STUDIA INFORMATICA 2013
Volume 34 Number 2B (112)

Krzysztof PATER, Tomasz TRACZYK
Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych

OPAKOWANIE ZASOBÓW CYFROWYCH NA POTRZEBY ARCHIWIZACJI DŁUGOTERMINOWEJ

Streszczenie. W artykule opisano koncepcję opakowania zasobów cyfrowych wraz z metadanymi, w sposób bezpieczny oraz dostosowany do archiwizacji długoterminowej. Przedstawiono wymagania i założenia dla opakowania. Porównano przedstawioną koncepcję do istniejących rozwiązań i przedstawiono dalsze perspektywy rozwoju.

Słowa kluczowe: archiwizacja długoterminowa, archiwum cyfrowe, metadane, opakowanie informacji

PACKAGING DIGITAL RESOURCES FOR LONG-TERM ARCHIVING

Summary. The paper describes a concept of packaging digital resources with their metadata in a safe way suitable for long-term archiving. The requirements and objectives for the package are shown. Presented concept is compared to existing solutions. Finally, further prospects are presented.

Keywords: digital long-term archives, metadata, information packaging, long-term archiving

1. Wprowadzenie

Archiwizacja długoterminowa, zwana również archiwizacją wieczystą, ma na celu przechowywanie dużych ilości danych cyfrowych przez okres istotnie długoterminowy (np. przez kilka pokoleń) lub bezterminowy [10]. Na tak długim horyzoncie czasowym nie ma możliwości zapewnienia trwałości nośnika elektronicznego ani zagwarantowania dostępności obecnie stosowanych technologii. Jednocześnie trzeba pamiętać o zabezpieczeniu danych przed utratą lub zniekształceniem, niepowołanym użyciem, a także należy zapewnić możliwość wyszuka-

nia, odczytu czy interpretacji danych w przyszłości. Dlatego też trwałe fizyczne zapisanie informacji jest niewystarczające, konieczne jest również odpowiednie jej opakowanie.

1.1. Pakiet – archiwizowany obiekt

Utworzona przez *Consultative Committee for Space Data Systems* (CCSDS) rekomendacja *Reference Model Open Archival Information System* (OAIS) [3], przyjęta przez ISO w normie 14721, przedstawia ogólne zalecenia odnośnie długoterminowego przechowywania zasobów cyfrowych, m.in. definiując pojęcie informacji w kontekście archiwum długoterminowego, a także przedstawiając model logiczny zasobu przetrzymywanego w takim archiwum.

Według standardu OAIS, archiwizowany obiekt (ang. *information package*) składa się z następujących składników.

- *Content Information* (CI) – właściwego zasobu, zawierającego:
  - *content data object* – treści właściwe w postaci ciągu bitów[1],
  - *representation information* – informację nadającą znaczenie tym bitom, tak aby mogły być one zrozumiałe przez użytkownika archiwum.
- *Preservation Description Information* (PDI) – informacji niezbędnych do zarządzania przechowywanym zasobem, w skład których wchodzą:
  - *provenance* – dokumentacja cyklu życia pakietu,
  - *context* – dokumentacja związku pakietu ze środowiskiem (np. dlaczego został stworzony, związek z innymi pakietami),
  - *reference* – zbiór unikalnych, globalnych identyfikatorów pakietu (np. ISBN, URN),
  - *fixity* – informacja zapewniająca integralność danych.
- *Packing Information* (PI) – metadanych, określających sposób powiązania CI z PDI, np. informacja o technologii, która została wykorzystana do stworzenia pakietu.
- *Descriptive Information* (DI) – metadanych, dzięki którym możliwe jest wyszukiwanie informacji zapisanych w archiwum.

W [16] propozycja OAIS została rozbudowana. Uszczegółowieniu zostały poddane elementy CI oraz PDI pakietu. Należy jednak zaznaczyć, że tych modeli nie należy traktować jako ostatecznych, gdyż w zależności od potrzeb archiwum, pewne elementy mogą, a nawet powinny ulec zmianom: archiwa powinno budować się z uwzględnieniem konkretnego celu archiwizacji.

[1] Według OAIS, *Content Data Object* może być reprezentowany przez obiekt fizyczny, np. budynek, lub poprzez dane cyfrowe w postaci ciągu bitów. W tej pracy rozpatrywany jest tylko drugi przypadek.

