XIV.263.3.2019 Załącznik nr 1a do SIWZ i nr 1 do OPZ

Projekt techniczny

System archiwizacji

Uwaga: niniejszy dokument powstawał w czasie tworzenia Systemu Archiwizacji. Stwierdzenia opisujące funkcjonalności i rozwiązania, sformułowane w czasie przyszłym lub z użyciem określenia „powinien” odnoszą się do elementów obecnie zrealizowanych i należy je traktować jako obowiązujące.

Spis treści

[2 Słownik 6](#_Toc514702105)

[3 Informacje ogólne 8](#_Toc514702106)

[3.1 Diagram poglądowy 8](#_Toc514702107)

[3.2 Przypadki użycia 10](#_Toc514702108)

[3.3 Obiekt jako podstawowa encja biznesowa 11](#_Toc514702109)

[3.3.1 Obiekt 11](#_Toc514702110)

[3.3.2 Kontent 12](#_Toc514702111)

[3.3.3 Segmentacja 13](#_Toc514702112)

[3.4 Wolumetria danych 13](#_Toc514702113)

[4 Procesy biznesowe 14](#_Toc514702114)

[4.1 Archiwizacja obiektu (proces P1) 14](#_Toc514702115)

[4.2 Ponowna archiwizacja obiektu (P2) 14](#_Toc514702116)

[4.3 Archiwizacja metadanych (P3) 15](#_Toc514702117)

[4.4 Weryfikacja stanu kasety (P4) 16](#_Toc514702118)

[4.5 Migracja kasety (P5) 17](#_Toc514702119)

[4.6 Przywrócenie zawartości obiektu (P6) 18](#_Toc514702120)

[4.7 Pozostałe procesy biznesowe (P7) 19](#_Toc514702121)

[4.8 Proces przywrócenia bazy danych (P8) 20](#_Toc514702122)

[5 Architektura 22](#_Toc514702123)

[5.1 Ogólne podsumowanie założeń architektonicznych systemu 22](#_Toc514702124)

[5.2 Przyjęty sposób replikacji 24](#_Toc514702125)

[5.3 Specyfikacja interfejsu programistycznego (API) 25](#_Toc514702126)

[5.4 Środowisko 25](#_Toc514702127)

[5.4.1 Środowisko testowe 25](#_Toc514702128)

[5.4.2 Środowisko produkcyjne 26](#_Toc514702129)

[5.5 Protokoły komunikacyjne 27](#_Toc514702130)

[5.6 Integracja z biblioteką taśmową 28](#_Toc514702131)

[5.7 Opis badań mechanizmu zapisu danych w systemie LTFS 28](#_Toc514702132)

[5.7.1 Partycja logiczna biblioteki taśmowej 28](#_Toc514702133)

[5.7.2 Numery kaset magnetycznych 29](#_Toc514702134)

[5.7.3 Sterowanie położeniem kaset 29](#_Toc514702135)

[5.7.4 Przygotowanie kasety do pracy w systemie plików LTFS 30](#_Toc514702136)

[5.7.5 Podłączenie kasety do systemu operacyjnego jako zasób 31](#_Toc514702137)

[5.7.6 Zapis kasety w systemie plików LTFS 31](#_Toc514702138)

[5.7.7 Przebieg procedury zapisu kasety 32](#_Toc514702139)

[5.8 Aplikacja administracyjna Systemu archiwizacji 32](#_Toc514702140)

[6 Wymagania funkcjonalne Systemu archiwizacji 33](#_Toc514702141)

[6.1 Elementy wspólne 33](#_Toc514702142)

[6.1.1 Dziennik zdarzeń biznesowych 33](#_Toc514702143)

[6.1.2 Parametry systemu 34](#_Toc514702144)

[6.1.3 Powiadomienia 34](#_Toc514702145)

[6.2 Archiwum 34](#_Toc514702146)

[6.2.1 Przyjęcie żądania zarchiwizowania paczki 35](#_Toc514702147)

[6.2.2 Skopiowanie paczki do Archiwum 36](#_Toc514702148)

[6.2.3 Przyjęcie żądania odczytu paczki/paczek archiwalnych 36](#_Toc514702149)

[6.2.4 Wybranie paczek do zapisania na kasetę 36](#_Toc514702150)

[6.2.5 Wybieranie kasety do odczytania 36](#_Toc514702151)

[6.2.6 Sterowanie dostępem do napędu 36](#_Toc514702152)

[6.2.7 Mechanizm pobierania i odkładania kaset 36](#_Toc514702153)

[6.2.8 Zapisanie kasety 36](#_Toc514702154)

[6.2.9 Odczyt paczek archiwalnych oraz weryfikacja sum kontrolnych 36](#_Toc514702155)

[6.2.10 Obsługa dodawania kaset do puli kaset do wykorzystania 36](#_Toc514702156)

[6.2.11 Wykrywanie powtarzalnego błędu zapisu kasety 37](#_Toc514702157)

[6.2.12 Mechanizm monitorujący błędy podczas odczytu kaset 37](#_Toc514702158)

[6.2.13 Mechanizmu ustawienia kasecie statusu PROBLEM 37](#_Toc514702159)

[6.3 Usługi pomocnicze 37](#_Toc514702160)

[6.3.1 Monitorowanie obiektów do zarchiwizowania i inicjowanie archiwizacji 37](#_Toc514702161)

[6.3.2 Uruchamianie procesu archiwizacji metadanych/bazy danych 38](#_Toc514702162)

