Pomiary Automatyka Robotyka, ISSN 1427-9126, R. 24, Nr 4/2020, 27–39, DOI: 10.14313/PAR_238/27

Problemy długoterminowej archiwizacji zasobów cyfrowych na przykładzie projektu CREDO

Piotr Pałka, Tomasz Śliwiński, Tomasz Traczyk
Politechnika Warszawska, Instytut Automatyki i Informatyki Stosowanej, Nowowiejska 15/19, 00-665 Warszawa

Streszczenie: Długoterminowe przechowywanie zasobów cyfrowych jest poważnym problemem, który nie znalazł jeszcze ani dostatecznej uwagi ze strony przemysłu IT, ani powszechnie dostępnych rozwiązań. Zachowanie użyteczności zasobów przechowywanych w archiwum cyfrowym wymaga nie tylko niezawodnego składowania plików z danymi, ale także możliwości skutecznego wyszukania informacji, weryfikacji jej autentyczności oraz jej poprawnej interpretacji, zarówno w sensie technicznym (format danych itd.), jak i semantycznym (zrozumienie informacji w odpowiednim kontekście itp.). Artykuł omawia te problemy i przedstawia ich rozwiązania przyjęte w projekcie CREDO.

Słowa kluczowe: archiwizacja długoterminowa, archiwizacja zasobów cyfrowych, repozytoria cyfrowe, przechowywanie danych, metadane

1. Wprowadzenie

Jednym z ważnych, lecz wciąż mało znanych problemów współczesności jest długoterminowe przechowywanie zasobów cyfrowych. Jest to problem nie tylko techniczny, lecz zgoła cywilizacyjny. Bez dobrych metod długotrwałego, a nawet wieczystego zachowania różnorodnych zasobów w postaci cyfrowej, ludzkość może utracić część zapisów współczesnej historii i kultury [12]. Dotyczy to zwłaszcza ostatnich lat, gdy znacząca część dorobku cywilizacyjnego powstaje od razu w tej postaci (ang. *born digital*) i nie ma innej, „analogowej" reprezentacji.

Tymczasem powszechnie stosowane sposoby przechowywania informacji cyfrowej wcale nie zapewniają jej długotrwałej dostępności z co najmniej czterech ważnych powodów:

- nietrwałości nośników – właściwie nie dysponujemy trwałymi metodami zapisu informacji cyfrowej o dużym wolumenie (więcej napisano w częściach 2.1 i 3);
- nietrwałości formatów – używane formaty zapisu informacji cyfrowej wcale nie są trwałe i podlegają ciągłej ewolucji bez gwarancji wstecznej zgodności na dłuższym horyzoncie, dotyczy to nawet powszechnie stosowanych formatów dokumentów;
- nietrwałości samego repozytorium – braku gwarancji długotrwałego działania od strony organizacyjnej, prawnej i finansowej;
- braku metadanych i dobrych metod wyszukiwania, o czym łatwo się przekonać, próbując odnaleźć konkretną fotografię sprzed kilku lat.

Autor korespondujący:
Tomasz Traczyk, t.traczyk@ia.pw.edu.pl

Artykuł recenzowany
nadesłany 15.10.2020 r., przyjęty do druku 10.11.2020 r.

1.1. Czym jest archiwum cyfrowe?

Pod pojęciem archiwum rozumie się zwykle organizację zajmującą się przechowywaniem informacji przeznaczonej dla określonej społeczności użytkowników [8]. Archiwum jest czymś więcej niż tylko technicznie zabezpieczonym trwałym repozytorium plików. Ma też inne zadania i cechy niż używane na bieżąco repozytoria plików czy też bazy danych.

W tzw. archiwum długoterminowym (wieczystym) głównym zadaniem jest przechowywanie zasobów przez bardzo długi czas – nawet wielu pokoleń – w sposób umożliwiający ich przyszłe odnalezienie, niezniekształcony odczyt i poprawną interpretację.

W archiwum płytkim dostęp jest realizowany *on-line*, możliwie szybko po otrzymaniu zadania. Archiwum głębokie [20] jest zaś takim rodzajem archiwum cyfrowego, do którego dostęp nie jest realizowany „na żądanie", lecz „na zamówienie" – odbywa się nie natychmiast, lecz według zaplanowanego wcześniej i zoptymalizowanego harmonogramu [24], dzięki czemu możliwe jest m.in. zapewnienie energetycznej efektywności działania archiwum. Takie podejście jest szczególnie właściwe dla archiwów długoterminowych ze względu na długi czas ich pracy. W tabeli 1 porównano cechy archiwów cyfrowych i baz danych.

W artykule opisano wymagania, jakie powinno spełniać długoterminowe archiwum cyfrowe. Pokazano także sposoby realizacji tych wymagań i związane z tym problemy na przykładzie archiwum cyfrowego CREDO, którego współtwórcami byli autorzy.

1.2. Długoterminowe archiwum cyfrowe CREDO

CREDO (Cyfrowe Repozytorium Dokumentów) jest repozytorium cyfrowym, będącym wynikiem projektu o tej samej nazwie, wykonanego w ramach programu Narodowego Centrum Badań i Rozwoju DEMONSTRATOR+ [22] przez konsorcjum złożone z Polskiej Wytwórni Papierów Wartościowych (lider projektu), Politechniki Warszawskiej oraz firmy Skytechnology Sp. z o.o.

Tab. 1. Porównanie cech baz danych i archiwów cyfrowych
Tab. 1. Comparison of the characteristics of databases and digital archives

	Baza danych	**Archiwum płytkie**	**Archiwum głębokie**
Główny cel	rejestracja i udostępnianie danych	przechowywanie i udostępnianie danych	przechowywanie danych
Czas eksploatacji	kilka – kilkadziesiąt lat	kilkadziesiąt lat	kilkadziesiąt – kilkaset lat
Objętość danych	gigabajty – petabajty	terabajty – exabajty	
Ładowanie danych	interaktywne (OLTP, ang. *OnLine Transaction Processing*) lub wsadowe (hurtownie)	wsadowe	
Odczyty danych	interaktywne lub wsadowe	interaktywne lub wsadowe	wsadowe *on request*
	częste	dość częste	rzadkie: WORO (ang. *Write Once, Read Occasionally*)
Typ treści	głównie tekstowa	głównie multimedialna	
Wierność odtworzenia	konieczna 100%	w niektórych przypadkach dopuszczalne określone błędy	
Modyfikacja zasobów	możliwa	niemożliwa (tylko ograniczone modyfikacje metadanych)	
Przeszukiwanie zasobów	możliwe	zwykle niemożliwe (tylko przeszukiwanie metadanych)	
Trwałość nośników	mało istotna	bardzo istotna	krytyczna
Trwałość technologii	wystarczająca	na ogół wystarczająca	niewystarczająca
Trwałość formatów	nie dotyczy	na ogół wystarczająca	niewystarczająca
Efektywność energetyczna	nieistotna	ważna	bardzo ważna

CREDO ma móc pełnić funkcję zarówno bezpiecznego repozytorium krótkoterminowego, jak i archiwum długoterminowego. W tym drugim przypadku jest tzw. archiwum głębokim. CREDO działa zgodnie z zasadami zawartymi w powszechnie przyjętym standardzie OAIS [8]. Z założenia repozytorium CREDO jest dość uniwersalne, jego rzeczywiste oraz potencjalne zastosowania obejmują różnego rodzaju archiwa (państwowe, zakładowe itp.), ale także potrzeby nadawców RTV i wytwórni filmowych, rejestry archiwów ksiąg wieczystych [17] czy służby zdrowia [26].

2. Wymagania wobec archiwów cyfrowych

Powszechnie uznane wymagania stawiane zbiorom zasobów cyfrowych organizacja ARMA International [2] zebrała w postaci tzw. *Generally Accepted Recordkeeping Principles* [7].

P1 *Principle of Accountability* – zasada odpowiedzialności wymaga istnienia osoby nadzorującej całość procesu zarządzania informacją.

P2 *Principle of Transparency* – zasada transparentności mówi, iż cały proces przechowywania informacji powinien być udokumentowany w sposób otwarty i dający się zweryfikować.

P3 *Principle of Integrity* – zasada integralności wymaga, by sposób przechowywania danych niezawodnie zapewniał ich autentyczność.

P4 *Principle of Protection* – zasada ochrony mówi, że informacja powinna być przechowywana w sposób zapewniający odpowiednią ochronę przed niepowołanym dostępem.

P5 *Principle of Compliance* – zasada zgodności żąda, by informacja była przechowywana w sposób zgodny z wymogami prawa oraz przepisami i politykami lokalnymi.

P6 *Principle of Availability* – zasada dostępności wymaga, by przechowywana informacja mogła być pozyskiwana na czas, w sposób efektywny i dokładnie.

P7 *Principle of Retention* – zasada trwałości żąda, by informacja była przechowywana przez właściwy czas, zgodnie z przepisami i innymi wymaganiami.

P8 *Principle of Disposition* – zasada dysponowania wymaga, by w sposób właściwy, zgodny z przepisami i lokalnymi politykami postępować z informacją, której nie trzeba już dłużej utrzymywać.

O ile zasady P1, P2, P5 i P8 mają właściwie charakter prawno-organizacyjny i co najwyżej mogą być wspierane przez odpowiednie oprogramowanie, o tyle zasady P3, P4, P6 oraz P7 mają już charakter wyraźnie techniczny, choć wymagają także odpowiedniego zaplecza prawno-organizacyjnego, np. odpowiednio zabezpieczonych fizycznie serwerowni.

W archiwum cyfrowym ww. techniczne zasady można zrealizować stawiając następujące wymagania [13, 27, 28].

A1 Trwałość informacji cyfrowej – odpowiada zasadzie P7.

A2 Weryfikowalność poprawności przechowywania – odpowiada zasadzie P3.

A3 Dostępność informacji – odpowiada zasadzie P6.

A4 Poufność informacji – odpowiada zasadzie P4.

Dodano także pewne dodatkowe wymagania, niewynikające wprost z wyżej wymienionych zasad, ale niezbędne dla funkcjonowania archiwum.

A5 Efektywność ekonomiczna przechowywania informacji – wymaganie niezbędne, by w długim horyzoncie można było realizować zasadę P7.

A6 Standaryzacja archiwum – użycie powszechnie przyjętych standardów zapewni realizację większości postulatów, w tym tych o charakterze nietechnicznym.

A7 Certyfikacja archiwum – upewnia że archiwum spełnia niezbędne postulaty i wymagania, w szczególności P2.

Te wymagania przedyskutowano w kolejnych podczęściach, pokazując też sposób ich realizacji w CREDO.

2.1. Trwałość informacji cyfrowej

Podstawową cechą zasobu archiwalnego powinna oczywiście być jego trwałość. Jak się jednak okazuje, cecha ta może być różnie rozumiana i jest przy aktualnym stanie technologii nadspodziewanie trudna do uzyskania.

