Title: Lodowcowa epoka archiwizowania zasobów

Author: Anna Małgorzata Kamińska

Citation style: Kamińska Anna Małgorzata. (2018). Lodowcowa epoka archiwizowania zasobów. “Nowa Biblioteka” (nr 4 (2018), s. 21-35)

Nowa Biblioteka. Usługi, Technologie Informacyjne i Media 2018, nr 4 (31), s. 21–35
ISSN 2451-2575 (wersja elektroniczna)
http://www.nowabiblioteka.us.edu.pl/
GICID 71.0000.1500.0677

Anna Małgorzata Kamińska
Zakład Bibliotekoznawstwa
Instytut Bibliotekoznawstwa i Informacji Naukowej
Uniwersytet Śląski w Katowicach
e-mail: anna.kaminska@us.edu.pl

Lodowcowa epoka archiwizowania zasobów

Abstrakt: Będąc uczestnikami społeczeństwa informacyjnego, stajemy się konsumentami i producentami coraz większej ilości informacji. Ich duża część ma ogromne znaczenie dla prawidłowej realizacji procesów gospodarczych i administracyjnych oraz kształtowania postaw społecznych, kulturowych czy estetycznych, i dlatego ich ochrona staje się sprawą wagi kluczowej. W artykule autorka wyjaśnia, dlaczego archiwizowanie zasobów w formach cyfrowych wiąże się z ochroną tych zasobów w długiej perspektywie czasu. Następnie omawia cechy, jakie powinny posiadać archiwa długoterminowej ochrony danych, i przedstawia technologie pozwalające na optymalne kosztowo budowanie tych archiwów. Mimo że technologie te są trudne do bezpośredniego zastosowania przez małe i średnie podmioty, to oferta kierowana przez dostawców usług chmurowych umożliwia przechowywanie danych w sposób bezpieczny i korzystny kosztowo. Artykuł kończy propozycja architektury archiwum długoterminowego wykorzystującego usługi dostawców trzecich, poświęcone składowaniu danych.

Słowa kluczowe: Amazon Glacier. Archiwa danych. Azure Archive Blob Storage. Bezpieczeństwo danych. Dostępność danych. Optymalizacja kosztów. Technologie składowania danych

Wstęp

Πάντα ῥεῖ καὶ οὐδὲν μένει[1] to stwierdzenie Heraklita z Efezu dające wyraz jego koncepcji zmiany jako stałej reguły otaczającej nas rzeczywistości. Powszechnie przyjmuje się, że głębsza filozofia ukryta za tymi słowami wpłynęła na kształtowanie poglądów takich myślicieli,

[1] Wszystko płynie, jest w ciągłym ruchu, nic nie stoi w miejscu.

jak Platon czy Protagoras, oraz zapoczątkowała wiele gałęzi filozofii zachodniej. Choć generalny porządek świata wciąż podlega ciągłym badaniom, to wydaje się, że wspomniana wypowiedź Heraklita znalazła ostateczne potwierdzenie w świecie makroskopowym dzięki badaniom Ludwiga Boltzmanna, wyjaśniającym II zasadę termodynamiki. Głosi ona, że entropia (miara nieuporządkowania układu) Wszechświata zawsze wzrasta, co jako pierwsze prawo fizyki nadaje wyróżniony kierunek upływowi czasu.

Wzrost nieuporządkowania cząstek składowych materii postrzegamy m.in. jako procesy starzenia się zarówno w sensie biologicznym, jak i fizycznym. Mimo że przebieg tych procesów dla niewielkich podzbiorów otaczającej nas rzeczywistości (lokalnie) jesteśmy w stanie opóźniać, to skutek opóźniania tych procesów powoduje dalszy wzrost entropii całkowitej (globalnie), co w dłuższej perspektywie musi doprowadzić do jej wzrostu również we wszystkich układach lokalnych.

Zasoby archiwalne, biblioteczne czy muzealne, stanowiące przedmiot rozważań w niniejszym wydaniu „Nowej Biblioteki. Usług, Technologii Informacyjnych i Mediów", rozumiane w ścisłym związku z ich fizyczną formą, nieuchronnie podlegają opisanym wcześniej mechanizmom. Warto zwrócić uwagę na istnienie zależności pomiędzy fizyczną reprezentacją formy tych artefaktów a możliwościami ochrony poszczególnych form przed zgubnymi skutkami upływu czasu. Obrazowym przykładem mogą być gliniane tabliczki z pismem klinowym, z których najstarsze datuje się na około 3000 lat p.n.e., w porównaniu z dokumentami papierowymi, o wiele bardziej podatnymi na fizyczne, chemiczne i biologiczne procesy degradacji.