[6.3.3 Monitorowanie obiektów do przywrócenia i inicjowanie przywrócenia 38](#_Toc514702163)

[6.3.4 Usuwanie plików z przywróconego lub niezredukowanego obiektu 39](#_Toc514702164)

[6.4 Archiwizator 41](#_Toc514702165)

[6.4.1 Przyjęcie żądania zarchiwizowania obiektu 42](#_Toc514702166)

[6.4.2 Przyjęcie żądania zarchiwizowania metadanych/bazy danych 43](#_Toc514702167)

[6.4.3 Przygotowywanie paczki archiwalnej z obiektem 43](#_Toc514702168)

[6.4.4 Przygotowywanie paczki archiwalnej ze zrzutem metadanych/bazy danych 50](#_Toc514702169)

[6.4.5 Przekazanie żądania archiwizacji paczki do instancji Archiwum 52](#_Toc514702170)

[6.4.6 Odebranie informacji o statusie zakończenia zadania archiwizacji paczki w instancji Archiwum wraz z podjęciem kolejnych kroków 53](#_Toc514702171)

[6.4.7 Obsługa zakończenia procesu archiwizacji paczki z obiektem 54](#_Toc514702172)

[6.4.8 Obsługa zakończenia procesu archiwizacji paczki z metadanymi 54](#_Toc514702173)

[6.4.9 Przyjęcie żądania przywrócenia obiektu / udostępnienia paczki archiwalnej 54](#_Toc514702174)

[6.4.10 Przyjęcie od instancji Archiwum żądania odczytu paczki archiwalnej 56](#_Toc514702175)

[6.4.11 Odebranie informacji o statusie zakończenia zadania odczytu paczki archiwalnej wraz z podjęciem kolejnych kroków 57](#_Toc514702176)

[6.4.12 Monitorowanie statusu przywrócenia obiektu 58](#_Toc514702177)

[6.5 Interfejs API Systemu archiwizacji 58](#_Toc514702178)

[6.5.1 Specyfikacja interfejsu API 58](#_Toc514702179)

[6.5.2 Uwierzytelnianie 59](#_Toc514702180)

[6.5.3 Zapisywanie zdarzeń wywołania interfejsu 59](#_Toc514702181)

[6.6 Przywracacz 60](#_Toc514702182)

[6.6.1 Przywrócenie plików do obiektu w Repozytorium 60](#_Toc514702183)

[6.7 Weryfikator 63](#_Toc514702184)

[6.7.1 Statusy kaset w Archiwum 63](#_Toc514702185)

[6.7.2 Wybranie kasety i zlecenie jej odczytu 63](#_Toc514702186)

[6.7.3 Ocena kondycji kasety podczas odczytu danych 63](#_Toc514702187)

[6.7.4 Rozpoczęcie i przeprowadzenie migracji kasety 63](#_Toc514702188)

[6.7.5 Procedura usuwania uszkodzonych kaset z Archiwum 64](#_Toc514702189)

[6.8 Administracyjna aplikacji GUI 64](#_Toc514702190)

[6.8.1 Widok listy napędów w instancji Archiwum 65](#_Toc514702191)

[7 Wymagania funkcjonalne zmian w Repozytorium 65](#_Toc514702192)

[7.1 Zmiany encji obiekt 65](#_Toc514702193)

[7.1.1 Status obiektu w kontekście archiwizacji 65](#_Toc514702194)

[7.1.2 Blokowanie obiektu (tzw. LOCK) 68](#_Toc514702195)

[7.1.3 Status streama w kontekście archiwizacji 68](#_Toc514702196)

[7.1.4 Modyfikacja obiektu w kontekście archiwizacji 69](#_Toc514702197)

[7.1.5 Data ostatniej próby redukcji 70](#_Toc514702198)

[7.1.6 Konfiguracja okresów karencji 70](#_Toc514702199)

[8 Monitorowanie usług Systemu archiwizacji 71](#_Toc514702200)

[8.1.1 Monitorowanie miejsca na macierzach pomocniczych przez Zarządcę 71](#_Toc514702201)

[8.1.2 Monitorowanie miejsca na dyskach twardych 71](#_Toc514702202)

[8.2 Monitorowanie usług 72](#_Toc514702203)

[9 Procedura wyłączenia i włączenia Systemu archiwizacji 72](#_Toc514702204)

[9.1 Wyłączenie Systemu archiwizacji 73](#_Toc514702205)

[9.1.1 Wyłączenie elementów podstawowych Systemu archiwizacji 74](#_Toc514702206)

[9.1.2 Przebieg wyłączenia elementów podstawowych Systemu archiwizacji 74](#_Toc514702207)

[9.1.3 Wyłączenie elementów wspomagających Systemu archiwizacji 75](#_Toc514702208)

[9.1.4 Wyłączenie usługi dodatkowej do monitorowania Systemu archiwizacji 75](#_Toc514702209)

[9.2 Włączenie Systemu archiwizacji 75](#_Toc514702210)

[9.2.1 Włączenie usługi dodatkowej do monitorowania Systemu archiwizacji 76](#_Toc514702211)

[9.2.2 Włączenie elementów wspomagających 76](#_Toc514702212)

[9.2.3 Włączenie elementów podstawowych 76](#_Toc514702213)

[10 Schemat i opis struktury bazy danych 77](#_Toc514702214)

[10.1 Baza danych Archiwizator 77](#_Toc514702215)

[10.2 Baza danych Archiwum 77](#_Toc514702216)

Cel dokumentu

Celem niniejszego dokumentu jest opisanie projektu technicznego projektowanego Systemu archiwizacji wraz ze specyfikacją zmian dot. rozbudowy i modyfikacji systemu Repozytorium Cyfrowe Biblioteki Narodowej związanych z wykorzystaniem Systemu archiwizacji.