Znaczenie terminu „trwałość" nie jest wcale oczywiste. Z punktu widzenia celu przechowywania najważniejsze jest to, by zasób mógł być w przyszłości prawidłowo zinterpretowany. Możliwość dokładnego odczytania przechowywanego ciągu bitów, np. w postaci pliku, nie wystarcza zaś do poprawnej interpretacji zasobu, gdyż może nie być znany format zapisu lub zabraknie informacji, jaką właściwie treść dany zasób reprezentuje. Zatem niezawodność tzw. *bitstream preservation* nie wystarcza, by uznać zasób za trwale użyteczny. Niezbędne jest także co najmniej zapewnienie trwałości formatu oraz istnienie trwałych metadanych pozwalających wyszukać zasób i prawidłowo zinterpretować jego treść (por. 2.3).

Z drugiej strony dla niektórych rodzajów zasobów wierność *bitstream preservation* wcale nie jest niezbędna, by osiągnąć zadowalającą jakość przechowania zawartej w zasobie treści, czyli tzw. *content preservation*. Na przykład zniekształcenie odosobnionych pikseli w fotografii czy też niektórych klatek w filmie nie przeszkadza istotnie we właściwym odbiorze zawartych w takim dziele treści.

W dalszym ciągu tego tekstu terminu „trwałość informacji cyfrowej" będziemy używać w węższym sensie, czyli *bitstream preservation*, pamiętając jednak, że nie jest to cecha wystarczająca, jednocześnie też nie musi być konieczna dla ogólnej poprawności przechowywania zasobów.

Ze względu na wagę problemu trwałości informacji cyfrowej, poświecono mu osobną część 3.

2.2. Weryfikowalność przechowywania

Weryfikowalność poprawności przechowywania jest cechą niezbędną dla zapewnienia trwałości. Musi istnieć metoda sprawdzenia, czy nie doszło do uszkodzenia informacji.

Najprostsza jest oczywiście weryfikacja samego przechowywanego strumienia bitów: można do tego użyć odpowiednio dobranej i przechowywanej, np. w metadanych, sumy kontrolnej. To jednak w wielu przypadkach nie wystarczy, potrzebne bywa także zapewnienie dodatkowych cech przechowywanej informacji:

- integralności – czyli pewności, że informacja pozostaje kompletna, np. w sensie wymagań użytego formatu, a także pewności, że nie dokonano nieuprawnionych modyfikacji informacji;
- autentyczności – czyli zgodności zawartości rzeczywistej z deklarowaną, np. w metadanych zasobu;
- niezaprzeczalności – czyli możliwości udowodnienia, że twórca informacji faktycznie ją utworzył.

Pierwsze dwie z tych cech można skontrolować w czasie zapisywania zasobu, a jeśli istnieje możliwość udowodnienia wierności przechowania strumienia bitów, to cechy te nie zostaną naruszone. W przypadku niezaprzeczalności rzecz nie jest tak prosta, gdyż niezawodne stwierdzenie autorstwa danego zasobu wymaga podpisu cyfrowego. Ten zaś dla swej pewności wymaga łańcucha instytucji certyfikujących. O ile w przypadku przechowywania krótkoterminowego nie powoduje to specjalnych trudności, to w archiwum długoterminowym problem jest poważny: nie można mieć pewności, że użyte instytucje certyfikujące będą istniały w odległej przyszłości.

2.3. Dostępność informacji

Dla użyteczności przechowywanej informacji kluczowe znaczenie ma możliwość jej odnalezienia. Do tego niezbędne jest istnienie odpowiednich i łatwo dostępnych metadanych oraz zapewnienie ich efektywnego przeszukiwania. Ze względu na wielkie znaczenie metadanych w przechowywaniu informacji poświęcono im osobną część 4.

Użytkowanie repozytorium ułatwia także odpowiednia jego organizacja logiczna. Najlepiej, jeśli jest ona zgodna z przyzwyczajeniami użytkowników i stosowanymi przez nich procedurami. Archiwum cyfrowe można np. logicznie podzielić na tzw. zespoły archiwalne, podobnie jak dzieje się to w archiwach „analogowych".

Po odnalezieniu danego zasobu powinno być możliwe jego pozyskanie w czasie dostosowanym do celu danego repozytorium. W repozytoriach bieżących oraz archiwach płytkich zwykle oczekuje się możliwie krótkiego czasu dostępu do zasobu. Inaczej jest w archiwach głębokich, gdzie dostęp z zasady nie jest możliwy natychmiast – na żądanie, lecz na zamówienie: zasób zamówiony dostarczany jest po pewnym czasie, niekiedy dość długim (dni, a nawet tygodnie). Wskazane jest, by po zamówieniu zasobu można było określić przewidywany czas jego pozyskania.

Nie mniej ważna jest możliwość prawidłowej interpretacji zasobu, szczególnie problematyczna przy przechowywaniu długoterminowym ze względu na starzenie się formatów zapisu cyfrowego. Jeśli bowiem nawet zachowano opis danego formatu, mogą już nie istnieć narzędzia służące do jego odczytu. Z tego punktu widzenia ważne jest, by informacje przeznaczone do długotrwałego przechowywania zapisywać w formatach możliwie prostych i samodokumentujących (świetnym przykładem jest tu XML) lub specjalnie przeznaczonych do celów archiwalnych (jak PDF/A). Niewskazane zaś jest przechowywanie w formatach prawnie zastrzeżonych (ang. *proprietary*), gdyż nie gwarantują one ani trwałości, ani zgodnej z prawem dostępności w odległej przyszłości.

Samo poprawne odczytanie formatu informacji nie zapewnia jego poprawnej interpretacji. W wielu przypadkach prawidłowe zrozumienie informacji możliwe jest tylko wtedy, gdy umiemy określić kontekst jej powstania, np. fotografia z jakiegoś wydarzenia może nie dać się właściwie zinterpretować, jeśli nie znamy czasu i miejsca jej wykonania i/lub przynajmniej zgrubnego opisu fotografowanego wydarzenia. Takich informacji dostarczają odpowiednie metadane skojarzone z zasobem, a niekiedy – a tak może być w przypadku fotografii cyfrowej – w nim zawarte.

2.4. Poufność informacji

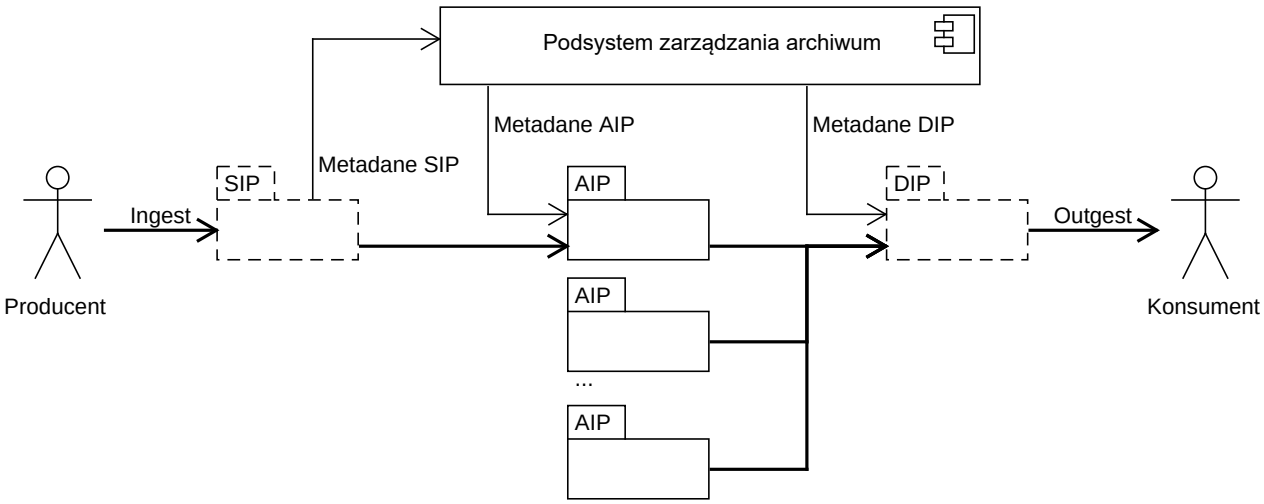
Niezbędna jest gwarancja udostępnienia informacji jedynie podmiotom uprawnionym. Oprócz odpowiedniej ochrony softwarowej samego repozytorium powinna ona obejmować także ochronę kanałów zasilania i dystrybucji informacji oraz ochronę fizyczną serwerowni, miejsc przechowywania nośników itp.

Powiązanym problemem jest ochrona prawna informacji: niektóre zasoby mogą stanowić własność prywatną lub być chronione prawnie z innych przyczyn.

2.5. Efektywność ekonomiczna

Archiwum – zwłaszcza długoterminowe – musi mieć akceptowalne koszty utrzymania. Archiwa cyfrowe mają znacznie mniejsze wymagania lokalowe od tradycyjnych, ale znaczącym składnikiem kosztów utrzymania są koszty energii. Obniżenie tych kosztów musi się jednak wiązać z pogorszeniem dostępności, gdyż oszczędności uzyskać można jedynie okresowo wyłączając zasilanie części urządzeń lub stosując nośniki niewymagające stałego zasilania, jak taśmy czy płyty CD. Takie nośniki jednak mają znacznie dłuższy czas dostępu od pamięci dyskowej i raczej nie są przeznaczone do częstego dostępu.

W przypadku repozytorium *on-line* możliwości ograniczenia kosztów energii są zatem dość iluzoryczne. Inaczej jest w przypadku archiwum głębokiego: tu dostęp na zamówienie można



Rys. 1. Przepływ informacji w archiwum cyfrowym (uproszczony)
Fig. 1. Information flow in the digital archive (simplified)

organizować tak, by minimalizować koszty energii. Efektywność energetyczna jest szczególnie ważna w przypadku repozytoriów opartych na pamięci dyskowej, gdyż stałe zasilanie dużego zespołu dysków, wraz z zarządzającymi nimi serwerami i chłodzeniem, jest bardzo kosztowne.

2.6. Standardy w archiwum cyfrowym

Tylko zgodność ze standardami może zapewnić długookresową możliwość poprawnej interpretacji zasobów zgromadzonych w archiwum cyfrowym. Jeśli bowiem archiwum nie korzysta z szeroko uznanych standardów, po dłuższym czasie wiedza o tym, jak poprawnie korzystać z jego zawartości, może zaginąć.

Standardy muszą dotyczyć zawartości archiwum, a zatem formatów przechowywanych zasobów, metadanych, organizacji danych itp. Także struktura archiwum i procedury jego działania powinny być zgodne ze standardami lub ogólnie przyjętymi dobrymi praktykami.

