STUDIA INFORMATICA 2009
Volume 30 Number 2B (84)

Krzysztof MARASEK
Polsko-Japońska Wyższa Szkoła Technik Komputerowych
Jerzy P. WALCZAK
ATM SA
Tomasz TRACZYK, Grzegorz PŁOSZAJSKI, Andrzej KAŹMIERSKI
Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych

KONCEPCJA ELEKTRONICZNEGO ARCHIWUM WIECZYSTEGO

Streszczenie. Opisano koncepcję elektronicznego archiwum wieczystego, przystosowanego do wiarygodnego przechowywania wielkich wolumenów informacji cyfrowej przez okres kilku pokoleń. Przedstawiono wymagania, które archiwum musi spełniać, i porównano je z problemami typowymi dla baz danych. Opisano koncepcję architektury archiwum opartego na pomyśle tzw. zasobników. Omówiono także zagadnienia przechowywania i udostępniania metadanych.

Słowa kluczowe: archiwa cyfrowe, VLDB, metadane, zasoby multimedialne

CONCEPT OF ELECTRONIC LONG-TERM ARCHIVE

Summary. The paper describes a concept of electronic long-term archive, designed for trustworthy storage of large volumes of digital information for a period of several generations. Requirements for the archive are shown and compared to problems typical for databases. A concept of archive architecture, based on special storage bin, is presented.

Keywords: digital long-term archives, VLDB, meta-data, multimedia resources

1. Wprowadzenie

Archiwizacja danych, a w szczególności ich długoterminowe przechowywanie, jest jednym z wyzwań stojących przed współczesną technologią informacyjną. Problematyka ta obejmuje stworzenie wiarygodnych mechanizmów i urządzeń samego przechowywania form cyfrowych, często o wielkiej objętości, oraz efektywnych i niezawodnych metod i narzędzi

przygotowania, opisywania i udostępniania informacji oraz zarządzania jej przechowywaniem i udostępnianiem. System archiwizacji powinien zapewniać długotrwałą ochronę i uprawnione użytkowanie cyfrowych kopii zasobów archiwalnych, zapewniając także efektywność ekonomiczną.

1.1. Cyfrowe zasoby archiwalne

Wiele współczesnych zasobów realizowanych jest już od początku w technologii cyfrowej, a nowe techniki digitalizacji pozwalają na coraz efektywniejsze i wierniejsze odwzorowywanie materialnych zasobów źródłowych.

Zasoby archiwalne, które winny być przechowywane w postaci cyfrowej, obejmują m.in.

- zasoby archiwalne przedsiębiorstw i organizacji, w tym także organów publicznych;
- zapisane w postaci cyfrowej dobra kultury (książki, gazety i czasopisma, obrazy, filmy, audycje radiowe i telewizyjne itd.), które oryginalnie istnieją lub istniały w postaci materialnej – niecyfrowej;
- wyniki bieżącej – cyfrowej od początku – produkcji wydawniczej, radiowej, telewizyjnej, filmowej itp.

W Polsce istnieje wiele instytucji, które są lub powinny być żywotnie zainteresowane archiwizacją zasobów postaci cyfrowej. Są to m. in. urzędy państwowe (np. ZUS), archiwa państwowe, muzea i biblioteki, Filmoteka Narodowa oraz firmy medialne (telewizje, radia, wydawcy itp.). Zasoby zgromadzone w postaci cyfrowej nie są jeszcze zbyt wielkie (z wyjątkiem telewizji, które już od dłuższego czasu produkują audycje w postaci cyfrowej), ale prace nad digitalizacją zgromadzonych zasobów trwają niemal we wszystkich potencjalnie zainteresowanych instytucjach. Wielkość archiwalnych zasobów cyfrowych, które powinny zostać zgromadzone w najbliższych latach, można zgrubnie oszacować na ok. 50 TB w przypadku Biblioteki Narodowej, 200 TB w przypadku Polskiego Radia i aż kilkadziesiąt petabajtów w przypadku Filmoteki Narodowej. Największe – w sensie objętości informacji cyfrowej – zasoby archiwalne mają jednak nadawcy telewizyjni: u jednego z czołowych nadawców prywatnych tylko nowe produkcje zajmują ok. 400 TB rocznie, a w telewizji państwowej mamy do czynienia ze znacznie większą produkcją bieżącą i wręcz gigantycznymi zasobami z lat minionych. Duża część tych zasobów stanowi dobra kultury narodowej, a w stosunku do wielu produkcji medialnych producenci mają prawny obowiązek długoterminowego przechowywania. W dodatku znaczna część tych zasobów, zwłaszcza nowszych, istnieje wyłącznie w postaci cyfrowej.

Konieczne jest zatem podjęcie wysiłków organizacyjnych i technicznych, które umożliwią wiarygodne długotrwałe przechowywanie cyfrowych zasobów archiwalnych. Niezbędne jest nie tylko odpowiednio trwałe fizyczne zapisywanie informacji, ale także jej odpowiednie

„opakowanie", umożliwiające odnalezienie i wykorzystanie owej informacji nawet w odległej przyszłości.

Archiwa cyfrowe mogą być tworzone dla różnych celów, szczególnie interesujące jednak wydają się archiwa obiektów audiowizualnych. Są one ważne dlatego, że są potencjalnie największe, ale także – a może przede wszystkim – dlatego, że znaczna część współczesnej twórczości audiowizualnej powstaje od razu w formie cyfrowej i nie ma żadnej innej trwałej reprezentacji, utrata zapisu cyfrowego oznacza zatem bezpowrotny przepadek takich dóbr.

Obiektem w archiwum cyfrowym jest zapis archiwalny wraz z przyporządkowanymi mu metadanymi. Obiekty takie można sklasyfikować na wiele sposobów według kryteriów, takich jak: sposób wytworzenia (cyfrowy, analogowy), cel przechowywania (informacyjny, do dalszej obróbki, kulturotwórczy itd.), wielkość obiektu, wartość obiektu z punktu widzenia właściciela archiwum czy społeczeństwa.

Już ze względu na sposób powstania mamy tu dwa rodzaje obiektów. Po pierwsze, są to obiekty, które zostały wytworzone od razu w technice cyfrowej. Obiektami tymi są przede wszystkim cyfrowe zapisy dźwiękowe, obrazy (fotografie) cyfrowe, rejestracje filmowe utworzone cyfrowo. Ich jakość nie może zostać podwyższona bez zmiany dzieła początkowego, a więc można je z punktu widzenia archiwalnego traktować jako oryginały, stąd też wynika specyficzny sposób ich traktowania. Każda kopia jest wiernym odtworzeniem oryginału i nie różni się od niego w żaden sposób, nie przechowuje się też zazwyczaj osobno oryginałów. Dla zachowania spójności należy jednak wskazać „kopie odniesienia" (*reference copies*), spełniające dla systemu rolę oryginałów. Obiekty te nie są zmieniane co do swojej treści; modyfikacji mogą ulegać jedynie ich metadane oraz miejsce przechowywania.

Kolejna grupa obiektów to te, które wytworzone zostały w sposób niecyfrowy, a następnie zostały poddane digitalizacji. Takie obiekty to cyfrowy zapis analogowo zarejestrowanego dźwięku, skany obrazów, fotografii, negatywów, skany analogowo zapisanych filmów, a także inne obiekty, jak skany oryginalnych wydawnictw, skany trójwymiarowe itp.

Jednym z możliwych podejść do digitalizacji jest próba zapewnienia „najlepszej technicznie możliwej" jakości. Innym jest zapewnianie jakości „odpowiedniej do celu digitalizacji". W tym pierwszym przypadku, wraz z rozwojem technologii kopiowania (skanowania), celowe może być ponowne wprowadzanie takiego obiektu do archiwum, w drugim zależy to od tego, czy jakość odpowiednia została osiągnięta, co w praktyce może oznaczać jakość niższą od „najlepszej technicznie możliwej" w danym czasie.

Wyższa jakość wyraża się przede wszystkim wyższą rozdzielczością obrazu i większą liczbą bitów reprezentujących poszczególne barwy składowe, a w przypadku dźwięku – większą częstotliwością próbkowania i liczbą bitów reprezentujących natężenie dźwięku. Do zapisania kopii o tak rozumianej wyższej jakości trzeba użyć większych plików cyfrowych, co może oznaczać większe koszty ich przechowywania w archiwum. Poza tym wyższa jakość

zależy od sprzętu (np. od jakości obiektywów, przetworników), co zazwyczaj wiąże się z jego ceną i pośrednio wpływa na koszty kopii cyfrowych. Stąd motywacja do zapewniania jakości odpowiedniej do celu, a nie zawsze najwyższej jakości możliwej.

Zwiększanie rozdzielczości ze wzrostem możliwości technologicznych może spotykać się z ograniczeniami wynikającymi z właściwości materiału analogowego. Taka granica w naturalny sposób jest wyznaczana np. przez osiąganie rozdzielczości skanowania dokładniejszej od rozmiarów ziarna na taśmie światłoczułej, a dla materiałów nagranych na płytach lub cylindrach Edisona (audio) – przez uzyskanie szczegółowego obrazu kształtu rowków. Inne granice dla zwiększania rozdzielczości bądź głębi bitowej mogą wynikać z właściwości oka i ucha ludzkiego. Zwiększanie rozdzielczości czy liczby bitów powyżej pewnych granic nie zapewni lepszej percepcji obrazu czy dźwięku. Jeszcze inne ograniczenia wynikają z niskiej jakości zapisu analogowego poddanego digitalizacji (np. na kasetach Compact). Co do celu digitalizacji, to inna rozdzielczość wystarczy do zapewnienia czytelnego obrazu tekstu książki, a inna (wyższa) może być potrzebna do prac naukowych, np. nad właściwościami papieru; inna do pokazania manuskryptu w sposób umożliwiający odczytanie jego treści czy zobaczenie podpisu znanej osobistości, a inna do badań grafologicznych. Na stosowaną rozdzielczość mogą mieć wpływ względy ekonomiczne, i to nie tylko w kierunku jej ograniczania. Na przykład, mogą one skłaniać muzealników do stosowania możliwie wysokiej technicznej jakości skanowania czy fotografowania obiektu, gdy koszt przygotowania go do digitalizacji jest znaczny (np. koszt demontażu średniowiecznego ołtarza).

Jeżeli jakość nie jest dostateczna albo gdy stosuje się do pewnych obiektów politykę zapewniania najwyższej możliwej jakości, trzeba liczyć się z koniecznością (celowością) powtarzania skanowania po pewnym czasie. Przy powtarzaniu skanowania napotykamy jednak na barierę: z upływem czasu wszystkie obiekty fizyczne w mniejszym lub większym stopniu zmieniają się – degradują technicznie. Dotyczy to np. taśmy światłoczułej. Choć stale rośnie jakość skanerów rejestrujących zapis światłoczuły w postaci cyfrowej, jednocześnie degradują się kolory zapisane na taśmie, powstają rysy i pęknięcia przypominające krakelurę tworzącą się na obrazach olejnych, a z czasem całe podłoże ulega trwałym zmianom chemicznym, tworząc plamy i czyniąc odczyt niemożliwy. Dla niektórych obiektów istnieją możliwości przedłużania życia w sposób analogowy (np. wykonanie kopii światłoczułej z oryginału), jednak i one wiążą się z utratą jakości, ponieważ kopia zawsze będzie zawierać błędy procesu analogowego kopiowania. Proces wprowadzania do systemu tego samego niecyfrowego obiektu należy uznać więc za ograniczony w czasie. Tempo degradacji niecyfrowego oryginału może jednak być stosunkowo nieduże w porównaniu z tempem doskonalenia metod digitalizacji. W przypadkach szybkiej degradacji analogowego oryginału kwestia jakości może być drugorzędna wobec celu, jakim jest w ogóle zachowanie kopii dzieła.

