EBIB Biuletyn EBIB, nr 9 (162)/2015,
IT w bibliotece
Artykuł

Tomasz Parkoła
Poznańskie Centrum Superkomputerowo-Sieciowe
tparkola@man.poznan.pl

Masowa digitalizacja obiektów z wykorzystaniem pakietu oprogramowania DInGO

Streszczenie: W niniejszym artykule opisano pakiet oprogramowania DInGO, który wspomaga realizację projektów masowej digitalizacji i jest od wielu lat rozwijany przez Poznańskie Centrum Superkomputerowo-Sieciowe. Pakiet DInGO realizuje funkcje związane z udostępnianiem i długoterminowym przechowywaniem obiektów dziedzictwa kulturowego i/lub naukowych, a także z zarządzaniem procesem digitalizacji. Oprogramowanie z pakietu DInGO umożliwia pełne monitorowanie i kontrolowanie przebiegu prac digitalizacyjnych, włączając w to: automatyzację niektórych czynności, zapewnianie jakości przetwarzanych informacji, czy ustandaryzowaną archiwizację danych. W skład pakietu DInGO wchodzą systemy dLibra, dMuseion, dLab oraz dArceo. Dwa pierwsze dostarczają funkcje pozwalające na udostępnianie obiektów cyfrowych poprzez strony internetowe. System dLab pozwala na zarządzanie procesem digitalizacji, natomiast system dArceo służy do długoterminowego przechowywania danych źródłowych.

Słowa kluczowe: digitalizacja; biblioteki cyfrowe, Pakiet oprogramowania DInGO

Wprowadzenie

Biblioteki cyfrowe w Polsce rozwijają się już od wielu lat. Po powstaniu pierwszej regionalnej biblioteki cyfrowej w Polsce w 2002 r. działania w zakresie digitalizacji i udostępniania obiektów cyfrowych nabierały coraz większego rozpędu. Kilka lat po uruchomieniu pierwszej biblioteki cyfrowej pojawiły się repozytoria naukowe, których celem jest m.in. udostępnianie efektów prac prowadzonych w ramach uczelni. W kolejnych latach, kiedy pojawiło się obszerne finansowanie digitalizacji z funduszy unijnych i krajowych, okazało się, że niezbędne są rozwiązania z zakresu profesjonalnej obsługi procesu digitalizacji, w szczególności tych prowadzonych w dużej skali (tzw. procesy masowej digitalizacji)[1]. W efekcie obecnie w Polsce funkcjonuje ponad 130 bibliotek cyfrowych, które udostępniają ponad 2 000 000 obiektów cyfrowych[2].

1 Comité des Sages, The New Rainessance [on-line]. Dostępny w: https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/final_report_cds_0.pdfdf.

2 LEWANDOWSKA, A., MAZUREK, C., WERLA, M. Enrichment of European Digital Resources by Federating Regional Digital Libraries in Poland. W: Research and Advanced Technology for Digital Libraries. 12th European Conference, ECDL 2008, Aarhus, 14–19 September 2008, Proceedings. LNCS vol. 5173, s. 256–259.

W obszarze digitalizacji mamy do czynienia z kilkoma kluczowymi elementami:

- Udostępnianie obiektów cyfrowych. To najważniejszy element procesu digitalizacji, który umożliwia użytkownikom korzystanie z informacji.
- Przechowywanie danych źródłowych. Działanie to jest niezmiernie istotne w kontekście zapewnienia dostępności do zasobów cyfrowych w długiej perspektywie czasu, bez względu na zachodzące zmiany technologiczne.
- Monitorowanie i kontrolowanie przebiegu prac, tak by zapewnić odpowiednią jakość i efektywność całego procesu digitalizacji.