Najważniejszym powszechnie uznanym standardem normującym sposób działania archiwów cyfrowych, zarówno pod względem technicznym jak i organizacyjnym, jest model referencyjny dla archiwów cyfrowych *Open Archival Information System* (OAIS) [8], stanowiący normę ISO 14721:2012.

Sesje i pakiety archiwalne Zgodnie z zaleceniami OAIS i szeroko przyjętą dobrą praktyką, przetwarzanie danych w archiwum powinno odbywać się w tzw. sesjach archiwalnych.

W czasie sesji *Ingest* z danych dostarczonych przez producenta zasobu archiwalnego w postaci tzw. *Submission Information Package* (SIP) tworzony jest pakiet archiwalny *Archival Information Package* (AIP), który zostaje zapisany w archiwum. W sesji *Outgest* pakiet AIP jest pozyskiwany z repozytorium i przekształcany na tzw. *Dissemination Information Package* (DIP), który jest udostępniany odbiorcy. Uproszczony przepływ informacji w archiwum przedstawiono na rys. 1. Warto zwrócić uwagę na fakt, że o ile każdy pakiet AIP powstaje z jednego pakietu SIP, to wyjściowy pakiet DIP może zawierać treści pozyskane z wielu pakietów AIP, także pochodzących od różnych producentów.

Inne rodzaje sesji służą do wyszukiwania informacji oraz do czynności administracyjnych, np. badania poprawności przechowywanych pakietów.

2.7. Certyfikacja archiwum cyfrowego

By użytkownik chcący przechowywać swoje zasoby w archiwum cyfrowym mógł mieć do tego archiwum zaufanie, musi ono nie tylko spełniać typowe wymagania w zgodności z uznanymi standardami, ale ten fakt musi być potwierdzony przez niezależne zaufane instytucje. Istnieje zatem potrzeba certyfikacji archiwów cyfrowych. Wspomniany wyżej model referencyjny OAIS dostarcza terminologii i struktury logicznej wymagań, zaś zasady certyfikacji określa dokument *Audit and certification of trustworthy digital repositories* [15], stanowiący normę ISO 16363:2012.

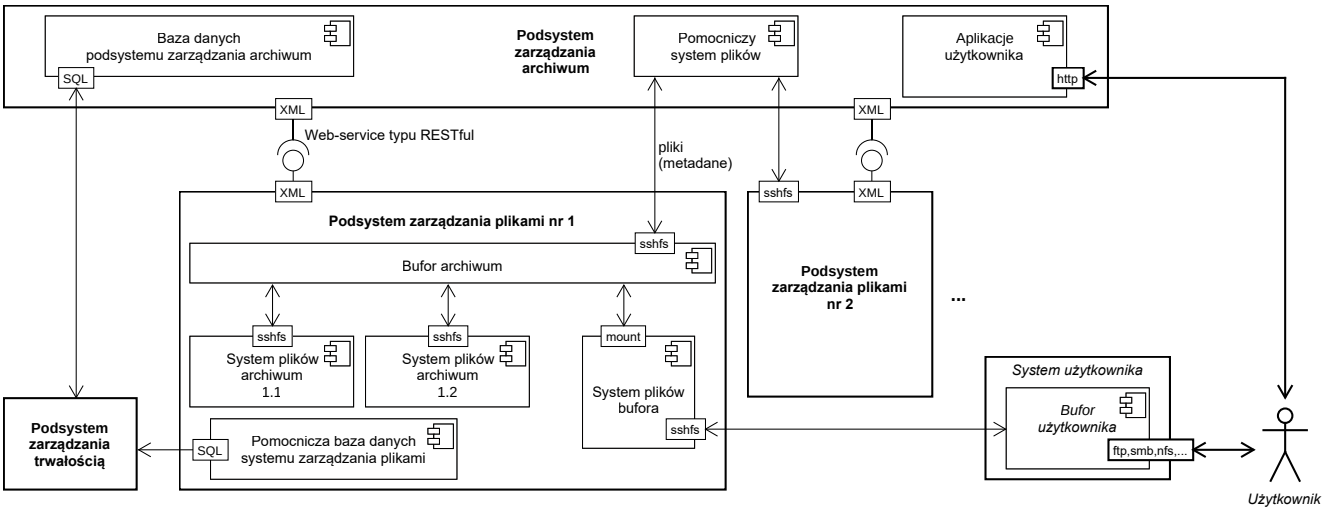
Co ciekawe, główne normy dotyczące przechowywania danych cyfrowych powstały pierwotnie na potrzeby badań kosmicznych. Być może powodem były szczególnie spektakularne przypadki utraty informacji, które zdarzyły się właśnie w tym środowisku, np. oryginalne taśmy z transmisją SSTV z lądowania Apollo 11 na Księżycu zostały wyczyszczone i użyte jako nośniki dla programu LANDSAT [23]!

3. Trwałe składowanie informacji cyfrowej

Głównym problemem w składowaniu informacji cyfrowej jest brak dostatecznie trwałych nośników danych cyfrowych. Istniejące nośniki, takie jak taśmy magnetyczne, dyski czy płyty CD, mają trwałość bardzo ograniczoną, od kilku do kilkudziesięciu lat, lecz by ją uzyskać, potrzebne są specjalne warunki przechowywania. Istnieją wprawdzie specjalne trwałe płyty M-DISK, dla których producent deklaruje około 1000 lat trwałości, dostępne są jednak w pojemności najwyżej BDXL, czyli 100 GB, co może wystarczać do archiwizacji zdjęć lub dokumentów, ale już nie do archiwizacji filmów czy produkcji telewizyjnej; póki co nie zdobyły też większej popularności. W przypadku najszerzej używanych nośników magnetycznych konieczne jest okresowe przepisywanie danych w celu odświeżenia zapisu, a także regularne „poruszanie" dyskami czy przewijanie taśm w celu uniknięcia problemów mechanicznych. Jest to kłopotliwe i kosztowne, więcej na ten temat napisano w części 3.1. Dość trwałym nośnikiem są pamięci *flash* w postaci kart pamięci, pendrive'ów czy też dysków SSD, ale ich cena jest ciągle wysoka.

Właściwie wszystkich typów pamięci cyfrowej dotyczy problem impulsu elektromagnetycznego, który może zniszczyć nośniki pamięci oraz czytniki, a zdarzyć się może np. na skutek zwiększonej aktywności Słońca.

Inny problem to zmienne technologie: stare nośniki mogą nie pasować do dostępnych napędów. Problem ten dotyka szczególnie taśm magnetycznych, także w najszerzej używanym stan-



Rys. 2. Architektura systemu CREDO
Fig. 2. Architecture of the CREDO system

dardzie LTO. W tym przypadku znormalizowano wprawdzie kasety od strony mechanicznej, ale standard zapisu zmienia się, zachowując wsteczną kompatybilność tylko w stosunku do kilku ostatnich wersji.

3.1. Trwałość nośników magnetycznych

Jednym z krytycznych wymagań przy projektowaniu archiwum jest zapewnienie, aby nośniki, na których będą przechowywane dane były bezawaryjne. Ponieważ jest to niemal niemożliwe, należy zapewnić odpowiednie monitorowanie ich stanu i wykorzystanie procedur służących do relokacji danych i wymiany nośników zagrożonych awarią.

Za realizację tych celów w CREDO odpowiada podsystem zarządzania trwałością (por. rys. 2). Jego zadaniem jest stworzenie wspólnego abstrakcyjnego mechanizmu służącego do dostarczania informacji na temat pojedynczego nośnika i zbiorów nośników, tzw. obszarów (najczęściej pochodzących z pojedynczego zakupu, od jednego producenta, a nawet z jednej serii produkcyjnej). Podsystem zarządzania trwałością daje podsystemowi zarządzania plikami wytyczne: z którego obszaru należy relokować pakiety, do którego obszaru należy relokować pakiety i które obszary należy uruchomić.

Szczegółowe rozwiązanie dotyczące wykorzystywanych w archiwum nośników powinno być dopasowane do ich charakterystyk i dostępnych technologii. Zrealizowane w demonstracyjnej wersji archiwum rozwiązanie wykorzystuje dyski twarde jako wówczas najbardziej rozpowszechnione nośniki pamięci masowej. Wykorzystano wytyczne NARA (ang. *U.S. National Archives and Record Administration*) [3] odnośnie archiwizowanych danych, teorie niezawodności i dane S.M.A.R.T. cyklicznie odczytywane z dysków.

3.2. Replikacja

Ponieważ nie dysponujemy technologiami dostatecznie niezawodnego długotrwałego przechowywania masowych danych cyfrowych, by wiarygodnie przechowywać zasoby cyfrowe, musimy uciec się do ich kopiowania i składowania wielu kopii. Na szczęście jedną z głównych zalet cyfrowej reprezentacji informacji jest możliwość wiernego kopiowania. Archiwum cyfrowe musi zatem przechowywać wiele kopii przechowywanych zasobów [20]. Kopie te powinny być oczywiście przechowywane na odrębnych nośnikach i okresowo weryfikowane.

Dyslokacja Niezawodność przechowywania znacząco poprawia dyslokacja, czyli rozproszenie lokalizacji przechowywania kopii. Uniezależnia funkcjonowanie archiwum od pojedynczego punktu awarii oraz chroni zasoby przed negatywnymi skutkami większości zdarzeń losowych. Ponieważ jest jedynym znanym sposobem zabezpieczenia zasobów przed skutkami katastrof i kataklizmów, należy uznać, że w profesjonalnym archiwum cyfrowym wykorzystanie dyslokacji zasobów jest niezbędne, a lokalizacje dyslokowanych kopii powinny być istotnie odległe i dobrze wybrane.

Wdrożenie repozytorium z dyslokacją może być zrealizowane nawet niskopoziomowo, w rozproszonym systemie plików. W przypadku archiwum cyfrowego zgodnego z wytycznymi OAIS lepszym rozwiązaniem wydaje się jednak wysokopoziomowe zarządzanie replikami całych pakietów archiwalnych.

Dywersyfikacja zapisu Dodatkowym postulatem, zwiększającym szanse poprawnego odczytania i interpretacji zasobu w odległej przyszłości, jest dywersyfikacja sposobu zapisu zasobów, tj. tzw. odrębność technologiczna kopii oraz zróżnicowanie formatów zapisu.

4. Metadane

Aby zasoby przechowywane w archiwum cyfrowym były użyteczne, zwłaszcza w dalekiej przyszłości, trzeba zapewnić możliwość sprawnego wyszukania informacji, weryfikacji jej autentyczności (ewentualnie stwierdzenia, jakie przechodziła przekształcenia) oraz jej poprawnej interpretacji, tak w sensie technicznym (format danych itd.) jak semantycznym (zrozumienie informacji w odpowiednim kontekście itp.). Zapewniają to metadane opisujące zarchiwizowane zasoby.