Z drugiej strony fakt, że poza systemem, będącym archiwum cyfrowym, występuje niecyfrowy oryginał, ma również znaczące implikacje – mianowicie stanowi on dodatkową kopię bezpieczeństwa, w dodatku wypełniającą postulat technologicznej odrębności przechowywania (patrz 2.3.1).

1.2. Cele archiwizacji

Planując archiwum cyfrowe, należy dokonać analizy celów archiwizacji. Konieczne jest określenie hierarchii tych celów, np: zachowanie dziedzictwa narodowego, łatwość dostępu, łatwość przetwarzania obiektów, możliwość wykorzystania do nowej produkcji itp. W praktyce, dla wielu archiwów cyfrowych czy też dużych projektów digitalizacji archiwów przeprowadza się rozbudowaną analizę ekonomiczną lub, częściej, analizę metodą BSC (*Balanced Scorecard*, czyli strategiczna karta wyników według Kaplan & Norton). W każdym przypadku projektowania archiwum cyfrowego wskazane jest przeprowadzenie takiej analizy przed rozpoczęciem planowania technologii, gdyż system priorytetów ma znaczący wpływ na rekomendowane rozwiązania techniczne.

Dla większości archiwów cyfrowych można rozdzielić (także technicznie) funkcje zachowania wieczystego i funkcje udostępniania/przetwarzania. Jest to odbiciem klasycznego archiwum, w którym przechowuje się oryginały, a do udostępniania stosuje się w zasadzie wyłącznie ich kopie. Wówczas mamy do czynienia z hierarchią archiwów: archiwum wieczyste u podstaw systemu, a archiwa pomocnicze (dystrybucyjne) w środkowej jego części. Szczyt tej „piramidy archiwów" stanowią archiwa podręczne użytkowników końcowych. Dla większości archiwów cyfrowych dodatkowe ułatwienie stanowi fakt, że wiele celów związanych z dystrybucją treści (kulturotwórczy, inspirujący, dużą część zastosowań emisyjnych, część zastosowań związanych z ponownym użyciem obiektów do celów produkcji) realizuje się na podstawieo zapisów o niższej rozdzielczości, a więc zajmując mniej miejsca w pamięci.

Zastosowanie takiej hierarchii archiwów jest o tyle użyteczne, że umożliwia jednoznaczne rozdzielenie priorytetów poszczególnych poziomów hierarchii. Dla archiwum wieczystego priorytetem jest niezawodność kosztem szybkości odczytu, a dla archiwum dystrybucyjnego nie ma konieczności podnoszenia poziomu niezawodności i można skoncentrować nakłady na osiąganiu szybkości odczytu i wyszukiwania. W razie utraty kopii obiektu istnieje bowiem możliwość odwołania się do jego oryginału przechowywanego w pamięci „głębokiej".

Wielkość obiektu archiwalnego (w zapisie cyfrowym) wpływa znacząco na wybór technologii przechowywania i koszty budowy archiwum, nie ma jednak wielkiego wpływu na zarządzanie systemem.

Bardzo istotna, choć niezmiernie trudna do oceny, jest wartość archiwalna obiektu. Tu wskazówką jest np. obowiązujące prawo, ustalające, jakie obiekty podlegają jakim obowiąz-

kowym działaniom archiwizacyjnym. Jednak przepisy prawa nie określają w pełni sposobów postępowania z kopiami (w przypadku archiwum cyfrowego wytworzonego z oryginałów analogowych), a także nie odpowiadają obecnemu stanowi wiedzy informatycznej. Dla obiektu powstałego w sposób cyfrowy, ale zapisanego w postaci fizycznego obiektu (jak np. zapis na taśmie cyfrowej *Digital BetaCam*), przewiduje się odrębne traktowanie oryginału od traktowania kopii, nawet gdy jest ona binarnie identyczna z oryginałem! Z punktu widzenia praktyki, większa część archiwów audiowizualnych nie jest w ogóle używana (StorageIO Group podaje, że w USA 90% treści zawartej w archiwach cyfrowych nigdy się nie odczytuje). Dotyczy to zresztą większości archiwów cyfrowych. Jednakże, m. in. z przyczyn prawnych, obiekty te muszą być przechowywane.

1.3. Czym jest elektroniczne archiwum wieczyste

Problemy, które trzeba rozwiązać archiwizując informację cyfrową, zależą od przewidywanej długotrwałości przechowywania tej informacji. Przechowanie tymczasowe (np. na czas przetwarzania danych) i krótkoterminowe (poniżej trzech lat) nie sprawia obecnie trudności. Trwałość informacji przechowywanej na dyskach twardych komputerów jest tu wystarczająca. Przechowywanie średnioterminowe (od 3 do 10 lat) także nie sprawia większych problemów technicznych; do tych celów wystarcza na ogół trwałość powszechnie używanych nośników wymiennych. Przechowywanie długoterminowe (powyżej 10 lat, ale z określonym końcowym terminem ważności) oraz bezterminowe (powyżej 10 lat i bez wyznaczonego terminu ważności) okazuje się być zagadnieniem złożonym i dotąd nieznajdującym kompleksowego rozwiązania.

Nie istnieją sprawdzone i satysfakcjonujące strategie długotrwałego przechowywania danych, nie opracowano także nośników w sposób pewny zapisujących dane na dłuższe okresy. Jedynym znanym rozwiązaniem jest dynamiczne przechowywanie danych, tj. okresowe przekopiowywanie ich, co niekiedy wymaga ingerencji w ich strukturę, np. dostosowania do nowych urządzeń [3], oprogramowania czy formatów. Taki sposób przechowywania danych wiąże się oczywiście z pewnym ryzykiem. Mamy tu więc do czynienia z paradoksem: dane cyfrowe dają się stosunkowo łatwo i bezstratnie kopiować, a jednak zaskakująco często dochodzi do ich utraty ([3] wymienia np. utratę danych z misji kosmicznych NASA).

Opracowanie niniejsze dotyczy archiwum wieczystego. Przez „wieczystość" rozumiemy tu przechowywanie istotnie długoterminowe (np. przez okres kilku pokoleń) lub bezterminowe. Na takim horyzoncie nie można zapewnić trwałości żadnego nośnika elektronicznego ani zagwarantować dostępności jakiejkolwiek obecnie stosowanej technologii (sprzętu, oprogramowania, metod zapisu, formatów danych itd.). „Archiwum" z kolei oznacza taki sposób przechowywania zasobów, który nie ma służyć do ich bieżącego wykorzystywania czy udo-

stępniania, lecz przede wszystkim zabezpieczać je przed wszelkimi zagrożeniami, takimi jak utrata lub zniekształcenie informacji, niepowołane użycie zasobu, niemożność wyszukania, odczytu czy interpretacji informacji itp.

W naszym przypadku, gdy rzecz dotyczy wieczystego przechowywania wielkich zasobów informacji, rozsądnie jest rozważać wykorzystanie tzw. archiwum głębokiego, tzn. takiego, które tworzone jest przy założeniu, że do składowanych w nim zasobów, a przynajmniej do ich znaczącej większości, dostęp jest bardzo rzadki, a w wielu przypadkach po zapisaniu nie nastąpi już nigdy. Interaktywność takiego archiwum jest zatem niska: można założyć, że w przeważającej części zleceń czas ich obsługi może być większy od tzw. latencji – nieusuwalnego opóźnienia wynikającego z powodów czysto technicznych (por. [10]). Konsekwencją tego założenia jest specyficzny sposób dostępu do zasobów, a mianowicie dostęp „na zamówienie", z czasem dostawy, który może być stosunkowo długi i nie musi być z góry gwarantowany (powinien co najwyżej być możliwy do oszacowania podczas przyjmowania zamówienia). Taki sposób dostępu znacząco różni się od powszechnego w systemach informatycznych (np. internetowych) modelu dostępu „na żądanie", gdzie oczekuje się możliwie wysokiej interaktywności systemu. Założenie dostępu „na zamówienie" jest ważne, ponieważ pozwala wypracować rozwiązania techniczne i organizacyjne znacząco obniżające koszty eksploatacji archiwum.

Oczywiście z istoty archiwum głębokiego wynika, że nie może ono pełnić funkcji podręcznego składu informacji, np. w celu jej przetwarzania czy udostępniania *on-line*. Cały system składowania i udostępniania informacji musi zatem – jak opisano wyżej – zawierać, oprócz archiwum wieczystego, także odpowiednie podsystemy buforujące, o wysokiej interaktywności, ale nieporównanie mniejszej pojemności i znacznie niższych wymaganiach co do trwałości i bezpieczeństwa zasobów.

1.4. Aktualne rozwiązania

Problem długotrwałego przechowywania informacji cyfrowej zaczyna być dostrzegany i uznawany za ważny, choć nie jest do końca jasne, co rozumie się przez przechowywanie wieczyste, nie ma nawet zgodności co do tego, czym są archiwa. Na świecie powstało jednak kilka organizacji, których celem jest uświadomienie tego problemu decydentom oraz znalezienie odpowiednich rozwiązań.

Jedną z takich organizacji jest SNIA (*Storage Networking Industry Association*) – zrzeszenie producentów i użytkowników technologii informatycznych, a w szczególności producentów pamięci masowych. Organizacja ta powołała grupę roboczą *100 Year Archive Task Force*, zajmującą się wieczystym przechowywaniem danych. Wnioski opublikowane w raporcie przygotowanym przez tę grupę są dość pesymistyczne [6]. Zidentyfikowane przez

SNIA zagrożenia procesu długotrwałego przechowywania danych wynikają z dwóch podstawowych przyczyn:

- długotrwałe przechowywanie danych wymaga ich regularnych migracji ze względu na starzenie się nośników i ze względu na starzenie się sprzętu i oprogramowania potrzebnego do odczytu informacji zapisanych na nośnikach;
- szybkość zmian w przemyśle informatycznym stoi w sprzeczności z podstawowym celem wieczystego przechowywania danych.

Aktualnie stosowane metody zapewnienia wieczystości zapisu polegają głównie na migracji danych na nowe nośniki (np. taśmy LTO[1] – obecnie najtańszy energetycznie sposób długotrwałego składowania danych cyfrowych) oraz na stosowaniu bieżących metod korekcji błędów (np. w systemach RAID). Metody te jednak zawodzą w przypadku bardzo dużych repozytoriów. W przypadku archiwów niewielkich okresowa migracja danych jest dość łatwa w realizacji, jednakże gdy archiwa zawierają znaczne zasoby, liczące nawet miliony odrębnych nośników, wówczas okresowa migracja nie wydaje się rozwiązaniem praktycznie realizowalnym, gdyż wymagałaby ogromnej liczby urządzeń kopiujących, dostosowanych zapewne do wielu różnych nośników i formatów, i byłaby niezwykle trudna logistycznie.

Zauważono, że do osiągnięcia celów wieczystego przechowywania informacji nie wystarczy zapewnienie trwałego zapisu danych, lecz potrzebne są liczne dodatkowe działania.

- Działania technologiczne: pozyskiwanie źródeł danych, zapewnienie odczytu fizycznego, zapewnienie odczytu logicznego, migracje wielkich repozytoriów danych, emulacje formatów, zapewnienie działania historycznych aplikacji i czytników, ochrona przed zmianami oraz przed utratą lub zniszczeniem, zapewnienie bezpieczeństwa fizycznego i logicznego danych, automatyzacja dostępu, wyszukiwanie i udostępnianie, testowanie oraz audyt.
- Działania organizacyjne: wspólne ustalenie wymagań, ustalenie ram prawnych, klasyfikacja danych, zapewnienie odpowiednich metadanych, standaryzacja.