Standard OAIS wyróżnia trzy typy pakietów, w zależności od ich przeznaczenia w archiwum:

- SIP (ang. *Submission Information Package*) – pakiet przesyłany przez producenta zasobu cyfrowego do archiwum,
- AIP (ang. *Archival Information Package*) – pakiet przetrzymywany w archiwum głębokim,
- DIP (ang. *Dissemination Information Package*) – pakiet przesyłany zainteresowanym użytkownikom końcowym.

Jeden pakiet AIP może obejmować jeden lub wiele SIP oraz jeden DIP może obejmować jeden lub wiele AIP. Jak podkreślono w [10], system archiwalny powinien móc transformować SIP na AIP, przetrzymywać długoterminowo AIP, a także transformować AIP na DIP. Nie jest jednoznacznie określone, czy transformacji ma dokonywać samo archiwum czy inne podsystemy zewnętrzne.

1.2. Metadane i metadane konserwatorskie

Aby długotrwałe przechowywanie informacji było sensowne, musi istnieć możliwość identyfikacji obiektów zapisanych w archiwum, zarządzania nimi i przede wszystkim dostępu do nich. Chcąc to zapewnić, konieczne jest zbudowanie odpowiedniej struktury metadanych. Istnieje wiele definicji mówiących, czym są metadane i jak z nich korzystać. Najprostsza i najczęściej stosowana określa je jako „dane o danych". Opracowanie [12] rozszerza tę definicję, określając metadane jako uporządkowaną informację, która opisuje, wyjaśnia, lokalizuje lub w inny sposób ułatwia pobieranie, wykorzystywanie lub zarządzanie zasobem źródłowym. Natomiast w [13] zdefiniowano metadane poprzez ich funkcjonalności: „Metadane umożliwiają identyfikację obiektów cyfrowych, zarządzanie nimi, dostęp, używanie, czynią też sensownym długotrwałe przechowywanie i przyszłe migracje na nowe nośniki".

Najczęściej spotykany podział metadanych wyróżnia trzy kategorie:

- opisowe (ang. *descriptive metadata*) – informacje służące identyfikacji, wyszukiwaniu oraz dokumentacji ewentualnych powiązań pomiędzy obiektami źródłowymi,
- administracyjno-techniczne (ang. *administrative metadata*) – informacje mające na celu wspieranie zarządzania zasobami,
- strukturalne (ang. *structural metadata*) – *informacje* określające strukturę zasobu, np. rozdziały w książce.

W kontekście długotrwałego przechowywania informacji pojawia się jeszcze czwarta grupa metadanych – metadane konserwatorskie (ang. *preservation metadata*). Według [7], metadane konserwatorskie są metadanymi opisowymi, strukturalnymi i administracyjnymi,

które wspierają długoterminowe przechowywanie materiałów cyfrowych poprzez działanie w pięciu głównych obszarach:

- *provenance* – historia składowania zasobu (od momentu utworzenia po wszystkie zmiany w jego fizycznej budowie lub prawach własności),
- *authenticity* – potwierdzenie tożsamości zasobu, zapewnienie, że nie został zmieniony (celowo lub przypadkowo) w nieudokumentowany sposób,
- *preservation activity* – dokumentacja wszelkich działań archiwizacyjnych, mających wpływ na postać zasobu archiwalnego,

technical environment – wymagania techniczne, mające na celu umożliwienie odczytu i używania obiektu cyfrowego w takim stanie, w jakim jest aktualnie przechowywany w archiwum,

- *right management* – wszelkie prawa własności intelektualnej, które ograniczają działania archiwizacyjne na tym obiekcie oraz jego udostępnianie aktualnym i przyszłym użytkownikom.

W [7] wymieniono również trzy główne powody, dla których metadane konserwatorskie są tak ważne.

- *Obiekty cyfrowe są zależne od technologii.* Dostęp do nich wymaga wielu rozwiązań technologicznych, umożliwiających zrozumienie zawartych treści, dlatego też nie jest wystarczające proste ich zapisanie, ale konieczne jest również przechowywanie informacji, w jaki sposób ją odczytać i korzystać z niej.
- *Obiekty cyfrowe są niestałe.* W łatwy sposób mogą być zmienione poprzez zamierzone lub niezamierzone działania. Poza tym nośniki informacji cyfrowej nie są wystarczająco trwałe dla potrzeb archiwizacji długoterminowej (w wyniku degradacji mogą pojawiać się błędy w danych, powodujące utratę informacji). Natomiast ze względu na duże tempo postępu technologicznego, obiekty cyfrowe muszą być przenoszone z jednego formatu na inny, w celu zapewnienia stałego dostępu do nich.
- *Obiekty cyfrowe są ograniczone poprzez prawa własności intelektualnej*. Ważne jest, aby archiwizacja informacji cyfrowej następowała na wczesnym etapie jej cyklu życia, m.in. ze względu na szybkie tempo starzenia się technologii cyfrowych, a także krótki czas trwałości nośników elektronicznych. Informacja niezabezpieczona w odpowiedni sposób we wczesnej fazie życia może zostać uszkodzona, a w konsekwencji utracona. W tej fazie życia archiwizowane zasoby są często jeszcze chronione prawami autorskimi, dlatego też archiwa długoterminowe powinny umożliwiać przechowywanie zasobów ograniczonych prawami autorskimi, definiując dopuszczalny sposób składowania i dostępu.

