie hostingowym[4]. Oznacza to, że jeśli dane dostępne w Sieci nie posiadają swojego odpowiednika w wersji papierowej lub nie zostały uprzednio zapisane, ulegają na zawsze zaprzepaszczeniu. Uporanie się z problemem zapisu danych nie rozwiązuje zresztą kłopotu, ponieważ postęp technologiczny powoduje tak szybkie zmiany form i standardów zapisu, że informacje dostępne niegdyś na takich nośnikach, jak dyskietka 5- czy 3,5-calowa powoli przestają być dostępne. Włodzimierz Gogołek powołuje w tym kontekście przykład British Broadcasting Corporation (BBC). Już w 1986 roku BBC, mając świadomość zachodzących na rynku mediowym zmian, dokonała na laserowych dyskach zapisu wielu wartościowych danych na temat Wielkiej Brytanii (takich jak teksty, mapy, zapisy video). Technologia uznawana podówczas za nowoczesną miała zagwarantować cyfrowym danym prawdziwą nieśmiertelność. Niestety, już na początku bieżącego millennium dane te stały się *de facto* nie do odczytania. W efekcie niezbędne stały się kolejne prace – rearchiwizacyjne, trwające przeszło dwa i pół roku[5].

Zadanie archiwizacji zasobów, w przypadku mediów tradycyjnych nienastręczające zbyt wielu kłopotów, w odniesieniu do Internetu wydaje się niemal niemożliwe do realizacji. Przy okazji procesu archiwizacji powstaje wiele pytań dotyczących kwestii zarządzania zbieranymi informacjami, sposobu udostępniania i gromadzenia danych oraz samego doboru stron, które temu procesowi mają podlegać. W polskiej literaturze przedmiotu, jak można sądzić po wstępnym rekonesansie, tematyka archiwizacji Internetu jest poruszana jedynie marginalnie, a sam temat nie jest powszechnie znany, ponadto Polska nie podjęła jeszcze instytucjonalnie usankcjonowanych, systematycznych prac nad archiwizacją zasobów sieciowych. Zarówno w Europie, jak i w innych częściach świata wypracowano natomiast narzędzia oraz międzynarodowe standardy w zakresie archiwizowania wirtualnego kontentu i zarządzania zgromadzonymi danymi. Celem zbadania tych zagadnień dokonano kwerendy

[2] The British Library, [online] <http://www.bl.uk>, dostęp: 1.10.2013.

[3] L. Derfert-Wolf, *Archiwizacja Internetu – wprowadzenie i przegląd wybranych inicjatyw*, „Biuletyn EBIB" 2012, nr 1(128), [online] <http://eprints.relis.org/17048/1/derfert_Web_archiving.pdf>, dostęp: 10.09.2013.

[4] B. Mrożewski, *Archiwa internetu*, „PC Format" 2012, nr 38, [online] <http://www.pcformat.pl/Archiwa-internetu,a,2374>, dostęp: 7.12.2013.

[5] Por. [online] <http://www.domesday1986.com>, za: W. Gogołek, *Komunikacja sieciowa. Uwarunkowania, kategorie, paradoksy*, Warszawa 2010, s. 158.

literatury przedmiotu oraz przeglądu międzynarodowych i krajowych inicjatyw mających w swym *spectrum* działania wspomniany obszar badawczy[6]. Podjęcie refleksji zarówno nad zasadnością procesu archiwizacji zasobów Internetu, jak i konsekwencjami braku należytego nadzoru nad przebiegiem tego procesu wydaje się nieodzowne. Wywołanie szeroko zakrojonej dyskusji jest obecnie potrzebne, ponieważ wspomniane kwestie trzeba będzie niebawem w Polsce omówić, zaś stosowne decyzje warto wypracować w oparciu o wielość międzynarodowych przykładów i doświadczeń[7]. Z tego też względu celem niniejszego artykułu jest przegląd najważniejszych dylematów związanych z problematyką zarządzania informacją w procesie archiwizowania zasobów Internetu. Ich klasyfikacji (aspekty techniczne, etyczne, ekonomiczne i prawne) dokonano w oparciu o doświadczenia i wnioski wynikające z już istniejących projektów oraz w odniesieniu do analizy krytycznej prezentowanej w literaturze przedmiotu.

Archiwizacja Internetu

Proces archiwizacji informacji dostępnej w Internecie polega, zdaniem M. Jankowskiej, na „poszukiwaniu, gromadzeniu i organizacji źródeł informacji w celu zabezpieczenia ich przed zniknięciem z WWW"[8]. Jest to więc proces wieloetapowy i bardzo złożony, ponieważ w jego ramach mieści się odpowiedzialność za dobór stron do archiwizowania, uzyskanie zgody na ich przechowywanie, prowadzenie zapisu w ustalonych, racjonalnie przyjętych odstępach czasu oraz takie zarządzanie zgromadzonymi danymi, aby były one dostępne właściwym osobom, w dogodnej formie i szybkim czasie. Czynność ta nie jest więc jedynie kwestią umiejętności wykorzystania technologii, lecz wymaga profesjonalnego przygotowania w zakresie kategoryzowania zdobytych danych, zapewnienia im fizycznego bezpieczeństwa oraz sprawności w stworzeniu dla realizowanego projektu odpowiedniego środowiska prawnego. Ogrom działań i ciążącej na pomysłodawcach odpowiedzialności spowodował, że w początkowej fazie procesu archiwizowania Internetu inicjatorami projektów były jednostki profesjonalne, czyli biblioteki narodowe. W ich działalności od początków nowego tysiąclecia zauważalna stała się tendencja odchodzenia od książek papierowych na rzecz udostępniania publikatorów naukowych w trybie *on-line*:

[6] Warto w nocie metodologicznej dodać, że obok przeglądu literatury przedmiotu dokonano także krytycznej analizy źródeł internetowych, będących w chwili obecnej ważnym (choć często trudnym do weryfikacji) źródłem materiału badawczego.

[7] Do analizy zastosowano metodę doboru próby zupełnej, co oznacza analizę wszystkich dotychczas (koniec 2013 roku) istniejących organizacji i stowarzyszeń (krajowych i o zasięgu międzynarodowym) zajmujących się archiwizacją stron WWW.

[8] M. Jankowska, *Biblioteki akademickie – trendy dotyczące zasobów elektronicznych*, w: *Informacja dla nauki a świat zasobów cyfrowych*, red. H. Ganińska, Poznań 2008, s. 168, [online] <http://www.library.put.poznan.pl/konf_idn/art/4_3.pdf>, dostęp: 1.12.2013.

popularność elektronicznych środków informacji wynika z faktu, że pracownicy naukowi oraz studenci zdecydowanie preferują elektroniczne źródła informacji, coraz rzadziej korzystają z drukowanych. Dowiodły tego wyniki badań przeprowadzone w 155 bibliotekach akademickich na świecie przez Publishers Communication Group, Inc. Potwierdziły one wzrost elektronicznych źródeł informacji w kolekcjach bibliotek aż o 91% oraz wykazały, że 84% bibliotek nie gromadzi już drukowanych odpowiedników elektronicznych czasopism[9].