# Słownik

Aplikacja webowa do konfiguracji biblioteki taśmowej – istniejąca aplikacja do konfiguracji biblioteki taśmowej.

Baza danych – (występuje na diagramie poglądowym w rozdziale 3.1) zasób służący do przechowywania informacji o obiektach, w tym metadanych.

Biblioteka taśmowa – urządzenie pozwalające na zapis danych na taśmach magnetycznych.

HCP – (występuje na diagramie poglądowym w rozdziale 3.1) sprzętowy klaster komputerów zapewniający funkcjonalność obiektowego przechowywania danych.

Klucz stream’a – oznaczenie/klasyfikacja pliku występująca w Repozytorium przy każdym stream’ie znajdującym się w ramach obiektu w kontencie. Jest nazywany w niniejszym opracowaniu również typem lub rodzajem stream’a/pliku w kontencie.

Macierz – (występuje na diagramie poglądowym w rozdziale 3.1) zasób służący do przechowywania zawartości wymagającej szybkiego dostępu w celach prezentacyjnych (kafelki, miniatury).

Macierz pomocnicza – (występuje na diagramie poglądowym w rozdziale 3.1) zasób służący do przechowywania plików wykorzystywanych przez System archiwizacji, np. w Archiwum do przechowywania odczytywanych/zapisywanych paczek archiwalnych.

Migracja kasety – operacja mająca na celu zapisanie każdej paczki archiwalnej znajdującej się na „migrowanej” kasecie, w dokładnie tej samej wersji, na inną kasetę. Do zapisu na nową kasetę może zostać wykorzystany plik z paczką archiwalną pochodzący z „migrowanej” kasety (o ile weryfikacja jego sumy kontrolnej jest pozytywna), lub też plik może pochodzić z innej kasety.

Napęd – fizyczne urządzenie służące do odczytu danych z kasety magnetycznej oraz zapisu danych na kasetę magnetyczną.

Obiekt – zasób w systemie będący cyfrową wersją Pozycji bibliotecznej, pochodzący z postaci analogowej (poprzez digitalizację) lub cyfrowej (tzw. BornDigital, jeżeli został przekazany od razu w formie cyfrowej). Składa się z metadanych i zawartości (kontentów).

Pozycja biblioteczna (dokument) – obiekt analogowy typu książka, czasopismo, druk lub inna pozycja z księgozbioru Biblioteki Narodowej.

Przywrócenie plików obiektu – proces wywoływany przez interfejs Systemu archiwizacji mający na celu przywrócenie do obiektu w Repozytorium stream’ów (plików) z kontentów.

Ramię robota – fizyczne urządzenie służące do przemieszczania kaset ze slotów do napędu oraz z napędu do slotu.

Redukcja obiektu – usunięcie plików z obiektu po jego zarchiwizowaniu. Wybór plików do usunięcia następuje na podstawie konfiguracji

Repozytorium Cyfrowe BN – (dalej: Repozytorium) system informatyczny wykorzystywany w Bibliotece Narodowej do wsparcia procesu digitalizacji, przechowywania oraz udostępniania pozycji bibliotecznych.

Slot – półka znajdująca się w bibliotece taśmowej, służąca do przechowywania kaset.

System archiwizacji – (występuje na diagramie poglądowym w rozdziale 3.1) produkt niniejszego projektu, narzędzie zapewniające przede wszystkim funkcjonalność archiwizacji oraz przywracania obiektów do Repozytorium.

Taśma magnetyczna – (dalej: taśma lub kaseta) nośnik danych wykorzystywany przez Bibliotekę taśmową. Przykładowymi rozmiarami taśmy magnetycznych: 1,5TB (klasa LTO-5) lub 6TB (klasa LTO-7).

Udostępnienie paczki archiwalnej – proces wywoływany przez interfejs Systemu archiwizacji mający na celu udostępnienie na macierz dyskową pliku paczki archiwalnej z obiektem lub ze zrzutem metadanych bez podejmowania jakichkolwiek dalszych czynności w szczególności przywrócenia danych do Repozytorium.

# Informacje ogólne

## Diagram poglądowy

Poniżej zamieszczono poglądowy diagram przedstawiający system Repozytorium Cyfrowe wraz z projektowanym Systemem archiwizacji.



W ramach Systemu archiwizacji można wyszczególnić następujące komponenty logiczne:

1. Archiwizator

Jeden z głównych komponentów Systemu archiwizacji, którego podstawowymi zadaniami są:

* 1. Przygotowywanie danych do procesu archiwizacji oraz zlecenie Archiwum zapisania archiwizowanych danych na kasetach magnetycznych.
  2. Zlecanie odczytu danych z Archiwum oraz przygotowanie ich do procesu przywracania do Repozytorium.

Korzysta z macierzy pomocniczej jako zasobu do tymczasowego przechowywania plików podczas wykonywania powierzonych mu zadań.

System archiwizacji będzie zawierał jedną instancję Archiwizatora.

1. Archiwum

Jeden z głównych komponentów Systemu archiwizacji, którego podstawowym zadaniem jest zapis oraz odczyt danych z kaset magnetycznych z wykorzystaniem biblioteki taśmowej. Korzysta z macierzy pomocniczej jako zasobu do tymczasowego przechowywania plików podczas wykonywania powierzonych mu zadań.