1.3. Aktualne rozwiązania

W literaturze można odnaleźć wiele inicjatyw mających na celu zbudowanie bezpiecznego formatu opakowania. Aby to osiągnąć, po pierwsze konieczne jest zbudowanie odpowiedniej struktury metadanych. Jedną z takich inicjatyw jest PREMIS (*Preservation Metadata: Implementation Strategies*) [14], która zrzesza kilka dużych organizacji, m.in. *OCLC*, *Library of Congress*. Alternatywą dla PREMIS jest zestaw metadanych Narodowej Biblioteki Nowej Zelandii [11], który został wykorzystany w praktyce m.in. przez Bibliotekę Narodową Niemiec.

Po drugie konieczne jest opracowanie sposobu powiązania danych z metadanymi. CCSDS opracowało standard opakowania informacji XFDU [2], który następnie został wykorzystany przez ESA w ramach projektu SAFE [15]. Natomiast w [5] pokazano, jak wykorzystać XFDU na potrzeby projektu CASPAR – ***Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval*** [1].

Po trzecie należy rozwiązać problem fizycznego zbudowania pakietu. Ze względu na fakt, iż w zasadzie we wszystkich rozwiązaniach metadane zapisywane są w postaci dokumentu XML, problem opakowania fizycznego danych sprowadza się zapisania plików binarnych w plikach XML. Swoją propozycję realizacji tego zagadnienia z wykorzystaniem powszechnie uznanego standardu MIME [6] przedstawiło konsorcjum W3C w standardzie XOP [22].

2. Wymagania dla opakowania

Projektując rozwiązanie, mające na celu opakowanie informacji w sposób bezpieczny oraz dostosowany do archiwizacji długoterminowej, należy sobie uświadomić wiążące się z tym problemy.

- Nietrwałość i zmienność standardów – nikt nie jest w stanie zagwarantować, że za sto lat ktokolwiek będzie w stanie odczytać plik w konkretnym formacie, np. PDF.
- Dyslokacja zasobów (występowanie tego samego zasobu w różnych miejscach) – zasoby archiwum długoterminowego powinny występować w przynajmniej dwóch kopiach, możliwie oddalonych od siebie, aby w razie poważnej katastrofy, np. powodzi, pożaru itp., nie utracić wszystkich danych.
- Trwała identyfikacja zasobów – zasób powinien być dostępny pod raz przydzielonym identyfikatorem przez cały cykl swojego życia. Dodatkowo, ze względu na fakt, iż może istnieć wiele kopii tego samego zasobu (redundantny zapis w różnych technologiach, nowa kopia po migracji, kopie w zapasowych archiwach), musi być zapewniona możliwość

identyfikacji każdej kopii, gdyż różne kopie mogą służyć do różnych celów, np. jedne od udostępniania a inne do rekonstrukcji.

- Zabezpieczenie przed możliwością utraty przez zasób swoich metadanych lub pomieszania metadanych różnych zasobów.
- Udostępnianie przynajmniej części metadanych *on-line*, podczas gdy sam zasób może być składowany w tzw. archiwum głębokim, będącym zwykle w stanie *off-line*.

W [10] znajdujemy podstawowe zalecenia odnośnie opakowania zasobów. Powinny być one przechowywane jako jeden obiekt-kontener (pakiet), tak aby uniknąć „zagubienia" metadanych przez zasób. Dodatkowe cechy, takie jak samoopisowość (ang. *self-describing*) oraz kompletność (ang. *self-contained*), mają zapewnić prawidłową interpretację zasobu w odległej przyszłości, „gdy kontekst jego istnienia może być nieznany, format niezrozumiały, a obiekty powiązane z zasobem – utracone".

Ponadto, aby zwiększyć prawdopodobieństwo poprawnej interpretacji danych, nawet w wieloletniej perspektywie, należy rozważyć przechowywanie tej samej informacji w więcej niż jednym formacie danych, z użyciem powszechnie uznanych, otwartych standardów zapisu i opakowania informacji.