Biblioteki są instytucjami posiadającymi niezbędne w procesie archiwizacji kompetencje, narzędzia i zasoby ludzkie. Cyfryzacja publikatorów naukowych, których liczba, choć olbrzymia, daje się jednak rozumowo ogarnąć, będąc w swej masie policzalna, stanowi zaledwie małą cząstkę bezkresu wirtualnego kontentu. Dlatego też od początku procesu archiwizacji Sieci biblioteki mogły liczyć na wsparcie organizacji non-profit, jednostek prywatnych, często wywodzących się ze środowiska programistów komputerowych, i w końcu rzeszy zwykłych internautów, zaangażowanych w mozolny, wręcz mrówczy proces kopiowania milionów stron WWW.

Pierwszą i najszerzej zakrojoną tego rodzaju inicjatywą na świecie było założone w 1996 roku cyfrowe archiwum Wayback Machine, należące do niekomercyjnej organizacji Internet Archive. Istotą tego amerykańskiego projektu jest skanowanie ponad 200 milionów wybranych stron w 40 językach świata[10], co łącznie daje liczbę 368 bilionów stron[11]. W tym samym roku powstały jeszcze trzy tego rodzaju inicjatywy: Australia's Web Archive, Tasmanian Web Archive i Sweden (Kulturarw3), chociaż zasięg ich działania jest względem amerykańskiego przedsięwzięcia znacznie skromniejszy. Do 2013 roku łącznie na świecie powstało niespełna 70 inicjatyw powołanych do archiwizacji danych cyfrowych (niestety, wśród nich nie znajduje się żaden polski projekt)[12]. W efekcie prowadzonych od lat prac Scott G. Ainsworth wraz z zespołem dowiedli, że 35–90% stron WWW pochodzących sprzed 2008 roku ma co najmniej jedną kopię archiwalną, 17–49% – od dwóch do pięciu kopii, 1–8% – od sześciu do dziesięciu kopii, a 8–63% – minimum dziesięć kopii. Jednocześnie tylko 14,6–31,3% stron jest archiwizowanych częściej niż raz w miesiącu[13]. Przyglądając się danym w innym ujęciu, warto zauważyć wartość zajętej przestrzeni dyskowej, która pomimo istotnych braków danych w bazie

[9] Publishers Communication Group, INC, *Global Electronic Collection Trends in Academic Libraries*, 2004, [online] <http://www.pcgplus.com/Resources/GlobalElTr.pdf>, za: M. Jankowska, dz. cyt.

[10] Zob. [online] <http://archive.org/projects>, dostęp: 7.04.2013.

[11] Dane na listopad 2013 roku, [online] <http://archive.org/web>, dostęp: 29.11.2013.

[12] Pełna lista inicjatyw i instytucji dostępna pod adresem: <http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives>, dostęp: 10.04.2013.

[13] Badania prowadzono w 2010 i 2011 roku na zasobach Internet Archive Wayback Machine, pamięciach podręcznych trzech wyszukiwarek (Google, Bing, Yahoo!) oraz w Diigo, Archive-It, UK National Archives i WebCite. Wyniki pokazały, że najwięcej kopii witryn znajduje się w Internet Archive Wayback Machine. Korzystanie z wielu różnych wyszukiwarek wyjaśnia także duże rozbieżności pomiędzy danymi. Zob. S.G. Ainsworth, A. AlSum, H. Salah-Eldeen, M.C. Weigle, M.L. Nelson, *How Much of the Web Is Archived?*, s. 5, [online] <http://arxiv.org/pdf/1212.6177v2.pdf>, dostęp: 1.12.2013.

Wikipedii oscyluje na koniec 2013 roku wokół 8692,6 TB, z czego 5,5 TB zajmuje Internet Archive (Wayback Machine). Jako ciekawostkę można przy okazji dodać, że TB (terabajt) jest równy bilionowi bajtów (10^{12} bajtów), tak więc gdyby w jednym kinie zgromadzić płyty z filmami, na których łącznie znajdowałoby się tyle terabajtów, ile zarchiwizowano danych, to seans taki trwałby nieprzerwanie około 430 lat. Liczby te, choć i tak duże, w niewielkiej tylko mierze oddają mnogość cyfrowej informacji. Główną bowiem wadą projektów archiwizacyjnych jest fakt, że przeczesywaniu i rejestrowaniu podlega jedynie internetowy *surface*, bez uwzględnienia zasobów Sieci głębokiej (zagadnienie to zostanie poruszone w dalszej części tekstu), co zresztą stanowi istotny argument w dyskusji osób będących przeciwnikami archiwizowania „wszystkiego"[14].

Wróciwszy raz jeszcze do historii projektów archiwizacji kolekcji zasobów dostępnych *on-line*, warto wskazać na trudności w realizacji partykularnych projektów. Internet bowiem tylko teoretycznie nie zna granic, w sytuacji zaś wymogu prawnego unormowania kwestii pobierania i zapisywania stron WWW istnieje konieczność prowadzenia współpracy międzynarodowej, która zapewnia bezpieczeństwo prawne realizowanym projektom. Pierwszym międzynarodowym konsorcjum, skupiającym obecnie 44 członków, jest założone w lipcu 2003 roku International Internet Preservation Consortium (IIPC)[15]. Organizacja ta koordynuje działania mające na celu wypracowanie:

> sprawnych narzędzi, standardów i najlepszych praktyk sieciowej archiwizacji, przy jednoczesnym wspieraniu współpracy międzynarodowej i propagowaniu szerokiego dostępu oraz możliwości korzystania z archiwów internetowych w dziedzinie badań i dziedzictwa kulturowego[16].

Członkowie konsorcjum regularnie publikują raporty, organizują szkolenia i warsztaty, udostępniają prezentacje oraz materiały. Działania te przyczyniają się do współpracy w zakresie wypracowania najlepszych praktyk, obniżenia kosztów działania oraz poprawy w dostępie do posiadanych przez podmiot archiwów, tak aby uniknąć tak zwanej pułapki spirali archiwizacji (polegającej na permanentnym przenoszeniu danych z nośnika na nośnik). Idei tej służy również Internet Memory Foundation, organizacja non-profit skupiająca od 2004 roku państwa europejskie[17].