Przechowuje się metadane wielu rodzajów [25], m.in.:
- opisowe – identyfikujące i opisujące zasób, używane np. do wyszukiwania;
- techniczne – opisujące sposób utworzenia zasobu, niezbędne do jego prawidłowego odczytywania i interpretacji;
- strukturalne – opisujące strukturę złożonych (np. wieloczęściowych) obiektów cyfrowych;
- konserwatorskie – opisujące proces archiwizacji i przechowywania zasobu, np. jego weryfikację i przekształcenia (migracje);
- prawne – określające prawa do zasobu i zakres jego dozwolonego udostępniania,
- administracyjne – służące do zarządzania zasobem.

Zalecane jest, by metadane opisujące zasób dostarczał jego producent w postaci osobnych ustandaryzowanych plików (reko-

mendowanym formatem jest XML), ewentualnie w postaci tzw. metadanych zagłębionych (ang. *embedded metadata*, patrz [27]).

Wiele typowych formatów plików multimedialnych umożliwia umieszczanie metadanych zagłębionych. Takie możliwości mają np. najpopularniejsze formaty graficzne (TIFF, JPEG) i audio (MP3). Przechowuje się tak zarówno metadane techniczne, zapisywane automatycznie przez urządzenia produkujące dane zasoby (np. metadane EXIF [4] tworzone przez aparaty fotograficzne i niektóre skanery), jak i metadane opisowe, wpisywane przez ludzi – twórców danego zasobu (np. metadane IPTC [6] opisujące zdjęcia). Metadane zagłębione mają tę zaletę, że nie istnieje ryzyko ich zagubienia czy też przyporządkowania do niewłaściwego zasobu. Dlatego chętnie się je wykorzystuje w obiegu informacji, np. przy przesyłaniu zdjęć. W przypadku archiwizacji zaleca się jednak, by kopia wyodrębnionych metadanych zagłębionych była osobno zapisywana w postaci czytelnych plików (np. w XML), co umożliwi odczytanie tych metadanych bez znajomości formatu samego zasobu i bez konieczności użycia specjalizowanego oprogramowania oraz ich użycie w wyszukiwaniu zasobów.

Oprócz metadanych pozyskanych od producenta zasobu, archiwum powinno także przechowywać wyprodukowane przez siebie metadane opisujące proces archiwizacji i przechowywania zasobu. Takie metadane zmieniają się w czasie przechowywania, gdyż odnotowywane są nie tylko wszelkie zmiany w przechowywanym zasobie, ale także kontrole poprawności zasobu, a niekiedy nawet wszystkie dostępy do niego.

5. Budowa i działanie systemu CREDO

CREDO jest repozytorium cyfrowym mogącym pełnić funkcje repozytorium *on-line* oraz archiwum cyfrowego działającego zgodnie z wytycznymi OAIS, szczególnie jako długoterminowe archiwum głębokie. Z założenia repozytorium korzysta przede wszystkim z pamięci dyskowych, choć – dzięki jego otwartej i elastycznej architekturze – możliwe jest zastosowanie innych rodzajów pamięci. W wersji demonstracyjnej archiwum cześć pamięci zrealizowano z użyciem biblioteki taśm LTO.

Jednym z ważniejszych postulatów realizowanych przez CREDO jest dostosowanie do zmienności technologii, m.in. dzięki modularności, wymienności nośników i systemów plików oraz wymienności technologii i komponentów systemu.

5.1. Architektura CREDO

Archiwum CREDO jest zbudowane z wyraźnie rozdzielonych podsystemów o dobrze określonych zadaniach, komunikujących się przez klarownie określone interfejsy, co przedstawiono na rys. 2.

Podsystemy komunikują się wywołując wzajemnie swoje usługi lokalnie przez wystawione interfejsy programistyczne (API) lub zdalnie przez usługi sieciowe typu RESTful. Wymiana danych między podsystemami następuje przez bazę danych archiwum lub przez komunikaty XML w usługach RESTful. Architektura ta sprzyja rozbudowie systemu i wymienności komponentów, w szczególności systemów plików.

Transmisja plików między podsystemami archiwum odbywa się przez zdalne katalogi zamontowane za pomocą protokołu sshfs. Do transmisji plików między archiwum a klientami używane są typowe protokoły sieciowe (np. ftp, scp). Klient ma dostęp wyłącznie do wydzielonego, przeznaczonego dla niego systemu plików, nie ma zaś żadnego dostępu do systemów plików samego archiwum.

Podsystem zarządzania archiwum steruje działaniem archiwum oraz zarządza sesjami. Zawiera własną bazę danych, zrealizowaną w technologii Oracle, która przechowuje informacje potrzebne do sterowania archiwum oraz kopie wybranych metadanych umożliwiające wydajne wyszukiwanie zasobów. Pomocniczy system plików służy do przetwarzania metadanych.

W tym podsystemie zawarte są także aplikacje udostępniające archiwum użytkownikowi. W wersji demonstracyjnej CREDO zrealizowano je w technologii Oracle Application Express, ale możliwe jest ich rozbudowywanie w innych technologiach. Aplikacje komunikują się z logiką podsystemu za pomocą API wystawionego przez bazę danych.

Podsystem zarządzania trwałością zajmuje się diagnostyką stanu nośników archiwum oraz optymalizacja dostępu do systemów plików archiwum pod względem efektywności energetycznej.

Podsystemy zarządzania plikami przechowują właściwą zawartość archiwum. Każdy z takich systemów zawiera pewną liczbę systemów plików, w których przechowuje się zasoby (w wersji demonstracyjnej zrealizowano dwa systemy plików, w osobnych lokalizacjach). Ma też pomocniczą bazę danych rejestrującą parametry pracy nośników (np. parametry dysków S.M.A.R.T.), służącą do diagnostyki. Bufor archiwum, z osobnym systemem plików, służy do bezpiecznej wymiany plików między archiwum a podsystemem zarządzania archiwum i systemem użytkownika.

Archiwum może zawierać wiele podsystemów zarządzania plikami, także zrealizowanych w różnych technologiach. Umożliwia to stworzenie osobnych podsystemów dla szczególnie wymagających klientów, na przykład potrzebujących szczególnej ochrony zasobów lub specyficznej ich lokalizacji. Całe sterowanie odbywa się za pomocą usług sieciowych typu RESTful. Dzięki takiej architekturze podsystem zarządzania plikami może być fizycznie odrębny od reszty archiwum i geograficznie od niego odległy. Można też łatwo łączyć lub dzielić istniejące archiwa i przenosić całe podsystemy między archiwami, bez ich fizycznego kopiowania (skopiowania lub przeniesienia wymagają tylko metadane w bazie danych archiwum). Luźne połączenie podsystemów przez klarowne, standardowe interfejsy umożliwia łatwe dołączenie do archiwum nowych podsystemów plikowych, być może zbudowanych inaczej i w obecnie jeszcze nieistniejących technologiach.

5.2. Przetwarzanie zasobów w CREDO

Sesje archiwalne Przetwarzanie danych odbywa się, zgodnie z wytycznymi OAIS, w sesjach archiwalnych. Użytkownik steruje sesjami za pomocą udostępnionej przez CREDO aplikacji.

Ingest W czasie sesji *Ingest* pliki przeznaczone do archiwizacji umieszcza się w wydzielonym systemie plików użytkownika, skąd zostają pobrane przez oprogramowanie archiwum. Pliki te stanowią pakiet SIP (patrz 2.6). Zalecane jest umieszczenie w takim pakiecie nie tylko samych zasobów, ale i plików z opisującymi je metadanymi. Część metadanych można także wprowadzić do systemu za pomocą aplikacji.

Pliki pakietu SIP są kopiowane do bufora podsystemu zarządzania plikami. Odczytywane i analizowane są pliki z metadanymi. Pliki z archiwizowanymi zasobami są sprawdzane co do zgodności formatów z deklaracjami zawartymi w metadanych oraz co do zgodności skrótów cyfrowych, jeśli były one podane przez producenta zasobów. Jeśli pliki zawierają metadane zagłębione (patrz 4), są one wyodrębniane. Wybrane metadane są zapisywane do bazy danych archiwum. Wyliczane są skróty cyfrowe wszystkich plików; będą one używane do sprawdzania poprawności przechowywania. Do pakietu dołą-

czane są pliki z ustandaryzowanymi metadanymi opisowymi (zawierającymi także informacje wyodrębnione z metadanych zagłębionych) oraz z metadanymi konserwatorskimi, zawierającymi m.in. spis plików pakietu z ich skrótami cyfrowymi oraz informacje o procesie archiwizacji. Tak uzupełniony pakiet staje się pakietem AIP i zostaje skopiowany do właściwego systemu plików archiwum. Jeśli pakiet ma być przechowywany w kilku replikach wysokopoziomowych (zarządzanych przez archiwum), takie repliki są tworzone w odpowiednich systemach plików. Na koniec sprawdzana jest poprawność wszystkich plików w docelowych lokalizacjach i – jeśli wszystko jest w porządku – bufory są opróżniane, a sesja *Ingest* kończy się.

Search W sesjach *Search* użytkownik może wyszukiwać zasoby z archiwum korzystając z metadanych zgromadzonych w bazie danych archiwum. Sesje te w ogóle nie potrzebują dostępu do systemów plików przechowujących zasoby archiwum. Możliwe jest m.in. wyszukiwanie konkretnych fraz w standardowej strukturze metadanych opisowych (DCMES, patrz 6.9), a także zadawanie dowolnych zapytań w języku XQuery do ustandaryzowanej XML-owej reprezentacji metadanych.

Wyniki wyszukania mogą być zapisane i stanowić punkt wyjścia do kolejnych wyszukań lub materiał dla sesji *Outgest*.

Outgest Sesje *Outgest* buduje się na podstawie wyników sesji *Search*. Wyszukane pakiety AIP są pozyskiwane z archiwum przez ich skopiowanie do bufora. Tworzony jest dodatkowy plik metadanych, opisujący strukturę pozyskanych pakietów oraz proces ich pozyskania. Wszystkie te pliki łącznie tworzą pakiet DIP. Sprawdzana jest poprawność wszystkich plików pakietu, a następnie pakiet DIP jest kopiowany do systemu plików użytkownika, który może go pobrać za pomocą typowych protokołów plikowych.

Inne sesje archiwalne są prowadzone wewnętrznie przez archiwum, bez udziału użytkownika. Mają one charakter konserwatorski: w czasie ich trwania wykonywane są okresowe sprawdzenia poprawności przechowywania zasobów, a w razie potrzeby dokonywana jest migracja na lepsze (bardziej niezawodne lub tańsze w eksploatacji) albo nowsze nośniki.

