EBIB Biuletyn EBIB, nr 9 (154)/2014,
Gromadzenie i zabezpieczanie danych cyfrowych
Artykuł

Aleksander Radwański
Zakład Narodowy im. Ossolińskich
aleksander.radwanski@ebib.pl

Konteksty długoterminowej archiwizacji

Streszczenie: *Autor porusza zagadnienie długotrwałego przechowywania zasobów i jego kontekstów. Dochodzi do wniosku, że długoterminowa archiwizacja powinna być powiązana z innymi zagadnieniami dostępu do zasobów cyfrowych, takimi jak powtórne wykorzystanie i łączenie danych. Sprzyja to racjonalnemu wykorzystaniu środków publicznych.*

Słowa kluczowe: *długoterminowa archiwizacja, ponowne wykorzystanie, łączenie danych, środki publiczne*

Długoterminowa archiwizacja danych cyfrowych jest coraz lepiej opracowanym zagadnieniem i powstaje coraz więcej narzędzi i usług z nią związanych. Istnieje też wiele nierozwiązanych problemów technicznych, organizacyjnych i finansowych, powodujących, że trudno uznać długotrwałą archiwizację za rutynowe działanie, dla którego istnieje ustalony tok postępowania. Aby nie zginąć w lawinie szczegółów, spróbuję wypunktować najważniejsze aspekty w kolejności wskazującej na logiczne powiązania. Osoby oswojone z tematem uznają być może kolejne definiowanie podstawowych pojęć za niepotrzebne, z tych definicji mają jednak wynikać konsekwencje nie zawsze uznawane za oczywiste.

Kto?

Po pierwsze interesować nas tu będą biblioteki, archiwa i muzea — więc instytucje budżetowe, konsumujące środki publiczne. Nie ma tu znaczenia przynależność resortowa czy forma prawna instytucji — jeśli konsumuje ona pieniądze podatników, to rezultaty tej konsumpcji powinny służyć podatnikom.

Co?

Dane cyfrowe powstają w znakomitej większości jako owoc procesu digitalizacji, czyli przetwarzania obiektów dwuwymiarowych (dokumenty, rysunki, mapy, rękopisy, druki) lub trójwymiarowych (rzeźby, zabytkowe przedmioty, architektura) na ich cyfrową reprezentację (obraz, odwzorowanie). Digitalizacji dokonują najczęściej skanery lub wyspecjalizowane kamery cyfrowe, produkując określone rodzaje plików (w znaczeniu jakiego używamy, mówiąc o plikach w komputerze), nazywane: plikami master, plikami źródłowymi, macierzą pierwotną, surowymi danymi i wieloma innymi określeniami, w zależności od kontekstu i przyjętej terminologii. Ustalmy „pliki master” jako termin, którym będziemy się posługiwać w tym artykule. Z plików master powstają publikacje i/lub obiekty cyfrowe (ustalmy analogicznie „pliki prezentacyjne” jako konsekwentną nazwę produktu otrzymanego z przetworzenia plików master, nie wnikając w charakter tych przetworzeń). Pliki prezentacyjne mają najczęściej inną postać niż pliki master, są mniejsze niż odpowiadające im pliki master i mniej liczne (na ogół jeden plik pre-

zentacyjny powstaje z wielu plików master). W przypadku kilkusetstronicowej książki plikami master będzie kilkaset obrazków z odwzorowaniem każdej strony (przeważnie w formacie TIFF lub JPEG), które złożą się na jeden plik prezentacyjny (przeważnie w formacie PDF lub DjVu) zawierający cyfrową wersję książki.

Nie ma stałej relacji pomiędzy rozmiarem plików master a rozmiarem plików prezentacyjnych, gdyż jest ona pochodną wielu zmiennych parametrów. Opierając się jednak na dotychczasowych doświadczeniach (np. w projektach digitalizacyjnych Ossolineum), przyjmijmy, że statystyczna średnia dla zbiorów bibliotecznych będzie wynosiła 1:100, czyli że pliki master będą wymagały 100 razy więcej miejsca do zapisu niż wersja prezentacyjna. Oczywiście te relacje w poszczególnych przypadkach mogą się wahać w bardzo szerokich granicach, nie chodzi tu jednak o ścisłość, ale o uzmysłowienie faktu, że pliki master generują inną skalę problemu długoterminowego przechowywania i technologie radzące sobie z plikami prezentacyjnymi wcale nie muszą sobie poradzić z plikami master.

Za co?