Niestety, mimo że Internet bez wątpienia stał się wiodącym medium naszych czasów, a jego powszechność użycia nie budzi już żadnych wątpliwości,

---

[14] Szerzej na ten temat zob. M. Wilkowski, *Trzy argumenty przeciwko archiwizowaniu Internetu*, „Historia i Media", [online] <http://historiaimedia.org/2011/10/04/trzy-argumenty-przeciwko-archiwizowaniu-internetu>, dostęp: 10.04.2013.

[15] Zob. International Internet Preservation Consortium, [online] <http://netpreserve.org>, dostęp: 19.09.2013. Rozmieszczenie członków IIPC zob. [online] <http://viewshare.org/views/abpo/iipc-member-archives-2>, dostęp: 1.12.2013. Reprezentantem Polski w IIPC jest Biblioteka Narodowa w Warszawie; dane na wrzesień 2013 roku.

[16] *About IIPC*, [online] <http://netpreserve.org/about-us>, dostęp: 1.12.2013.

[17] Rozmieszczenie uczestników projektu Internet Memory Foundation w Europie zob. [online] <http://internetmemory.org/images/uploads/Carte_partenaires.pdf>, dostęp: 1.12.2013.

Polska uczestniczy w międzynarodowych projektach w bardzo ograniczonym zakresie. Obok Biblioteki Narodowej w Warszawie, zajmującej się digitalizacją dorobku naukowego[18], 1 kwietnia 2009 roku do życia powołano Narodowe Archiwum Cyfrowe (NAC). Instytucja ta, jak podano na jej stronie internetowej, „aby zabezpieczyć informacje o historycznym znaczeniu dla państwa polskiego, archiwizuje strony internetowe najważniejszych urzędów państwowych"[19]. Ściśle rzecz ujmując, NAC archiwizowała wspomniane strony jedynie w latach 2009–2011, dokonując przeglądu 46 instytucji i zapisując ich kopie od kilku do kilkunastu razy. A zatem przyjąć można, że proces archiwizacji zasobów Internetu to w Polsce kwestia przyszłości. Warto więc, biorąc pod uwagę wieloletnie doświadczenia parudziesięciu krajów i projektów, zastanowić się nad kluczowymi kwestiami związanymi z problemami zarządzania informacją w procesie archiwizacji zasobów Internetu, tak aby wchodząc kiedyś w międzynarodowe struktury i porozumienia oraz podejmując wysiłek digitalizacji polskiego dziedzictwa *on-line*, wykorzystać najlepsze wzorce i rozwiązania.

Zarządzanie informacjami zarchiwizowanymi

Zarządzanie informacją rozumieć można jako zespół działań podejmowanych w celu kontrolowania przepływu informacji. Dokonywane jest ono bądź to z chęci optymalizacji przebiegu procesów informacyjnych, polegających na pozyskiwaniu, gromadzeniu, generowaniu, przechowywaniu, przetwarzaniu i dystrybucji informacji, bądź też w celu kontroli przepływu strumieni informacji. W tym też sensie proces zarządzania informacjami archiwizowanymi polega na wielości i wielopłaszczyznowości działań, takich jak: wybór stron WWW, ustalenie częstotliwości i głębokości dokonywanych zapisów, przyjęcie metod gromadzenia informacji zgodnych z obowiązującym stanem prawnym, gwarantującym równocześnie bezpieczeństwo przechowywania i spójność zapisu danych, tworzenie przejrzystych systemów klasyfikacji zgromadzonego materiału, określenie zasad dostępu do danych oraz pozyskiwanie środków na działalność niewpływających na obiektywizm w realizacji poszczególnych elementów wspomnianego procesu. Na każdym jednak etapie podejmowane decyzje mogą budzić zastrzeżenia, co uzasadnione jest uznaniowością

[18] W kwestii działalności Biblioteki Narodowej w Warszawie należy dodać, że „od stycznia 2007 roku Biblioteka Narodowa jest partnerem projektu Biblioteka Europejska, w którym prezentowane są zasoby katalogowe i cyfrowe europejskich bibliotek narodowych. Biblioteka Europejska od 2006 roku przygotowywała się do zbudowania portalu obejmującego dokumenty cyfrowe z różnego rodzaju instytucji kultury i dającego możliwość jednoczesnego przeszukiwania zasobów europejskiego dziedzictwa kulturowego, dostępnego w językach krajów Unii Europejskiej. Portal ten nazwany został Europeana" (*Europeana – pomyśl o kulturze*, [online] <http://www.bn.org.pl/zasoby-cyfrowe-i-linki/europeana>, dostęp: 1.12.2013).

[19] Archiwum Internetu Narodowego Archiwum Cyfrowego, [online] <http://www.archiwum-internetu.nac.gov.pl>, dostęp: 1.12.2013.

w zakresie ich podejmowania. W dalszej części pracy postarano się zebrać i omówić problemy techniczne, prawne, etyczne i ekonomiczne mające, zdaniem autorki, istotny wpływ na ostateczny kształt procesu archiwizacji zasobów Internetu oraz zarządzanie zgromadzonymi danymi.

Aspekty techniczne

Z technicznego punktu widzenia, archiwizowanie zasobów Internetu (ang. *harvesting*) jest:

> zautomatyzowanym procesem gromadzenia zbiorów i metadanych, które są następnie indeksowane i składowane w archiwum cyfrowym według ściśle określonych parametrów. Służą do tego specjalne oprogramowania bazujące na pracy robotów (ang. *web crawlers*) „przeczesujących" wybrane obszary Sieci zgodnie z zadanymi wymaganiami danego archiwum[20].

Wydawać by się więc mogło, że sam proces zbierania materiału źródłowego jest prosty i w pełni zautomatyzowany[21], pozwalający na przeczesanie zadanych robotom obszarów. Niestety, pomimo zaawansowania technologicznego wciąż pojawiają się problemy z archiwizowaniem stron zbudowanych w oparciu o instrumentarium Flash, co stanowi istotny hamulec w pozyskiwaniu całości materiału ze strony. Fragmentaryzacji podlega nie tylko zapis stron, lecz także Internetu rozumianego jako całość, jego zasoby podzielić można bowiem na te znajdujące się na jego powierzchni (ang. *surface web*) oraz w tak zwanej Sieci głębokiej (ang. *deep web*). Możliwości tradycyjnych wyszukiwarek są ograniczone jedynie do Sieci powierzchniowej, a zasoby Sieci głębokiej, będącej zasobem około 500 razy bogatszym, pozostają poza zasięgiem możliwości rejestracji i pobierania z nich danych[22]. Podane więc wartości liczbowe, ukazujące trudne wprost do rozumowego ogarnięcia wartości pamięci, są w świetle przytoczonych danych jedynie wierzchołkiem góry lodowej.