Wartość zasobów podlegających ochronie oraz wartość przekazywanych przez nie informacji można rozważać na trzech poziomach abstrakcji:

– przekaz treści – odnoszący się jedynie do warstwy semantycznej zasobu (przykładowo treść tekstu czy scena obrazu);
– przekaz formy – odnoszący się do warstwy formy przekazu semantyki zasobu (przykładowo forma linii zakreślona stalówką oraz kolor lub nawet skład chemiczny użytego atramentu w przypadku tekstów, czy wypukłości i ślady pędzla na wyschniętych farbach oraz struktura płótna w przypadku malarstwa olejnego);
– materializacja – odnosząca się do wykorzystania w danym miejscu, czasie i przez daną osobę konkretnych półproduktów fizycznych, których zastosowanie zaowocowało powstaniem zasobu niepowtarzalnego, często nazywanego oryginałem.

O ile ochrona tego ostatniego jako bytu fizycznego, w świetle wcześniej przedstawionych tez polegać może jedynie na opóźnianiu skutków procesów degradacyjnych, o tyle zarówno treść, jak i forma jako byty koncepcyjne/idee mogą zostać opisane informacjami, których istnienie jest niezależne od istnienia ich oryginalnej materializacji. Informacje te, aby zachować trwałość przekazu, muszą oczywiście również zostać zmaterializowane, ale ich nośnikiem nie musi już być analog oryginału, lecz jego bardziej trwały substytut.

Trzeba zauważyć, że informacje o zasobach, gromadzone w postaci cyfrowej (czyli ich digitalizacje), przy zastosowaniu powszechnie już wykorzystywanych algorytmów detekcji, a nawet korekcji błędów, w przeciwieństwie do informacji gromadzonych w postaci analogowej, są odporne na procesy przekłamania czy zniekształcania zapisów. Daje to możliwość wielokrotnego czytania i zapisywania informacji bez utraty jej jakości, co w przypadku ograniczonej trwałości ich nośnika (podlegającego przecież takiemu samemu prawu wzrostu entropii co oryginały) oznacza kopiowanie informacji pomiędzy kolejnymi generacjami nośników, a co za tym idzie – praktycznie nieograniczone w czasie przedłużanie życia chronionych idei.

Archiwa cyfrowe

Spostrzeżenia dotyczące możliwości przedłużania życia informacji cyfrowej, jak również fakt, że współcześnie coraz więcej informacji powstaje od razu w postaci cyfrowej (ang. *born digital*), legły u podstaw koncepcji tworzenia długoterminowych archiwów cyfrowych (ang. *long-term digital archive*).

Podstawowe cechy, jakie powinny posiadać takie systemy, to m.in.:

– pojemność – digitalizacja zasobów, w zależności od wymaganej wierności odwzorowania zarówno w wymiarze ilościowym, jak i jakościowym, może generować duże strumienie danych. Archiwa długoterminowe powinny być przygotowane na gromadzenie dużych wolumenów danych, a ich architektura powinna być zdefiniowana w sposób pozwalający na łatwe i korzystne kosztowo rozbudowywanie systemu w celu zwiększania jego pojemności;
– bezpieczeństwo – dane gromadzone w repozytoriach powinny być przechowywane w sposób uniemożliwiający ich przypadkową utratę na skutek zarówno błędów ludzkich, jak i czynników losowych, tak aby pojedyncza awaria systemu nie była w stanie przerwać łańcucha życia informacji cyfrowej;

– indeksowanie – w przypadku archiwów długoterminowych najczęściej wykorzystywaną funkcją jest prawdopodobnie składowanie zbiorów danych, jednak przynajmniej część informacji o zawartości tych zbiorów powinna być dostępna bez konieczności czasochłonnego przeszukiwania całych zbiorów. Jest to możliwe dzięki zastosowaniu technik indeksowania według wcześniej ustalonego opisowego zbioru metadanych oraz poprzez użycie mechanizmów indeksowania pełnozawartościowego (ang. *full content index*). Znanych jest wiele technik budowania indeksów pełnej zawartości dla dokumentów tekstowych, graficznych czy dźwiękowych. W przypadku metadanych zagłębionych (ang. *embedded*) w archiwach warto rozważyć ich ekstrakcję do poziomu danych indeksowych;
– interpretacja – w zależności od rodzaju digitalizowanych zasobów lub typu informacji opisywanych danym zasobem cyfrowym archiwa mogą być przechowywane w plikach o różnych formatach. W związku z ciągłym rozwojem technologii informatycznych zmieniają się również formaty danych. Interoperacyjność jest ważnym aspektem każdego archiwum danych, a w przypadku archiwów długoterminowych nabiera ona kluczowego znaczenia. Można wymienić wiele już obecnie zapomnianych formatów danych, które były popularne zaledwie kilka lub kilkanaście lat temu. Skłania to do refleksji nad możliwościami interpretacji gromadzonych danych w dłuższej perspektywie czasu. Wydaje się, że minimalnym wymogiem dla archiwów długoterminowych powinno być gromadzenie specyfikacji formatów danych, które są używane w archiwizowanych zasobach. Pozwoli to przy mniejszym lub większym nakładzie pracy na odtworzenie treści zapisanej za pomocą danego pliku. W razie potrzeby szybkiego dostępu do zawartości archiwów wykorzystujących nieaktualne formaty platformy repozytoriów należy wyposażyć w mechanizmy konwersji archiwów do formatów używanych współcześnie.