Również w dziedzinie zapisu metadanych konserwatorskich możemy znaleźć kilka ogólnych zaleceń. W [10] uwzględniono następujące:

- zapis niezależny od źródłowego i docelowego formatu danych, z konwersją przy zapisie i odczycie,
- zapis metadanych w sposób ogólny, samoopisujący i szeroko stosowany, np. z wykorzystaniem standardu RDF [21],
- stosowanie ogólnie przyjętych standardów co do treści metadanych, np. Dublin Core [4],
- opieranie systemu zapisu metadanych na ontologiach, zapisanych w powszechnie zrozumiały sposób, np. za pomocą OWL [20], i przechowywanych w archiwum.

Z kolei poniżej przedstawiono zalecane wg [7] cechy schematu metadanych konserwatorskich:

- Comprehensive – możliwie wszechstronny, rozważając nawet dodanie na początku kilku zbędnych elementów, z myślą o ich wykorzystaniu w przyszłości.
- Oriented toward implementation – projektowany z myślą o dobrych praktykach implementacyjnych, np. powinien dostarczać, gdzie to jest możliwe, słownik do wypełniania elementów, a nie polegać na „wolnym tekście", oraz być przystosowany do automatycznego przepływu kolekcji metadanych oraz zarządzania nimi.
- Interoperable – zaprojektowany tak, aby radził sobie z wymaganiami wszystkich potencjalnie zainteresowanych stron. Musi istnieć możliwość transformacji schematu z lub do innych postaci, bliższych środowisku zewnętrznemu archiwum. Przykłady zaczerpnięte

z OAIS to transformacja SIP do AIP oraz AIP do DIP, a także transfer obiektów do innych repozytoriów (być może z informacją zapisaną w innym schemacie metadanych lub też w innej technologii).

3. Koncepcja opakowania zasobów cyfrowych do celów archiwizacji długoterminowej

W artykule zaproponowano sposób budowy pakietu na potrzeby przechowywania w archiwum długoterminowym. Opisano rozwiązanie, które zawiera propozycję schematu metadanych konserwatorskich oraz sposób opakowania metadanych wraz z zasobami cyfrowymi w jeden pakiet fizyczny. Pakiet taki, przez zestaw odnośników do innych pakietów, może wraz z nimi tworzyć większy obiekt logiczny.

3.1. Założenia

Budowany pakiet jest przeznaczony do składowania w archiwum głębokim, a więc jest to pakiet AIP (ang. *Archival Information Package*) w sensie terminologii OAIS. Powinien on posiadać mechanizmy pozwalające sprostać przedstawionym wyżej problemom, związanym z długoterminowym przechowywaniem danych cyfrowych.

Szczególny nacisk w opracowanej koncepcji położono na elastyczność, a więc możliwość dostosowania pakietu do różnych celów archiwizacji. Pokazano jak to zrobić na przykładzie archiwizacji fotografii zabytków architektonicznych.

3.2. Wybór technologii

Budowa pakietu archiwalnego obejmuje trzy zasadnicze obszary: metadanych konserwatorskich, opakowania logicznego i opakowania fizycznego. W każdym z tych obszarów należało wybrać technologię w celu stworzenia kompletnego rozwiązania.

3.2.1. Metadane konserwatorskie

W koncepcji opakowania wykorzystany został schemat metadanych Narodowej Biblioteki Nowej Zelandii (NLNZ – ang. *National Library of New Zealand*). Wyboru tego dokonano na podstawie zaleceń dla metadanych do długotrwałego przechowywania, ukazanych w [13], które stwierdzają między innymi, że wprowadzenie standardu nowozelandzkiego w Polsce przebiegałoby łatwiej niż przyjęcie PREMIS. Za wykorzystaniem schematu NLNZ przemawiają również jego praktyczne wykorzystanie, między innymi przez Bibliotekę Narodową Niemiec, oraz prostsza forma niż schematu PREMIS.

3.2.2. Opakowanie logiczne

Zaproponowana koncepcja opiera się na standardzie XFDU. Wyboru tego dokonano na podstawie doświadczeń uczestników projektu CASPAR [1]. W artykule [5] zwrócono uwagę na takie zalety standardu XFDU, jak:

- ustandaryzowanie oraz dobre udokumentowanie przez *Consultative Committee for Space Data Systems* (CCSDS),
- zaprojektowanie od początku zgodnie z terminologią OAIS,
- praktyczne wykorzystanie, między innymi przez *The European Space Agency* (ESA) podczas tworzenia formatu opakowania *Standard Archive Format for Europe* (SAFE) [15],
- istniejące otwarte narzędzia oraz API zrealizowane w języku Java.