Problemów nastręcza także zapisywanie treści dostępnych na portalach społecznościowych i portalach blogowych. W pierwszym ze wspomnianych

---

[20] L. Derfert-Wolf, dz. cyt. Głównie wykorzystuje się narzędzia stworzone przez IIPC, na przykład robota Heritix.

[21] Jak słusznie zauważa Julien Masanès, rozwój wirtualnego kontentu stwarza użytkownikom Internetu wiele nowych możliwości, równocześnie powodując znaczne utrudnienia w pracy związanej z archiwizowaniem zasobów sieciowych. W literaturze przedmiotu opisano wielość podejść do procesu archiwizacji. Lekturę artykułu *Web Archiving Methods and Approaches: A Comparative Study*, w którym autor proponuje podejście porównawcze i zintegrowane, polecam czytelnikom zainteresowanym systematycznymi studiami nad metodologią procesu archiwizacji zasobów sieciowych. Zob. J. Masanès, *Web Archiving Methods and Approaches: A Comparative Study*, [online] <https://www.ideals.illinois.edu/bitstream/handle/2142/2452/Masanes.pdf?sequence=2>, dostęp: 1.03.2014. Zob. również: tenże, *Web Archiving*, Berlin – Heidelberg 2006, [online] <http://stevejones.me/pubs/2006/WebUseWebStudies.pdf>, dostęp: 7.03.2014.

[22] E. Sweeney, K. Curran, E. Xie, *Automating Information Discovery within the Invisible Web*, w: *Web-Based Support System*, red. JingTao Yao, London 2010, s. 171.

przypadków powstaje pytanie, jaką przyjąć metodę gromadzenia danych, ponieważ statyczna metoda skanowania danych z jednego profilu nie ukazuje w żadnej mierze społecznościowego charakteru zamieszczanych treści, powiązanie zaś wpisów na wielu kontach różnych użytkowników nie zawsze jest prawnie możliwe (o czym w dalszej części tekstu)[23]. Dokonywanie zatem zapisu danego konta bez szerszego odniesienia do pozostałych to jak rejestracja słów tylko jednej osoby uczestniczącej w grupowej polemice. W tym więc przypadku i sens, i pożytek z takiego materiału wydaje się być istotnie ograniczony.

Kolejną kwestią niepodlegającą automatyzmowi jest praca koncepcyjna, wytwór pracy intelektualnej człowieka odpowiedzialnego za decydowanie w sprawie doboru stron do archiwizowania oraz częstotliwości robienia kopi[24]. Ta z pozoru prosta czynność dotyczy miliardów stron, co w znacznym stopniu komplikuje i wydłuża proces decyzyjny i kwalifikacyjny. Dodatkowo należy mieć świadomość, że puszczona w ruch maszyna musi być poddawana permanentnej kontroli człowieka. Po pierwsze dlatego, że narzędzia i metody zawsze mogą być zawodne, a więc istnieje konieczność poprawiana jakości filtra archiwizującego, po drugie, że napotkawszy na złośliwe oprogramowanie robot nie posuwa się dalej, i w końcu po trzecie, że w wirtualnej przestrzeni publicznej niektóre strony, początkowo bez większej wartości badawczej, w wyniku zmiany społecznej nabierają ogromnej wagi. Zdarza się również, że w związku z ważkimi wydarzeniami niemalże w ciągu doby powstają nowe serwisy, stanowiące bogate źródło danych o dziejącym się w danym momencie historycznym wydarzeniu[25]. Brak należytej elastyczności w działaniu oraz opieszałość w reagowaniu na zmieniającą się sytuację powoduje więc wyparcie z Sieci informacji ważnych przez potok nic niewartych newsów[26].

Trzeba też brać pod uwagę kwestię tyle oczywistą, co trudną do realizacji, a odnoszącą się do konieczności zapewnienia zgromadzonym danym odpowiedniego poziomu bezpieczeństwa: oto przekazanie baz danych w nieodpowiednie ręce może skutkować nienależytym ich wykorzystaniem. W tym kontekście wspomnieć można chociażby o osobie Edwarda Snowdena, byłego

[23] W 2010 roku Biblioteka Kongresu Stanów Zjednoczonych ogłosiła rozpoczęcie działań archiwizacyjnych nad portalem Twittera. Zapisowi miały podlegać wszystkie publiczne wpisy począwszy od 2006 roku. Zob. M. Wilkowski, *Archiwum Twittera w Bibliotece Kongresu*, „Historia i Media", [online] <http://historiaimedia.org/2011/06/14/archiwum-twittera-w-bibliotece-kongresu>, dostęp: 2.12.2013.

[24] Więcej na ten temat zob. S. Abiteboul, *Issues in Monitoring Web Data*, w: *Database and Expert Systems Applications*, red. A. Hameurlain, R. Cicchetti, R. Traunmüller, Berlin – Heidelberg 2002, [online] <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4562&rep=rep1&type=pdf>, s. 5, dostęp: 1.03.2014.

[25] Jako przykład można podać stronę poświęconą wydarzeniom związanym z przejściem nad USA w 2005 roku huraganów Katrina i Rita. Zob. [online] <http://www.hurricanearchive.org>, dostęp: 30.11.2013.

[26] British Library jako argument za archiwizowaniem zasobów w Sieci podaje fakt, że wiele cennych danych na temat zamachu terrorystycznego w Londynie w 2005 roku zniknęło z Sieci bezpowrotnie. Zob. *British Library chce zarchiwizować internet*, [online] <http://www.tvn24.pl/british-library-chce-zarchiwizowac-internet,316939,s.html>, dostęp: 1.12.2013.

pracownika Central Intelligence Agency (CIA) i National Security Agency (NSA), który na łamach prasy ujawnił informację o programie inwigilacji prowadzonym przez NSA[27], polegającym na monitorowaniu na międzynarodową skalę skrzynek mailowych, serwisów społecznościowych oraz połączeń telefonicznych zwykłych obywateli i wpływowych europejskich dygnitarzy. Dane te, przechowywane w bazie danych, zostały przez Snowdena upublicznione, dyskredytując dyplomację Stanów Zjednoczonych. Problematyka bezpieczeństwa wiąże się zresztą nie tylko z kwestią zasadnego udostępniania danych, lecz także, o czym warto na marginesie wspomnieć, z zarządzaniem strumieniami informacji, ponieważ właściwy, logiczny i spójny podział na odpowiednie kategorie modeluje szerokość dostępu do zgromadzonych treści. Tak więc i w tym przypadku czynnik ludzki ma największe znaczenie.

Aspekty etyczne

Po podjęciu tematu archiwizacji zasobów Internetu w aspekcie kwestii technicznych niemal automatycznie nasuwa się pytanie o to, czy chcąc w pewnym sensie zatrzymać czas, przechowywać będziemy wszystko, a więc i to, co najbardziej nieludzkie, czarne, krwią spływające i wulgarne. Czy może jednak czynić będziemy starania, aby w oczach przyszłych pokoleń wybielić karty historii i archiwizacji poddawać tylko rzeczy piękne, wzniosłe i wartościowe?