Poznańskie Centrum Superkomputerowo-Sieciowe (PCSS), jako jednostka badawczo-rozwojowa, wspiera wyżej wymienione działania poprzez implementację i utrzymywanie narzędzi wspomagających prace w obszarze udostępniania i przechowywania danych cyfrowych. Narzędzia te wchodzą w skład pakietu oprogramowania Digitise and GO" (DInGO), który jest kompleksowym rozwiązaniem dla instytucji prowadzących prace w zakresie digitalizacji, przechowywania lub udostępniania danych. Poszczególne elementy pakietu DInGO odpowiadają za określone aspekty funkcjonowania procesu digitalizacji, a w szczególności:

- Oprogramowanie dLibra służy do udostępniania obiektów cyfrowych na potrzeby bibliotek publicznych i uniwersyteckich, zarówno w formie tradycyjnej biblioteki cyfrowej, jak i w formie repozytorium cyfrowego, które działa, np. w trybie zielonej ścieżki Open Access. System dLibra może być również wykorzystywany jako repozytorium danych w firmach komercyjnych czy na potrzeby archiwów.
- Oprogramowanie dMuseion służy do udostępniania obiektów cyfrowych na potrzeby muzeów i galerii, gdzie istotne są wizualne aspekty udostępnianych materiałów oraz specyficzne rozwiązania w zakresie integracji z zewnętrznymi systemami informatycznymi, np. tymi do inwentaryzacji zabytków[3].
- Oprogramowanie dArceo służy do długoterminowego przechowywania danych źródłowych, w szczególności obiektów graficznych, tekstowych oraz audiowizualnych. dArceo przygotowane jest w zgodzie z modelem OAIS i może współpracować z różnymi systemami składowania danych, np. usługą powszechnej archiwizacji w sieci PIONIER[4] [4]. dArceo odpowiedzialne jest za przygotowanie pakietu archiwalnego, jego archiwizację w systemie składowania danych oraz dalsze zarządzanie, np. w kontekście zmiany formatu plików (migracji danych).
- Oprogramowanie dLab służy do zarządzania procesem digitalizacji. Odpowiada za przypisywanie użytkowników do wykonywania poszczególnych etapów prac związanych z digitalizacją, automatyzację niektórych czynności (np. weryfikacją formatów, konwertowanie, kopia bezpieczeństwa) oraz generowanie podsumowań i statystyk związanych z realizowanymi pracami.

3 CZYŻ, P.P., ROMEYKO-HURKO, M. dMuseion: od bazy danych do muzeum cyfrowego. Konferencja „Polskie Biblioteki Cyfrowe", 9 grudnia 2009, Poznań. Materiały konferencyjne, s. 21–29. ISBN 978-83-7712-020-0.

4 BRZEŹNIAK, M. i in. Popular Backup/Archival Service and its Application for the Archival of the Network Traffic in the PIONIER Academic Network. W: Computational Methods in Science and Technology, 2010, spec. iss., p. 109–118.

Wyżej opisane systemy tworzą wspólnie kompleksowe rozwiązanie dla instytucji, które chcą profesjonalnie i efektywnie realizować zadania związane z digitalizacją, archiwizacją i udostępnianiem obiektów cyfrowych. W typowej konfiguracji oprogramowanie dLab odpowiedzialne jest za kontrolowanie procesu digitalizacji, włącznie z wykonywaniem interakcji z pozostałymi elementami pakietu DInGO, np. archiwizacją plików master w systemie dArceo czy wprowadzaniem wersji prezentacyjnej obiektu do systemu dLibra.

Warto podkreślić, że PCSS rozwija pakiet oprogramowania DInGO na bazie doświadczeń zdobytych przy wdrażaniu tychże systemów, doświadczeń z realizacji krajowych i międzynarodowych projektów badawczo-rozwojowych, a także z doświadczeń w obszarze współpracy w ramach europejskiego Centrum Kompetencji IMPACT w zakresie digitalizacji (http://digitisaiton.eu/), gdzie PCSS jest członkiem-założycielem, oraz inicjatywy Open Preservation Foundation (http://openpreservation.org/), w której PCSS jest członkiem stowarzyszonym.