Jeśli CREDO działa jako archiwum głębokie, które z założenia nie gwarantuje dostępu *on-line*, sesja archiwalna może trwać długo, nawet wiele dni. Sesja zainicjowana przez użytkownika nie wymaga oczywiście jego stałego udziału; aktualny stan sesji użytkownik może w każdej chwili sprawdzić za pomocą aplikacji. Jednak czas oczekiwania na zamówione przez użytkownika rezultaty może być znaczny. Wynika to głównie z optymalizacji zużycia energii przez archiwum (co opisano w części 6.6).

Nie dotyczy to wyszukiwania zasobów w sesjach *Search* – to jest zawsze szybkie, ponieważ taka sesja korzysta wyłącznie z metadanych zgromadzonych w bazie danych archiwum, a ta jest stale *on-line*.

6. CREDO a wymagania stawiane archiwom cyfrowym

Repozytorium CREDO spełnia wymagania techniczne stawiane archiwom cyfrowym, ma też mechanizmy ułatwiające spełnienie wymagań o charakterze prawno-organizacyjnym. Szczegóły opisano niżej.

6.1. Trwałość informacji cyfrowej

W obecnej wersji CREDO podstawowym nośnikiem danych są dyski magnetyczne. Zbudowano w tej technologii dwa repozytoria o objętości 1 PB. Gdy stanie się to ekonomicznie opłacalne, można będzie bez modyfikacji systemu użyć dysków SSD. Niewielki fragment repozytorium funkcjonuje – głównie do celów doświadczalnych – w oparciu o bibliotekę taśm LTO.

Budowa archiwum cyfrowego na pamięciach dyskowych ma ważne zalety w porównaniu najczęściej spotykanych do archiwów taśmowych:
- repozytorium może pełnić jednocześnie rolę szybkiego archiwum płytkiego (np. podręcznego) i archiwum głębokiego;
- nawet w przypadku archiwum głębokiego łatwo jest zapewnić sprawny dostęp do metadanych potrzebnych do wyszukiwania informacji oraz do zarządzania archiwum;
- weryfikacja poprawności zapisu oraz jego konserwacja, czyli okresowe poruszanie nośnikami oraz przepisywanie danych, nie nastręcza problemów technicznych ani organizacyjnych i jest szybka;
- sprawna i łatwa jest także migracja na nowe nośniki, np. w celu wymiany nośników zużytych.

Takie rozwiązanie ma też jednak wady:
- przechowywanie porównywalnej wielkości danych jest droższe niż w archiwach taśmowych; częściowo jest to jednak równoważone przez znacznie mniejsze koszty obsługi;
- trzeba rozwiązać problem kosztów energii, której zużycie przez działające *on-line* archiwum dyskowe jest znacznie większe niż w archiwach taśmowych (zastosowane rozwiązanie opisano w części 6.6).

Otwarta architektura CREDO pozwoli bez większych problemów użyć w przyszłości innych, nowych nośników i włączyć do CREDO oprogramowanie optymalizujące sposób ich użycia, np. inne metody zarządzania energią, inne algorytmy badania niezawodności czy dodatkowe zabezpieczenia. CREDO potrafi też automatycznie migrować dane na nowe nośniki.

Systemy plików w CREDO W obecnej wersji CREDO stosowany jest rozproszony system plików SZPAK, zbudowany na bazie otwartego systemu plików MooseFS [21]. Ten system plików pozwala na tworzenie niskopoziomowych replik, a nawet na ich dyslokacje. Zawiera też pewne potrzebne w archiwum mechanizmy pomocnicze, np. obliczanie sum kontrolnych plików.

Można jednak w CREDO użyć standardowych systemów plików. Repozytorium obsłuży bez większych problemów dowolny system plików zgodny z POSIX. Nie musi to nawet być rozwiązanie natywne danego systemu plików, zgodność z POSIX można bowiem uzyskać dzięki dodatkowej warstwie abstrakcji, np. *FUSE over FUSE*.

Relokacja w CREDO Repozytorium automatycznie wykonuje potrzebne relokacje danych, w tym automatyczną „ucieczkę" z nośników niepewnych lub oznaczonych przez operatora jako przestarzałe. Optymalizacja alokacji i relokacji następuje z uwzględnieniem danych statystycznych dotyczących awaryjności (patrz 6.2), mając za cel umieszczenie danych na najpewniejszych dostępnych nośnikach.

Replikacja w CREDO Zastosowano dwa poziomy replikacji. Replikacja niskopoziomowa wykonywana jest na poziomie systemu plików. Replikacja wysokopoziomowa jest zarządzana przez archiwum na poziomie replik całych pakietów archiwalnych. Kopie pakietów są binarnie identyczne, nie można zatem automatycznie zrealizować dywersyfikacji formatów. Możliwe jest natomiast tworzenie replik odrębnych technologicznie, np. na różnych systemach plików lub na różnych nośnikach (dyski + taśmy). Każda z replik wysokopoziomowych powinna być zapisana w wielu kopiach niskopoziomowych lub korzystać

z innych metod wspomagania niezawodności, np. kodów korekcyjnych.

Dyslokacja w CREDO Założono, że w ramach archiwum zasoby będą dyslokowane w co najmniej dwóch odległych lokalizacjach. Dyslokację zrealizowano jako replikację wysokopoziomową zarządzaną przez archiwum. Replikę pakietu archiwalnego można umieścić w konkretnym systemie plików. Systemy plików mieszczą się w odrębnych repozytoriach, znajdujących się w odległych od siebie lokalizacjach.

W planach rozwojowych CREDO przewidziano także możliwość dyslokacji w ramach federacji archiwów, ze wzajemną świadomością posiadania kopii i stanu ich poprawności oraz z koordynacją działań związanych z ryzykiem uszkodzenia kopii.

6.2. Monitorowanie trwałości informacji

Aby zapewnić trwałość archiwizowanych zasobów, należy stale monitorować zarówno je jak i sprzęt, na którym są one składowane.

W CREDO mamy do czynienia z regularnym dwupoziomowym monitorowaniem stanu zasobów archiwalnych. Na poziomie systemu plików są cyklicznie kontrolowane sumy kontrolne niskopoziomowych porcji informacji (tzw. *chunks*). Natomiast na poziomie archiwum również cyklicznie sprawdzana jest kompletność pakietów oraz poprawność skrótów cyfrowych dla poszczególnych plików należących do każdego pakietu. W obecnej implementacji wykorzystywana jest funkcja skrótu SHA-256, ale możliwe jest dostosowanie algorytmu służącego do obliczania skrótu do potrzeb, a także równoczesne wykorzystanie wielu standardów.

Należy pamiętać, że w archiwum głębokim nośniki są przez większość czasu wyłączone. Powoduje to potrzebę planowania, często z dużym wyprzedzeniem, operacji zarówno konserwacyjnych jak i tych związanych z sesjami *Ingest* czy *Outgest*, które zapisują lub odczytują odpowiednie pakiety. Na podstawie danych zapewnianych przez podsystem zarządzania trwałością jest obliczane prawdopodobieństwo awarii dla każdego nośnika i, kiedy przekroczy ono progową wartość, nośnik jest wprowadzany do harmonogramu operacji monitorowania. Także całe obszary mają zagregowaną miarę prawdopodobieństwa awarii. Ponadto dane te są wykorzystywane do określenia czy dany nośnik należy wyznaczyć jako cel relokacji dla potencjalnych pakietów, czy też należy określić nośnik jako zagrożony awarią i zacząć planować przeniesienie pakietów, które się na nim znajdują.

Zapewnienie trwałości na poziomie nośników polega na ich przemagnesowaniu (w wypadku nośników magnetycznych), przewinięciu (taśmy LTO), czy też użyciu innych metod, specyficznych dla danego sprzętu.

6.3. Weryfikowalność przechowywania

CREDO zapewnia weryfikację zarówno integralności jak i autentyczności zapisanych w repozytorium zasobów. Okresowe sprawdzenia są wykonywane automatycznie.

Integralność zasobów można sprawdzić dzięki temu, że zasoby są opatrzone metadanymi, a dodatkowo kopia wybranych metadanych jest przechowywana osobno w bazie danych archiwum. Weryfikacja integralności obejmuje sprawdzenie kompletności pakietów oraz niezmienności zapisu na podstawie zawartych w metadanych skrótów cyfrowych.

Autentyczność zasobów może być zweryfikowana na podstawie metadanych. Ponieważ metadane są zapisane w XML, czyli w formacie otwartym i samodokumentującym, poprawna ich interpretacja będzie możliwa nawet po wielu latach. Kopie metadanych w bazie danych archiwum są z kolei zapisane w elastycznych strukturach, które pozwalają na zapis metadanych w różnych standardach, także jeszcze nieistniejących. Wymagane przez archiwum metadane umożliwiają zaś kontrolę zgodności zawartości pakietu oraz formatu plików deklaracjami.

Co do niezaprzeczalności, to do jej zapewnienia potrzebna jest infrastruktura podpisu cyfrowego, a to wymaga trwałego istnienia odpowiedniego łańcucha instytucji certyfikujących. Samo archiwum oczywiście nie może tego zapewnić, może jedynie przechowywać odpowiednie certyfikaty. Trzeba jednak pamiętać, że w kontekście przechowywania wieczystego możliwość zagwarantowania trwałości instytucji certyfikujących jest bardzo wątpliwa.

6.4. Dostępność informacji

Wyszukiwanie zasobów w CREDO jest wykonywane efektywnie dzięki kopiom kluczowych metadanych przechowywanym w bazie danych archiwum. Ponieważ jest to wysokiej klasy relacyjna baza danych Oracle, zapytania są w niej wykonywane z wysoką wydajnością. W tej bazie danych przechowywane są między innymi metadane opisowe, zrzutowane do standardu Dublin Core [9] i zapisane w strukturze relacyjnej, oraz wskazane metadane w XML. Baza danych przechowuje też różnorodne identyfikatory zasobów (DOI, URI itp.), których można użyć do wyszukiwania. Oracle oferuje różne mechanizmy wyszukiwania w metadanych: zapytania do danych relacyjnych w SQL, wyszukiwanie w XML za pomocą zapytań w języku XQuery oraz wyszukiwanie pełnotekstowe. Baza danych archiwum jest stale dostępna *on-line*, a przeszukiwanie zapisanych w niej metadanych nie wymaga dostępu do głównego systemu plików archiwum, nie powoduje zatem dodatkowych kosztów energii związanych z takim dostępem.

Korzystanie z archiwum CREDO ułatwia jego organizacja logiczna, odpowiadająca organizacji klasycznych archiwów. Pakiety archiwalne zapisywane w archiwum są podzielone na tzw. zespoły archiwalne. Każdy zespół archiwalny ma swojego właściciela.