Wybór nie jest oczywisty, a ocena jednoznaczna, albowiem przyjąwszy, że chronimy dziedzictwo kulturowe, sprawa wydaje się bezdyskusyjna, jeśli zaś dla przyszłych pokoleń chcemy zachować zapis naszego życia społecznego, to warto dyskutować nad tym, co zapisywać i jak te dane potem udostępniać. Warto również zauważyć, że dokonywanie oceny, arbitralne przyporządkowywanie oraz kwalifikacja stron do projektu dokonuje się jako proces myślowy jednostki posiadającej pewne przekonania i system wartości. Z tego też względu pozostawienie tak ważnych decyzji w rękach wąskiej grupy osób może budzić zastrzeżenia co do tendencyjności w zakresie stosowanych kryteriów oceny. Wychodząc najprawdopodobniej z podobnego założenia, twórcy angielskiego projektu UK Web Archive (UKWA) na swojej stronie internetowej zamieścili informacje o stosowanych, w związku ze wspomnianymi problemami, kryteriach. Zgodnie z zamieszczoną informacją, do archiwizacji kieruje się materiały, które powinny być przydatne do badań naukowych przez dłuższy okres i odzwierciedlać różnorodność zainteresowań, działań oraz ogólnie życia w Zjednoczonym Królestwie, a także powinny ukazywać innowacyjność stron internetowych[28]. W praktyce jednak, jak zauważa Katarzyna Gmerek,

[27] Anglojęzyczna strona Wikipedii udostępnia bardzo dokładny biogram E. Snowdena oraz opis informacji będących w jego posiadaniu. Zob. [online] <http://en.wikipedia.org/wiki/Edward_Snowden>, dostęp: 1.12.2013.

[28] K. Gmerek, *Archiwa internetowe po obu stronach Atlantyku – Internet Archive, Wayback Machine oraz UK Web Archive*, „Biuletyn EBIB" 2012, nr 1(128), s. 7, [online] <http://www.ebib.pl/images/stories/numery/128/128_gmerek.pdf>, dostęp: 1.12.2013.

UKWA zarchiwizowało bardzo niewiele stron irlandzkich, choć w rzeczywistości na serwerach posadowionych nie tylko w Republice, lecz także w USA znajduje się dużo stron WWW dotyczących Irlandii Północnej. Uznaniowość ta jest możliwa, ponieważ stanowi pokłosie przyjętej – autorytarnej – metody zarządzania procesem archiwizowania zasobów Sieci. W przeciwieństwie do niego wskazać można na model demokratyczny funkcjonujący w strukturach Internet Archive (IA), zasadzający się na współpracy specjalistów z rzeszą wolontariuszy. Dzięki przyjęciu takiego rozwiązania, jak pisze dalej K. Gmerek, „liczba samych zdigitalizowanych tekstów sięgnęła we wrześniu 2011 roku trzech milionów, z czego dwa wprowadzili pracownicy IA, zaś pozostały milion stowarzyszeni ochotnicy z całego świata, którym udostępniono odpowiednie oprogramowanie i miejsce na serwerach w USA”[29].

Wybór metody, należy podkreślić z całą mocą, ma ogromny wpływ na proces archiwizacji w wielu jego aspektach, to bowiem ludzie podejmują decyzje o tym, co będzie zapisane, jak dokładnie zapis ten zostanie zrealizowany i w końcu ile z zasobów Sieci uda się zarchiwizować. Kolejną istotną w tym kontekście kwestią jest sprawa szacunku i ochrony godności osób. Wśród potoku zamieszczanych w Sieci danych dostępne stają się także wizerunki osób prywatnych, ich dane osobowe, informacje godzące w ich dobre imię czy w końcu treści powszechnie uznawane za wulgarne. W procesie archiwizacji dokonywanej przez roboty zapisywane jest wszystko, bez dokonywania wstępnej oceny aksjologicznej. Osoba pomówiona, pomimo uzyskania po latach satysfakcjonującego ją wyroku sądu, nie jest więc w stanie także w wirtualnym świecie oczyścić się z postawionego jej zarzutu. Społeczność sieciowa wyrok wydaje bez rozpoznania sprawy, a przebieg społecznego osądu jeszcze latami jest dostępny w archiwum Sieci.

Na koniec jeszcze jeden istotny, zasygnalizowany przez Gmerek fakt. Autorka ta wskazuje na niepokojące zjawisko manipulacji w procesie archiwizacji. Polega ono na próbach wywierania wpływu na przyszły kształt i zawartość internetowych materiałów źródłowych. Dla przykładu:

> IA usunęło na żądanie scjentologów stronę poświęconą polemice z nimi, choć nikt nie usuwał stron propagujących tę religię. Ostatnio wiele kontrowersji wzbudza wzmożona aktywność radykalnych użytkowników islamskich, umieszczających w Internet Archive tak dużo materiałów, jak to tylko możliwe. Dotyczy to wszelkich dokumentów – audiowizualnych i tekstów oraz stron www[30].

Czyszczenie kart historii, a także wpływanie na nadreprezentowanie niektórych treści stanowiących w przyszłości obraz czasów, w istotny sposób może wpłynąć na rzetelną ocenę czasów nam współczesnych.

[29] Tamże.
[30] Tamże, s. 5.

Aspekty ekonomiczne

Obok refleksji nad zasadnością realizacji procesu archiwizacji należy równocześnie podjąć namysł nad ekonomicznie rozumianą opłacalnością przedsięwzięcia. Postęp technologiczny w sposób oczywisty wymusza digitalizację treści, ale to, co dla odbiorców wydaje się naturalne, dla instytucji wiąże się z dużymi nakładami finansowymi. Tymczasem z badania „inicjatyw europejskich prowadzonych przez Internet Memory Foundation wynika, że 52,7% projektów nie dysponuje specjalnym budżetem, 5,5% posiada budżet mniejszy niż 10 tys. euro, a 16,4% ma do dyspozycji ponad 200 tys. euro"[31]. Okazuje się więc, że bez posiadania istotnych nakładów finansowych można, choć w ograniczonym stopniu, realizować plan digitalizacji posiadanych zasobów. Możliwość ta istnieje w związku z faktem, że choć archiwizacja materiałów cyfrowych jest procesem kosztownym, to ich przechowywanie jest znacznie tańsze. Pokazuje to przykład badań wybranych światowych bibliotek, w których zestawiono ze sobą wydatki poniesione w latach 2007–2011 na utrzymanie źródeł elektronicznych w całości poniesionych kosztów. Stosunek wspomnianych wydatków zdecydowanie potwierdza przekonanie o niskich kosztach przechowywania zasobów w postaci elektronicznej[32].