3.2.3. Opakowanie fizyczne

Podążając za [19] oraz [17], stwierdzono, że najwłaściwszym wyborem co do opakowania fizycznego jest format ZIP. Do jego zalet należą następujące cechy:

- jest bezpłatny,
- jest dobrze rozpoznany i powszechnie używany,
- ma indeks, dzięki czemu możliwy jest swobodny dostęp do plików w pakiecie,
- pliki w pakiecie mogą być kopiowane z jednego pakietu do drugiego, bez konieczności dekompresji i ponownej kompresji,
- możliwość kompresji pozwala zmniejszyć rozmiar plików,
- możliwe jest tworzenie pakietów samorozpakowujących się.

Ze względu na ograniczenia, na wielkość pliku formatu ZIP (maksymalny rozmiar to niespełna 4 GB) wykorzystano jego rozszerzenie, a mianowicie format ZIP64.

3.3. Budowa pakietu

Wzorując się na [18], w tej sekcji przedstawiono proponowaną budowę pakietu AIP z czterech różnych perspektyw: fizycznej, logicznej, strukturalnej oraz systemowej. Należy zaznaczyć, że całość jest zgodna z terminologią OAIS oraz modelem zapisu metadanych, będącym połączeniem XFDU i NLNZ oraz metadanych związanych z archiwizacją fotografii zabytków architektonicznych.

3.3.1. Perspektywa fizyczna

Z perspektywy fizycznej proponowany pakiet AIP jest plikiem zapisanym w formacie ZIP64 i składa się z trzech części: danych w formie binarnej, metadanych (zapisanych w manifeście XFDU oraz ewentualnie w plikach z danymi) oraz elementów opakowania (związanych ze standardem ZIP64). W pakiecie znajduje się dokładnie jeden manifest, nato-

miast plików binarnych może być wiele. Poprzez zastosowanie standardu XOP, pliki binarne są połączone z manifestem XFDU i dzięki temu ze swoimi metadanymi. Jednakże połączenie to odbywa się wyłącznie na podstawie identyfikatorów, a manifest i dane cyfrowe mają charakter oddzielnych plików, dlatego też istnieje ryzyko zagubienia przez zasób swoich metadanych. Problem ten został ograniczony przez wykorzystanie standardu ZIP64, dzięki któremu pliki binarne i manifest zostają spakowane do jednego pliku archiwum.

3.3.2. Perspektywa logiczna

Z perspektywy logicznej pakiet jest zdefiniowany poprzez manifest XFDU, który może zostać rozszerzony o dodatkowe elementy, mające na celu zbudowanie takiego zestawu metadanych, który umożliwi odczyt i interpretację danych w odległej przyszłości oraz będzie odpowiedzią na postawione cele archiwizacji. Można to zrealizować poprzez stworzenie specjalnych przestrzeni nazw (*XML namespaces*) i dodanie ich w sekcji *metadataObject* pakietu XFDU. W ramach opisanej tu pracy zostały dodane elementy zgodne ze standardem metadanych konserwatorskich Narodowej Biblioteki Nowej Zelandii oraz metadane niezbędne do realizacji wybranego celu archiwizacji, np. dla archiwizacji fotografii zabytków architektonicznych wybrano standardy *Dublin Core* – zawierające metadane do opisu fotografii, oraz *Object ID*, *Core Data Standard, Core Data Index* – zwierające metadane do opisu zabytków architektonicznych w zależności od obszaru, który opisują [9].

Wykorzystując mechanizm odniesień zewnętrznych, można połączyć archiwizowany pakiet z innymi pakietami, wspólnie tworzącymi jedną całość. Dodatkowo poza pakiet można wydzielić metadane mające charakter powszechnych standardów, w celu uniknięcia redundantnego zapisu. Rozwiązanie takie wprawdzie podnosi ryzyko zagubienia przez zasób części swoich metadanych, ale znacznie zmniejsza ilość miejsca niezbędnego do przechowywania zasobu w archiwum, co z kolei może znacząco obniżyć koszty utrzymania i eksploatacji archiwum oraz ułatwia ewentualną modyfikację metadanych, np. po wprowadzeniu nowej wersji danego standardu. Zalety te zostały również zauważone w [5].