System archiwizacji będzie zawierał od jednej do wielu instancji Archiwów.

1. Przywracacz

Funkcjonalność, której głównym zadaniem jest przywrócenie plików zarchiwizowanych obiektów do Repozytorium.

Logicznie ta funkcjonalność będzie częścią Archiwizatora.

1. Weryfikator

Funkcjonalność, której głównym zadaniem jest weryfikacja stanu kaset z danymi, przechowywanych w ramach Archiwów.

Logicznie ta funkcjonalność będzie częścią Archiwum.

1. Integracja

Komponent obejmujący niezbędne zmiany w Repozytorium oraz dodatkowe usługi udostępniające Repozytorium funkcjonalności zapewnione przez System archiwizacji.

## Przypadki użycia

Poniżej zamieszczono następujące diagramy przypadków użycia Systemu archiwizacji:

1. Diagram przypadków użycia Systemu archiwizacji.
2. Diagram przypadków użycia zmian w Repozytorium w zakresie obsługi zamówień na pliki.

Z uwagi na to, że System archiwizacji będzie pełnił rolę niezależnego mechanizmu odpowiedzialnego za archiwizację obiektów i metadanych, część przypadków użycia będzie wywoływana przez wewnętrzne zdarzenia w Systemie archiwizacji. Jako że te przypadki użycia będą miały wpływ na Repozytorium, zostały one uwzględnione na poniższym diagramie i oznaczone strzałką skierowaną do Repozytorium.

Poniżej zamieszczono diagram przypadków użycia Systemu archiwizacji.



Poniżej zamieszczono diagram przypadków użycia zmian w Repozytorium w zakresie obsługi zamówień na pliki.

Z uwagi na to, że obsługa zamówień użytkowników będzie asynchronicznie realizować pewne zadania, część zaprezentowanych przypadków użycia będzie wywoływana przez wewnętrzne zdarzenia. Te przypadki użycia będą miały wpływ na użytkownika, dlatego zostały one uwzględnione na poniższym diagramie i oznaczone strzałką skierowaną do użytkownika.



## Obiekt jako podstawowa encja biznesowa

W niniejszym rozdziale opisano aktualną strukturę obiektu w Repozytorium.

### Obiekt

Obiekt jest odzwierciedleniem w systemie konkretnego utworu/publikacji (takiego jak książka, rękopis, pocztówka, zeszyt czasopisma) wchodzącego w skład zasobów Biblioteki Narodowej. Każdy obiekt składa się z:

1. metadanych – istotnych informacji określających cechy danego utworu/produktu,
2. kontentów – zawierających zdigitalizowaną wersję utworu/produktu.

Obiekt może być powiązany relacją z wieloma innymi obiektami. Występuje kilka rodzajów relacji m.in. „następny obiekt”, „poprzedni obiekt”, „cytowany obiekt”.

Obiekt może posiadać wiele obiektów podrzędnych (tzw. podobiekty).

W strukturze obiektu nie są przechowywane miniatury i pliki prezentacyjne piramid tiff wygenerowane na jego podstawie – są one jedynie z nim powiązane.

Poniżej zamieszczono koncepcyjny diagram struktury obiektu.



### Kontent

Każdy kontent może składać się z jednego lub wielu tzw. stream’ów, z czego pierwszy dodany stream jest zawsze tzw. mainStream’em, kolejne streamy są formami pochodnymi mainStreamu, zazwyczaj utworzonymi automatycznie przez system. Każdy stream ma swój jasno określony typ, np. tekst, download\_alto (plik pochodzący z procesu OCR), dowlonad\_fullJPG (wersja JPG tworzona automatycznie dla plików graficznych w formatach bez kompresji).

W szczególnych przypadkach kontent może nie mieć stream’ów, wówczas kontent jest wizualizowany w Repozytorium z użyciem ikony „duszka”. Kontenty takie mogą powstać np. w przypadkach, gdy kontroler stwierdzi brak jednej strony w zdigitalizowanej wersji utworu – tworzy wtedy pusty kontent, który zostanie później uzupełniony przez odpowiednią osobę.

Kontenty, pod względem logicznym, dzielą się na:

1. postaci zbiorcze, ich kolejność zawsze jest zgodna z kolejnością dodawania ich do systemu i nie może być modyfikowana przez użytkownika,
2. kontenty znajdujące się w strukturze paginacyjnej, ich kolejność może być dowolnie zmieniana przez użytkownika; kontenty zbiorcze zawierają dodatkowe atrybuty takie jak układ strony, numer strony, opis strony.

Każdy kontent znajdujący się w strukturze paginacyjnej może posiadać powiązane z nimi kontenty alternatywne. Kontenty te zawierają zazwyczaj tę samą informację, co kontent podstawowy, są jednak ich alternatywną wersją (np. strona książki zawierająca dodatkowo na marginesie ręczne dopiski autora).

W ramach obiektu każdorazowo wybrany musi być dokładnie jeden kontent główny, który używany jest do prezentacji danego obiektu w wynikach wyszukiwania (np. dla książki kontentem głównym zazwyczaj będzie skan okładki). Kontentem głównym nie może być kontent oznaczony jako kontent ukryty.

W szczególnych przypadkach kontent może być przypisany do więcej niż jednego obiektu.

### Segmentacja

Dla kontentów, które są w postaci graficznej system ma możliwość przeprowadzenia automatycznej segmentacji. Dotyczy to np. skanów stron gazet, które dzielone są na artykuły.