Pogłębione rozważania dotyczące dobrych praktyk składowania danych cyfrowych znaleźć można w opracowaniu *Dobre praktyki publikowania danych badawczych* (Kamińska, 2017), natomiast na temat cech długoterminowych archiwów pisali: Beth Plale, Robert H. McDonald, Kavitha Chandrasekar, Inna Kouper, Stacy Konkiel, Margaret L. Hedstrom, James Myers i Praveen Kumar, prezentując koncepcje programowej warstwy integracyjnej celem stworzenia sfederowanych repozytoriów danych badawczych (Plale i in., 2013). Szersze spojrzenie na zagadnienie źródeł potrzeb tworzenia archiwów długoterminowych w kontekście ich potencjalnych użytkowników oraz wpływu tych potrzeb na wymagania i architekturę przedstawiły Kristin R. Eschenfelder

i Kalpana Shankar (2016). Kwestie długoterminowej archiwizacji danych stanowią przedmiot zainteresowania międzynarodowej organizacji standaryzacji i zostały opisane w dokumencie ISO 14721:2012 (2012). Jednak pracom standaryzacyjnym podlegać mogą jedynie zagadnienia o dobrze zdefiniowanych wymaganiach i wykorzystujące dobrze poznane technologie. W przypadku archiwizacji długoterminowej wyniki tych prac z natury rzeczy mają mocno ograniczony horyzont czasowy. Ciekawe, choć czasem kontrowersyjne spostrzeżenia przedstawia w swoim artykule Wasim Ahmad Bhat (2018).

Długoterminowa archiwizacja danych staje się ważna zarówno w zakresie bieżącego funkcjonowania struktur społecznych, gdzie przepisy prawa często wręcz nakazują długoterminowe przechowywanie w celach dowodowych dokumentacji dotyczącej poszczególnych dziedzin życia, jak i w zakresie ochrony dorobku kulturowo-cywilizacyjnego. Bez skutecznego zachowania różnorakich zasobów w postaci cyfrowej ludzkość narażona jest na bezpowrotną utratę części swojej historii i kultury.

O ile w pierwszym z wymienionych przypadków minimalny czas przechowywania zasobów jest ściśle określony przez czynniki legislacyjne (przykładowo dokumentacja księgowa i bankowa powinny być przechowywane przez pięć lat, zaś dokumentacja medyczna przez dwadzieścia lat), o tyle w przypadku ochrony dziedzictwa kulturowego z przyczyn oczywistych nie zakłada się z góry okresu, po którym dane zasoby „ulegną przedawnieniu".

Wobec powyższego oraz wobec faktu, że wraz z ciągłym postępem technologicznym produkujemy coraz więcej informacji, oczywiste jest, że wolumen danych podlegający archiwizacji będzie przyrastał w znaczącym tempie. Skłania to do refleksji nad metodami i technologiami wykorzystywanymi w ich długoterminowej archiwizacji, tak aby w przyszłości metody te nie stanowiły zagrożenia dla środowiska naturalnego, a jednocześnie były optymalne kosztowo.

Technologie składowania danych

Zagadnienia konstruowania pamięci systemów komputerowych to obszerna dziedzina nauk o technologiach komputerowych. Od czasów powstania pierwszych systemów obliczeniowych postęp technologiczny sprawił, że możliwe stało się konstruowanie pamięci o wiele rzędów wielkości bardziej pojemnych i szybszych oraz obniżenie kosztów ich wytworzenia w przeliczeniu na jednostki pojemności.