Czas dostępu do odnalezionego zasobu zależy od trybu pracy CREDO. Jeśli system lub odpowiednia jego część pracuje jako repozytorium *on-line* czy archiwum płytkie, dostęp do zasobu jest szybki, ponieważ zasoby są składowane na dyskach. Czas dostępu odpowiada wówczas praktycznie czasowi dwukrotnego kopiowania zasobu: z systemu plików archiwum do bufora dostępowego oraz z tego bufora na nośnik użytkownika. Jeśli mamy do czynienia z archiwum głębokim, czas dostępu zależy od polityki zarządzania energią i może być długi, liczony nawet w dniach czy tygodniach. System optymalizuje bowiem dostęp tak, by możliwie rzadko włączać zasilanie zespołów dysków.

Poprawność odczytu i interpretacji zasobów można zapewnić przechowując je wyłącznie w odpowiednich formatach, szeroko używanych i znormalizowanych. Archiwum rekomenduje użycie takich formatów, a próba zapisu danych w formatach niezalecanych wywołuje odpowiednie ostrzeżenia. W przypadku formatów rekomendowanych, archiwum CREDO przechowuje ich specyfikacje jako chronione zasoby systemowe.

Dodatkowe informacje potrzebne do interpretacji zasobu, np. opisowe czy techniczne, mogą być pozyskane z metadanych przechowywanych w pakiecie archiwalnym wraz z zasobem. Szczegóły opisano w części 6.9.

6.5. Poufność informacji

Repozytorium CREDO zapewnia poufność powierzonej mu do przechowania informacji. Ochrona fizyczna i zabezpieczenia techniczne serwerowni są zgodne z najwyższymi standardami przemysłowymi, co wynika ze specyfiki podstawowej działalności PWPW – lidera projektu. Dostęp do interfejsów systemu CREDO jest możliwy wyłącznie w chronionej sieci VPN. Użytkownicy systemu nigdy nie mają bezpośredniego dostępu do systemu plików archiwum. System plików archiwum jest też

chroniony przed nieprawidłowymi działaniami samego oprogramowania archiwum: wydzielony podsystem bezpieczeństwa uprawnia programy CREDO do operowania na plikach archiwum tylko w niezbędnym zakresie i na niezbędny czas.

Z zasady repozytorium udostępnia zasoby jedynie ich właścicielowi oraz użytkownikom przez niego upoważnionym. Dla zasobów wymagających szczególnych zabezpieczeń można w repozytorium stworzyć osobne systemy plików, podlegające szczególnej ochronie, np. fizycznie umieszczone w specjalnych odrębnych lokalizacjach.

6.6. Efektywność energetyczna archiwum

Długi okres przechowywania danych w archiwum CREDO narzuca szczególnie ostre wymagania dotyczące zużycia energii. Dostęp do danych wymaga uruchomienia odpowiedniego nośnika lub załadowania kasety z taśmą do czytnika. Kluczowe jest takie zarządzanie przechowywaniem informacji, dostępem do niej i działaniami konserwatorskimi aby, uwzględniając bezpieczeństwo przechowywania i dostępu, brać pod uwagę całkowity koszt działania archiwum, w tym koszt zużycia energii.

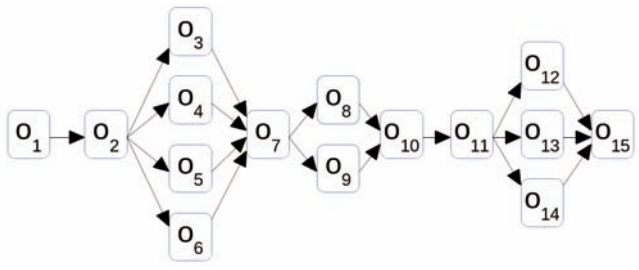
Opracowanie odpowiednich algorytmów zarządzania wymaga właściwej identyfikacji źródeł kosztów i ryzyk związanych z bezpieczeństwem. Koszty mieszczą się w jednej z dwóch kategorii: obsługi bieżącej (w tym energii) oraz zużycia sprzętu.

W przypadku składowania danych na klasycznych dyskach twardych (HDD) koszt energii zużytej w trakcie całego okresu użytkowania dysku (kilka lat) jest porównywalny do kosztu zakupu dysku, ale przy założeniu, że dysk jest cały czas aktywny. Koszt ten spada znacząco jeśli w okresie bezczynności dysk jest na pewien czas wyłączany. Koszt energii zużywanej na ponowne włączanie dysku jest zaniedbywalny.

Składowanie taśm magnetycznych wymaga znacznie mniejszego zużycia energii, związanego głównie z operacjami odczytu lub zapisu oraz klimatyzacją magazynów.

Koszt zużycia sprzętu wynika wprost z kosztu zakupu i czasu użytkowania. Typowy dysk twardy pracujący w trybie ciągłym ma deklarowany średni czas między awariami (MTBF) na poziomie 500 do 1000 tys. godzin, jednak rzeczywiste dane [16] wskazują na prawdopodobieństwo awarii w ciągu roku na poziomie 1% do 10%, zależnie od modelu dysku. Również liczba uruchomień dysku jest ograniczona i w przypadku większości napędów oscyluje wokół 300 tys. Nie powinno to jednak stanowić większego problemu, o ile dostęp do danych zostanie rozsądnie zaplanowany. Niestety, brak jest danych opisujących trwałość nośnika dyskowego w scenariuszach zakładających jego okresowe wyłączanie.

W przypadku taśm magnetycznych producenci deklarują trwałość na poziomie 30 lat, jednak w praktycznych zastosowaniach spada ona do około 10 lat, a chęć zapewnienia maksymalnego bezpieczeństwa danych powoduje, że taśmy nie powinny być używane dłużej niż 4 lata. Najbardziej ograniczającą cechą taśm magnetycznych w zastosowaniach archiwalnych jest niewielka maksymalna liczba przewinięć taśmy, czyli w praktyce liczba operacji zapisu/odczytu. Wynika ona z fizycznej degradacji nośnika, skutkującej częstszymi błędami. Przyjmuje się, że liczba przewinięć taśmy nie powinna przekraczać 150. Kolejnym ograniczeniem tego medium jest maksymalna liczba taśm, które mogą być jednocześnie odczytywane/zapisywane. Jest to równoważne liczbie czytników zainstalowanych w systemie taśmowym. Generalna zasada jest, że koszt zakupu sprzętu i nośników taśmowych jest znacząco większy od kosztu użytkowania, a zwłaszcza od kosztu energii.



Rys. 3. Przykładowa procedura
Fig. 3. Sample procedure

Znając szczegółowe charakterystyki użycia energii i zużycia sprzętu można prawidłowo zaprojektować algorytmy zarządzające archiwum długoterminowym, zapewniające bezpieczeństwo i niskokosztowe zarządzanie danymi. Faktyczne zadanie, które zostało postawione przed projektantami CREDO, może zostać przedstawione następująco. Dany jest zbiór planowanych operacji, np. odczytu, zapisu czy prac konserwatorskich. Operacje te są pogrupowane w procedury. Każda procedura jest sekwencją zbiorów operacji wykonywanych równolegle (rys. 3). Dana operacja ma zdefiniowany przedział czasowy, w którym musi się rozpocząć i zakończyć. Chwila rozpoczęcia operacji może być narzucona lub pozostawiona do decyzji algorytmowi zarządzania/harmonogramowania. Niektóre z operacji wymagają dostępu do obszarów przechowywania danych. W tabeli 2 wyszczególniono wszystkie istotne parametry operacji.

Tab. 2. Podstawowe parametry operacji
Tab. 2. Basic parameters of an operation

Pred	Zbiór operacji poprzedzających (w ramach procedury)
Odst	Obowiązkowy odstęp czasu po poprzedniej operacji
T^e, T^l	Przedział czasu, w którym operacja może być wykonywana
Src	Zbiór obszarów do odczytu (tylko jeden zostanie wybrany)
$[T_o^r]$	Czas odczytu z każdego obszaru $o \in$ Src
Dst	Zbiór obszarów do zapisu (tylko jeden zostanie wybrany)
$[T_o^w]$	Czas zapisu na każdy z obszarów $o \in$ Dst
$[S_o^w]$	Rozmiar zapisu na każdy z obszarów $o \in$ Dst (może być różny)
Lock	Rodzaj blokady zasobu (obszaru): brak, tylko zapis, pełna
T^s	Predefiniowany czas startu (nie musi być ustawiony)

Tab. 3. Podstawowe parametry obszaru
Tab. 3. Basic parameters of an area

Size	Rozmiar obszaru
Used	Zajęty rozmiar (bajty)
R, W	Maksymalna liczba operacji odczytu i zapisu wykonywanych jednocześnie
T^{off}	Minimalny czas bezczynności skutkujący wyłączeniem
C^{up}	Koszt uruchomienia obszaru
C^{on}	Jednostkowy koszt działania obszaru
Rel	Współczynnik niezawodności

Obszar przechowywania danych jest najmniejszą niepodzielną częścią archiwum o znanej pojemności, która może zostać tymczasowo wyłączona. Obszar może składać się z pewnej liczby nośników danych, które z punktu widzenia algorytmu harmonogramowania dostępu są nierozróżnialne. Parametry tych nośników mogą jednak mieć wpływ na niektóre zagregowane parametry obszaru, jak np. trwałość i pewność przechowywanych informacji.

Obszary stanowią zasoby krytyczne. W szczególności mają ograniczony rozmiar oraz liczbę równoległych odczytów/zapisów. Dodatkowo, niektóre operacje wymagają dostępu do obszarów na zasadzie wyłączności. Każdy obszar ma przypisane parametry niezbędne dla wyznaczenia prawidłowego harmonogramu operacji, takie jak koszt uruchomienia obszaru, koszt działania, minimalny czas bezczynności pozwalający wyłączyć obszar czy współczynnik niezawodności danych. Najważniejsze parametry obszarów zostały zebrane w tabeli 3.

Dla każdej operacji jest zdefiniowany zbiór obszarów, z których jeden zostanie wybrany do odczytu, i zbiór obszarów, z których jeden zostanie wybrany do zapisu. Ostateczny harmonogram operacji zawiera informacje nie tylko o tym, w jakim momencie dana operacja się rozpoczyna i kończy, ale także o tym, do jakich obszarów uzyskuje dostęp.

Moduł zarządzania, a w nim algorytm harmonogramowania, dostarcza zoptymalizowanego harmonogramu spełniającego kilka kryteriów/celów: minimalizacja całkowitego kosztu użycia archiwum w horyzoncie czasowym, maksymalizacja bezpieczeństwa i spójności składowanych danych, równoważenie zajętości obszarów, równoważenie operacji zapisu/odczytu między obszarami.