W sytuacji, gdy kontent został poddany segmentacji do obiektu dodawany jest dodatkowy podział logiczny na artykuły, a dla każdego artykułu wydzielane są sekcje. Zarówno artykuły jak i sekcje posiadają powiązane metadane opisujące ich zawartość, np. każda sekcja wskazuje konkretne kontenty i koordynaty danej sekcji w plikach graficznych tych kontentów.

## Wolumetria danych

Aktualnie produkcyjna baza danych metadanych zrzuconych w postaci SQL zajmuje około 600GB.

Aktualnie wykorzystywane taśmy typu LTO-5 oraz LTO-7 (LTO - Linear Tape Open).

# Procesy biznesowe

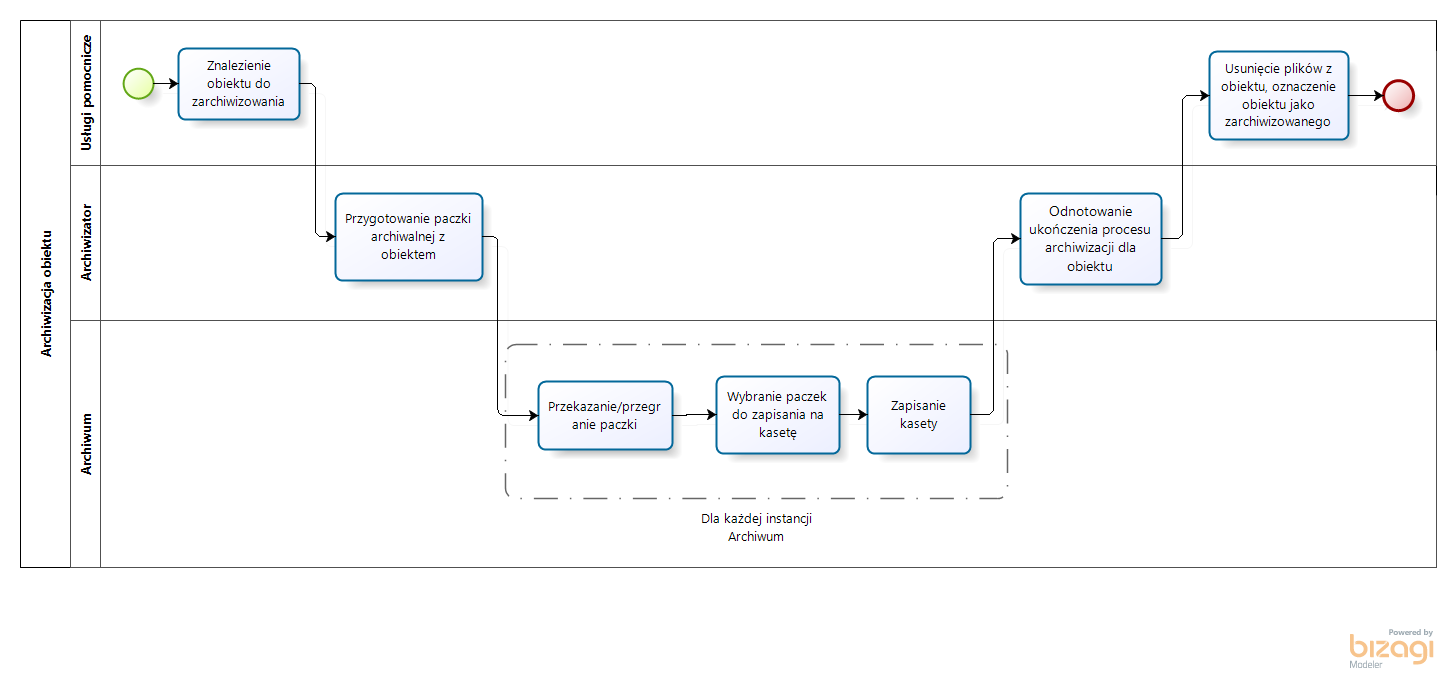
Niniejszy rozdział opisuje procesy biznesowe związane z Systemem archiwizacji.

## Archiwizacja obiektu (proces P1)

Poniżej zamieszczono metrykę procesu.

|  |  |
| --- | --- |
| Nazwa: | Archiwizacja obiektu |
| Kod: | P1 |
| Cel: | Zarchiwizowanie obiektu we wszystkich instancjach Archiwum. |
| Warunki początkowe: | Dostęp do bazy Repozytorium z danymi obiektów |
| Warunki końcowe: | Paczka archiwalna z obiektem zapisana na kasecie.  Usunięte pliki z obiektu. |
| Wykorzystywane zasoby | Macierz pomocnicza aby przechować dane obiektów i przygotować paczkę archiwalną.  Biblioteka taśmowa aby zapisać paczkę archiwalną na kasetę. |
| Uwagi: | Brak. |

Poniżej zamieszczono diagram procesu.