O ile finansowanie procesu digitalizacji uzyskało w miarę stabilną postać (projekty resortowe, europejskie/unijne, samorządowe), to finansowanie długoterminowej archiwizacji i udostępniania jest wciąż kwestią przyszłości. Oczywiście łatwiej sfinansować wytworzenie określonej liczby plików w określonej technologii niż zaplanować i wdrożyć infrastrukturę do bezterminowego archiwizowania niewiadomej liczby plików o niewiadomych właściwościach. Jedynym do tej pory projektem, który zawiera dedykowaną temu tematowi część, jest PLATON i jego usługa powszechnej archiwizacji (U4). Czy finansowanie PLATON-a będzie rosło razem z powierzanymi mu zasobami? Kto powinien być gwarantem ciągłości finansowania archiwizacji? Czy koszty archiwizacji nie wrócą rykoszetem do właścicieli plików? Czy uruchomione zostanie finansowanie projektów archiwizacyjnych analogicznych do projektów digitalizacyjnych? Dziś nie znamy odpowiedzi na te pytania.

Archiwizacja

Generalnie mamy do zarchiwizowania pliki master i pliki prezentacyjne. Może się jednak zdarzyć, że powstają też pliki pośrednie, które warto archiwizować, np. z powodu długotrwałości przetwarzania z plików master na pliki prezentacyjne, co przy setkach tysięcy lub milionach powtórzeń może oznaczać np. rok pracy wielu komputerów oraz obsługujących je ludzi. Taki półprodukt warto przechować, nawet jeśli zwiększa to rozmiary archiwum. Są też podejścia skrajne — archiwizujemy tylko pliki master, ponieważ wszystkie inne są pochodne, zaś pliki prezentacyjne trafiają do bibliotek cyfrowych lub na inne platformy, które mają swoje kopie bezpieczeństwa.

Drugie skrajne podejście polega na archiwizacji wyłącznie plików prezentacyjnych, ponieważ to one stanowią cel digitalizacji. Pliki master pozostają wtedy wyłącznie na nośnikach off-line i degradują się wraz z nimi. Większość projektów digitalizacyjnych nie precyzuje, jaki rodzaj plików ma być udostępniany, poprzestając na ogólnym wymogu udostępnienia wyników digi-

talizacji. Znacząca część projektów nie zawiera też żadnego planu archiwizacji, ponieważ wykracza to poza zakres finansowania (zwykle ze środków na digitalizację nie można kupować ani sprzętu, ani oprogramowania, które nie służy bezpośrednio samemu procesowi digitalizacji, czyli skanowaniu lub fotografowaniu). Bywa też, że brak planu archiwizacji w projekcie wynika z niedostatków wiedzy autorów projektu.

Nie ma w tym zakresie wypracowanej praktyki, która polegałaby na tworzeniu całościowych projektów zawierających wszystkie etapy: digitalizację, inwentaryzację (tworzenie inwentarzy, katalogów i metadanych), archiwizację i udostępnianie. O ile kwestie związane z plikami prezentacyjnymi i ich udostępnianiem znalazły naturalne rozwiązanie w bibliotekach cyfrowych, to kwestia udostępniania plików master jest wciąż sprawą przyszłości. Archiwa master (jeśli istnieją, bo często są to tylko zapisane nośniki, składowane off-line) są traktowane jako ściśle techniczne zaplecze danej instytucji i nie są przygotowane do udostępniania, ani organizacyjnie (odpowiednia struktura i metadane), ani technicznie (wydajność).

Powtórne wykorzystanie (re-use)

Postulat budowania aktywnych archiwów udostępniających pliki master wydaje się wart przemyślenia w kontekście dyskusji o powtórnym wykorzystaniu materiałów cyfrowych (re-use). Sens powtórnego wykorzystania jest oczywisty — po co wydawać pieniądze na to, co już raz opłaciliśmy jako podatnicy. Niech każdy skorzysta z możliwości własnego przetworzenia uzyskanego raz materiału cyfrowego. Wszystko wskazuje jednak na to, że w polskiej praktyce legislacyjnej powtórnym wykorzystaniem zostaną objęte raczej pliki prezentacyjne (gdyż mówi się tam o zbiorach cyfrowych) niż archiwa zawierające pliki master. Logika powtórnego wykorzystania jest natomiast dokładnie odwrotna — to pliki master są bardziej atrakcyjne, gdyż zawierają możliwość przetworzeń alternatywnych, wyzyskujących aspekty pominięte lub utracone w toku wytwarzania pierwszych plików prezentacyjnych. Taki jest też sens archiwizacji plików master — dla przyszłych, nieznanych jeszcze technologii, które pozwolą uzyskać lepsze sposoby prezentacji niż dzisiejsze. Wymaga to nie tylko archiwizacji długoterminowej, ale również budowania archiwów aktywnych, udostępniających pliki on-line w czasie rzeczywistym.