W celu opisu przepływu i transformacji informacji w archiwum cyfrowym w wielu pracach przyjmuje się warstwowy model systemu przechowywania. Na przykład, w [1] założono, że do właściwego odczytu zasobów archiwalnych potrzebne są trzy warstwy: *Bit Layer* – warstwa zapewniająca odczyt danych binarnych z nośników, *Data Layer* – warstwa formująca podstawowe jednostki informacji (np. rekordy, kontenery czy obiekty) oraz *Application Layer* – warstwa interpretująca dane. Format zwracany przez *Data Layer* powinien być samoopisujący, niezależny od nośnika, dostawcy i platformy, móc zawierać wewnętrzne łączniki i odwołania, zapewniać wysoką szybkość przepływu danych także dla dużych zbiorów, umożliwiać równoległy zapis i odczyt, umożliwiać migrację danych pomiędzy różnymi sys-

temami bez ich utraty i być interpretowalny w przyszłości, być rozszerzalny, tani, zrozumiały przez ludzi i maszyny oraz wspierać dodatkowe funkcje (kompresję, kodowanie, kryptografię itp.).

W dodatku do rekomendacji OAIS [2] podano nieco inny warstwowy model informacyjny archiwum wieczystego, składający się z pięciu warstw.

- *Media Layer* to warstwa nośnika (dyski, taśmy itp.) wraz ze sprzętem potrzebnym do jego odczytania.
- *Stream Layer* to warstwa, przekazująca dane odczytane z nośnika w postaci strumieni bajtów, które warstwy wyższe mogą pozyskiwać w sposób niezależny od natury nośnika, posługując się pewną nazwą. Warstwa ta odpowiada np. systemowi plików.
- *Structure Layer* to warstwa, w której następuje konwersja strumieni bajtów na adresowalne prymitywne struktury, złożone z prostych typów danych zorganizowanych w proste formy, takie jak rekordy czy tablice. Ta warstwa odpowiada strukturom danych wykorzystywanym w językach programowania.
- *Object Layer* jest warstwą, która łączy prymitywne struktury w obiekty właściwe dla dziedziny zastosowań, odpowiadające np. obiektom multimedialnym. W tej warstwie pojawiają się kontenery gromadzące powiązaną informację oraz metadane opisujące obiekty. Rolę tę w systemach informacyjnych pełnią odpowiednie biblioteki lub repozytoria.
- *Application Layer* obejmuje programy, które służą do analizowania i prezentacji obiektów.

W celu uzyskania cech, oczekiwanych od wyższych warstw tych modeli, postuluje się opakowywanie archiwizowanych treści cyfrowych w kontener zawierający także metadane, w formie możliwie łatwej do interpretacji i samoopisującej. Użyć można tu [6] struktur i metod zaproponowanych przez *Consultative Commitee for Space Data Systems* w rekomendacji *Reference Model for Open Archival Information System (OAIS)* [2] i przyjętych przez ISO w normie 14721. OAIS definiuje podstawowe koncepcje i model referencyjny archiwów długoterminowych. W modelu tym archiwizowany obiekt (zwany *information package*) składa się z:

- *Content Information* – właściwego zasobu złożonego z:
  - *content data object* – właściwych treści podlegających przechowywaniu,
  - *representation information* – informacji niezbędnych do prawidłowej interpretacji danych;
- *Preservation Description Information* (PDI) – informacji o przechowywaniu zasobu, składającej się z:

[1] LTO (*Linear Tape-Open*) – otwarty standard formatów magnetycznych taśm do przechowywania danych cyfrowych.

- *provenance* – opisu historii i pochodzenia danych oraz historii operacji dokonywanych na danych,
- *reference* – unikalnego globalnie i trwałego identyfikatora zawartości pakietu,
- *context* – dokumentacji celu stworzenia pakietu i jego odniesienia do otoczenia,
- *fixity* – informacji wykazujących, że treść danych nie została zmieniona w sposób nieudokumentowany.

OAIS proponuje rozróżnienie różnych typów obiektów-pakietów, w zależności od ich przeznaczenia w systemie archiwum:

- *Submission Information Package* (SIP) – pakiet informacji przesyłany do archiwum przez producenta-dostarczyciela zasobu cyfrowego;
- *Archival Information Package* (AIP) – pakiet informacji przeznaczony do długotrwałego przechowywania w archiwum;
- *Dissemination Information Package* (DIP) – pakiet przesyłany z archiwum do odbiorcy informacji.

Pakiety te powinny zapewne różnić się zawartością i szczegółami budowy, w szczególności przewiduje się, iż jeden AIP może powstawać z wielu SIP, a DIP może zawierać treść wielu AIP. System archiwum powinien mieć zdolność do transformacji SIP na AIP, długotrwałego przechowywania AIP oraz transformacji AIP na DIP.

Alternatywną propozycję opakowywania danych stanowi XSet [1], wykorzystujący format XOP (*XML-binary Optimized Packaging*), będący standardem W3C [9]. W tym wypadku dane przechowywane są w paczkach złożonych z dokumentu XML, opisującego atrybuty i cechy paczki danych, strumieni danych binarnych (XStreams), przechowywanych jako załączniki MIME, oraz spisu treści dla każdego z XStreams. Przykład takiego formatu zamieszczono w pracy [1]. Zaproponowano tam także połączenie enkapsulacji OAIS AIP z XOP. Celem jest stworzenie *Self-Describing Self-Contained Data Format* (SD-SCDF), czyli formatu przechowywania informacji możliwie niezależnego od zewnętrznych opisów i definicji.

W studium [5] przygotowanym przez grupę roboczą Nestor (*Network of Expertise in Long-Term Storage*) poddano analizie kryteria niezbędne do nadania cyfrowym archiwom cechy wiarygodności (patrz też [4]). Kryteria oceny repozytorium elektronicznego sformułowano w zgodzie z modelem OAIS.

Podkreślić należy, że – pomimo rosnącej świadomości zagrożeń – na podstawie przeglądu literatury można stwierdzić, iż problem długotrwałego przechowywania danych w postaci cyfrowej nie doczekał się jeszcze pełnego rozwiązania. Inicjowane są jednak duże projekty w tej dziedzinie; przykładem może być kontrakt z 2005, na podstawie którego konsorcjum firm zebranych wokół Lockheed Martin ma do roku 2011 za sumę 308 milionów USD przygotować system archiwizacji wieczystej na zlecenie amerykańskiej agencji NARA (*National*

Archives and Records Administration). Należy mieć nadzieję, że podobne poważne projekty będą wkrótce zainicjowane i w naszym kraju.

2. Wymagania dla elektronicznego archiwum wieczystego

By możliwe było opracowanie koncepcji archiwum wieczystego, trzeba sformułować wymagania, jakie taki system ma spełniać. Jak się okazuje, wymagania te są dość złożone, a niektóre z nich są bardzo trudne do spełnienia.

Do precyzyjnego opisania i przeanalizowania niektórych wymagań potrzebny jest dość dokładny model warstw systemu przechowywania informacji. Użyjemy tu modelu warstwowego OAIS (patrz wyżej), rozwijając go jednak przez wprowadzenie pewnych warstw cząstkowych oraz dodatkowego poziomu:

- *Media Layer* podzielimy na dwie warstwy:
  - nośnika fizycznego,
  - sygnału (analogowego lub ciągu bitów);
- *Stream Layer* i *Structure Layer* pozostają bez zmian;
- *Object Layer* podzielimy na warstwy:
  - kontenera (pakietu), mogącego zawierać wiele struktur danych i metadanych,
  - modelu logicznego;
- *Application Layer* podzielimy na warstwy:
  - semantyczną,
  - aplikacji,
  - funkcjonalną;
- dodamy dodatkowy poziom – nazwijmy go umownie *Institutional Layer* – podzielony na warstwy:
  - organizacyjno-instytucjonalną,
  - prawną.

2.1. Wiarygodność archiwizacji

Podstawowym postulatem, jaki stawiamy przed archiwami, w tym archiwami cyfrowymi, jest wiarygodność (ang. *trustworthiness* [4]). Dla archiwum cyfrowego oznacza ona autentyczność, integralność, poufność i dostępność informacji [5].

Dla zapewnienia autentyczności, integralności i dostępności informacji niezbędne jest oczywiście jej przechowywanie w sposób pewny, tj. trwały i dający się zweryfikować.

2.1.1. Trwałość informacji

Zagadnienie trwałości informacji rozważać należy analizując poszczególne warstwy opisane wyżej, gdyż samo pojęcie trwałości ma różne interpretacje na różnych warstwach, a i możliwości osiągnięcia tej trwałości także istotnie się różnią.

Na poziomie nośnika fizycznego trwałość należy interpretować jako pewność istnienia dokładnie tego samego stanu zapisu po zadanym czasie. Osiągnięcie długoterminowej trwałości nośnika jest obecnie niemożliwe, a trwałość bezterminowa jest w ogóle nieosiągalna ze względu na powszechnie obowiązujące prawa fizyki (w szczególności II zasadę termodynamiki). Na szczęście trwałość na tym poziomie nie jest niezbędna do budowy archiwów wieczystych.

Na poziomie sygnału trwałość interpretować można jako powtarzalność odczytu, tj. pewność uzyskania dokładnie tego samego sygnału po zadanym czasie. W przypadku sygnału analogowego taka trwałość jest w ogóle nieosiągalna (można osiągnąć tylko przybliżoną powtarzalność). Wielką przewagą sygnału cyfrowego jest to, że może on być powtarzany dokładnie. Ta właściwość pozwala oddzielić informację od nośnika jako fizycznego obiektu, na którym jest ona rozpinana, umożliwia także bezstratne kopiowanie informacji, co – jak zobaczymy później – umożliwia konstrukcję cyfrowych archiwów długoterminowych. Choć sygnał cyfrowy sam w sobie jest powtarzalny, to – ze względu na niemożność dostatecznie trwałego zapisu danych – na tym poziomie nie jesteśmy w stanie zapewnić trwałości długoterminowej. Na szczęście, jak się okazuje, długoterminowa trwałość na tym poziomie także nie jest niezbędna.

Na poziomie strumienia danych trwałość winna być interpretowana jako pewność uzyskania po zadanym czasie „spod" tej samej nazwy dokładnie tego samego strumienia bajtów, który został tam zapisany. Trwałość da się uzyskać zarówno długoterminowo, jak i bezterminowo. W archiwach wieczystych będziemy postulowali możliwość uzyskania właśnie takiej trwałości dla właściwych danych archiwalnych. Choć cecha ta nie jest bezwzględnie konieczna dla wieczystego funkcjonowania archiwum, znacznie ułatwia jego konstrukcję i użytkowanie. W przypadku metadanych nie musimy gwarantować całkowitej trwałości na tym poziomie. Dla sformalizowanych metadanych bowiem treść odczytywana nie musi być co do bajtu identyczna z zapisaną, a jedynie musi być zgodna na poziomie struktury lub nawet modelu logicznego. Dlatego nie musimy postulować trwałości metadanych w warstwie strumienia danych. Dobrym przykładem jest zapis metadanych w języku XML: do poprawnej interpretacji informacji całkowicie wystarcza zgodność na poziomie tzw. *DOM fidelity*, czyli zgodności drzew DOM danych zapisanych i odczytanych. Strumień informacji odtworzonej nie musi być identyczny z zapisaną, np. może się różnić stroną kodową czy wcięciami.

Na poziomie struktury i kontenera trwałość rozumiemy jako pewność odtworzenia dokładnie tej samej informacji, która została zapisana, i w tym samym formacie (choć nieko-

niecznie ze zgodnością co do bajtu). Trwałość taką najłatwiej jest zapewnić gwarantując ją poziom niżej, tj. na poziomie strumienia danych, dlatego tak właśnie postulowano. Trwałość na poziomie struktury nie jest bezwzględnie uwarunkowana trwałością strumienia danych, nawet jeśli niezbędna jest wierność co do bajtu, jeśli bowiem strumień danych zmienił się w ograniczonym zakresie, dokładne odtworzenie informacji może być możliwe dzięki odpowiedniej rekonstrukcji. W archiwach wieczystych postulujemy – choć znów nie jest to bezwzględnie konieczne, lecz stanowi pewien kompromis – trwałość bezterminową na poziomie struktury i kontenera.