3.3.3. Perspektywa strukturalna

Perspektywa strukturalna określa hierarchię elementów w pakiecie. Można wyróżnić trzy zasadnicze poziomy tej hierarchii: *Primary Node*, czyli archiwizowany obiekt, np. jeden wybrany detal architektoniczny, w skład którego wchodzą *Intermediate Nodes,* np. zdjęcia, które z kolei mogą być zapisane w różnych wersjach jako *Terminal Nodes*, np. w zależności od przeznaczenia: *master* do długoterminowego przechowywania, *service* do udostępniania czy *preview* do przeglądania [18]. W standardzie XFDU można to osiągnąć w ramach elementu *informationPackageMap*, zagnieżdżając w odpowiedni sposób elementy *contentUnit*.

3.3.4. Perspektywa systemowa

Z perspektywy systemowej pakiet jest wynikiem transformacji danych otrzymanych z systemów źródłowych do docelowej postaci, przeznaczonej do składowania w archiwum. Transformacja ta odbywa się poprzez opisany pokrótce w następnym rozdziale system opakowania. Należy tutaj podkreślić, że pakiet został zaprojektowany w taki sposób, aby był niezależny od infrastruktury, a więc może być wykorzystany w różnych środowiskach.

3.4. System opakowania

System informatyczny, dokonujący opakowania powinien zawierać bazę danych, przechowującą metadane wykorzystywane zarówno w samym archiwum (metadane konserwatorskie), jak i te, dostarczane z systemów źródłowych – wykorzystywane np. do wyszukiwania zasobów. Ze względu na dużą różnorodność metadanych, zdecydowano, że taka baza powinna mieć strukturę generyczną.

W czasie projektowania bazy zwrócono szczególną uwagę na następujące problemy:

- występowanie tych samych metadanych w różnych standardach;
- różnice w zestawach metadanych, pomimo zgodności z tym samym standardem (np. Dublin Core), definiuje zestaw elementów obowiązkowych, które muszą występować, i opcjonalnych, z których można zrezygnować; dodatkowo istnieje również możliwość dodania własnych elementów, adekwatnych do celu archiwizacji; w ten sposób możemy otrzymać wiele różnych kombinacji metadanych, w różnych systemach źródłowych, które dalej są zgodne z głównym standardem, gdyż zawierają wszystkie elementy obowiązkowe);
- możliwość występowania tych samych danych z systemów źródłowych w różnych pakietach archiwalnych;
- konieczność stworzenia, gdzie to możliwe, słownika dopuszczalnych wartości dla określonych elementów metadanych.

Tworzenie pakietu archiwalnego odbywa się wg schematu:

1. System źródłowy dostarcza dane w postaci dokumentu XML (w wersji testowej w dostarczanym do systemu opakowania dokumencie XML znajdują się odnośniki umożliwiające pobranie plików binarnych).
2. Dokument zostaje sparsowany i następuje zapisanie metadanych do bazy.
3. Metadane otrzymane z systemu źródłowego są uzupełniane o brakujące elementy, m.in. z zakresu metadanych konserwatorskich.
4. Na podstawie zawartości bazy danych generowany jest manifest XFDU, do którego jest dołączany sam zasób (w postaci danych binarnych) z wykorzystaniem standardu XOP.

5. Całość zostaje opakowana, z wykorzystaniem standardu ZIP64, w jeden plik fizyczny i zapisana w systemie plików archiwum.

Zauważono również, że taki system informatyczny będzie przechowywał pełen zestaw metadanych pakietów zapisanych w archiwum, dlatego też może on zostać wykorzystany przy tworzeniu systemu wyszukującego informacje znajdujące się w archiwum.

4. Podsumowanie

W opracowaniu tym przedstawiono problematykę opakowania zasobów cyfrowych na potrzeby archiwizacji długoterminowej oraz zaproponowano koncepcję rozwiązania opartą na znanych i powszechnie uznanych standardach.

4.1. Cechy proponowanego rozwiązania

Proponowana budowa pakietu AIP opiera się na standardzie XML, który charakteryzują ważne w kontekście archiwizacji długoterminowej cechy, takie jak: niezależność od platformy, szerokie wsparcie przemysłowe, długowieczność, możliwość migracji, narzędzia do przetwarzania oraz możliwość walidacji [8]. Dzięki zastosowaniu formatu XFDU otrzymuje się rozwiązanie zgodne z terminologią OAIS oraz mające narzędzia, które można wykorzystać w fazie implementacji. Zastosowanie schematu metadanych konserwatorskich Narodowej Biblioteki Nowej Zelandii dostarcza stosunkowo prostego rozwiązania, które jednocześnie jest w stanie sprostać wymaganiom archiwizacji długoterminowej. Przedstawiona budowa pakietu jest również na tyle elastyczna, że może zostać dostosowana do dowolnego celu archiwizacji, poprzez odpowiednie dobranie metadanych właściwych dla tego celu, co sprawdzono na przykładzie archiwizacji fotografii zabytków architektonicznych. Ważną cechą zaproponowanego rozwiązania jest również jego kompletność. Na koniec otrzymuje się jeden plik fizyczny, gotowy do zapisu w docelowym środowisku, które oczywiście musi być odpowiednio dostosowane do archiwizacji długoterminowej, np. opierając się na koncepcji zasobnika, którą przedstawiono w [10].