W kontekście architektury długoterminowych archiwów danych interesujące są rozważania na temat technologii pamięci nieulotnych (ang. *non-volatile memory*), a w szczególności pamięci pozwalających na gęste upakowanie danych, przy jednoczesnej przystępności kosztów. Kryterium to spełniają systemy z mechaniczną adresacją. Odczyt interesującego w danym momencie fragmentu pamięci (pliku), zapisanego w takim systemie, polega na odpowiednim mechanicznym pozycjonowaniu elementu odczytującego (ang. *head*) nad konkretnym fragmentem poruszającego się pod nim nośnika informacji. Do najbardziej popularnych pamięci tego typu możemy zaliczyć:

– taśmy magnetyczne przechowujące dane na zwijanych wstęgach z tworzywa sztucznego z napylonym na nim materiałem ferromagnetycznym;
– dyski magnetyczne przechowujące dane na powierzchni giętkiego dysku wyciętego z tworzywa sztucznego z napylonym na nim materiałem ferromagnetycznym;
– dyski twarde przechowujące dane na powierzchni sztywnych dysków szklanych pokrytych warstwą metalu z napylonym materiałem ferromagnetycznym i zamkniętych w pyłoszczelnej obudowie zintegrowanej z elementem odczytującym (głowicą);
– dyski optyczne przechowujące dane na powierzchni poliwęglanowych dysków pokrytych materiałem odbijającym/rozpraszającym promienie laserowej głowicy pod różnymi kątami, kodującymi w ten sposób informacje cyfrowe[2].

Obserwując rozwój poszczególnych technologii na przestrzeni lat dojrzewania technologii informatycznych, można stwierdzić, że zarówno dyski magnetyczne, jak i dyski optyczne czasy swojej popularności mają już za sobą, ustępując obecnie miejsca tzw. pamięciom flash (ang. *flash drive*), wykonanym w technologiach półprzewodnikowych (przestały więc to już być urządzenia z mechaniczną adresacją).

Również dyski twarde, mające główne zastosowanie jako szybka, pojemna, ale nieulotna pamięć wewnętrzna większości systemów komputerowych, są powoli zastępowane przez półprzewodnikowe „dyski SSD" (ang. *state solid drive*), gdzie określenie „dysk" posiada już jedynie konotacje historyczne i wynika z przeznaczenia samego urządzenia.

Zupełnie inaczej przebiega natomiast rozwój technologii taśm magnetycznych. Jako najstarsze z omawianych tu pamięci (technologia ta

[2] W rzeczywistości budowa dysku optycznego jest bardziej skomplikowana i może się zmieniać w zależności od zastosowania konkretnej technologii. Więcej na ten temat np. w (Milster, 2003).

ma już ponad 60 lat) towarzyszyły początkom rozwoju systemów superkomputerów, jako pierwsze stanowiły też podstawowy nośnik informacji dla pionierskich ośmiobitowych komputerów domowych.

Szybko wyparte z zastosowań domowych i biurowych przez dyski elastyczne (ang. *floppy disc*) o rozmiarach 8, 5¼ i wreszcie 3½ cala, na długo zniknęły z oczu przeciętnych użytkowników systemów komputerowych. Technologia ta nie została jednak zapomniana, a jej rozwój zagwarantowało zastosowanie w dziedzinie, w której nie miała sobie równych – w tzw. kopiach bezpieczeństwa dużych systemów komputerowych. Obecnie najbardziej popularny i wciąż rozwijany standard taśm magnetycznych to LTO[3] (ang. Linear Tape-Open). Konstrukcję kasety drugiej generacji, zawierającej nośnik, pokazano na fot. 1. Aktualnym standardem, ogłoszonym w ostatnim kwartale 2017 r., jest LTO-8. Urządzenia z nim zgodne umożliwiają zeskładowanie 12TB[4] danych na pojedynczym nośniku, którego żywotność szacuje się na 30 lat przy korzystnych warunkach przechowywania.

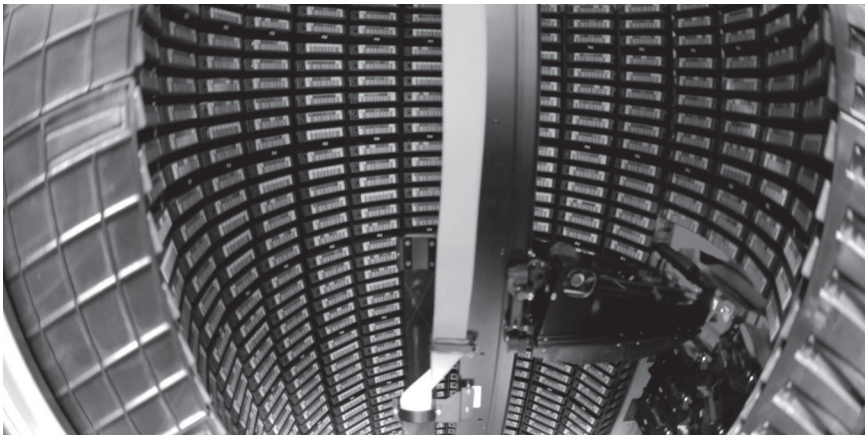
Fot. 1. Taśma magnetyczna w obudowie LTO drugiej generacji ze zdjętą pokrywą
Źródło: (*Austinmurphy*, 2008).