Łączenie i zagnieżdżanie (linking and embedding)

Łączenie i zagnieżdżanie, to terminy znane każdemu informatykowi, wynikające z realizacji dwóch istotnych zasad informatycznej ekonomiki — jednokrotnego wprowadzania danych oraz pozostawienia ich w miejscu wytworzenia. Łatwiej bowiem zmienić algorytm niż pozyskać lub modyfikować dane, zaś wielokrotne wprowadzanie lub niekontrolowane powielanie prędzej czy później prowadzą do utraty spójności i rodzą całą lawinę niekorzystnych konsekwencji. Jeśli zatem jakaś informacja w systemie już istnieje, to powinna zostać przywołana. W internecie realizacją tej idei jest technologia stron WWW, dla których łączenie i zagnieżdżanie jest cechą konstytutywną. Aby archiwa i zawarte w nich pliki mogły stać się częścią tego systemu, muszą zostać w odpowiedni sposób udostępnione, zaś niezawodność dostępu

i bezterminowa trwałość są warunkiem bogactwa odesłań i relacji, jakie będą się budować w sieci. Zniknięcie lub przemieszczenie pliku jest w tym kontekście prawdziwą katastrofą, ponieważ dotyczy nie tylko oryginalnego zasobu, ale też wszystkich zasobów z nim połączonych. Świadomość tego jest, niestety, bardzo nikła, co w rezultacie torpeduje szersze stosowanie łączenia i zagnieżdżania. Dlaczego jest to ważne?

W realizowanych w ostatnich latach projektach (szczególnie dotyczących dużych platform informacyjnych) można zaobserwować marnotrawstwo środków, związane z wielokrotnym digitalizowaniem tych samych obiektów i wielokrotnym kopiowaniem zasobów. To drugie wydawałoby się pozytywne, ale niestety nie jest, ponieważ każda z kopii funkcjonuje autonomicznie, jest więc odrębnie archiwizowana, inwentaryzowana i udostępniana, co pociąga za sobą kolejne koszty. W rezultacie współistnieją w sieci lepsze i gorsze kopie, kopie mniej lub bardziej kompletne, lepiej lub gorzej opisane, bez żadnej gwarancji, że te lepsze są bardziej dostępne. Nie dość zatem, że wytwarzany jest bałagan, to za ten bałagan płacimy z własnej kieszeni, i to wielokrotnie. Łączenie i zagnieżdżanie nie jest antidotum na każdy rodzaj bałaganu, ale eliminując niepotrzebne powielanie, sprzyja zachowaniu większego porządku i podnosi pozycję zasobów wartościowych w rankingach wyszukiwawczych. Łączenie i zagnieżdżanie plików prezentacyjnych jest stosunkowo łatwe i szeroko stosowane, ponieważ posiadają one stabilne adresy URL. Niestety, nie dotyczy to archiwów i plików master, co sprzyja niepotrzebnym powtórzeniom digitalizacji.

Postulaty

W podsumowaniu przedstawiam postulaty, które wydają się ważne dla instytucji GLAM (galerie, biblioteki, archiwa, muzea). Jeszcze raz trzeba zaznaczyć, że postulaty te dotyczą instytucji budżetowych, konsumujących środki publiczne (podatników polskich i europejskich). Jakie one są?

1. Długoterminowe archiwa nie powinny realizować jedynie swojej funkcji podstawowej, ale również umożliwiać udostępnianie do sieci plików master w celu ich ponownego wykorzystania lub automatycznego łączenia z innymi zasobami i systemami.
2. Koszty infrastruktury służącej takim funkcjom przekraczają możliwości finansowe i techniczne instytucji budżetowych z uwagi na duże rozmiary archiwów i wymaganą wydajność udostępniania. Powinny zatem powstać mechanizmy stabilnego finansowania takich archiwów, bez względu na ich formę organizacyjną.
3. Niezbędna jest koordynacja budowy długoterminowych archiwów wraz z opracowaniem podstawowych standardów merytorycznych i technicznych oraz mechanizmów egzekwowania tych standardów dla utrzymania rzeczywistej dostępności. Instytucje deponujące zasoby nie mogą się martwić ani o ich fizyczną trwałość, ani kompatybilność ze zmieniającymi się technologiami odczytu.

Skoordynowana budowa archiwów powinna zracjonalizować projekty digitalizacyjne, zoptymalizować koszty zabezpieczenia ich dorobku oraz zapobiec degradacji lub utracie zasobów z powodu upływu czasu i zmian technologicznych. Oczywiście te cele nie zostaną osiągnięte

ani szybko, ani łatwo, jednak należałoby do nich zmierzać, zdając sobie sprawę, że częściowe rozwiązania dadzą częściowe efekty, zaś wiele zagadnień warunkuje się wzajemnie w zasadniczy sposób (bez dostępnych archiwów nie ma mowy ani o skutecznym re-use, ani o stosowaniu dynamicznego łączenia danych).

Radwański, A. Konteksty długoterminowej archiwizacji. *Biuletyn EBIB* [on-line] 2014, nr 9 (154), Gromadzenie i zabezpieczanie danych cyfrowych. [Dostęp 20.01.2015]. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/304. ISSN 1507-7187.