4.2. Porównanie z istniejącymi rozwiązaniami

Podobnie jak większość rozwiązań z literatury, również to opiera się na pliku manifestu zapisanym w postaci XML. Zastosowany tutaj standard XFDU stanowi również istotny element w projektach CASPAR i SAFE.

Wykorzystano tutaj schemat metadanych Narodowej Biblioteki Nowej Zelandii, który znalazł zastosowanie między innymi w Narodowej Bibliotece Niemiec i został wybrany jako proponowany punkt wyjścia w [13].

Do połączenia manifestu zapisanego w postaci pliku XML z plikami binarnymi użyto standardu XOP, przy czym zdecydowano się nie wykorzystywać standardu MIME jako formatu opakowania, a zastąpić go standardem ZIP64, którego główną zaletą w stosunku do MIME jest możliwość kompresji, pozwalająca zmniejszyć rozmiar pakietu.

To co wyróżnia przedstawione rozwiązanie od innych znanych z literatury, to jego kompletność, począwszy od wyboru zestawu metadanych konserwatorskich, poprzez dobór metadanych odpowiednich ze względu na wybrany cel archiwizacji, aż do logicznego opakowania informacji i zamknięcia całości w jeden pakiet fizyczny, gotowy do zapisu na nośniku danych.

4.3. Dalsze prace i perspektywy rozwoju

Przedstawiony w niniejszym artykule system opakowania został zaimplementowany w wersji testowej. W przyszłości powinna powstać implementacja profesjonalna, mogąca sprawnie pakować zasoby o wielkiej objętości, np. wieloterabajtowe (w przypadku szczególnie wielkich zasobów być może powinny być użyte także rozwiązania sprzętowe). W procesie opakowania należy również uwzględnić zapis części metadanych do podsystemu umożliwiającego wyszukanie informacji. W koncepcji przedstawionej w [10] rolę takiego systemu pełnią elektroniczna kartoteka i witryna udostępniająca. Można również rozważać rozbudowę systemu opakowującego, tak by on sam mógł pełnić rolę takiej kartoteki.

Struktura zapisu metadanych obecnie opiera się na dokumencie XML, zdefiniowanym z użyciem schematu *XML Schema*. W przyszłości warto rozważyć zapis części metadanych, oparając się na ontologiach.

Istotnym problemem z punktu widzenia archiwizacji jest również identyfikacja pakietów. Identyfikator powinien mieć dwie podstawowe cechy: niepowtarzalność – każdą kopię zasobu musi się dać jednoznacznie identyfikować, oraz trwałość – obiekt powinien być dostępny pod raz przydzielonym identyfikatorem przez cały cykl swojego życia. Aby to zapewnić, nie wystarczy samo nadanie unikalnego identyfikatora. Konieczne jest stworzenie systemu trwałej identyfikacji, uwzględniającego złożoność zagadnienia, m.in. problemy wynikające z istnienia kopii zasobów. Prace nad takim systemem są prowadzone.