Na poziomie modelu logicznego trwałość rozumiemy jako pewność odtworzenia dokładnie tej samej informacji, która została zapisana, ale niekoniecznie w tym samym formacie. Trwałość taką znowu najłatwiej jest oczywiście zapewnić gwarantując ją niżej, tj. na poziomie struktury danych i kontenera. Należy jednak zauważyć, że trwałość modelu logicznego także nie jest bezwzględnie uwarunkowana trwałością na poziomie struktury czy kontenera, jeśli bowiem struktura danych zmieniła się w sposób kontrolowany, dokładne odtworzenie informacji może być możliwe, np. dzięki odpowiedniej konwersji. Na przykład dla danych tekstowych możemy dopuścić zmianę strony kodowej (byle nie na uboższą), gdyż nie ogranicza to możliwości odtworzenia informacji. Dla archiwów wieczystych postulujemy trwałość na poziomie modelu logicznego.

Zasadniczym wymaganiem, które stanowi istotę archiwizacji informacji, zatem musi być celem każdego archiwum, jest trwałość na poziomie semantycznym. Trwałość ta oznacza pewność, że po zadanym czasie informacja zostanie tak samo zinterpretowana, jak była interpretowana w chwili zapisu. Niestety, wymaganie to w miarę łatwo można spełnić jedynie w przypadku danych tekstowych, znaczenie tekstu – o ile jest on kompletny – daje się bowiem zwykle odtworzyć, jeśli tylko jest on zapisany w przynajmniej częściowo znanym języku. Wynika to z faktu, że informacja tekstowa jest na ogół w mniejszym lub większym stopniu samoobjaśniająca. W przypadku danych multimedialnych to wymaganie jest znacznie trudniejsze do spełnienia. Postulować będziemy jednak na pewno trwałość bezstratną, tj. taką, by pewne było, że z danych odczytanych da się wyinterpretować całą informację, która dawała się wyinterpretować z danych zapisywanych. Zauważyć znowu należy, że trwałość na poziomie semantycznym w zasadzie nie wymaga trwałości formatu ani nawet modelu logicznego danych, ale postulowanie trwałości na owych niższych poziomach znacznie upraszcza problem.

Dla archiwów wieczystych oczywiście postulujemy bezterminową trwałość w warstwie semantycznej. Wymaganie to, choć ma zasadnicze znaczenie, nie jest niestety łatwe do spełnienia; jak się nawet wydaje, nie jesteśmy w stanie zagwarantować jego spełnienia w stu procentach. Nie możemy bowiem być całkiem pewni, że po upływie kilku pokoleń jakkolwiek zapisana informacja będzie poprawnie interpretowana przez istniejące wówczas oprogramo-

wanie (ba, nie możemy nawet być pewni, czy będzie istniało jakieś „oprogramowanie" w dzisiejszym sensie). To, co zrobić można, to zapis informacji zgodny z powszechnie obecnie używanymi standardami. Taki zapis powiększa prawdopodobieństwo poprawnej interpretacji danych w przyszłości, gdyż im popularniejszy jest standard, tym większa pewność, że długo będą istnieć środki do jego odczytywania lub możliwa będzie konwersja na nowsze formy zapisu. Dlatego dla właściwych danych archiwalnych uzyskanie trwałości w warstwie semantycznej rozumiane być może jako zapewnienie trwałości w warstwie modelu logicznego połączone z użyciem formatu danych zgodnego z jednym z powszechnie uznanych standardów. Prawdopodobieństwo poprawnej interpretacji znacznie powiększa także redundantny zapis tej samej informacji w więcej niż jednym formacie.

Przez trwałość w warstwie aplikacji rozumieć można, że po zadanym czasie będziemy dysponowali aplikacjami, które umożliwią takie same działania na odczytanej informacji, jakie mogliśmy wykonywać w chwili jej zapisu, i z takimi samymi rezultatami. Oczywiście, określenie „takie same działania" nie jest precyzyjne, ale trudno tu o ogólną definicję. Lepiej można zdefiniować trwałość na poziomie funkcjonalnym: po zadanym czasie powinna istnieć taka sama funkcjonalność, jaka była dostępna w chwili zapisu informacji, a poszczególne funkcje aplikowane do odczytanej informacji powinny dawać wyniki zgodne z uzyskiwanymi w chwili zapisu.

Z punktu widzenia archiwizacji wieczystej trwałość na poziomie aplikacyjnym i funkcjonalnym wydaje się jednak być postulatem dość dyskusyjnym. Zagwarantowanie trwałości na poziomie aplikacji jest długoterminowo bardzo trudne, a bezterminowo wydaje się niemożliwe. Trwałość na poziomie funkcjonalnym jest zapewne osiągalna, ale być może nie jest warta wysiłku. Trudno bowiem przewidzieć, czy wszystkie obecnie dostępne funkcjonalności będą potrzebne w przyszłości, nie sposób też zgadnąć, czy niezbędne nie staną się jakieś nowe funkcje, obecnie nieznane. Dlatego, zamiast postulowania trwałości w sensie opisanym wyżej, lepsze wydaje się postulowanie trwałości w sensie szerszym: rozumianej jako neutralność aplikacyjna informacji zapisanej w archiwum. Taka neutralność – podobna do postulowanej w przypadku struktur baz danych – oznaczałaby, że informacja jest przechowywana w archiwum w sposób możliwie niezależny od jej zastosowań i w formie takiej, by możliwe było wykonywanie na niej jak najszerszego wachlarza działań zarówno znanych w chwili obecnej, jak i mogących się pojawić w przyszłości. Praktycznie wymaganie to sprowadzać się może do stosowania powszechnie uznanych otwartych standardów zapisu i opakowywania informacji.

Wprowadzenie dodatkowej warstwy „instytucjonalnej" wiąże się z doświadczeniem uzyskanym w tradycyjnych archiwach (niecyfrowych), które uczy, że samo istnienie niezniszczonego zasobu archiwalnego nie gwarantuje jeszcze możliwości jego pełnego wykorzysta-

nia. Pojawić się bowiem mogą problemy wynikające z braku odpowiedniej instytucji lub jej złego funkcjonowania oraz z przeszkód prawnych.

W przypadku tradycyjnych archiwów przynajmniej trwałość zasobów nie musi zależeć od prawidłowego funkcjonowania utrzymującej je instytucji (np. gliniane tabliczki z pismem klinowym przetrwały w wielkiej liczbie mimo upadku całej wytwarzającej je cywilizacji). W archiwach cyfrowych niestety nawet sama trwałość danych musi być gwarantowana przez sprawnie działającą instytucję, do utrzymania informacji niezbędne są bowiem okresowe działania „odświeżające", np. migracja. Informacja prawidłowo utrzymana może zaś nie być dostępna ze względu na przeszkody prawne, np. związane z prawami autorskimi: spory prawne (wystarczy przypomnieć słynną sprawę „Bolka i Lolka") lub brak informacji co do praw lub ich właścicieli. Dostępność i użyteczność informacji może też być ograniczona ze względu na niedoskonałość rozwiązań prawno-organizacyjnych (tu przykładem mogą być archiwa IPN).

Przez trwałość w warstwie organizacyjnej rozumieć więc można istnienie w zadanym okresie odpowiednich instytucji, zapewniających utrzymanie trwałości informacji oraz jej sprawne i wiarygodne udostępnianie. Przez trwałość w warstwie prawnej z kolei rozumieć należy z jednej strony trwałe istnienie odpowiednich przepisów, które umożliwiają, a w wybranych przypadkach zapewniają utrzymanie informacji archiwalnej oraz jej udostępnianie, z drugiej strony odpowiednie powiązanie zasobów archiwalnych z jasnymi i zweryfikowanymi prawami do tych zasobów rozumianych jako własność intelektualna.

2.1.2. Weryfikowalność poprawności przechowywania

Z pewnym przechowywaniem informacji mamy do czynienia tylko wtedy, gdy możliwe jest zweryfikowanie, czy informacja zapisana i odczytana jest całkowicie zgodna. Osiągnąć to można za pomocą powszechnie używanych środków technicznych, takich jak sumy kontrolne. W elektronicznym archiwum wieczystym postulujemy możliwość tego typu weryfikacji w warstwach strumienia danych oraz struktury.

2.1.3. Integralność informacji

Integralność informacji w tym kontekście oznacza [4], że informacja jest kompletna (także w sensie spełniania wymagań narzuconych przez format) oraz że nie dokonano w niej nieuprawnionych modyfikacji (np. jako skutku nietrwałości zapisu, błędu systemu, pomyłki ludzkiej lub celowego nieuprawnionego działania). Integralność dla zasobów archiwalnych można zapewnić np. przez jej kontrolę przed zapisem, połączoną z pewnością przechowywania. Integralność postulować powinniśmy w warstwach strumienia danych, struktury oraz obiektu (kontenera i modelu logicznego).

2.1.4. Autentyczność informacji

Autentyczność informacji rozumiana jest tu [4] jako zgodność rzeczywistej zawartości zasobu informacyjnego z zawartością deklarowaną (np. w metadanych). Aby ta cecha mogła być zweryfikowana, niezbędne jest istnienie zapisów dokumentujących wszelkie uprawnione zmiany w zasobie. Ta cecha powinna być postulowana w warstwie obiektu, gdyż w warstwach niższych nie występują deklaracje zawartości informacji, a jedynie czysto techniczne nazwy (identyfikatory) pewnych zbiorów informacji.

2.1.5. Dostępność informacji

Przez dostępność (ang. *availability*) informacji rozumiemy [4] możliwość uzyskania żądanego zasobu przez potencjalnego użytkownika oraz to, że pozyskiwane zasoby są interpretowalne. Pierwsza z tych cech wymaga trwałości przechowywania zasobów oraz możliwości ich odnalezienia. Druga z cech wynika wprost z trwałości w warstwie semantycznej, która oczywiście musi być bezwzględnie postulowana (choć w perspektywie wieczystej może być trudna do osiągnięcia).

2.1.6. Poufność informacji

Poufność informacji jest rozumiana [4] jako gwarancja, że zasoby będą udostępniane jedynie uprawnionym użytkownikom. Poufność musi być zapewniona we wszystkich warstwach od warstwy strumienia w górę, a dla archiwów przechowujących szczególnie wrażliwe dane należy rozważyć poufność także na poziomie nośnika i sygnału – zapewnić ją można np. przez szyfrowanie zapisywanych danych.

2.2. Struktura zasobów i dostęp do nich

2.2.1. Treść zasobów

Treścią zasobu archiwalnego powinna być dowolna forma zapisu cyfrowego, zalecać jednak należy, by w miarę możliwości stosowane były formaty możliwie powszechnie uznane. Dodatkowymi pożądanymi cechami formatu są zwartość i samokontrola.

Choć wielkość zasobów jest istotna, postuluje się stosowanie wyłącznie bezstratnych algorytmów kompresji, gdyż wierność odtworzenia zasobu jest celem pierwszoplanowym.

2.2.2. Metadane

Gromadzenie zasobów w archiwum ma sens jedynie wtedy, gdy będą one możliwe do odnalezienia. Niezbędny jest zatem opis zasobów w postaci metadanych. Istnienie rozbudowanego systemu metadanych, możliwość ich przeszukiwania i zachowanie ich spójności z obiektami archiwalnymi na każdym poziomie hierarchii archiwów jest jednym z najważ-

niejszych, z użytkowego punktu widzenia, wymagań wobec archiwów cyfrowych. Istotna jest szczegółowość metadanych, umożliwiająca odnalezienie wybranych obiektów, a nawet ich fragmentów. Z drugiej strony, zbyt szczegółowe metadane przy wielkich rozmiarach archiwów w praktyce utrudniają lub nawet uniemożliwiają przeszukiwanie (*casus* Google). Metadane powinny być fizycznie i nierozerwalnie związane z obiektem archiwalnym, ale z drugiej strony powinny pozostać w systemie nawet po usunięciu zeń tego obiektu.