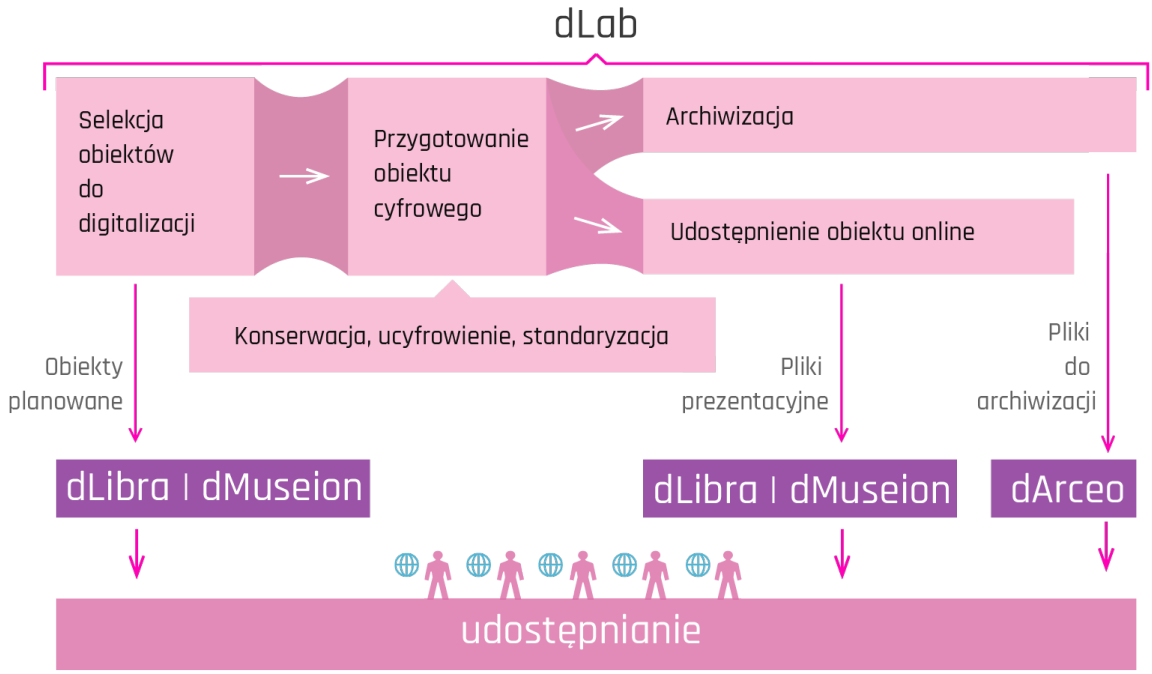
W dalszej części niniejszego artykułu przedstawiono główne założenia systemu dLab oraz zasady jego interakcji z pozostałymi systemami z pakietu oprogramowania DInGO.

1. Zarządzanie procesem digitalizacji w systemie dLab

Zasadniczym celem systemu dLab jest usprawnienie prac związanych z digitalizacją poprzez pomoc pracownikom w wykonywaniu przydzielonych im obowiązków. Cel ten realizowany jest z wykorzystaniem wielu funkcjonalności wbudowanych w system dLab, a w szczególności poprzez:

- przydział pracowników do określonych czynności w ramach procesu digitalizacji poszczególnych obiektów,
- definiowanie przepływu prac związanych z digitalizacją poszczególnych typów obiektów, czyli kto, co i w jakiej kolejności ma wykonać,
- zapewnianie jakości w obszarze poszczególnych etapów realizacji prac digitalizacyjnych, np. poprzez weryfikację manualną lub automatyczną,
- monitorowanie postępów prac z wykorzystaniem systemu raportowania, np. raporty związane z postępem prac w ramach poszczególnych zadań, raporty dotyczące zadań nieukończonych, raporty wydajności poszczególnych etapów przepływu prac,
- automatyzację czynności wykonywanych w ramach procesu digitalizacji, np. generowanie wersji prezentacyjnych, konwertowanie plików, standaryzacja nazewnictwa plików, archiwizacja danych, wykonywanie kopii bezpieczeństwa, wykonywanie OCR, nakładanie znaku wodnego, import danych z zewnętrznego systemu informatycznego, wprowadzenie danych do zewnętrznego systemu informatycznego, kontrola jakości poszczególnych etapów prac.

W kontekście procesu digitalizacji system dLab funkcjonuje jako element zarządzający przepływem prac, który umożliwia integrację zewnętrznych narzędzi informatycznych. Ilustracja nr 1 przedstawia ideę funkcjonowania systemu dLab w obszarze digitalizacji z wykorzystaniem pozostałych komponentów pakietu oprogramowania DInGO.