BIBLIOGRAFIA

1. CASPAR – Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval, http://www.casparpreserves.eu/caspar-project.html.
2. Consultative Committee for Space Data Systems: XML Formatted Data Unit (XFDU) Structure and Construction Rules, Draft Recommended Standard, CCSDS 661.0-B-0 Blue Book, 2008, http://sindbad.gsfc.nasa.gov/xfdu/pdfdocs/xfduspec.pdf.
3. Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1 Blue Book, 2002, http://public.ccsds.org/ publications/archive/650x0b1s.pdf.
4. Dublin Core Metadata Initiative: Dublin Core Metadata Element Set. DCMI Recommendation, 2012, http://dublincore.org/documents/dces/ (patrz też Metadata Terms …/dcmi-terms).
5. Dunckley M., Ronen S., Henis E. A., Rabinovici-Cohen S., Reshef P., Conway E., Giaretta D.: Using XFDU for CASPAR information packaging. OCLC Systems & Services: International digital library perspectives, Vol. 26, No. 2, 2010, s. 80÷93.
6. Levinson E.: The MIME Multipart/Related Content-type. RFC-2387, 1998, http://www.ietf.org/rfc/rfc2387.txt.
7. Lavoie B., Gartner R.: Technology Watch Report. Preservation Metadata, 2005, http://www.dpconline.org/component/docman/doc_download/88-preservation-metadata-preservation-metadata.
8. Lou Reich/CSC: XML information Packaging Standards for Archives, Long Term Knowledge Retention Workshop, 2006, http://edge.cs.drexel.edu/LTKR/nistconf.pdf.
9. Rybiński M.: Zaprojektowanie i wykonanie bazy danych i aplikacji wspierających dokumentowanie zagrożonego detalu architektury w ramach projektu Ginący detal. Praca dyplomowa inżynierska, kierownik pracy T. Traczyk. Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Warszawa 2009.
10. Marasek K., Walczak J., Traczyk T., Płoszajski G., Kaźmierski A.: Koncepcja elektronicznego archiwum wieczystego. Studia Informatica, Vol. 30, No. 2B (84), 2009, s. 275÷307.
11. National Library of New Zealand: Metadata standards framework [electronic resource]: preservation metadata, 2002, http://natlib.govt.nz/records/20714436?search%5Bpath%5D=items&search%5Btext%5D=Metadata+standards+framework.
12. NISO Press National Information Standards Organization: Understanding Metadata, DPC Technology Watch Series Report 05-01, 2004, http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

13. 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.
14. Preservation Metadata: Implementation Strategies (PREMIS), http://www.loc.gov/standards/premis/.
15. The European Space Agency (ESA): Standard Archive Format for Europe (SAFE), 2011, http://earth.esa.int/SAFE/.
16. The OCLC/RLG Working Group on Preservation Metadata: Preservation Metadata and the OAIS Information Model. A Metadata Framework to Support the Preservation of Digital Objects, 2002, www.oclc.org/research/pmwg/pm_framework.pdf.
17. The OpenOffice.org XML Project team: XML Packages, http://www.openoffice.org/xml/package.html.
18. UTA: Archival Information Package (AIP) Design Study, Identification Number: LC-DAVRS-07, 2001, http://www.loc.gov/rr/mopic/avprot/AIP-Study_v19.pdf.
19. World Wide Web Consortium (W3C): Report on XML Packaging. 1999, http://www.w3.org/1999/07/xml-pkg234/Overview.
20. World Wide Web Consortium (W3C): OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation, 2012, http://www.w3.org/TR-/2012/ REC-owl2-overview-20121211/.
21. World Wide Web Consortium (W3C): RDF/XML Syntax Specification (Revised). W3C Recommendation, 2004, http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.
22. World Wide Web Consortium (W3C): XML-binary Optimized Packaging. W3C Recommendation, 2005, http://www.w3.org/TR/2005/REC-xop10-20050125/.

Wpłynęło do Redakcji 16 stycznia 2013 r.

Abstract

The purpose of the long-term archiving is to store large amounts of digital data over a significantly long period (e.g. a few generations). This leads to many problems such as instability of the electronic media, obsolescence of technology, ensuring economic efficiency. At the same time we cannot forget basic functionalities that long-term archive should meet: to provide data security and access to the data. To achieve these objectives it is not enough to

permanently store the information, an appropriate packaging of the data along with metadata is also necessary.

The object of this study is to analyse the problem of packaging digital data along with metadata in a safe way, suitable for long-term archiving. An exemplary solution is also presented.

The first part introduces the goal of packaging resources in a digital archive, presents the general model of resource stored in the archive, and discusses the nature of preservation metadata.

The second part proposes a package structure for digital long-term archiving. It shows standards, which are selected to design the structure, and presents the designed package from four different perspectives: logical, physical, structural and system. Packaging system is also described, which could become a part of long-term archive infrastructure.

At the end the work is summarized by presenting features of the proposed solution, comparing the presented concept to existing solutions and identifying further opportunities, that could be a next steps in building a fully functional long-term archive.

Adresy

Krzysztof PATER: Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, ul. Nowowiejska 15/19, 00-665 Warszawa, Polska, K.Pater@stud.elka.pw.edu.pl.
Tomasz TRACZYK: Politechnika Warszawska, Instytut Automatyki i Informatyki Stosowanej, ul. Nowowiejska 15/19, 00-665 Warszawa, Polska, T.Traczyk@ia.pw.edu.pl.