[3] Więcej informacji można znaleźć na stronie internetowej (*Ultrium*, 2018).
[4] Informacja zamieszczona na stronie internetowej producenta (*IBM*, 2018).

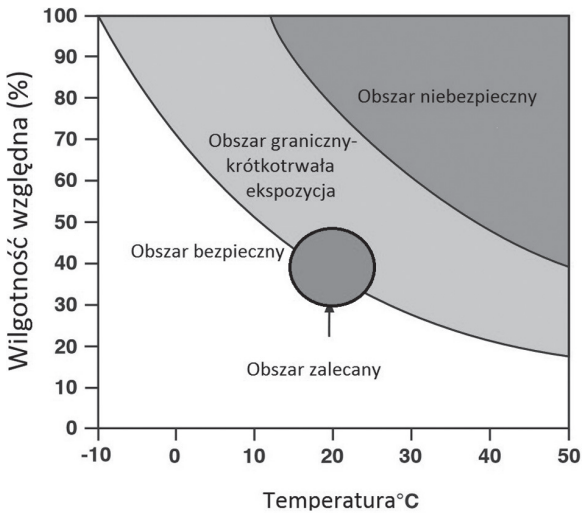
O szczególnej przydatności taśm magnetycznych w przedstawionych zastosowaniach decydują ich właściwości, wynikające przede wszystkim z prostoty konstrukcji:

– cena – to podstawowy czynnik mający wpływ na tak dużą popularność tego nośnika informacji; technologie potrzebne do budowy taśm magnetycznych są tanie, dobrze znane i sprawdzone; opracowywanie nowych technologii ukierunkowane jest głównie na zwiększanie gęstości zapisu, czyli pojemności informacyjnej pojedynczego urządzenia;
– odporność na awarie – prosta konstrukcja gwarantuje o wiele większą odporność na awarie mechaniczne niż np. w przypadku dysków twardych wyposażonych w delikatne elementy pozycjonowania głowicy;
– możliwość łączenia w biblioteki – w profesjonalnym użyciu stosuje się urządzenia bibliotek taśmowych pełniące funkcję automatycznie zarządzanych magazynów nośników taśmowych; biblioteki te przejmują odpowiedzialność za organizację danych archiwalnych, ich rozmieszczenie na poszczególnych nośnikach, kontrolę jakości zapisywanych informacji i szanse ich ponownego odczytu, duplikację danych w celu zapewnienia ich maksymalnej dostępności i wiele innych.

W rozbudowanych zastosowaniach biblioteki taśmowe mogą zajmować całe odpowiednio chronione pomieszczenia o kontrolowanym dostępie i klimatyzacji zapewniającej maksymalny czas życia urządzeń. Przykładowe zdjęcie takiego pomieszczenia przedstawia fot. 2, a orientacyjne charakterystyki temperatury i wilgotności powietrza dla maksymalizacji żywotności nośników na podstawie poradnika (Bogart, 1995) zobrazowano na rys. 1.



Fot. 2. Przykład rozbudowanej i całkowicie zautomatyzowanej biblioteki taśmowej Źródło: (*ChrisDag*, 2009).



Rys. 1. Orientacyjne charakterystyki temperatury i wilgotności powietrza dla taśm magnetycznych

Źródło: Opracowanie własne.

Więcej informacji na temat poszczególnych technologii składowania danych oraz wad, zalet i zagrożeń wynikających ze stosowania tych technologii w przystępny sposób podaje w swojej książce Nijaz Bajgoric (2008), natomiast modele kosztowe pozwalające optymalizować koszty budowanego archiwum wykorzystującego technologie taśmowe lub dyskowe przedstawiają w swym artykule Richard L. Moore, Jim D'Aoust, Robert H. McDonald i David Minor (2007).

Choć, jak wspomniano, pojemności pojedynczych urządzeń standardowych pamięci taśmowych osiągają wielkości rzędu kilkunastu TB przy jednoczesnym zachowaniu kompaktowych rozmiarów, to zapowiedzi najnowszych dokonań już teraz wskazują na osiągnięcie granicy 330TB[5] dla pojedynczego nośnika.

Wymieniając liczne zalety taśm magnetycznych, warto zauważyć, że niezawodność, prostota konstrukcji i wynikająca z niej niska cena okupione zostały długim czasem dostępu do archiwizowanych danych. Ponieważ konstrukcja urządzenia nie integruje w sobie głowicy czytającej oraz implikuje sekwencyjny dostęp do zapisanych danych, to ich odtworzenie wiąże się z odszukaniem właściwej taśmy, załadowaniem nią napędu odczytującego i oczekiwaniem, aż interesujący nas plik zostanie odnaleziony i skopiowany na dysk lokalny.