Algorytm bazuje na rozbudowanej heurystyce konstrukcyjnej. Harmonogram jest tworzony sekwencyjne dla każdej nowo pojawiającej się operacji, która wymaga zaplanowania. Spośród możliwych rozwiązań jest wybierane takie, dla którego koszt krańcowy wartości funkcji celu jest najmniejszy (najmniejszy wzrost wartości funkcji celu). Wartość funkcji celu jest przy tym agregacją kryteriów dokonaną z uwzględnieniem preferencji decydenta. Uwzględniane są jednak tylko te rozwiązania, dla których nie są przekroczone żadne ograniczenia, takie jak kolejność wykonania w ramach procedury, pojemność obszaru, maksymalna liczba równoległych operacji odczytu/zapisu w obszarze itp.

Opracowany algorytm jest wysoce parametryzowalny, a przy tym efektywny zarówno ze względu na czas wyznaczania harmonogramu, jak i na jego jakość. Bierze pod uwagę wiele rzeczywistych ograniczeń i wymagań, a decydentom pozwala wyrażać różne preferencje dotyczące bezpieczeństwa, niezawodności czy efektywności energetycznej.

6.7. Standardy w CREDO

Tworząc repozytorium CREDO starano się w maksymalny sposób wykorzystać istniejące normy. Zapewniono więc zgodność „filozofii" i działania systemu ze standardem OAIS. Samo repozytorium zapewnia oczywiście jedynie zgodność techniczną ze standardem. Standard obejmuje też zagadnienia prawno-organizacyjne, których nie można zrealizować technicznie, lecz powinny być zapewnione przez instytucję zarządzającą archiwum. W przypadku metadanych tworzonych i przechowywanych przez archiwum użyto standardowych formatów (patrz punkt 6.9). Techniczna konstrukcja repozytorium także gdzie to możliwe wykorzystuje standardy, m.in. POSIX [14], FUSE [11] oraz wiele standardów związanych z XML.

Producentom archiwizowanych zasobów zaleca się korzystanie z formatów zasobów oraz metadanych zgodnych z otwartymi i powszechnie uznanymi standardami. Użycie takich formatów zapewni poprawną interpretację zasobów także w odległej przyszłości. Archiwum CREDO rekomenduje stosowanie właściwych formatów oraz ostrzega w przypadku użycia niezalecanych. Dokumentacja użytych formatów powinna być dostępna w archiwum i powiązana z zasobami. CREDO ma wspierające to mechanizmy.

6.8. Certyfikacja archiwum

Budując repozytorium CREDO założono, że korzystające z niego archiwum musi być zdatne do certyfikacji. Jest to możliwe dzięki zgodności z modelem referencyjnym OAIS i przejrzystej architekturze systemu z dobrze określonym podziałem zadań. Zapewniono niezbędne do certyfikacji szczegółowe rejestrowanie wszelkich zdarzeń w archiwum w dziennikach (logach). Ponieważ nie ma krajowych instytucji certyfikujących, całą dokumentację systemu przygotowano w języku angielskim.

6.9. Metadane w CREDO

Archiwum CREDO umożliwia przechowywanie wszelkiego rodzaju metadanych. Zapewnia też specjalne przetwarzanie metadanych konserwatorskich oraz wybranych metadanych opisowych i technicznych.

W pakiecie SIP można zawrzeć manifest w formacie METS [18], w którym deklaruje się m.in. dostarczone pliki, ich skróty cyfrowe i formaty; mogą tam także znajdować się metadane opisowe. Jeśli manifestu nie dostarczono, archiwum tworzy listę plików, którą użytkownik weryfikuje i uzupełnia, np. o deklaracje formatów.

W czasie sesji *Ingest* archiwum weryfikuje zgodność dostarczonego pakietu z taką deklaracją. Do sprawdzenia formatu pliku i jego wersji nie tylko na podstawie rozszerzenia, ale także na podstawie zawartości, zastosowano narzędzie DROID [5]. O ile dany format na to pozwala, archiwum odczytuje z pliku metadane zagłębione za pomocą narzędzia Apache Tika [1]. Metadane można też wprowadzić lub uzupełnić ręcznie przy pomocy aplikacji archiwum.

Wybrane metadane użyteczne do wyszukiwania informacji, głównie opisowe i techniczne, są zapisywane w bazie danych archiwum. Metadane o prostej budowie klucz-wartość są zapisywane w strukturze relacyjnej, zapewniającej bardzo efektywne wyszukiwanie. Zastosowano tu elastyczną strukturę generyczną, co umożliwia przechowywanie metadanych pochodzących z różnych standardów i łatwe uwzględnienie standardów nowych. Metadane bardziej złożone mogą być zapisane w formacie XML; takie metadane mogą być przeszukiwane za pomocą zapytań w języku XQuery. Ponieważ baza danych zapisuje dokumenty XML nie jako tekst, ale w postaci tzw. drzew DOM, przeszukiwanie takie może także być dość wydajne. W bazie danych przechowywane są także metadane konserwatorskie, rejestrujące wszystkie procedury wykonywane na zasobach przez archiwum.

Wszystkie pliki metadanych dostarczone oryginalnie w pakiecie SIP, są bez zmian zapisywane w pakiecie archiwalnym (AIP). Archiwum dodaje także własny plik manifestu w formacie METS, definiujący zawartość pakietu AIP, oraz plik metadanych konserwatorskich w formacie PREMIS [19], opisujący proces archiwizacji. Po każdym działaniu na pakiecie AIP, np. po okresowym sprawdzeniu poprawności pakietu, plik PREMIS jest wymieniany na nową wersję, uzupełnioną o opisy ostatnich czynności.

Przechowywanie metadanych w oryginalnej formie dostarczonej przez producenta zasobów jest potrzebne ze względu na zachowanie oryginalności informacji i poprawną interpretację zasobów. Nie sprzyja jednak efektywnemu wyszukiwaniu informacji, nie pozwala bowiem na formułowanie prostych i ujednoliconych kryteriów wyszukiwania. Dlatego w CREDO

wprowadzono unifikację metadanych opisowych, które najczęściej wykorzystuje się do wyszukiwania, przez rzutowanie dostarczonych metadanych na powszechnie używany standard *Dublin Core Metadata Element Set* [10]. Sposób rzutowania, bazujący na wyszukiwaniu w XML-owej reprezentacji metadanych za pomocą języka XQuery, jest nadspodziewanie prosty, elastyczny i łatwy do rozszerzenia.

W czasie sesji *Outgest* tworzony jest pakiet DIP, który może zawierać zasoby z wielu pakietów archiwalnych. Dlatego oprócz metadanych zawartych w pakietach AIP archiwum dodaje do pakietu DIP dodatkowy plik manifestu w formacie METS, definiujący zawartość tego pakietu, oraz plik metadanych konserwatorskich w formacie PREMIS, opisujący czynności, jakie na dostarczanych zasobach zostały wykonane przez archiwum.

Zrealizowane funkcjonalności dotyczące metadanych umożliwiają m.in. weryfikację poprawności przechowywania pakietów archiwalnych, wydajne wyszukiwanie zasobów według zróżnicowanych kryteriów oraz możliwość kontroli wszystkich operacji wykonywanych przez archiwum na przechowywanych zasobach. Dzięki temu możliwe jest spełnienie wymagań stawianych archiwom cyfrowym, m.in. co do trwałości, weryfikowalności, integralności, autentyczności i dostępności informacji przechowywanej w archiwum.

7. Podsumowanie

Długoterminowa archiwizacja zasobów cyfrowych staje się coraz ważniejszym, choć słabo uświadomionym, problemem naszej „cywilizacji cyfrowej". Nie ma powszechnie uznanych i dostępnych rozwiązań, które przy rozsądnych kosztach pozwoliłyby przedsiębiorstwom, urzędom czy twórcom medialnym bezpiecznie przechowywać tworzone zasoby cyfrowe. Niezwykle szybki i zwykle cieszący nas rozwój technologii cyfrowych w kontekście przechowywania długoterminowego stanowi raczej źródło problemów, gdyż sposoby zapisu informacji (rozwiązania sprzętowe oraz formaty danych) zmieniają się bardzo szybko i już nawet po kilku latach może być bardzo trudno odczytać zasoby przechowywane na przestarzałych nośnikach czy w formatach, które wyszły z użycia. Tymczasem chcielibyśmy móc przechowywać informacje przez kilkadziesiąt czy nawet kilkaset lat, zachowując gwarancję możliwości ich odczytania i poprawnej interpretacji.

Ponieważ jednak problem istnieje od lat, dopracowano się przynajmniej zbioru zasad i standardów, które powinny być uwzględnione przy przechowywaniu informacji cyfrowej. W tym artykule starano się przedstawić główne problemy oraz najważniejsze z owych zasad i standardów.

Tekst oparto na doświadczeniach zdobytych przy tworzeniu Cyfrowego Repozytorium Dokumentów CREDO, które powstało jako tzw. demonstrator, czyli rodzaj rozwiniętego prototypu, mającego stanowić *proof-of-concept* dla zaproponowanej technologii.

Na podstawie doświadczeń zdobytych przy tworzeniu systemu CREDO sformułować można następujące wnioski.