Metadane opisujące zasób archiwalny pełnią jeszcze inne funkcje i muszą zawierać informacje:

- statyczne – opisujące sam zasób:
  - techniczne – opisujące sposób, w jaki zasób został wytworzony (np. szczegóły procesu digitalizacji) oraz format zapisu,
  - opisowe – wyjaśniające „merytoryczną" zawartość zasobu,
  - strukturalne – przy obiektach złożonych (np. książka złożona ze stron, czasopismo z serii, roczników i zeszytów; prezentacja złożona np. z powiązanych w czasie plików dźwiękowych i graficznych),
  - behawioralne – opisujące sposób prezentacji obiektu przy jego udostępnianiu (np. zdjęcie pionowo czy poziomo),
  - prawne – dane o prawach autorskich, ograniczenia udostępniania;
- dynamiczne – opisujące zmiany dokonywane w zasobie, ewentualnie także historię wykorzystywania zasobu, np. dane o umieszczeniu fragmentu obiektu w innych obiektach archiwalnych (cytaty, montaże, wykorzystanie twórcze).

Aby metadane były użyteczne, muszą być zapisane w sposób, który umożliwi ich interpretację także w odległej przyszłości. Ułatwiają to samoopisujące formaty tekstowe; jak się wydaje obecnie, zalecać należy użycie XML.

Dla różnych rodzajów archiwizowanej treści istnieją różne specyficzne standardy metadanych; w wielu dziedzinach istnieje nawet po kilka konkurencyjnych standardów i toczą się spory o wyższość jednych nad drugimi, należy też spodziewać się pojawiania się nowych propozycji. Właściwym rozwiązaniem wydaje się zapis niezależny od źródłowego i docelowego formatu, z konwersją przy zapisie i odczycie. Aby taka konwersja była możliwa także w dalekiej przyszłości, zalecić należy zapisywanie metadanych w sposób bardzo ogólny, samoopisujący i szeroko stosowany, np. w postaci trójek RDF[2]. Co do treści metadanych radzić można jedynie stosowanie bardzo ogólnych standardów (np. *Dublin Core*), ale oczywiście nie rozwiązuje to problemu interpretacji informacji bardziej wyspecjalizowanej. Dobrą gwarancję prawidłowej interpretacji metadanych uzyskać można byłoby opierając system zapisu

[2] RDF (*Resource Description Framework*) – specyfikacja modelu metadanych, określona przez *World Wide Web Consortium*, zazwyczaj implementowana w języku XML.

metadanych na ontologiach, które oczywiście musiałyby także być zapisane w powszechnie zrozumiały sposób (np. w OWL[3]) i przechowywane w archiwum.

Metadane muszą dać się pozyskać z archiwum w formatach zgodnych z obowiązującymi w danej dziedzinie standardami zarówno międzynarodowymi, jak i krajowymi [7]. Jest kwestią dyskusyjną, czy konwersji ma dokonywać samo archiwum, czy jakiś system zewnętrzny, natomiast koniecznością jest przechowywanie w archiwum wszystkich informacji wymaganych w danym standardzie. Niektórzy użytkownicy żądają przechowywania i udostępniania metadanych w ich „ulubionym" formacie, który może nie być ogólnie przyjęty. W takiej sytuacji rozważyć należy podwójną reprezentację metadanych w archiwum: w formacie użytkownika oraz w formacie ogólnym, łatwym do konwersji.

W odróżnieniu od zasobów archiwalnych, które dostępne są na zamówienie, metadane muszą być dostępne na żądanie, tj. z zapewnieniem znacznej interaktywności. By jednak uzyskać pewność dostępu do zasobu, metadane (przynajmniej te statyczne) powinny być przechowywane razem z zasobem, czyli w archiwum głębokim. Stwarza to konieczność duplikacji zapisu metadanych: „oryginał" powinien być przechowywany razem z opisywanym zasobem, a „kopia robocza" – w interaktywnie dostępnym buforze. System archiwizacji musi zapewniać zgodność obu egzemplarzy metadanych.

2.2.3. Opakowanie zasobów

Ponieważ zasadniczym celem istnienia cyfrowego zasobu archiwalnego jest umożliwienie jego odnalezienia, odczytu i interpretacji po upływie bardzo długiego czasu, niezbędne jest wspólne przechowywanie danych i metadanych, najlepiej opakowanych w jeden obiekt-kontener (pakiet), tak by nie było możliwe „zgubienie" metadanych opisujących dany zasób. Struktura takiego obiektu powinna być łatwa w interpretacji, możliwie samoopisująca (*self-describing*) i kompletna (*self-contained*). Cechy te są niezbędne, by zasób mógł być prawidłowo zinterpretowany w odległej przyszłości, gdy kontekst jego istnienia może być nieznany, format niezrozumiały, a obiekty powiązane z zasobem – utracone.

2.2.4. Dostęp do zasobów archiwum

Samo pewne przechowywanie zasobów archiwalnych nie jest oczywiście celem istnienia archiwum, lecz jedynie środkiem do celu, którym jest długoterminowe lub bezterminowe zapewnienie możliwości dostępu do zgromadzonych obiektów. Warunkiem dostępu jest oczywiście dostępność (ang. *availability*) informacji rozumiana tak, jak opisano to wyżej. Nie wystarcza ona jednak do zapewnienia efektywnego korzystania z informacji. Niezbędne

[3] OWL (*Web Ontology Language*) – język służący do zapisu ontologii, o składni opartej na XML, standard *World Wide Web Consortium*.

są jeszcze mechanizmy wyszukiwania informacji, oparte na wiedzy zgromadzonej w metadanych oraz mechanizmy buforowania.

Zaznaczyć jednak trzeba wyraźnie, że bezpośrednie udostępnianie informacji użytkownikom końcowym (*on-line* czy w innej formie) nie jest w ogóle zadaniem archiwum. Powinny to robić systemy zewnętrzne, które z archiwum czerpią zasoby i pośredniczą w ich udostępnianiu końcowemu odbiorcy.

2.2.5. Wyszukiwanie informacji

Zasoby archiwalne muszą dać się efektywnie wyszukiwać według różnych kryteriów. By było to możliwe, niezbędne jest gromadzenie, obok właściwych zasobów, także metadanych opisujących ich zawartość.

2.2.6. Buforowanie informacji

W archiwum głębokim dostęp do informacji, zwłaszcza o większym rozmiarze, może zajmować dość znaczny czas, a w dodatku nie można gwarantować, iż cały zasób (który – zwłaszcza w przypadku multimediów – może mieć znaczny rozmiar) będzie dostępny jednocześnie; poszczególne części zasobu mogą być osiągalne sukcesywnie (takie założenie jest ważne m.in. ze względu na optymalizację zasilania). Dlatego niezbędne jest buforowanie informacji na styku archiwum głębokiego z otoczeniem zarówno przy pozyskiwaniu zasobów z archiwum, jak i przy ładowaniu ich do archiwum.

2.3. Inne wymagania

2.3.1. Dyslokacja i technologiczna odrębność przechowywania

Urządzenia, na których zapisane są dane, mogą, mimo wszelkich zabezpieczeń, ulec zniszczeniu na skutek jakiejś awarii (zalania, pożaru, przepięcia itp.) lub przypadkowego bądź celowego działania człowieka (jak błędy obsługi, kradzież, dywersja, zamieszki, działania wojenne itp.). Jedynym sposobem uniknięcia tego typu zagrożeń jest zastosowanie dyslokacji, tj. wymaganie, by każdy zasób był zapisany przynajmniej dwukrotnie, przy czym aparatura, na której znajduje się każda z kopii, powinna być oddalona od siebie na znaczną odległość (w przypadku szczególnie cennych zbiorów powinny być to setki kilometrów).

Z powodu potencjalnych trudności z poprawnym działaniem urządzeń i prawidłowym odczytem formatu danych w dalekiej przyszłości postuluje się także technologiczną odrębność przechowywania, tj. wymaganie, by dane były przechowywane w kilku różnych formatach i na różnych nośnikach.

2.3.2. Niezależność warstw

System informacyjny archiwum wieczystego powinien być tak skonstruowany, by możliwe było niezależne zmienianie reprezentacji poszczególnych warstw (patrz część 2) , np. sposobu zapisu na nośniku, systemu „plików", formatów danych, sposobu opakowania obiektów itd., w całym systemie lub w jego części, bez zaburzenia poprawnego działania całego systemu. Taka niezależność jest niezbędna, gdyż w długim okresie eksploatacji archiwum wszelkie stosowane obecnie standardy, formaty itp. przestaną być aktualne, a archiwum – by zachować „żywotność" – musi być gotowe do przyjęcia nowych rozwiązań. Jednocześnie – ze względu na rozmiar i bezpieczeństwo zasobów – trudno sobie wyobrazić przeprowadzenie *upgrade'u* całego archiwum, wraz z zawartością, z powodu np. zmiany obowiązujących standardów.

2.3.3. Niezależność od infrastruktury technicznej

Ze względu na długookresowość planowanej eksploatacji archiwum postuluje się taką jego budowę, by poszczególne komponenty systemu informatycznego archiwum – zarówno sprzętowe jak i programowe – mogły być bez większych trudności wymieniane, także (a może zwłaszcza) na komponenty wykonane w innej technologii. Wskazane jest też, by proste komponenty mogły być zastępowane przez pełniące takie same zadania, ale bardziej złożone podsystemy, czyli by możliwa była hierarchiczna rozbudowa archiwum.

2.3.4. Wymagania ekonomiczne

Archiwa wieczyste przechowywać będą ogromne ilości informacji, a eksploatowane powinny być przez dziesięciolecia lub nawet stulecia. Dlatego szczególnie ważna jest ich efektywność ekonomiczna, a w szczególności osiągnięcie możliwie małego kosztu przechowywania jednostki objętości danych oraz minimalizacja kosztów stałych, niezależnych od wykonywanych w danym okresie zadań. Wśród czynników umożliwiających ograniczenie kosztów szczególnie ważna wydaje się oszczędność energii potrzebnej do utrzymania archiwum.

2.3.5. Certyfikacja archiwów

By poważne instytucje odpowiedzialne za powierzone im zasoby archiwalne mogły i chciały wykorzystać system elektronicznego archiwum wieczystego, niezbędne jest udowodnienie, że posiada ono pożądane cechy, a w szczególności, że zapewnia wiarygodność przechowywania zasobów. Takiego dowodu może dostarczyć proces certyfikacji konkretnego rozwiązania. Postuluje się zatem, by elektroniczne archiwum wieczyste umożliwiało przeprowadzenie procesu certyfikacji oraz by spełniało warunki certyfikacji.

Oczywiście certyfikacja musi opierać się na pewnym standardzie wymagań w stosunku do archiwum. Aktualnie najpełniejszą propozycją takiego standardu wydaje się opracowanie

grupy Nestor [5, 4]. Wymagania przedstawione w części 2.1 w dużej mierze oparte są na tym właśnie opracowaniu.

3. Archiwa wieczyste a bazy danych

Na pierwszy rzut oka może się wydawać, że długoterminowe archiwum cyfrowe jest po prostu specyficznym rodzajem bazy danych. Jest to jednak wrażenie mylne – zarówno cele archiwum długoterminowego, jak i potrzebne środki różnią się znacznie. Niektóre problemy są jednak podobne, więc doświadczenia zdobyte przy konstruowaniu systemów z bazami danych mogą być pomocne i w konstrukcji archiwów wieczystych.