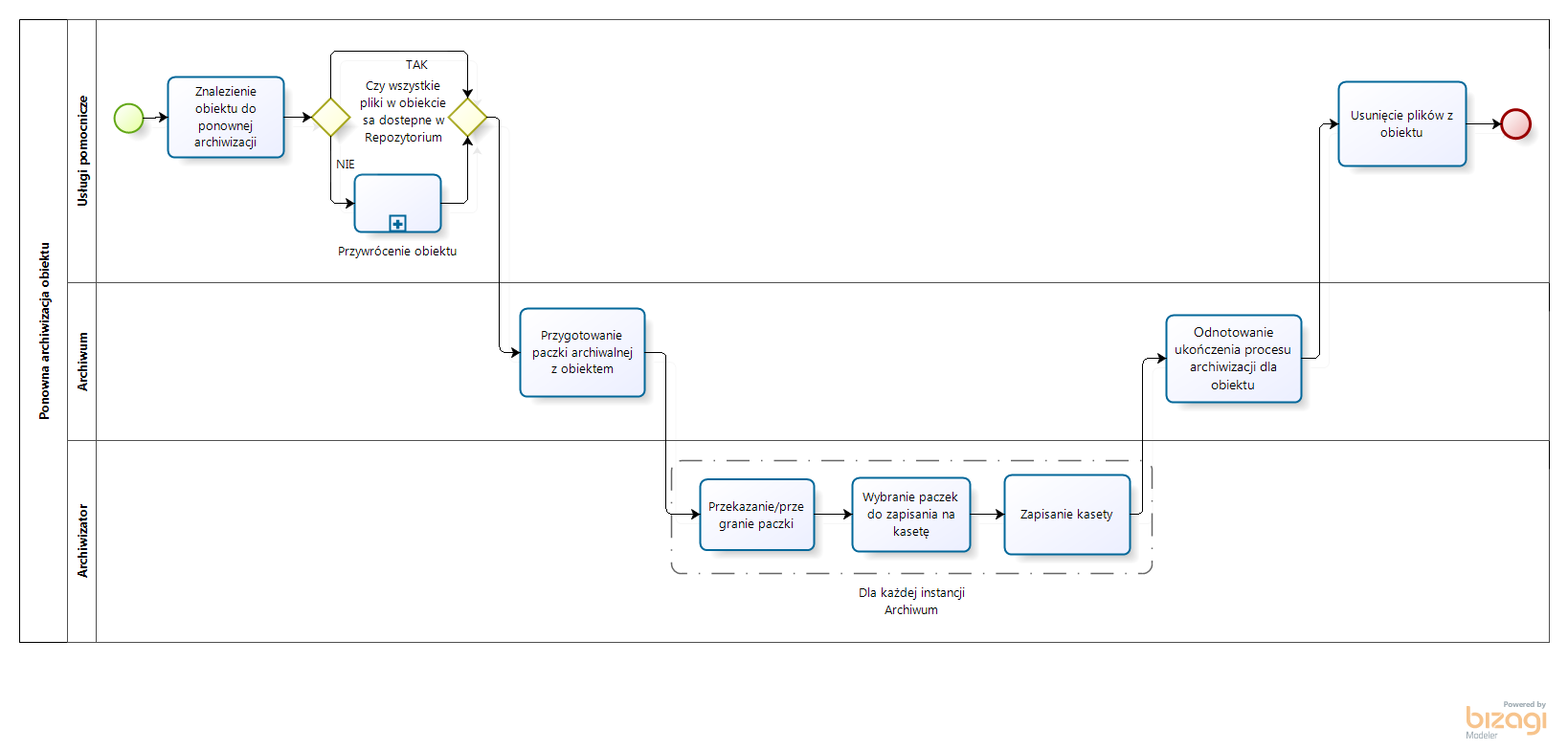


## Ponowna archiwizacja obiektu (P2)

Poniżej zamieszczono metrykę procesu.

|  |  |
| --- | --- |
| Nazwa: | Ponowna archiwizacja obiektu |
| Kod: | P2 |
| Cel: | Ponowne zarchiwizowanie obiektu we wszystkich instancjach Archiwum. |
| Warunki początkowe: | Dostęp do bazy Repozytorium z danymi obiektów |
| Warunki końcowe: | Paczka archiwalna z obiektem zapisana na kasecie.  Usunięte pliki z obiektu. |
| Wykorzystywane zasoby | Macierz pomocnicza aby przechować dane obiektów i przygotować paczkę archiwalną.  Biblioteka taśmowa aby zapisać paczkę archiwalną na kasetę. |
| Uwagi: | Brak. |

Poniżej zamieszczono diagram procesu.

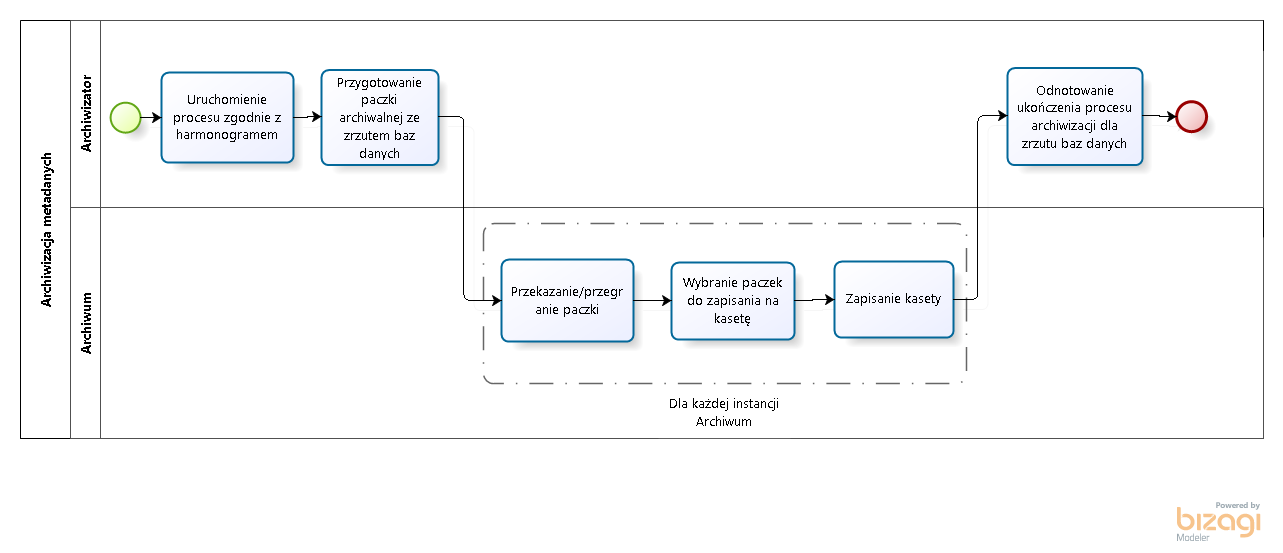


## Archiwizacja metadanych (P3)

Poniżej zamieszczono metrykę procesu.