Il. 1. Zarządzanie procesem digitalizacji – system dLab
Źródło: Poznańskie Centrum Superkomputerowo-Sieciowe

Jak widać na rysunku proces digitalizacji został schematycznie podzielony na cztery elementy, mianowicie: selekcję obiektów do digitalizacji, przygotowanie obiektu cyfrowego, archiwizację oraz udostępnienie obiektu on-line. Całością zarządza system dLab, komunikując się przy realizacji poszczególnych elementów z innymi systemami informatycznymi. Przykładowo etap selekcji obiektów do digitalizacji może zostać zrealizowany z wykorzystaniem tzw. obiektów planowanych w systemie dLibra lub dMuseion[5]. Ten tryb zakłada, że użytkownik definiuje w systemie dLibra obiekty planowane, które w efekcie dają automatycznie początek odpowiadającym im zadaniom digitalizacji w systemie dLab. Zadanie można również utworzyć niezależnie od systemu dLibra. Po utworzeniu zadań digitalizacji w dLab użytkownik może rozpocząć z nimi prace. Proces digitalizacji może wyglądać inaczej w przypadku różnych typów obiektów, np. obiekty wymagające konserwacji mają dodatkową czynność z tym związaną, a te niewymagające jej są przetwarzane bez konserwacji. Sam etap przygotowania obiektu cyfrowego może składać się z wielu czynności związanych z np. skanowaniem, standaryzacją nazewnictwa plików, weryfikacją jakości danych i formatu plików, konwersją plików master do plików prezentacyjnych. Gdy obiekt cyfrowy jest już gotowy, mogą zostać wykonane kolejne czynności, np. archiwizacja oraz udostępnienie obiektu on-line. Warto zauważyć, że etapy te mogą być wykonywane równolegle, tzn. nie ma znaczenia, który z etapów zostanie wykonany pierwszy. Jednak dla zakończenia zadania konieczne jest wykonanie wszystkich etapów i występujących w ich ramach czynności. W typowym wdrożeniu

5 MAZUREK, C., WERLA, M., Digital Object Lifecycle in dLibra Digital Library Framework. Proceedings of DELOS 9th Thematic Workshop: Digital Repositories, 11–3 May, 2005, Heraklion.

całego pakietu DInGO czynność archiwizacji delegowana jest do systemu dArceo, natomiast pliki prezentacyjne obiektu cyfrowego przesyłane są automatycznie przez system dLab do oprogramowania dLibra lub dMuseion. W tym miejscu może nastąpić również automatyczne opublikowanie obiektu na stronach systemu dLibra lub dMuseion. W takim podejściu użytkownik pośrednio steruje zachowaniem zewnętrznych systemów (np. dLibra) z poziomu oprogramowania dLab. Usprawnia to prace i ułatwia weryfikację zakończenia poszczególnych etapów digitalizacji, powiązanych z konkretnymi obiektami cyfrowymi.

2. Kluczowe funkcje systemu dLab

W ramach wdrożenia pakietu oprogramowania DInGO system dLab jest dostosowywany do potrzeb danej instytucji. Oznacza to, że funkcjonalność systemu może być odpowiednio zmodyfikowana do warunków pracy w danej instytucji. Niemniej oprogramowanie dLab charakteryzuje się określonymi właściwościami, które determinują możliwości konfiguracyjne i elastyczność w zakresie dostosowania docelowego wdrożenia. W systemie dLab funkcjonuje kilka kluczowych pojęć związanych z użytkowaniem systemu:

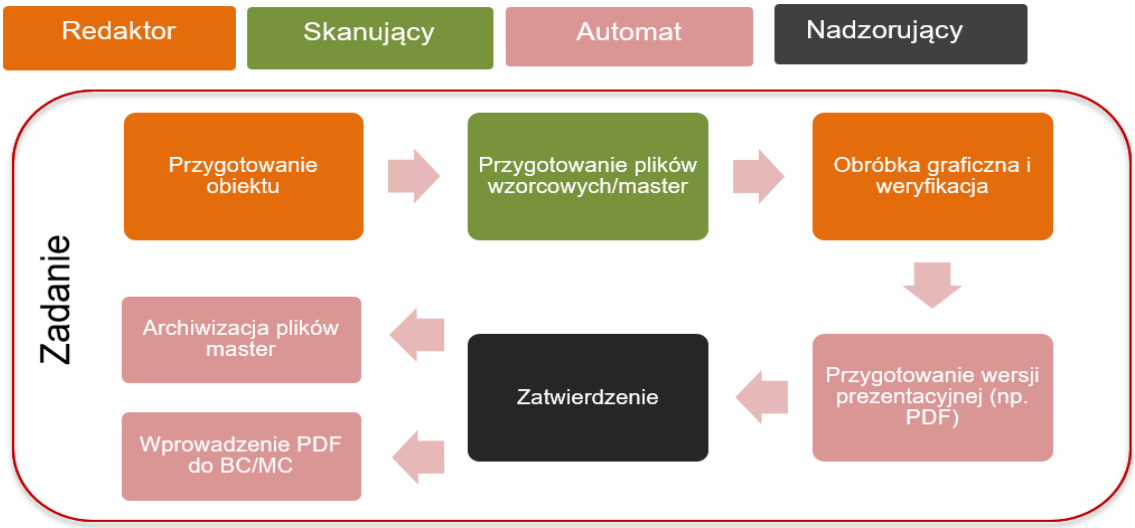
- **Zadanie** to zasadniczy element w systemie, który jest bezpośrednio powiązany z digitalizacją pojedynczego obiektu. Zadanie składa się z zestawu czynności przeznaczonych do realizacji w ramach digitalizacji pojedynczego obiektu (np. książki, numeru czasopisma, nagrania audio). Obiekt rozumiemy tutaj jako element, który po digitalizacji staje się obiektem cyfrowym. System dLab monitoruje realizację (postęp) każdego zadania poprzez wyznaczanie użytkowników i/lub automatyczne narzędzia do wykonywania poszczególnych czynności. Przykładowo czynność „Przygotowanie plików master" wykonywana może być przez użytkowników z grupy „Skanujący", a czynność „Archiwizacja plików master" wykonywana jest przez automat składowania danych w systemie dArceo. Ponieważ niektóre czynności mogą być ograniczone kolejnościowo, wykonywanie zadania przebiega zależnymi od siebie etapami, np. najpierw wykonywana jest czynność „Przygotowanie plików master", a dopiero po niej może być wykonana czynność „Archiwizacja plików master".
- **Czynność** to część składowa zadania, która identyfikuje konkretne działanie do wykonania w ramach całego zadania digitalizacji. Czynność może być wykonana przez uprawnionego użytkownika (człowieka) korzystającego z systemu dLab lub przez automat (narzędzie informatyczne). Przykładem czynności wykonywanej przez użytkownika jest „Przygotowanie plików master". Jest to czynność wykonywana przez użytkownika, ponieważ to użytkownik konfiguruje urządzenie do digitalizacji (np. skaner), a następnie wyniki digitalizacji wprowadza do systemu dLab. Przykładem czynności wykonywanej przez automat może być „Przygotowanie plików prezentacyjnych". Automat może przekonwertować pliki master do wersji prezentacyjnych przy użyciu wskazanego przez użytkownika profilu konwersji. Automat może wykorzystać w tym celu zewnętrzne narzędzie, np. silnik FineReader w celu przygotowania plików PDF (wraz z warstwą tekstową) lub np. oprogramowanie ImageMagick w celu utworzenia galerii JPG. Warto podkreślić, że w zakresie automatyzacji istnieją bardzo szerokie możliwości konfiguracyjne oprogramowania

dLab, tzn. system dLab można zintegrować w zasadzie z dowolnym narzędziem, które można uruchomić w trybie automatycznym czy wsadowym.