Podstawowy cel archiwum, czyli pewne przechowywanie informacji, jest oczywiście podobny jak w przypadku baz danych. Jak jednak pokazano wyżej, trwałość informacji jest w przypadku archiwów wieczystych rozumiana szerzej niż w bazach danych.

Cykl życia informacji w archiwum nieco przypomina hurtownię danych: dane są ładowane w trybie wsadowym, a kasowanie danych jest sporadyczne. Pozyskiwanie danych jest za to wyraźnie inne: w hurtowniach ma charakter w dużej mierze interaktywny (analizy typu OLAP), a w archiwum głębokim – wyłącznie wsadowy. Archiwum można też uznać za system typu WORO (*Write Once, Read Occasionally*), a więc taki, w którym poszczególne dane są używane bardzo rzadko, podczas gdy w hurtowni dane mogą być odczytywane stosunkowo często.

Co do natury danych, archiwum przypomina systemy typu *content management* – zawiera różne treści, w dużej części multimedialne. Objętość danych w archiwum jest jednak nieporównanie większa. Jednak problemy związane z wyszukiwaniem informacji są w obu typach systemów bardzo podobne. Ze względu na rodzaj przechowywanych zasobów oraz ich wielkość, archiwa przypominają też nieco multimedialne bazy danych, jest jednak wiele różnic: wielkość zasobów w archiwach i pożądana trwałość jest wielokrotnie większa, co uniemożliwia stosowanie tej samej technologii, inny jest też sposób operowania na danych: bazy multimedialne zwykle umożliwiają dostęp na żądanie, a w wielu przypadkach posiadają wewnętrzną możliwość manipulowania na składowanych obiektach multimedialnych, np. ich dekompozycji, analizowania czy przeszukiwania. Archiwa oferują dostęp jedynie na zamówienie, a możliwości manipulacji i przeszukiwania ograniczone są do metadanych.

Zupełnie egzotyczne z punktu widzenia specjalistów od baz danych są – zasadnicze dla archiwów – problemy z trwałością nośnika. Konstruując system zarządzania bazą danych, zakłada się po prostu trwałość nośników, zaś w praktyce, ze względu na stały rozwój technologii, dane są dość często migrowane na nośniki nowe. Dodatkowo trwałość danych w bazach danych wspomagają macierze typu RAID, a w razie kłopotów można dane odtworzyć z

kopii rezerwowych. Takie podejście w przypadku archiwów jest oczywiście – ze względu na wielkość zasobów – niemożliwe do zaakceptowania.

Podobnie nieznane w „świecie" baz danych są problemy wynikające ze względnej nietrwałości technologii. W stosunku do typowego czasu życia bazy danych (kilkanaście lat) technologie są bowiem stosunkowo trwałe, a wybór technologii, mającej dobrą pozycję rynkową w chwili tworzenia nowego systemu, z dużym prawdopodobieństwem zapewnia aktualność i rozwój tej technologii oraz dostępność wsparcia producenta przez cały czas życia systemu. Po kilkunastoletnim czasie eksploatacji stare systemy są zwykle zastępowane nowymi, a dane są – przynajmniej częściowo – migrowane.

Jedynie w przypadku szczególnie wielkich baz danych, gdzie migracja byłaby bardzo kosztowna i ryzykowna, ten sam system utrzymuje się siłą „przy życiu" przez wiele dziesięcioleci (dobry przykład to egzystowanie systemów IBM IMS – przedstawicieli przebrzmiałej z górą trzydzieści lat temu technologii baz hierarchicznych). Problemy występujące w takich systemach są podobne jak w archiwach wieczystych, a kłopoty z nimi pokazują wyraźnie, że postulat niezależności od technologii jest kluczowy dla przechowywania długoterminowego.

Choć archiwum nie jest systemem transakcyjnym, jednak pewne problemy podobne do systemów transakcyjnych będą w nim występować. W szczególności niezbędne jest zapewnienie atomowości operacji zapisu obiektów do archiwum, a w systemach z dyslokacją danych (patrz 4.2.4) występują problemy podobne jak w rozproszonych transakcjach. Konieczność duplikacji metadanych (patrz 2.2.2) rodzi zaś dobrze znane w świecie baz danych problemy ze zgodnością replik.

4. Koncepcja architektury elektronicznego archiwum wieczystego

Po analizie wymagań i istniejących rozwiązań uznać należy, że wymagania stawiane przez archiwizację wieczystą nie mogą być w dostatecznym stopniu spełnione przez typowy sprzęt (np. macierze dyskowe) ani oprogramowanie (np. bazy danych). Zaproponować zatem trzeba rozwiązania specjalne. Poniżej opisana zostanie taka propozycja architektury elektronicznego archiwum wieczystego, bazująca na koncepcji zasobnika, będącego opartym na strukturze sieci danych (*data raster*) specjalizowanym rozwiązaniem sprzętowo-programowym.

4.1. Założenia

Budowane archiwum jest archiwum głębokim w sensie opisanym wcześniej. Powinno ono spełniać wszystkie postulaty opisane wyżej, a także dawać możliwość zastosowania standardów wypracowanych przez odpowiednie organizacje, takich jak model OAIS.

Jeśli proponowana architektura i rozwiązania szczegółowe spełnią w znacznym stopniu opisany wyżej postulat niezależności od technologii, system może stać się rzeczywiście wieczysty, gdyż będzie mógł być w czasie całego okresu eksploatacji stopniowo regenerowany przez wymianę przestarzałych komponentów na współczesne. Zaproponowana architektura pozwala uniknąć problemu *technology flow*, który – ze względu na bardzo długi okres eksploatacji – ma zasadnicze znaczenie w konstruowaniu archiwów elektronicznych.

Szczególny nacisk w opracowanej propozycji położono na ekonomiczną opłacalność przedsięwzięcia, a w szczególności minimalizację kosztów związanych z zasilaniem oraz konstrukcję potrzebnych urządzeń z typowych (a zatem niezbyt drogich) komponentów. Jednocześnie wymiana starych komponentów niezbędna jest tylko wtedy, gdy nie są one już zdatne do dalszej pracy albo gdy dalsza praca nie jest celowa z powodów technicznych (np. brak części zamiennych) lub ekonomicznych (np. wysokie koszty eksploatacji czy serwisu). Konieczności wymiany nie powoduje zaś samo „moralne" zestarzenie się danej technologii czy też jej niekompatybilność z najnowszymi rozwiązaniami.

Proponowana bazowa architektura systemu elektronicznego archiwum wieczystego jest przedstawiona na rysunku 1 i składa się z następujących obiektów:

- specjalnego zasobnika do długoterminowego składowania danych;
- elektronicznej kartoteki udostępniającej metadane, z funkcjami wyszukiwania oraz adresacji położenia żądanego zasobu;
- witryny udostępniającej, czyli podsystemu udostępniania zasobów użytkownikom (z odpowiednim buforowaniem);
- zespołu przygotowania treści do ich składowania w archiwum (z odpowiednim buforowaniem);
- podsystemu administracji archiwum, nadzorującego działanie procesów wykonywanych automatycznie oraz procedur wymagających ścisłego nadzoru;
- zewnętrznych podsystemów realizujących usługi specjalne, czyli pewne specyficzne funkcje archiwum trudne do realizacji technicznej, a wymagane przez użytkowników (zostały one wydzielone do zewnętrznego podsystemu dla uproszczenia konstrukcji zasobnika i uniezależnienia jej od specyfiki konkretnego archiwum).

W kontekście modelu warstwowego przedstawionego wcześniej, poszczególne części architektury można przypisać tak: zasobnik odpowiada za warstwę strumienia danych, tj. w sposób pewny przechowuje wprowadzone informacje zorganizowane podobnie do systemu

plików, witryna udostępniająca (przy odczycie) i zespół przygotowania treści (przy zapisie) odpowiadają zaś warstwie obiektu, odpowiednio obsługując złożoną strukturę kontenerów opakowanych zasobów.

Zaznaczyć należy, że choć w podstawowej konfiguracji archiwum składa się z jednego zasobnika i jednej witryny udostępniającej z kartoteką, możliwe są także inne konfiguracje, gdzie w archiwum funkcjonuje kilka zasobników i/lub kilka wyspecjalizowanych kartotek czy witryn.

4.2. Koncepcja zasobników

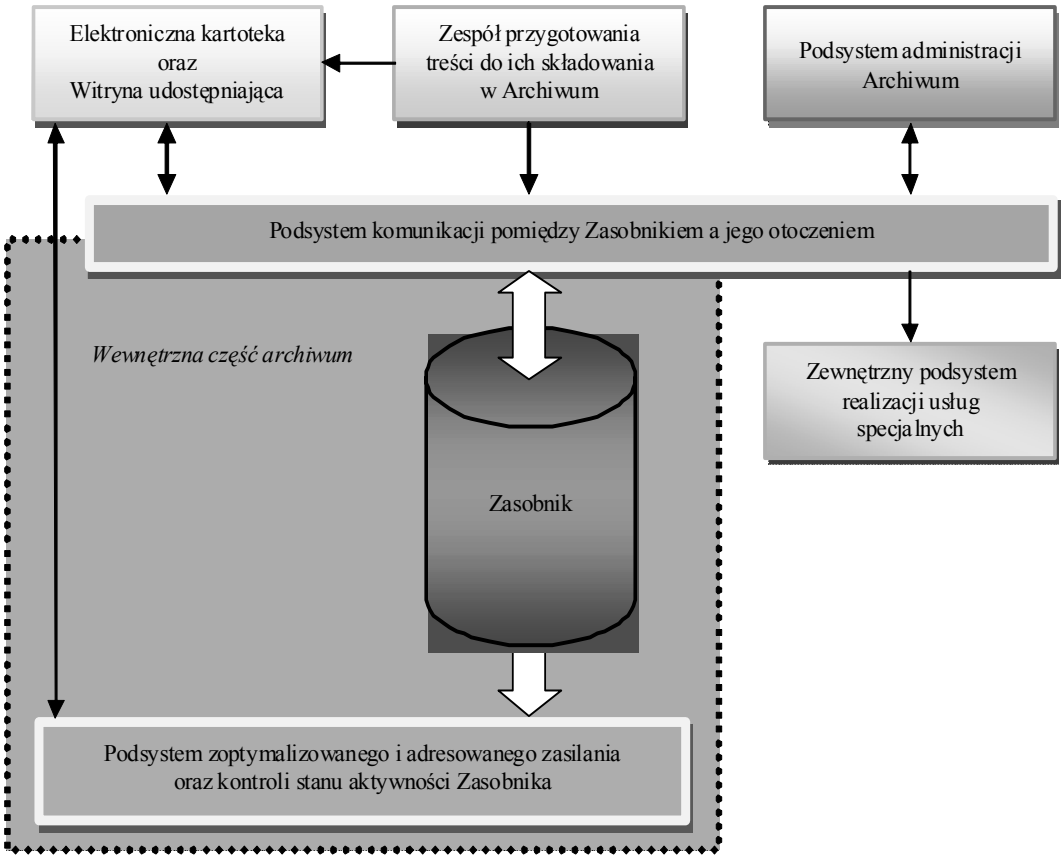
Kluczowy element proponowanej architektury archiwum stanowi specjalny zasobnik do długoterminowego składowania danych [10, 11]. Pełni on rolę pamięci masowej i zasadniczo może być wykorzystywany we wszystkich zastosowaniach, w których aplikuje się tego typu pamięci, ale przede wszystkim jest dostosowany do budowy archiwum głębokiego o wielkiej pojemności. W celu znalezienia odpowiedniej struktury takiego zasobnika określono następujące założenia:

- Rozszerzenie przestrzeni zasobnika powinno być łatwe w realizacji oraz powinno być możliwe powiększenie go do praktycznie dowolnego rozmiaru.
- Powinno być możliwe aktywowanie oraz dezaktywowanie dowolnej określonej części przestrzeni zasobnika, gdy całość zasobnika nie jest potrzebna.
- Każdy fragment przestrzeni zasobnika powinien być niezależny funkcjonalnie od innych fragmentów oraz jego całości.
- Adresowanie przestrzeni zasobnika powinno być jednoznaczne i możliwe powinno być adresowanie liniowe dla całej przestrzeni adresowej zasobnika.
- Adresowanie przestrzeni zasobnika w miarę możliwości powinno zachowywać chronologię jego rozbudowy.
- Architektura zasobnika w miarę możliwości powinna pozwalać, by każdy funkcjonalny fragment miał analogiczną architekturę jak całość, tzn. że można go wydzielić i odseparować od całości, a następnie potraktować jako samodzielny zasobnik.
- Architektura powinna dopuszczać możliwość rozbudowy zasobnika w dłuższym okresie z technologicznie niehomogenicznych fragmentów.