– Istniejące i powszechnie uznane zasady i standardy dotyczące przechowywania oraz archiwizacji zasobów cyfrowych wydają się stanowić wystarczającą podstawę do budowy archiwów cyfrowych, w tym archiwów długoterminowych.
– Szeroko stosowane formaty plików w niewielkim stopniu odpowiadają potrzebom archiwizacji długoterminowej. Zastrzeżeń nie można mieć właściwie tylko do prostych plików tekstowych, do dokumentów w XML, o ile mają prostą strukturę lub istnieje dokumentacja tej struktury, oraz do formatu PDF/A, specjalnie dostosowanego do celów archiwizacji. Powszechne użycie dokumentów w formatach prawnie zastrzeżonych (ang. *proprietary*), w dodatku na ogół szybko i nie zawsze w dobrze kontrolowany sposób ewoluujących, stanowi duży problem w kontekście archiwizacji.
– Istniejące nośniki danych cyfrowych mają zdecydowanie zbyt małą trwałość w stosunku do oczekiwań, zwłaszcza związanych z archiwizacją długoterminową. W dodatku niemal żadne nośniki nie są odporne na impuls elektromagnetyczny. Powstały wprawdzie „kamienne" płyty optyczne o potencjalnie wielusetletniej trwałości, ale ich pojemności są mizerne wobec potrzeb, zwłaszcza w kontekście przechowywania multimediów, np. produkcji telewizyjnej czy filmowej. Niezbędnym sposobem zabezpieczenia zasobów cyfrowych jest więc ich wielokrotne kopiowanie i dyslokacja.
– Obecnie stosowane technologie podpisywania dokumentów cyfrowych, które mogą także służyć do zapewnienia niezaprzeczalności zasobów, bazują na tzw. infrastrukturze klucza publicznego. Ta zaś zależy od istnienia zaufanych instytucji certyfikujących. W przypadku archiwów długoterminowych to rozwiązanie nie sprawdzi się, gdyż trudno od instytucji, zwykle komercyjnych, oczekiwać wieczystego trwania.
– Choć większość archiwów cyfrowych wykorzystuje taśmy LTO ze względu na stosunkowo niski koszt samych nośników i ich utrzymania, archiwum bazujące na dyskach okazało się mieć wiele zalet, w tym możliwość łatwego wykorzystania jako repozytorium dostępnego *on-line*, nieporównanie łatwiejsze prowadzenie czynności konserwatorskich (okresowe sprawdzenia i odświeżanie zapisu, migracje itp.) oraz uniknięcie problemów z kompatybilnością nowych nośników ze starymi napędami. Koszty budowy takiego rozwiązania są wprawdzie wyższe, ale – dzięki optymalizacji zużycia energii – koszty eksploatacji mogą być porównywalne lub nawet niższe.
– Dominująca obecnie tendencja, by metadane zasobów cyfrowych zapisywać w specjalnych dialektach XML, wydaje się bardzo korzystna. Dobrze skonstruowane dokumenty XML są samoopisujące, zatem nawet w bardzo odległej przyszłości mogą być poprawnie interpretowane. Przetwarzanie XML jest relatywnie łatwe, a narzędzia temu służące są rozwinięte i dostępne, co pozwala wygodnie i efektywnie przeszukiwać, przetwarzać i wytwarzać metadane składowanych w archiwum zasobów.
– Tworzenie archiwum cyfrowego dla instytucji, która ma zamiar nie tyle sama z niego korzystać, ile wynajmować przestrzeń w archiwum innym podmiotom, nie okazało się pomysłem szczęśliwym. Lepiej byłoby, jak się wydaje, tworzyć archiwum bezpośrednio dla instytucji mającej w nim przechowywać swoje zasoby. W takim przypadku zaangażowanie przyszłego właściciela archiwum byłoby zapewne znacznie większe, łatwiejsze byłoby także uzyskanie informacji o rzeczywistych potrzebach przyszłych użytkowników systemu.
– Nie sprawdził się pomysł, by wytworzony w wyniku prac badawczo-rozwojowych projekt przechodził na własność partnera przemysłowego. Takie rozwiązanie uniemożliwia bowiem samodzielne kontynuowanie prac badawczych przez partnera naukowego, zaś partner przemysłowy może nie być zainteresowany dalszymi badaniami po zakończeniu ich finansowania przez państwowego sponsora.

Mimo opisanych trudności i problemów, projekt CREDO dowiódł, że stworzenie spełniającego ogólnie przyjęte wymagania cyfrowego archiwum długoterminowego w oparciu o pamięć dyskową i uznane technologie informatyczne jest możliwe przy zaangażowaniu rozsądnych środków i w stosunkowo krótkim czasie.

Podziękowania

Projekt *Cyfrowe repozytorium dokumentów – CREDO* realizowany był w latach 2013–2016 w ramach przedsięwzięcia pilotażowego NCBiR „Wsparcie badań naukowych i prac rozwojowych w skali demonstracyjnej DEMONSTRATOR+" oraz był współfinansowany z działania 1.5 POIG. Umowa nr UOD-DEM-1-385/001.

Bibliografia

1. Apache Tika. http://tika.apache.org. Dostęp: 2020-09-02.
2. ARMA international – Association of Records Managers and Administrators. https://www.arma.org/. Dostęp: 2020-07-15.
3. National Archives and Records Administration (NARA). http://www.archives.gov/. Dostęp: 2020-00-29.
4. Exchangeable image file format for digital still cameras: Exif version 2.3. http://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf, 2012. Dostęp: 2020-08-11.
5. DROID: file format identification tool. https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/, 2013. Dostęp: 2020-08-11.
6. IPTC photo metadata. http://www.iptc.org/site/Photo_Metadata/, 2014. Dostęp: 2020-08-11.
7. ARMA International. Generally accepted recordkeeping principles. https://www.arma.org/page/principles. Dostęp: 2020-07-15.
8. Consultative Committee for Space Data Systems. Reference model for an open archival information system (OAIS). Recommended practice. https://public.ccsds.org/pubs/650x0m2.pdf, June 2012. Dostęp: 2020-08-11.
9. Dublin Core Metadata Initiative. http://dublincore.org/. Dostęp: 2020-08-11.
10. Dublin Core Metadata Initiative. Dublin core metadata element set, version 1.1. http://dublincore.org/documents/dces, 2012. Dostęp: 2020-08-11.
11. Filesystem in Userspace (FUSE). https://github.com/libfuse/libfuse. Dostęp: 2020-08-20.
12. Ghosh P., *Google's Vint Cerf warns of 'digital Dark Age'*. http://www.bbc.com/news/science-environment-31450389, Luty 2015. BBC News. Dostęp: 2020-08-11.
13. Huhnlein D., Korte U., Langer L., Wiesmaier A., *A comprehensive reference architecture for trustworthy long-term archiving of sensitive data*. 3rd International Conference on New Technologies, Mobility and Security, 2009, 1–5, IEEE, DOI: 10.1109/NTMS.2009.5384830.
14. POSIX.1-2017. The Open Group Base Specifications Issue 7. http://pubs.opengroup.org/onlinepubs/9699919799, 2018. Dostęp: 2020-08-20.
15. International Standard Organization. Space data and information transfer systems – audit and certification of trustworthy digital repositories ISO 16363:2012.
16. Klein A., One billion drive hours and counting: Q1 2016 hard drive stats. http://www.backblaze.com/blog/hard-drive-reliability-stats-q1-2016. Dostęp: 2020-09-29.
17. Lemieux V.L., *Evaluating the use of blockchain in land transactions: An archival science perspective*. "European Property Law Journal", Vol. 6, No. 3, 2017, 392–440, DOI: 10.1515/eplj-2017-0019.
18. Library of Congress. Metadata encoding & transmission standard. http://www.loc.gov/standards/mets. Dostęp: 2020-08-11.
19. Library of Congress. PREMIS preservation metadata maintenance activity. http://www.loc.gov/standards/premis. Dostęp: 2020-09-29.
20. Marasek K., Walczak J., Traczyk T., Płoszajski G., Kazmierski A., *Koncepcja elektronicznego archiwum wieczystego*. „Studia Informatica", T. 30, Nr 2B, 2009, 275–307.
21. MooseFS. http://moosefs.com. Dostęp: 2020-08-07.
22. Narodowe Centrum Badań i Rozwoju. Demonstrator+ Wsparcie badań naukowych i prac rozwojowych w skali demonstracyjnej. https://www.ncbr.gov.pl/programy/programy-krajowe/demonstrator-wsparcie-badan-naukowych-i-prac-rozwojowych-w-skali-demonstracyjnej. Dostęp: 2020-08-11.
23. National Aeronautics and Space Administration. The Apollo 11 telemetry data recordings: A final report. https://www.nasa.gov/pdf/398311main_Apollo_11_Report.pdf. Dostęp: 2020-08-02.
24. Pałka P., Śliwiński T., Traczyk T., Ogryczak W., *Persistence management in digital document repository*. Koziel-ski S. i in., redaktorzy, *Advanced Technologies for Data Mining and Knowledge Discovery: 12th International Conference BDAS, Ustroń, Poland*, 2016, 668–682. Springer International Publishing, DOI: 10.1007/978-3-319-34099-9_52.
25. Płoszajski G. (ed.), *Standardy techniczne obiektów cyfrowych przy digitalizacji dziedzictwa kulturowego*. Biblioteka Główna Politechniki Warszawskiej, Warszawa 2008.
26. Teng C.-C., Mitchell J., Walker C., Swan A., Davila C., Howard D., Needham T., *A medical image archive solution in the cloud*. Software Engineering and Service Sciences (ICSESS), 2010 IEEE International Conference on, 2010, 431–434. IEEE, DOI: 10.1109/ICSESS.2010.5552343.
27. Traczyk T., Ogryczak W., Pałka P., Śliwiński T., *Digital Preservation: Putting It to Work*, Vol. 700, Studies in Computational Intelligence. Springer International Publishing, 2017, DOI: 10.1007/978-3-319-51801-5.
28. Wallace C., Pordesch U., Brandner R., *Long-term archive service requirements*. http://www.ietf.org/rfc/rfc4810.txt, March 2007. Dostęp: 2020-08-11.

Problems of Long-Term Archiving of Digital Resources on the Example of the CREDO Project

Abstract: Long-term archiving of digital resources is a serious problem that has not yet found sufficient attention from the IT industry, nor widely available solutions. Preservation of usability of stored resources in the digital archive requires not only reliable storage of data files, but also the possibility of efficient searching, as well as verification of data authenticity and its correct interpretation both in the technical (data format, etc.), and semantic sense (information understanding in an appropriate context, etc.). The paper discusses these problems and presents solutions adopted in the CREDO project.

Keywords: long-term archiving, archiving of digital resources, digital repositories, data storage, metadata

dr inż. Piotr Pałka
p.palka@ia.pw.edu.pl
ORCID: 0000-0002-0006-363X

Jest adiunktem na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej, gdzie pracuje od 2009 r. Jego zainteresowania naukowe obejmują oprócz archiwizacji wieczystej m.in. problematykę rozproszonej sztucznej inteligencji, systemów wieloagentowych, agentowych modeli symulacyjnych, badań operacyjnych i wspomagania decyzji, w szczególności zagadnienia kooperacyjnej teorii gier oraz modelowanie problemów infrastrukturalnych za pomocą matematycznych modeli liniowych i mieszanych, relacyjne bazy danych.

dr inż. Tomasz Śliwiński
t.sliwinski@ia.pw.edu.pl
ORCID: 0000-0002-5111-1830

Jest adiunktem na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej, gdzie pracuje od 2006 r. Zainteresowania naukowe i zawodowe skupione wokół optymalizacji – ciągłej, dyskretnej, liniowej i nieliniowej – w takich dziedzinach jak: harmonogramowanie, tworzenie harmonogramu czasu pracy personelu, sprawiedliwy przydział zasobów, optymalizacja wielokryterialna, systemy wspomagania decyzji.

dr inż. Tomasz Traczyk
t.traczyk@ia.pw.edu.pl
ORCID: 0000-0002-6602-4094

Jest docentem na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej, gdzie pracuje od 1985 r. Sprawuje tam funkcję zastępcy dyrektora Instytutu Automatyki i Informatyki Stosowanej oraz kierownika studiów podyplomowych. Jego zainteresowania naukowe i kompetencje techniczne obejmują m.in. projektowanie systemów informacyjnych z zastosowaniem baz danych i języka XML, budowę systemów informatycznych wspierających wielkie eksperymenty fizyki wysokich energii oraz archiwizację długoterminową zasobów cyfrowych.