|  |  |
| --- | --- |
| Nazwa: | Archiwizacja metadanych |
| Kod: | P3 |
| Cel: | Zarchiwizowanie metadanych w postaci paczki archiwalnej zawierającej zrzuty baz danych we wszystkich instancjach Archiwum |
| Warunki początkowe: | Dostęp do bazy Repozytorium z metadanymi obiektów |
| Warunki końcowe: | Paczka archiwalna z obiektem zapisana na kasecie. |
| Wykorzystywane zasoby | Macierz pomocnicza aby przygotować paczkę archiwalną.  Biblioteka taśmowa aby zapisać paczkę archiwalną na kasetę. |
| Uwagi: | Brak. |

Poniżej zamieszczono diagram procesu.



## Weryfikacja stanu kasety (P4)

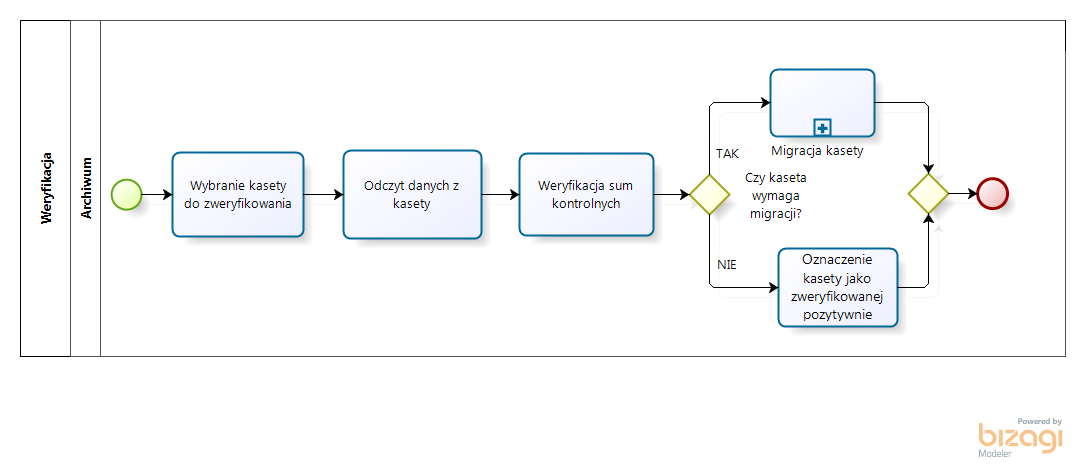
Poniżej zamieszczono metrykę procesu.

|  |  |
| --- | --- |
| Nazwa: | Weryfikacja stanu kasety |
| Kod: | P4 |
| Cel: | Weryfikacja stanu kasety w celu oceny czy kaseta wymaga migracji |
| Warunki początkowe: | Brak. |
| Warunki końcowe: | Jeżeli kaseta wymaga migracji – przekazanie jej do zmigrowania.  Jeżeli kaseta nie wymaga migracji – oznaczenie jako zweryfikowanej. |
| Wykorzystywane zasoby | Biblioteka taśmowa aby odczytać paczki archiwalne.  Macierz pomocnicza aby przechować paczki archiwalne po odczytaniu z kasety. |
| Uwagi: | Brak. |

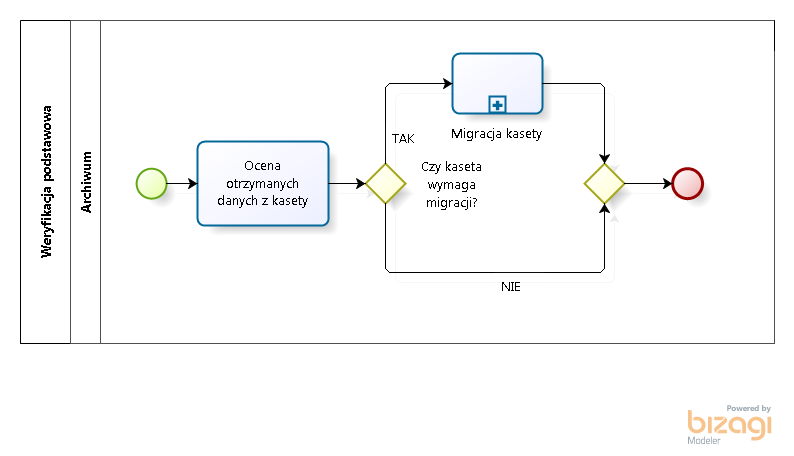
Można wyróżnić dwa rodzaje weryfikacji:

1. Weryfikacja pełna – polegająca na weryfikacji sum kontrolnych plików na kasecie.
2. Weryfikacja podstawowa – polegająca na ocenie stanu kasety w oparciu o metadane otrzymane z napędu przy okazji odczytu kasety.