Konieczność poniesienia wysokich nakładów finansowych w procesie pozyskiwania materiału jest związana w efekcie finalnym z logiką udostępniania danych odbiorcy końcowemu. Zarządzanie zgromadzonymi zasobami odbywać się zatem może na warunkach komercyjnych, czyli odpłatnie, lub też dostęp do zasobów może mieć charakter całkiem wolny. Paradoksalnie jednak żadne z tych dwóch rozwiązań nie jest możliwe do wprowadzenia w całości. Całkowite skomercjalizowanie dostępu do archiwalnych zasobów Internetu stanęłoby w sprzeczności z prawem w zakresie choćby Ustawy o dostępie do informacji publicznej[33], godziłoby w interesy niektórych instytucji i organizacji (partii politycznych czy organizacji pożytku publicznego), jak również w znacznym stopniu mogłoby wpłynąć na obniżenie motywacji cyberwolontariuszy do tworzenia cyfrowego kontentu. *A contrario*, zniesienie odpłatności za dostęp do wszystkich zarchiwizowanych danych pozwoliłoby na swoiste obejście konieczności poniesienia opłaty za dostęp do wybranych treści, takich jak na przykład archiwa gazet, dodatki tematyczne, pełne wersje raportów i analiz firm badawczych czy choćby dostęp do wybranych baz danych. Iście salomonowym rozwiązaniem jest powiązanie obu metod, jednakże stosowanie

[31] *Web Archiving in Europe. A Survey Provided by the Internet Memory*, [online] <http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf>, za: L. Derfert-Wolf, dz. cyt.

[32] Wydatki na zasoby elektroniczne *versus* ogół wydatków w latach 2007–2011 w wybranych bibliotekach świata zob. Y. Sung, *The Rapid Growth of Electronic Resources in East Asian Library Collections*, [online] <http://www.arl.org/storage/documents/publications/lcdp-2012-poster-sung-yunah.pdf>, dostęp: 1.12.2013.

[33] Ustawa z dnia 6 września 2001 roku o dostępie do informacji publicznej (Dz.U. 2001, nr 112, poz. 1198).

dostępów mieszanych generuje dodatkowe koszty i trudności w zakresie ściślejszego nadzoru nad procesem zarządzania danymi zarchiwizowanymi.

Aspekty prawne

Problematyka prawnych zawiłości Sieci jest już dość dobrze opisana w literaturze przedmiotu. Nieco gorzej wygląda natomiast sama infrastruktura prawna. W opisie do książki Piotra Waglowskiego *Prawo w sieci. Zarys regulacji internetu* można przeczytać:

> Dynamiczny rozwój internetu sprawił, że organy ustawodawcze wielu państw nie nadążały z przygotowywaniem regulacji prawnych dotyczących korzystania z niego. Początkowy okres rozwoju sieci przypominał kolonizację nowego lądu – pełna bezkarność, chaos i brak jakichkolwiek ograniczeń. Ten stan rzeczy doprowadził do sytuacji, w której nawet najgorętsi orędownicy wolności w sieci zaczęli się zastanawiać, jak uniknąć płynących z niej zagrożeń. Zaczęto więc tworzyć prawa, regulacje i przepisy, które jednak nie zawsze spełniają swoją rolę. Na całym świecie toczy się dyskusja nad kształtem regulacji prawnych społeczeństwa informacyjnego[34].

Brak jasnych i spójnych przepisów prawa regulujących sposób funkcjonowania w ramach Sieci generuje zatem trudności związane z archiwizacją internetowych zasobów już u samych jej początków. Fundamentalną kwestią jest bowiem problem samego prawa do zapisu strony WWW. Przyczyn tego doszukiwać się można w sytuacji, w której w ramach jednej strony są zamieszczane utwory (teksty, obrazy, muzyka) podlegające prawu autorskiemu – i takie, które spod tego prawa zostały „wyjęte", nie zapomniawszy o fakcie, że sam projekt strony internetowej może także być chroniony przywołanym przepisem[35]. W realizacji projektu UKWA przyjęto zasadę, że archiwizacji nie podlegają zasoby, których właściciele nie wyrazili na ten proces zgody (poprzez na przykład zamieszczenie informacji o licencji Creative Commons lub innej deklaracji zgody), albo też nie zostały one ustawowo wyodrębnione (tak jak przykładowo strony publiczne). Odmienną logiką kieruje się kierownictwo IA, które archiwizuje wszystko, a na prośbę zainteresowanej strony usuwa z posiadanych zasobów strony i treści, których właściciele tego sobie zażyczą. Przyjęcie tej, w gruncie rzeczy najprostszej, metody ma także swoje słabe strony. Otóż Wayback Machine miała wiele spraw sądowych, w których powodowie domagali się zaprzestania archiwizowania ich stron oraz likwidacji danych już zapisanych[36].

---

[34] [Online] <http://prawo.vagla.pl/prawo_w_sieci>, dostęp: 1.12.2013.

[35] Dodatkowo sama strona może być utworem, stanowiąc przejaw czyjejś unikatowej działalności artystycznej, czyli „o ile istotnie jest przejawem działalności twórczej i ma charakter indywidualny, ustalony w jakiejkolwiek postaci, niezależnie od wartości, przeznaczenia i sposobu wyrażania" (P. Waglowski, *Prawo w sieci. Zarys regulacji internetu*, Gliwice 2005, s. 123).

[36] K. Gmerek, dz. cyt., s. 5.

Na gruncie polskich przepisów prawa funkcjonuje wiele zapisów, które pomimo dużej łącznie liczby nadal nie wyjaśniają sprawy. Zgodnie bowiem z Zaleceniami Komisji Europejskiej z dnia 27 października 2011 roku w sprawie digitalizacji i udostępniania w Internecie dorobku kulturowego oraz w sprawie ochrony zasobów cyfrowych (2001/711/UE)[37] wprowadzono nowelizację Ustawy o obowiązkowych egzemplarzach bibliotecznych[38]. Ustawa ta miała uporządkować kwestię prawa do zapisu oraz doprecyzować zapisy Ustawy o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 roku[39]. Z drugiego z przywołanych zapisów wynika, że „egzemplarzowi obowiązkowemu podlegają nie tylko utwory rozpowszechniane drukiem, lecz również egzemplarze utworów rozpowszechnianych za pośrednictwem sieci"[40]. A zatem uproszczenie procedury archiwizacji polegało w intencjach legislatorów na stworzeniu prawa do zapisu i archiwizowania utworów, które podlegałyby przechowaniu nie tylko w formie materialnej, ale również cyfrowej. Ustawa ta była dobrym krokiem w kierunku uporządkowania zapisów, ale dotyczy jedynie kwestii utworów.