- **Grupa zadań** - służy do grupowania zadań, które przy wykonywaniu niektórych czynności powinny być traktowane jako całość. Przykładowo wysyłanie kilkudziesięciu zadań (np. książek) do pracowni digitalizacji może zostać zrealizowane dla wszystkich zadań równocześnie – stratą czasu jest wykonywanie tej czynności kolejno dla każdego zadania. Dlatego możliwe jest utworzenie w tym celu grupy z zadań, które planujemy wysłać w danym transporcie do pracowni digitalizacji i obsługiwać czynność wysłania zadań do pracowni na poziomie zdefiniowanej grupy. Ponieważ, tak jak wspomniano, grupy zadań mają sens tylko w przypadku niektórych czynności, każda grupa zadań powiązana jest z konkretnymi typami czynności.
  W domyślnej konfiguracji systemu dLab istnieją dwa rodzaje grup zadań:
  - **Klocek** – reprezentuje klocek introligatorski, który może zawierać kilka pozycji w jednej oprawie. Pracując z systemem dLab, tworzymy w tym przypadku zadania dla konkretnych pozycji z klocka, a następnie grupujemy te zadania w grupę typu klocek. Z grupą typu klocek powiązana jest czynność przygotowania plików master (skanowanie). Zatem klocek skanowany jest w całości, a po skanowaniu poszczególne strony z wyniku skanowania mogą być przydzielone do poszczególnych zadań zgrupowanych w klocku.
  - **Lista przewozowa** – to lista zadań, które przesyłane są z magazynu do pracowni digitalizacji, a następnie z pracowni digitalizacji do magazynu. W domyślnej konfiguracji istnieją cztery czynności powiązane z listą przewozową: przesłanie listy przewozowej do pracowni digitalizacji, przyjęcie listy przewozowej w pracowni, zwrot listy przewozowej z pracowni digitalizacji do instytucji oraz przyjęcie zwrotu. Wyszczególniono aż cztery czynności, aby móc w pełni monitorować przepływ dokumentów między magazynem dokumentów/obiektów a pracownią digitalizacji.

W kontekście powyższych definicji przedstawiony został rysunek 2. Na rysunku przedstawiono czynności, które składają się na zadanie digitalizacji w systemie dLab. Dla uproszczenia przyjęto, że zadanie składa się z siedmiu czynności. Ograniczenia kolejnościowe wyznaczają strzałki między czynnościami. Oznacza to, np. że czynności „Obróbka graficzna i weryfikacja” musi być wykonana przed czynnością „Przygotowanie plików wzorcowych/master”. Równocześnie oznacza to, że czynność „Archiwizacja plików master” oraz czynność „Wprowadzenie PDF do BC/MC” są od siebie niezależne i jak tylko czynność „Zatwierdzenie” zostanie wykonana, to mogą być wykonywane w dowolnej kolejności. Poza ograniczeniami kolejnościowymi system dLab zapewnia, że poszczególne czynności będą wykonywane przez wcześniej przypisane do nich osoby lub automaty. Przykładowo, jak wskazują kolory na rysunku, można sobie wyobrazić sytuację w której redaktorzy odpowiadają za wykonanie czynności „Przygotowanie obiektu” oraz „Obróbka graficzna i weryfikacja”, osoby skanujące odpowiedzialne są za wykonywanie czynności „Przygotowanie plików wzorcowych/master”, natomiast automat za wykonywanie czynności automatycznych, tzn. „Przygotowanie wersji prezentacyjnej (np. PDF)”, „Archiwizacja plików master” oraz „Wprowadzenie PDF do BC/MC”. Do tego osoba nadzorująca prace odpowiedzialna jest za wykonanie czynności „Zatwierdzenie”. Takie rozłożenie odpowiedzialności w syste-

mie dLab skutkuje tym, że poszczególne osoby mają swoje czynności do wykonania w procesie digitalizacji i możliwe jest monitorowanie stopnia zaawansowania każdego z przetwarzanych zadań. Odpowiedzialność za wykonywanie poszczególnych czynności spoczywa na użytkownikach, którzy mają uprawnienia do jej wykonywania.



II. 2. Przykład zadania digitalizacji i jego składowych czynności
Źródło: Poznańskie Centrum Superkomputerowo-Sieciowe

Poza ograniczeniami kolejnościowymi czynności oraz przydziałem użytkowników lub automatów do ich wykonywania możliwe są również:

- oznaczanie czynności określonymi właściwościami, np. czy czynność jest wymagana w każdym procesie digitalizacji, czy czynność jest możliwa do pominięcia w szczególnych przypadkach lub czy czynność jest półautomatyczna,
- określenie stanu poszczególnych czynności w kontekście ich wykonania w ramach procesu digitalizacji. Dozwolone są następujące stany: W oczekiwaniu, Do wykonania, W trakcie wykonania, Wykonana, Zaakceptowana, Do ponownego wykonania, Odmowa wykonania, Pominięta, Usunięta, Cofnięta.