Proponowany system zasobnika składa się z:

- właściwego zasobnika, tj. sprzętowo-programowego urządzenia zapewniającego długoterminową pamięć trwałą;
- podsystemu komunikacji pomiędzy zasobnikiem a jego otoczeniem zewnętrznym;
- podsystemu zoptymalizowanego i adresowanego zasilania oraz kontroli stanu aktywności zasobnika.

Podstawowym założeniem dla zasobnika jest żądanie, by samoistnie zapewniał on długoterminową trwałość przechowywanej informacji. Oznacza to, że zasobnik musi być wyposażony w automatyczne mechanizmy kontroli i odświeżania zapisu.



Rys. 1. Architektura elektronicznego archiwum wieczystego
Fig. 1. Architecture of electronic long-time archive

4.2.1. Regeneracja danych

Jak wiemy, długoterminowej trwałości zapisu nie da się osiągnąć wyłącznie przez zastosowanie możliwie trwałego nośnika, trwałość najlepszych obecnie dostępnych nośników nie jest bowiem wystarczająca dla zadań archiwizacji wieczystej. Pozostaje zatem dynamiczne przechowywanie informacji. Spodziewane uszkodzenia nośnika mogą mieć charakter nieodwracalny i wówczas jedyną metodą jest migracja. W przypadku powszechnie dziś stosowanych nośników magnetycznych najczęstsze uszkodzenia nie są nieodwracalne, lecz polegają na odwracalnym samoistnym rozmagnesowaniu. W takim przypadku nie jest konieczna migracja, a jedynie regeneracja zapisu. Regeneracja jest oczywiście znacznie tańsza od migracji i nie nastręcza większych problemów logistycznych. Także wielkie rozmiary informacji nie prowadzą tu do znaczących problemów, ponieważ regeneracja może być przeprowadzana sukcesywnie na kolejnych stosunkowo niewielkich fragmentach. Zauważyć należy, że by regeneracja była możliwa, konieczne jest stosowanie nośników wielokrotnie zapisywalnych

(R/W), a nie jednokrotnego zapisu. Do takiego użycia nadają się wszelkie typy nieulotnych pamięci wielokrotnego zapisu, takie jak dyski magnetyczne, inne nośniki wielokrotnego zapisu, np. CD-RW, DVD-RW, MD (*Minidisc*) itp., pamięci *flash*, a także – z pewnymi ograniczeniami – taśmy magnetyczne. Koncepcja odświeżania bazującego na regeneracji parametrów stanu zapisu jest rozwiązaniem oryginalnym, zgłoszonym do opatentowania [12].

Oczywiście proces odświeżania zapisu powinien być stosowany zapobiegawczo, tj. zanim dane ulegną uszkodzeniu. Zasobnik zatem powinien automatycznie wykonywać wymagane czynności serwisowe z odpowiednią, wynikającą z cech nośnika, częstotliwością. Jeśli jednak dane zostały nieznacznie uszkodzone, możliwa powinna być ich lokalna, tj. bez sięgania do zewnętrznej kopii, rekonstrukcja, np. na podstawie redundantnych zapisów. Jeśli stan nośnika nie pozwala na dalszą regenerację, konieczna jest migracja na inny nośnik. Także ten proces powinien być zautomatyzowany, co oznacza konieczność istnienia automatycznych procedur diagnostycznych.

Proces odświeżania wiąże się oczywiście z pewnym ryzykiem, ale może ono być znacznie zredukowane dzięki dyslokacji danych i takiemu zsynchronizowaniu odświeżania, by nigdy nie następowało ono jednocześnie na więcej niż jednej kopii tego samego zasobu.

4.2.2. Budowa zasobnika

Wewnętrzna budowa zasobnika opiera się na koncepcji tzw. sieci danych (*data grid*) [8]. Zasobnik złożony jest zatem z niezależnych programowalnych elementów nazywanych węzłami. Zadaniem węzłów jest przechowywanie i opieka nad stanem zapisanej do ich pamięci treści w formie plików cyfrowych. Zasobnik musi też zawierać wewnętrzną kartotekę techniczną, pełniącą funkcje zbliżone do FAT.

Zakłada się, że konstrukcja zasobnika powinna być możliwie najprostsza i ograniczona do jednego tylko serwisu: przekazu plików cyfrowych (danych i metadanych) z i do zasobnika. Dopuszczalne są jednak dodatkowe zadania, które mogą być wykonywane lokalnie w węzłach zasobnika, tzn. w ramach ich oprogramowania, pozwalające na przetwarzanie danych, pod warunkiem że produktem przetworzenia będzie nowy plik, a nie strumień danych.

Zaproponowana architektura zasobnika jest skalowalna, czyli pozwala na swobodną rozbudowę przestrzeni składowania. Dopuszczalna jest również rozbudowa lokalnej pamięci związanej z każdym węzłem. Węzły mogą być wyposażone w pamięć o różnej pojemności, nie ma też koncepcyjnych przeszkód, by mogły być wykonane w różnych technologiach. Jest to ważne, gdyż rozbudowa zasobnika może trwać latami, zatem zapewnienie homogeniczności technologii nie jest możliwe. Architektura taka, przy zachowaniu odpowiedniego adresowania, pozwala też na pełną optymalizację zasilania zasobnika.

Jeśli chodzi o typ pamięci użytych do budowy zasobnika, to oprócz omówionego wcześniej wymagania, by były to pamięci wielokrotnego zapisu, brać trzeba pod uwagę także

aspekty ekonomiczne oraz łatwość serwisowania. Do budowy zasobnika należy zatem zastosować podzespoły najbardziej popularne w czasie budowy konkretnej części zasobnika, unikać zaś rozwiązań egzotycznych. Dlatego obecnie postuluje się wykorzystanie do konstrukcji zasobnika zwykłych dysków magnetycznych. Wraz z popularyzacją innych typów dużych pamięci masowych wielokrotnego zapisu, np. macierzy typu MAID (*Massive Array of Idle Disks*), czy pamięci półprzewodnikowych typu *flash*, możliwe stanie się konstruowanie zasobników (lub ich węzłów) opartych na tych nowych technologiach.

4.2.3. Optymalizacja zasilania

Elektroniczne archiwum wieczyste składać się może z wielu zasobników, a każdy z nich z wielu węzłów. Wobec olbrzymiej wielkości zasobów multimedialnych liczba węzłów może wynosić setki lub tysiące. Gdyby cała ta aparatura musiała być uruchomiona w sposób ciągły, jej zapotrzebowanie na energię byłoby bardzo duże, a utrzymanie takiego systemu byłoby niezwykle kosztowne (doliczyć trzeba jeszcze koszty klimatyzowania pomieszczeń).

Na szczęście z założenia, że budujemy archiwum głębokie, w którym nie jest konieczny interaktywny dostęp do zasobów, wynika możliwość uśpienia elementów systemu i budzenia ich tylko wówczas, gdy są potrzebne. Dlatego istotnym składnikiem architektury zasobnika jest podsystem optymalizacji zasilania, odpowiednio sterujący aparaturą i utrzymujący w stanie włączenia tylko te jej elementy, które są w danej chwili niezbędne dla wykonania bieżących zadań zasobnika. Takie zarządzanie energią znakomicie redukuje koszty eksploatacji całego systemu.

Podsystem zasilania został powiązany ze strukturą adresową sieci danych zasobnika [10]. Oznacza to, że odgrywa on aktywną rolę w adresowaniu przestrzeni zasobnika. To rozwiązanie zostało zgłoszone do opatentowania [13]. Dzięki odpowiedniemu adresowaniu w podsystemie zasilania można precyzyjnie określić zapotrzebowanie energetyczne na wykonanie zadań zleconych, np. transportu danych czy regeneracji.

4.2.4. Dyslokacja danych

Mimo zastosowania automatycznej diagnostyki i odświeżania informacji nie można zagwarantować całkowitej pewności przechowywania, jeśli jest ono dokonywane wyłącznie na urządzeniach znajdujących się w fizycznej bliskości; w razie poważniejszej katastrofy grozi bowiem jednoczesne zniszczenie całego zapisu. W systemie archiwum wieczystego trzeba zatem stosować dyslokację kopii zasobów. Dyslokacja może dotyczyć węzłów jednego zasobnika lub kilku zasobników (zasób jest wówczas powielony w kilku zasobnikach odległych geograficznie).

Wprowadzenie wymagania dyslokacji powoduje oczywiście pewne problemy w zarządzaniu zasobami. Przede wszystkim trzeba zapewnić stałą zgodność kopii zarówno w cza-

sie zapisu, jak i w czasie składowania (tu pojawią się problemy podobne do występujących w rozproszonych transakcjach). Synchronizacji wymaga odświeżanie zapisu (co opisano wyżej), a w razie uszkodzenia którejś z kopii potrzebne są procedury odtwarzania jej na podstawie kopii zachowanych. Istnienie kilku kopii wymaga też wypracowania strategii optymalizacji dostępu do nich, by zamówiony zasób był pozyskiwany w sposób najefektywniejszy.

4.3. Witryna udostępniająca i kartoteka

Witryna udostępniająca jest portalem dla użytkowników systemu archiwum. W wersji podstawowej pozwala na udostępnianie i przeszukiwanie metadanych o zasobach oraz realizuje zlecenia pobierania zasobów cyfrowych z zasobnika i ich przekazu do określonych urządzeń.

O ile metadane są udostępniane użytkownikom wprost przez witrynę udostępniającą, o tyle nie przewiduje się przekazywania w ten sposób samych zasobów cyfrowych archiwum, zwłaszcza za pomocą sieci publicznej (jedną z przyczyn jest tu ogromna objętość zasobów multimedialnych). Jako podstawowy model przekazu przyjmuje się architekturę źródło – cel, gdzie źródłem jest archiwum, a celem przekazu jest zwykle odpowiednie określone w zleceniu urządzenie, np. zespół emisyjny, macierz dyskowa zespołów edycyjnych itp.

Jedną z ważnych funkcji witryny jest dokonywanie „rozpakowania" złożonych opakowanych obiektów z archiwum i prezentowanie poszczególnych składników: danych i metadanych. Witryna powinna także wykonywać odpowiednie kontrole spójności przekazywanych zasobów na poziomie obiektu (kontenera).

Jak zauważono wyżej, metadane muszą być składowane nie tylko w archiwum głębokim, ale także w miejscu umożliwiającym interaktywny dostęp, wyszukiwanie itp.; takim miejscem jest elektroniczna kartoteka archiwum. Realizuje ona dwa główne zadania: jest bazą metadanych (stałych i zmiennych) o zasobach oraz uczestniczy w uruchamianiu węzłów.