[5] Pisał o tym np. David Grossman na stronach portalu *Popular Mechanics* (2017).

Trzeba również zwrócić uwagę, że choć istnieją rozwiązania produktowe pozwalające na zastosowanie omawianych technologii w celach prywatnych czy biurowych, to uzasadnienie wyboru tej technologii do takiego przeznaczenia jest już mniej oczywiste. Pojedyncze urządzenia taśmowe są tanie, lecz sprzęt niezbędny do ich zapisu i odczytu będzie stanowił znaczny dodatkowy koszt, a brak możliwości odpowiedniego zabezpieczenia pomieszczeń, w których składuje się taśmy, i tak nie pozwoli na osiągnięcie maksymalnych parametrów bezpieczeństwa i trwałości składowanych danych. Nie oznacza to jednak zupełnego braku szans wykorzystania koncepcji funkcjonalnych leżących u podstaw tej technologii, gdyż najwięksi dostawcy usług chmurowych prześcigają się w opracowywaniu najkorzystniejszych ofert udostępniających długoterminowe archiwizowanie danych.

Usługi długoterminowego archiwizowania danych

Dostawcy usług przetwarzania danych w chmurze proponują całą gamę usług składowania danych różniących się od siebie funkcjonalnościami i rodzajami zastosowań. Są to przykładowo usługi składowania danych binarnych, danych w postaci relacyjnej czy w formie klucz – wartość. Specyficzną formę usług składowania danych stanowią usługi dedykowane archiwizacji długoterminowej, wymagające najczęściej wyprzedzającej subskrypcji na odczyt wcześniej zdeponowanych danych. Ów opóźniony dostęp, jako cecha najbardziej charakterystyczna dla tej grupy usług, stał się dla pracowników działów marketingu poszczególnych dostawców inspiracją do proponowania chwytliwych nazw własnych usług – nazwy te nawiązują często do czynnika chłodu jako zabezpieczającego materię przed rozkładem, ale i wymagającego czasu na jej „rozmrożenie" celem przywrócenia do stanu używalności.

Glacier (Lodowiec) to usługa składowania danych korporacji Amazon, oferowana w ramach grupy usług AWS (ang. Amazon Web Services) od roku 2012. Technologia zastosowana do jej implementacji owiana jest tajemnicą, a w zasobach sieci Internet znaleźć można wiele spekulacji sugerujących wykorzystanie bibliotek taśmowych, niskokosztowych wolnoobrotowych dysków twardych czy pamięci optycznych. Transfer danych do usługodawcy jest bezpłatny, a miesięczne opłaty za składowanie 1GB danych wynoszą zaledwie 0,004 dolara. Przygotowanie danych przez usługodawcę do natychmiastowego pobrania trwa (w ramach standardowego pakietu) od 3 do 5 godzin i w ramach miesięcznego limitu 5% całości wolumenu składowanych danych jest

bezpłatne. Warto jednak zauważyć, że warunki umowy rozkładają ten limit na interwały godzinne, co oznacza, iż bezpłatnie w ciągu jednej godziny można pobrać około 0,006944% całości danych, a przekroczenie tej wartości skutkować będzie naliczeniem dodatkowych opłat zgodnych z aktualnym cennikiem[6].

Cool Blob Storage (zimny skład wielkich danych binarnych) jest konkurencyjną do opisanej wcześniej usługą z grupy Azure, czyli usług chmurowych korporacji Microsoft. Udostępniona w roku 2016, charakteryzuje się o wiele krótszym czasem przygotowania danych do pobrania, co, niestety, wiąże się ze znacznie wyższymi opłatami taryfowymi. Dlatego w grudniu 2017 r. korporacja Microsoft, dywersyfikując swoją ofertę, udostępniła usługę o nazwie Archive Blob Storage, jeszcze bardziej ukierunkowaną na długoterminowe składowanie danych, i zaproponowała cenę rzędu 0,002 dolara za 1GB składowanych miesięcznie danych. Podobnie jak w przypadku Glacier, tak i w przypadku tej usługi dostępnych jest kilka planów taryfowych, których szczegółowy cennik znaleźć można na stronach WWW informujących o usługach Microsoft Azure[7].

Przedstawione usługi, choć niewątpliwie oferowane przez jednych z największych dostawców rozwiązań chmurowych, stanowią czubek góry lodowej usług w tym zakresie. Wymienić tu można jeszcze Coldline (firmy Google), Backblaze i wiele innych proponowanych przez dostawców o zasięgu lokalnym. Wybór konkretnej usługi powinna poprzedzać dogłębna analiza potrzeb dotyczących rozmiaru składowanych danych, czasu ich odtwarzania oraz bezpieczeństwa/dostępności powierzonych danych.