W powyższym kontekście warto wspomnieć o jeszcze jednym zapisie prawnym – Ustawie o ochronie danych osobowych[41]. Wątpliwości budzi archiwizowanie stron pochodzących z portali społecznościowych, na łamach których pojawiają się informacje o danych osobowych dostępnych w formie pozwalającej na zidentyfikowanie osoby fizycznej (art. 6 ust. 1), przy czym:

> osobą możliwą do zidentyfikowania jest osoba, której tożsamość można określić bezpośrednio lub pośrednio, w szczególności poprzez powołanie się na numer identyfikacyjny albo jeden lub kilka specyficznych czynników określających jej cechy fizyczne, fizjologiczne, umysłowe, ekonomiczne, kulturowe lub społeczne[42].

Warto zwrócić uwagę, że te wymienione powyżej dane dość standardowo zamieszcza się na portalach społecznościowych, takich jak Nasza Klasa (NK.pl) czy Facebook. Firmy będące właścicielami tych portali przechowują

[37] *Commission Recommendation of 27 October 2011 on the Digitization and Online Accessibility of Cultural Material and Digital Preservation* (2001/711/EU), [online] <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:283:0039:0045:EN:PDF>, dostęp: 7.12.2013.

[38] Ustawa z dnia 7 listopada 1996 roku o obowiązkowych egzemplarzach bibliotecznych, [online]<http://www.bn.org.pl/zbiory/egzemplarz-obowiazkowy/ustawa-o-obowiazkowych-egzemplarzach-bibliotecznych>, dostęp: 7.12.2013.

[39] Ustawa z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. 1994, nr 24, poz. 83).

[40] K. Ślaska, A. Wasilewska, *Archiwizacja internetu – sytuacja w polskim prawie z punktu widzenia bibliotekarzy*, „Biuletyn EBIB" 2012, nr 1(128), s. 3, [online] <http://www.ebib.pl/images/stories/numery/128/128_slaska.pdf>, dostęp: 19.09.2013.

[41] Ustawa z dnia 29 sierpnia 1997 roku o ochronie danych osobowych (Dz.U. 1997, nr 133, poz. 883), [online] <http://isip.sejm.gov.pl/DetailsServlet?id=WDU19971330883>, dostęp: 9.12.2013.

[42] Tamże, art. 6 ust. 2.

i przetwarzają dane osobowe osób fizycznych[43], a w przypadku archiwizowania ich zasobów dane te zostaną przekazane do podmiotu trzeciego, co spowoduje sytuację utrudniającą jednostce wgląd, zmianę, aktualizację oraz prawo żądania usunięcia swoich danych z rejestru.

Zagarnianie czy *de facto* prywatyzacja własności jednostek składających się w swej masie na zasób dziedzictwa narodowego stało się głośną i kontrowersyjną kwestią za sprawą koncernu Google. Otóż od 2005 roku firma ta, udostępniając usługę Google Books (pierwotnie Google Print), pozwala na wyszukiwanie książek znajdujących się w największych światowych księgarniach oraz bibliotekach. Część z pozycji dostępnych z poziomu wyszukiwarki można przeczytać w całości, część z nich jedynie przeglądnąć (za pomocą słowa kluczowego). W ten sposób koncern mający w Polsce pozycję monopolisty na rynku wyszukiwarek[44] w odpowiedzi na zapotrzebowanie klientów poszerzył swoją ofertę w sposób legalny (stając się jedynie pośrednikiem dostępnych w Sieci zasobów), choć bardzo kontrowersyjny (nie wszystkie zamieszczone w Sieci książki zostały udostępnione za zgodą autora lub/i wydawnictwa).

Zakończenie

Proces archiwizacji zasobów globalnej Sieci trwa nieprzerwanie od 1996 roku. W różnych częściach świata proces ten przybiera odmienne formy i natężenie. Kwestia konieczności ochrony wirtualnego kontentu rozumianego jako dziedzictwo kulturowe oraz materiał historyczny nie podlega już dzisiaj dyskusji. Sporna natomiast wydaje się kwestia wyboru modelu realizacji tego procesu. Ze względów metodologicznych i w zakresie oceny materiału źródłowego lepszy jest model autokratyczny. Znając jednak realia polskiego budżetu, można przewidywać, że fundusze przeznaczane na naukę nie pozwolą na działalność tego typu. Nieodzowne stanie się zatem posiłkowanie się grupą wolontariuszy. Krok ten spowoduje zwiększenie tempa przyrostu treści zdigitalizowanych, zmniejszając równocześnie kontrolę nad cyfrowym zasobem. Pojawiające się w tym kontekście, a opisane w artykule problemy są immanentną cechą procesu. Ważne jest więc, aby odpowiedzi na newralgiczne pytania udzielić sobie zawczasu, zanim proces ten wymknie się spod kontroli, i zanim zarządzanie zarchiwizowanymi informacjami rozumiane będzie nie jako proces porządkowania utrwalonych zasobów, lecz jako walka o przejęcie kontroli nad informacjami.

[43] Warto przywołać głośną sprawę sporu o prawo portalu Nasza Klasa do przetwarzania danych osobowych osób posiadających na portalu profile prywatne. Zob. P. Waglowski, *GIODO: imię, nazwisko, zdjęcie, szkoła, klasa i rocznik – łącznie – nie są danymi osobowymi...*, [online] <http://prawo.vagla.pl/node/8090>, dostęp: 1.12.2013.

[44] Z silnika Googla korzysta w Polsce 97,5% internautów. Zob. Ł. Szewczyk, *Wyszukiwarki i katalogi*, [online] <http://media2.pl/badania/100174-Megapanel-grudzien-2012-kategorie-tematyczne/20.html>, dostęp: 12.10.2013.

Bibliografia

Abiteboul S., *Issues in Monitoring Web Data*, w: *Database and Expert Systems Applications*, red. A. Hameurlain, R. Cicchetti, R. Traunmüller, Berlin – Heidelberg 2002, [online] <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4562&rep=rep1&type=pdf>, dostęp: 1.03.2014.

Ainsworth S.G., AlSum A., SalahEldeen H., Weigle M.C., Nelson M.L., *How Much of the Web Is Archived?*, [online] <http://arxiv.org/pdf/1212.6177v2.pdf>, dostęp: 1.12.2013.

British Library chce zarchiwizować internet, [online] <http://www.tvn24.pl/british-library-chce-zarchiwizowac-internet,316939,html>, dostęp: 1.12.2013.

Cho J., Garcia-Molina H., *The Evolution of the Web and Implications for an Incremental Crawler*, [online] <http://ilpubs.stanford.edu:8090/376/1/1999-22.pdf>, dostęp: 10.12.2013.