4.3.1. Metadane

Kartoteka archiwum gromadzi i – przez witrynę udostępniającą – udostępnia metadane opisujące zasoby (stałe), a także zawiera metadane zmienne, opisujące dzieje zmian zasobów i ich wykorzystania. Metadane gromadzone przez kartotekę muszą być utrzymywane w stanie zgodności z metadanymi składowanymi w zasobniku, nie ma jednak konieczności, by były identycznie reprezentowane: w kartotece winny się one znajdować w odpowiedniej bazie danych, w zasobniku zaś powinny mieć raczej postać dokumentów XML. Zgodność kopii metadanych powinna być sprawdzana przynajmniej przy każdej manipulacji na zasobie, tj. przed każdym zapisem zasobu do zasobnika archiwum oraz po każdym odczytaniu zasobu z zasobnika.

Metadane opisujące archiwum muszą oczywiście być udostępniane użytkownikom przez witrynę udostępniającą tego archiwum, ale pożądane jest, by były także dostępne za pomocą bardziej uniwersalnych metod (np. związanych z architekturą SOA), by umożliwiać integrację metadanych wielu archiwów i budowanie zintegrowanych informacyjnie sieci archiwów.

4.4. Przepływ danych w archiwum

Napełnianie zasobników archiwum odbywać się ma wyłącznie przez wydzielony zespół przygotowania treści do ich składowania w archiwum. Podsystem ten powinien dokonywać odpowiedniego opakowania zasobu oraz sprawdzenia spójności obiektu przed jego zapisaniem w zasobniku. Jednocześnie z zapisem zasobu do zasobnika jego metadane powinny być przekazane do kartoteki archiwum.

Modyfikacja zgromadzonych zasobów powinna być oczywiście możliwa, ale w ograniczonym zakresie. Ponieważ zakłada się, że w archiwum głębokim autentyczność i spójność treści musi być bezwzględnie chroniona, a czas dostępu do obiektów nie jest krytyczny, w archiwum nie przewiduje się możliwości cząstkowej edycji (*piecewise update*) zgromadzonych zasobów. Każda modyfikacja wymaga zatem pobrania całego obiektu, rozpakowania go, dokonania potrzebnej zmiany, ponownego zapakowania i zapisu w archiwum.

W przypadku większości archiwizowanych zasobów modyfikacja właściwej treści zasobu nie powinna pociągać za sobą utraty wersji dotychczasowej, konieczne jest więc odpowiednie zorganizowanie procedur modyfikacji, by zmieniony zasób był zapisywany obok dotychczasowego, ale były one powiązane odpowiednią metainformacją. Ze względu na wielkie rozmiary zasobów multimedialnych nie wydaje się wskazane, by różne wersje zasobu były zawarte w tym samym obiekcie-kontenerze. Należy zaznaczyć, że zapisywanie kolejnych wersji archiwizowanej treści może być dla niektórych rodzajów zasobów działaniem dość rutynowym. Tak może się dziać w przypadku zasobów, których oryginał nie ma postaci cyfrowej i jest dobrze zachowany (patrz 1.1); wtedy celowe może być okresowe powtarzanie digitalizacji i umieszczanie w archiwum kolejnych coraz doskonalszych cyfrowych wersji tego samego zasobu.

Modyfikacja samych metadanych statycznych obiektu, gdy nie ulega zmianie opisywana treść, nie musi nieść ze sobą konieczności zachowania poprzedniej wersji, ale powinna istnieć możliwość prowadzenia i przechowywania dziennika zmian.

Kasowanie zasobów z archiwum jest działaniem podejmowanym wyjątkowo i może odbywać się jedynie komisyjnie, np. z jednoczesnym udziałem co najmniej dwóch upoważnionych operatorów.

W każdym przypadku archiwum musi stale zapewniać zgodność metadanych zapisanych w zasobniku (razem z zasobem) z metadanymi przechowywanymi w kartotece archiwum. Niezbędne jest także zapewnienie stałej identyczności dyslokowanych kopii obiektu.

5. Podsumowanie

Problematyka długo- i bezterminowej archiwizacji zasobów cyfrowych, choć wydaje się mieć dość zasadnicze znaczenie, do niedawna nie znajdowała należnego miejsca w badaniach, wysiłkach inżynierów ani w świadomości decydentów. Sytuacja ta wydaje się na szczęście zmieniać, a szybkie znalezienie rozwiązań jest niezbędne wobec eksplozji wytwarzanej informacji cyfrowej i postępującej degradacji zasobów niecyfrowych (np. taśmy filmowej).

W tym opracowaniu przedstawiono wymagania dla elektronicznych archiwów wieczystych, opierając się na obecnie najbardziej zaawansowanych propozycjach standaryzacyjnych. Zaprezentowano też koncepcję architektury archiwum wieczystego, bazującą na pomyśle tzw. zasobnika, będącego rozwiązaniem sprzętowo-programowym, wewnętrznie zorganizowanym jako sieć danych, zapewniającym wiarygodność przechowywania informacji m. in. dzięki dyslokacji danych i odświeżaniu zapisów przez ich regenerację.

5.1. Cechy proponowanego rozwiązania

Należy podkreślić zgodność zaproponowanego rozwiązania z postulatami długoterminowego przechowywania danych cyfrowych [5]: pewność przechowywania danych (zapewnioną przez ustawiczną kontrolę stanu zapisu i jego odświeżanie), możliwość zapewnienia weryfikacji autentyczności danych (przez odpowiednie metadane), zapewnienie integralności danych (w szczególności ich enkapsulacja zgodnie z modelem OAIS), niezależność od infrastruktury (w szczególności niezależność przechowywania od fizycznej metody zapisu), efektywność energetyczną.

Archiwum o zaproponowanej konstrukcji jest możliwe do zbudowania w sposób rozsądnie prosty, z użyciem typowych komponentów o niewygórowanych cenach. Umożliwi długoterminowe przechowanie danych cyfrowych, w sposób efektywny ekonomicznie, gwarantując także możliwość wieloletniego utrzymania, przebudowy i rozbudowy technicznej infrastruktury systemu.

W proponowanej koncepcji szczególnie interesujące i nowatorskie wydają się:

- mechanizmy samokontroli, zapewniające regenerację zapisu danych (automatyzacja opieki nad zgromadzonymi zasobami),

- efektywne energetycznie rozwiązania przechowywania danych,
- rozdzielenie funkcji wewnętrznych i zewnętrznych archiwum,
- zintegrowane zarządzanie zasobami archiwalnymi i metadanymi,
- skalowalność zarówno co do pojemności archiwum, jak i zarządzania zasobami.

5.2. Dalsze prace i plany implementacji

Instytucje, w których afiliowani są autorzy tego opracowania, rozpoczęły projekt, którego celem jest zaproponowanie programu kompleksowego rozwiązania problemu archiwizacji wieczystej przez opracowanie metod, narzędzi i – o ile to możliwe – rozwiązań prawnych i organizacyjnych, umożliwiających i zapewniających długoterminowe wiarygodne przechowywanie zasobów cyfrowych. Wynikiem projektu ma być m. in. prototyp sprzętu i oprogramowania, dający się praktycznie użyć w instytucjach mających wielkie archiwalne zasoby multimedialne. Potencjalnymi użytkownikami zaprojektowanego systemu mają być przede wszystkim instytucje odpowiedzialne za zachowanie narodowych dóbr kultury oraz nadawcy radiowo-telewizyjni.

Problem długoterminowej wiarygodnej archiwizacji zasobów cyfrowych staje się coraz pilniejszy do rozwiązania. Proponowana w tym opracowaniu koncepcja konstrukcji archiwum może – o ile dalsze prace nad nią znajdą odpowiednie wsparcie – przyczynić się do efektywnego ekonomicznie rozwiązania tego problemu, w szczególności w instytucjach już dysponujących wielkimi zasobami informacji cyfrowej i przechowujących ją w sposób daleki od doskonałości. Rozwiązanie problemu wieczystej archiwizacji zasobów cyfrowych jest bardzo ważne, gdyż pozwoli zachować dla przyszłych pokoleń dobra kultury i zasoby informacji, które obecnie zagrożone są zniszczeniem.

BIBLIOGRAFIA

1. Cohen S.: Towards Self-Describing Self-Contained Data Format (SD-SCDF). http://www-.snia.org/forums/dmf/programs/ltacsi/SNIA-DMF_Towards_%20SDSCDF_2007-0412.pdf.
2. Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf.
3. Coy W.: Perspektiven der Langzeitarchivierung multimedialer Objekte, Nestor Working Group, Berlin, 2006. http://edoc.hu-berlin.de/series/nestor-materialien/5/PDF/5.pdf.

4. Dobratz S., Schroger A., Strathmann.: The nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification. Journal of Digital Information, Vol. 8, No. 2 (2007).
5. Nestor Working Group: Catalogue of Criteria for Trusted Digital Repositories, 2006. http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf.
6. Peterson M.: White Paper – The Coming Archive Crisis. SNIA, 2006. www.snia.org-/forums/dmf/programs/ltacsi/100_year/SNIA-DMF_The-%20Coming-Archive-Crisis-_20061130.pdf.
7. Płoszajski G. (red.): Standardy techniczne obiektów cyfrowych przy digitalizacji dziedzictwa kulturowego. Biblioteka Główna Politechniki Warszawskiej, Warszawa 2008 (w druku). Dostępna także pod adresem http://bcpw.bg.pw.edu.pl/dlibra/docmetadata?id=1262.
8. Venugopal S., Buyya R., Ramamohanarao K.: A Taxonomy of Data Grids for Distributed Data Sharing, Management, and Processing. ACM Computing Surveys, Vol. 38, 2006.
9. W3C: XML-binary Optimized Packaging. W3C Recommendation, 2005. http://www.w3-.org/TR/xop10/.
10. Walczak J. P., Marasek K.: A Basic Concept of the Electronic System for Long Term Storage of Digital Data. SMPTE Journal 2009 (w druku).
11. Walczak J. P., Marasek K.: Elektroniczne Archiwum Wieczyste: streszczenie. Manuskrypt, 2008.
12. Walczak J. P.: Zgłoszenie patentowe, w UE nr 08460028.7/EP08460028, w USA nr US 12/166,923.
13. Walczak J. P.: Zgłoszenie patentowe, w UE nr 08460029.5/EP08460029; w USA nr US 12/166,894.

Recenzent: Dr inż. Adam Duszenko

Wpłynęło do Redakcji 20 stycznia 2009 r.

Abstract

The paper describes a concept of electronic long-term archive, designed for trustworthy storage of large volumes of digital information for a period of several generations. Requirements for the long-time archive are shown. They are quite specific, due to the fact, that no known electronic or IT technology can be expected to survive such a long period, in particular no electronic digital media can guarantee data fidelity after dozens of years. A layered

information model, brought from OAIS [2], has been appropriately extended and used for precise definition of trustworthiness and – in particular – persistence of digital information. The requirements are compared to problems typical in database systems. A concept of the archive architecture is presented, based on special hardware + software solution, called storage bin, which is internally organized as a data raster. Several basic issues concerning long-term archive construction are discussed, including data dislocation, metadata storage and flow, and optimization of power consumption.

Adresy

Krzysztof P. MARASEK: Polsko-Japońska Wyższa Szkoła Technik Komputerowych, ul. Koszykowa 86, 02-008 Warszawa, Polska, kmarasek@pjwstk.edu.pl.

Jerzy P. WALCZAK: ATM SA, ul. Grochowska 21a, 04-186 Warszawa, Polska, Jerzy.Walczak@atm.com.pl.

Tomasz TRACZYK, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, ul. Nowowiejska 15/19, 00-665 Warszawa, Polska, T.Traczyk@ia.pw.edu.pl.

Grzegorz PŁOSZAJSKI, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, ul. Nowowiejska 15/19, 00-665 Warszawa, Polska, G.Ploszajski@ia.pw.edu.pl.

Andrzej KAŹMIERSKI: Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, ul. Nowowiejska 15/19, 00-665 Warszawa, Polska, A.Kazmierski@elka.pw.edu.pl.