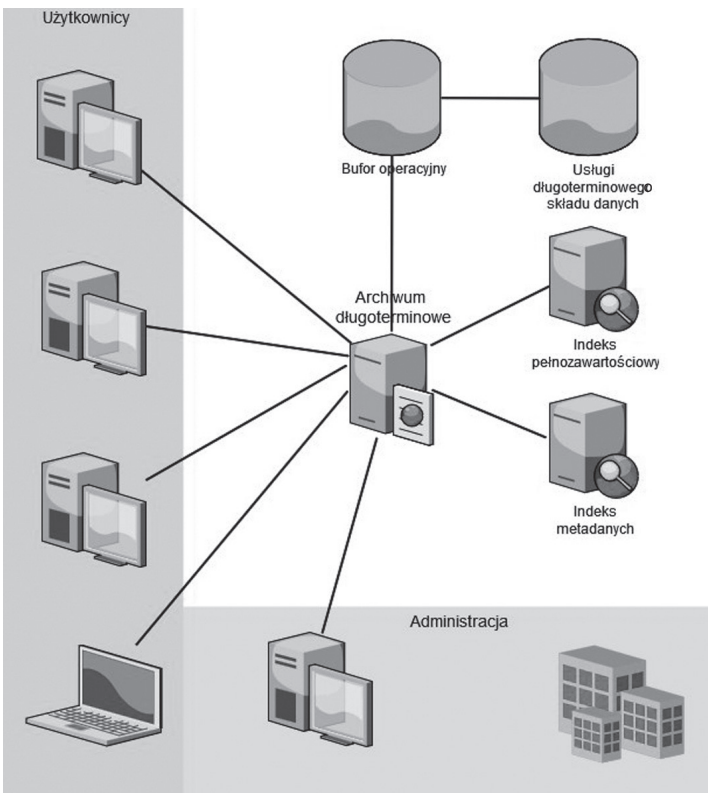
Podsumowanie

Wraz z ciągłym rozwojem technologii informatycznych możemy obserwować coraz bardziej zdecydowany kierunek migracji przetwarzania i składowania danych w stronę rozwiązań chmurowych. Dla większości osób oczywiste jest składowanie kopii bezpieczeństwa zawartości telefonu w bliżej nieznanych lokalizacjach nazywanych chmurą. W przypadku aplikacji biurowych, multimedialnych czy inżynierskich coraz częściej wykorzystuje się zasoby obliczeniowe nie lokalnego systemu

[6] Cennik usług dotyczących Glaciera można znaleźć na stronie Amazona (*Amazon*, 2018).

[7] Cennik usług Archive Blob Storage znajduje się na stronie Microsoft Azure (*Archive*, 2018).

komputerowego, ale udostępniane z poziomu chmur obliczeniowych. Brak wiedzy o fizycznej lokalizacji ośrodków udostępniających te zasoby nie powinien być dla nas niepokojący, tak samo jak nie jest niepokojący brak wiedzy o lokalizacji elektrowni dostarczającej energię do naszych mieszkań. Często w jednym i w drugim przypadku trudno wręcz jednoznacznie określić konkretną lokalizację, gdyż zasoby zazwyczaj są rozpraszane na wiele fizycznych ośrodków w celu ochrony danych przed awariami pojedynczych węzłów obliczeniowych czy kataklizmami przyrodniczymi. Równie często tajemnicą objęte są lokalizacje konkretnych ośrodków w celu ochrony przed potencjalnymi aktami terrorystycznymi. Wszystko to sprawia, że usługi dostarczane z poziomu chmur obliczeniowych realizowane są bezpieczniej, wydajniej i bardziej niezawodnie, a jednocześnie niewiele drożej lub wręcz taniej dzięki efektowi skali, niż usługi implementowane w tradycyjnych serwerowniach poszczególnych podmiotów.



Rys. 2. Logiczny model cyfrowego archiwum zasobów
Źródło: Opracowanie własne.

Spostrzeżenia te są szczególnie istotne w kontekście długotrwałej cyfrowej archiwizacji zasobów. Przykładowy logiczny model takiego systemu archiwizacji przedstawia rys. 2. Widzimy tutaj, że choć usługi długoterminowego przechowywania danych nie mogą stanowić jedynej składnicy systemu, który pełni przecież również funkcję systemu informacyjno-wyszukiwawczego, to przejmują one rolę składnicy podstawowej, zapewniając wysoką dostępność i bezpieczeństwo gromadzonych danych.