Commission Recommendation of 27 October 2011 on the Digitization and Online Accessibility of Cultural Material and Digital Preservation (2001/711/EU), [online] <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:283:0039:0045:EN:PDF>, dostęp: 7.12.2013.

Derfert-Wolf L., *Archiwizacja Internetu – wprowadzenie i przegląd wybranych inicjatyw*, „Biuletyn EBIB" 2012, nr 1(128), [online] <http://eprints.relis.org/17048/1/derfert_Web_archiving.pdf>, dostęp: 10.09.2013.

Europeana – pomyśl o kulturze, [online] <http://www.bn.org.pl/zasoby-cyfrowe-i-linki/europeana>, dostęp: 1.03.2014.

Gmerek K., *Archiwa internetowe po obu stronach Atlantyku – Internet Archive, Wayback Machine oraz UK Web Archive*, „Biuletyn EBIB" 2012, nr 1(128), [online] <http://www.ebib.pl/images/stories/numery/128/128_gmerek.pdf>, dostęp: 10.09.2013.

Gogołek W., *Komunikacja sieciowa. Uwarunkowania, kategorie, paradoksy*, Warszawa 2010.

Jankowska M., *Biblioteki akademickie – trendy dotyczące zasobów elektronicznych*, w: *Informacja dla nauki a świat zasobów cyfrowych*, red. H. Ganińska, Poznań 2008.

Masanès J., *Web Archiving*, Berlin – Heidelberg 2006, [online] <http://stevejones.me/pubs/2006/WebUseWebStudies.pdf>, dostęp: 1.03.2014.

Masanès J., *Web Archiving Methods and Approaches: A Comparative Study*, [online] <https://www.ideals.illinois.edu/bitstream/handle/2142/2452/Masanes.pdf?sequence=2>, dostęp: 1.03.2014.

Mrożewski B., *Archiwa internetu*, „PC Format" 2012, nr 4, [online] <http://www.pcformat.pl/Archiwa-internetu,a,2374>, dostęp: 7.12.2013.

Sung Y., *The Rapid Growth of Electronic Resources in East Asian Library Collections*, [online] <http://www.arl.org/storage/documents/publications/lcdp-2012-poster-sung-yunah.pdf>, dostęp: 1.12.2013.

Sweeney E., Curran K., Xie E., *Automating Information Discovery within the Invisible Web*, w: *Web-Based Support System*, red. JingTao Yao, London 2010.

Szewczyk Ł., *Wyszukiwarki i katalogi*, [online] <http://media2.pl/badania/100174-Megapanel-grudzien-2012-kategorie-tematyczne/20.html>, dostęp: 12.10.2013.

Ślaska K., Wasilewska A., *Archiwizacja internetu – sytuacja w polskim prawie z punktu widzenia bibliotekarzy*, „Biuletyn EBIB" 2012, nr 1(128), [online] <http://www.ebib.pl/images/stories/numery/128/128_slaska.pdf>, dostęp: 19.09.2013.

Waglowski P., *GIODO: imię, nazwisko, zdjęcie, szkoła, klasa i rocznik – łącznie – nie są danymi osobowymi...*, [online] <http://prawo.vagla.pl/node/8090>, dostęp: 1.12.2013.

Waglowski P., *Prawo w sieci. Zarys regulacji internetu*, Gliwice 2005.

Wilkowski M., *Archiwum Twittera w Bibliotece Kongresu*, „Historia i Media", [online] <http://historiaimedia.org/2011/06/14/archiwum-twittera-w-bibliotece-kongresu>, dostęp: 2.12.2013.

Wilkowski M., *Trzy argumenty przeciwko archiwizowaniu Internetu*, „Historia i Media", [online] <http://historiaimedia.org/2011/10/04/trzy-argumenty-przeciwko-archiwizowaniu-internetu>, dostęp: 10.04.2013.

Dokumenty prawne

Ustawa z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. 1994, nr 24, poz. 83).

Ustawa z dnia 7 listopada 1996 roku o obowiązkowych egzemplarzach bibliotecznych, [online] <http://www.bn.org.pl/zbiory/egzemplarz-obowiazkowy/ustawa-o-obowiazkowych-egzemplarzach-bibliotecznych>, dostęp: 7.12.2013.
Ustawa z dnia 29 sierpnia 1997 roku o ochronie danych osobowych (Dz.U. 1997, nr 133, poz. 883).
Ustawa z dnia 6 września 2001 roku o dostępie do informacji publicznej (Dz.U. 2001, nr 112, poz. 1198).

Streszczenie

Zadanie archiwizacji zasobów, które w przypadku mediów tradycyjnych nie nastręcza zbyt wielu kłopotów, w odniesieniu do Internetu wydaje się niemal niemożliwe do realizacji. Przy okazji procesu archiwizacji powstaje wiele pytań dotyczących kwestii zarządzania zbieranymi informacjami, sposobu udostępniania i gromadzenia danych oraz samego doboru stron, które temu procesowi mają podlegać. Warto więc badać wspomniany obszar, ponieważ w polskiej literaturze przedmiotu tematyka archiwizacji Internetu jest poruszana jedynie marginalnie, a sam temat nie jest powszechnie znany, ponadto Polska nie podjęła jeszcze instytucjonalnie usankcjonowanych, systematycznych prac nad archiwizacją zasobów sieciowych. Wydaje się przy tym, że zarówno w Europie, jak i w innych częściach świata wypracowano już narzędzia oraz międzynarodowe standardy w zakresie archiwizowania wirtualnego kontentu i zarządzania zgromadzonymi danymi. Z tego też względu celem niniejszego artykułu jest przegląd głównych kwestii związanych z problematyką zarządzania informacją w procesie archiwizowania zasobów Internetu. Ich klasyfikacji (aspekty techniczne, etyczne, ekonomiczne i prawne) dokonano w oparciu o doświadczenia z już istniejących projektów oraz w odniesieniu do analizy krytycznej prezentowanej w literaturze przedmiotu.

Summary

Information management in the process of web archiving

In the case of traditional media, the archiving of contents does not cause many problems. However, when it comes to the Internet it seems almost impossible to proceed. While the process of web archiving a number of questions concerning the information management occur, such as how to select and share data and which the sites should be chosen to archive. Therefore, it is quite important to explore the area.

Moreover, in the Polish literature, the issue of web archiving is mentioned very little and this problem is not well known. In addition, Poland has not yet enforced systematic work on the archiving of network resources. It seems that in Europe and in the other parts of the world, the tools and international standards regarding web archiving and collected information management have been already developed. Therefore, the aim of this article is to review the main dilemmas related to the issues of information management in the process of web archiving. The presented classification is based on the experience of existing projects and the critical analysis presented in professional literature.