W przypadku narzędzi automatycznych w systemie dLab domyślnie dostępnych jest kilkanaście automatów, które mogą być wykorzystane w ramach wdrożenia systemu dLab. Możliwe jest również opracowanie nowych automatów, które będą integrowały z systemem dLab specyficzne narzędzia lub systemy informatyczne. W ramach domyślnych narzędzi wymienić można m.in.:

- automat generujący z plików master (lub z plików po obróbce graficznej) obiekt prezentacyjny w postaci galerii obrazów JPG. Automat wykorzystuje w tym celu oprogramowanie ImageMagick,
- automat generujący z plików master (lub z plików po obróbce graficznej) obiekt prezentacyjny w postaci pliku PDF lub DjVu wraz z wersją tekstową. Dostępne są tutaj automaty integrujące się z oprogramowaniem FineReader oraz Document Express,

- automat generujący z plików master w formacie ODT obiekt prezentacyjny w postaci pliku PDF. Automat ten wykorzystuje w tym celu oprogramowanie LibreOffice,
- automat, który nanosi określony znak wodny na pliki graficzne, np. na pliki po obróbce graficznej, które w dalszej kolejności są wykorzystywane do tworzenia wersji prezentacyjnych.

3. Podsumowanie

Szeroko zakrojone prace w zakresie digitalizacji wymagają odpowiednich procedur, sprzętu i oprogramowania, aby w profesjonalny sposób zrealizować wyznaczone cele. Pakiet oprogramowania DInGO tworzony jest dla instytucji prowadzących prace związane z digitalizacją. W skład pakietu DInGO wchodzą systemy dLibra, dMuseion, dLab oraz dArceo. Oprogramowanie dLab służy do zarządzania procesem digitalizacji i może wykorzystywać zewnętrzne narzędzia informatyczne w celu automatyzacji niektórych czynności. dLab posiada wiele możliwości konfiguracyjnych, w tym możliwość definiowania niezależnych ścieżek digitalizacji, dostosowanych do potrzeb danej instytucji. Oprogramowanie dArceo umożliwia długoterminową archiwizację danych źródłowych, natomiast systemy dLibra i dMuseion umożliwiają udostępnianie wersji prezentacyjnych obiektów cyfrowych.

Obecnie w Polsce pełen pakiet oprogramowania DInGO jest wykorzystywany przez kilkadziesiąt różnych instytucji. Instytucje te, choć realizują podobne cele, mają bardzo zróżnicowane wewnętrzne procedury digitalizacji. Świadczy o tym fakt, że do tej pory nie było dwóch takich samych wdrożeń pakietu oprogramowania DInGO. Oznacza to również, że podejście zastosowane przy projektowaniu oprogramowania DInGO, a w szczególności systemu dLab, było słuszne, a indywidualne traktowanie procesów digitalizacji w różnych instytucjach lub projektach uzasadnione.

Bibliografia:

1. Comité des Sages, The New Rainessance [on-line, dostęp 13.01.2016]. Dostępny w: https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/final_report_cds_0.pdf.
2. LEWANDOWSKA, A., MAZUREK, C., WERLA, M. Enrichment of European Digital Resources by Federating Regional Digital Libraries in Poland. W: Research and Advanced Technology for Digital Libraries. 12th European Conference, ECDL 2008, Aarhus, 14–19 September 2008, Proceedings. LNCS vol. 5173, s. 256–259.
3. CZYŻ, P.P., ROMEYKO-HURKO, M. dMuseion: od bazy danych do muzeum cyfrowego. Konferencja „Polskie Biblioteki Cyfrowe", 9 grudnia 2009, Poznań. Materiały konferencyjne, s. 21–29. ISBN 978-83-7712-020-0.
4. BRZEŹNIAK, M. i in. Popular Backup/Archival Service and its Application for the Archival of the Network Traffic in the PIONIER Academic Network. W: Computational Methods in Science and Technology, 2010, spec. iss., p. 109–118.
5. MAZUREK, C., WERLA, M., Digital Object Lifecycle in dLibra Digital Library Framework. Proceedings of DELOS 9th Thematic Workshop: Digital Repositories, 11–3 May, 2005, Heraklion.

Parkoła, T. Masowa digitalizacja obiektów z wykorzystaniem pakietu oprogramowania DInGO. *Biuletyn EBIB* [on-line] 2015, nr 9 (162), *IT w bibliotece.* [Dostęp 25.01.2016]. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/397. ISSN 1507-7187.