Literatura

Amazon Glacier Pricing (2018). Pobrane z: https://aws.amazon.com/glacier/pricing/ (21.01.2018).

Austinmurphy (2008). LTO-2 cartridge with the top shell removed, showing the internal components. W: Wikipedia. Pobrane z: https://en.wikipedia.org/wiki/Linear_Tape-Open#/media/File:LTO2-cart-wo-top-shell.jpg, CC BY-SA 3.0 (28.01.2018).

Archive Storage General Availability Pricing Microsoft Azure (2018). Pobrane z: https://azure.microsoft.com/en-us/pricing/details/storage/blobs/archive-ga/ (21.01.2018).

Bajgoric, N. (2008). *Continuous Computing Technologies for Enhancing Business Continuity*. Hershey PA: IGI Global.

Bhat, W.A. (2018). Long-term preservation of big data: prospects of current storage technologies in digital libraries. *Library Hi Tech*, *36*(3), 539–555. Pobrane z: https://doi.org/10.1108/LHT-06-2017-0117 (6.08.2018).

Bogart, J.W.C. van (1995). Magnetic Tape Storage and Handling : A Guide for Libraries and Archives. Pobrane z: https://clir.wordpress.clir.org/wp-content/uploads/sites/6/2017/02/pub54.pdf (14.11.2018).

ChrisDag (2009). Robotic tape library guts (fisheye). Pobrane z: https://www.flickr.com/photos/chrisdag/3202766292/, CC BY 2.0 (19.11.2018).

Eschenfelder, K.R., Shankar, K. (2016). Designing sustainable data archives: comparing sustainability frameworks. *iConference 2016 Proceedings*, 1–7. doi: 10.9776/16243.

Grossman, D. (2017). New 330 TB Magnetic Tape Is a Data Storage Monster. In: *Popular Mechanics*. Pobrane z: http://www.popularmechanics.com/technology/a27602/330-tb-magnetic-tape/ (21.01.2018).

IBM TS2280 Tape Drive. Pobrane z: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=TSD03243USEN& (21.01.2018).

ISO 14721:2012(en). (2012). Space data and information transfer systems. Open archival information systems (OAIS). Reference model. In: *Online Browsing Platform (OBP)*. Pobrane z: https://www.iso.org/obp/ui/#iso:std:iso:14721:ed-2:v1:en (6.08.2018).

Kamińska, A.M. (2017). Dobre praktyki publikowania danych badawczych. *Biuletyn EBIB*, nr 6 (176). Pobrane z: http://open.ebib.pl/ojs/index.php/ebib/article/view/569. (21.01.2018).

Milster, T.D. (2003). Optical data storage. In: C. Webb, J. Jones (Eds.), *Handbook of Laser Technology and Applications* (vol. 3, pp. 2391–2420). Bristol and Philadelphia: Institute of Physics Publishing.

Moore, R.L., D'Aoust, J., McDonald, R.H., Minor, D. (2007). Disk and Tape Storage Cost Models. In: *Archiving 2007*, pp. 29–32. Pobrane z: https://libraries.ucsd.edu/chronopolis/_files/publications/dt_cost.pdf (21.01.2018).

Plale, B., McDonald, R.H., Chandrasekar, K., Kouper, I., Konkiel, S., Hedstrom, M.L., Myers, J., Kumar, P. (2013). SEAD virtual archive: building a federation of institutional repositories for long-term data preservation in sustainability science. *The International Journal of Digital Curation*, *8*(2), 172–180. doi:10.2218/ijdc.v8i2.281.

Ultrium LTO. Pobrane z: https://www.lto.org (21.01.2018).

Tekst w wersji poprawionej wpłynął do redakcji 25 kwietnia 2018 r.

Anna Małgorzata Kamińska
Department of Library Studies
Institute of Library and Information Science
University of Silesia in Katowice
e-mail: anna.kaminska@us.edu.pl

The ice age of data archiving

Abstract: Being a part of the information society, we become consumers and producers of more and more information. A large part of it is crucial for the proper implementation of economic and administrative processes, as well as for the shaping of social, cultural or aesthetic attitudes, and therefore data protection becomes a matter of basic importance. In this article, the author explains why archiving resources in digital formats gives good opportunities to protect them in the long term. It defines the features that archives of long-term data preservation should have, and technologies that enable cost-effective building of these systems. Although these technologies are difficult to directly implement by small and medium-sized entities, the offer

directed by cloud service providers enables data storage in a safe and cost-effective manner. The article concludes with proposal of the archiving system architecture using third party services dedicated to long-term data storage.

Keywords: Amazon Glacier. Azure Archive Blob Storage. Cost optimization. Data archives. Data availability. Data security. Data storage technologies