Poniżej zamieszczono diagram procesu pełnej weryfikacji.



Poniżej zamieszczono diagram podstawowej weryfikacji.

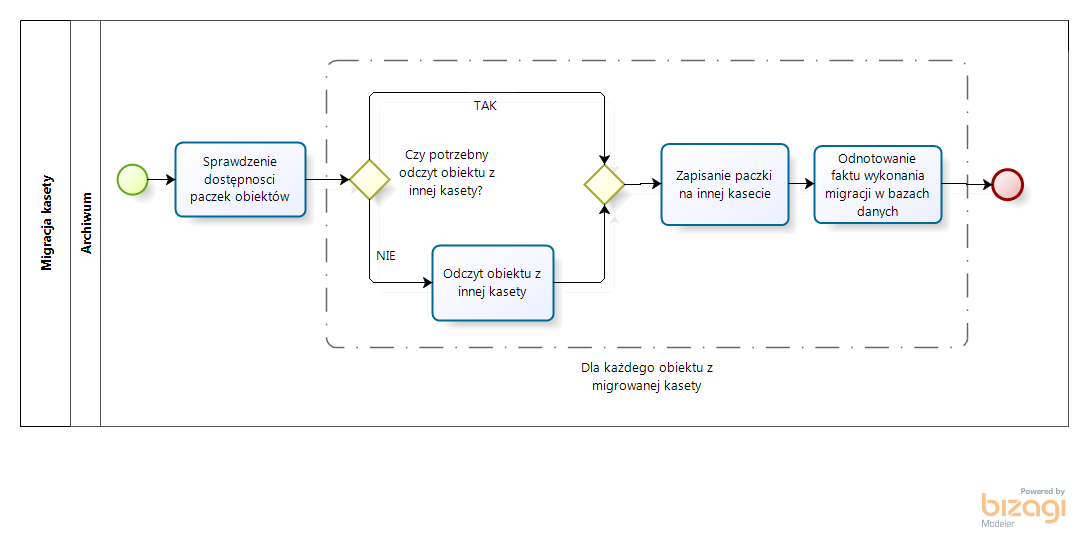


## Migracja kasety (P5)

Poniżej zamieszczono metrykę procesu.

|  |  |
| --- | --- |
| Nazwa: | Migracja kasety |
| Kod: | P5 |
| Cel: | Migracja paczek znajdujących się na kasecie zakwalifikowanej do migracji na nowe kasety. |
| Warunki początkowe: | Brak. |
| Warunki końcowe: | Paczki archiwalne zmigrowane na nowe kasety.  Migrowana kaseta oznaczona jako „do usunięcia” z partycji biblioteki taśmowej. |
| Wykorzystywane zasoby | Biblioteka taśmowa aby odczytać migrowane paczki archiwalne.  Macierz pomocnicza aby przechować paczki archiwalne po odczytaniu z kasety.  Biblioteka taśmowa aby zapisać migrowane paczki archiwalne na kasetę. |
| Uwagi: | Brak. |

Poniżej zamieszczono diagram procesu. *Uwaga na diagramie należy zamienić położenie oznaczeń „TAK”, „NIE”*.

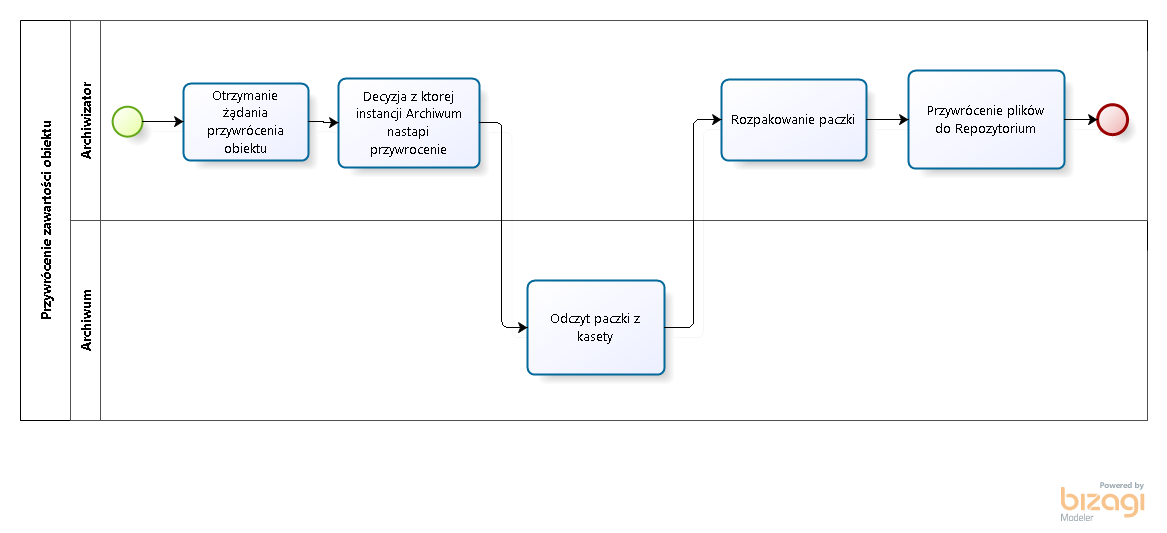


## Przywrócenie zawartości obiektu (P6)

Poniżej zamieszczono metrykę procesu.

|  |  |
| --- | --- |
| Nazwa: | Przywrócenie zawartości obiektu |
| Kod: | P6 |
| Cel: | Przywrócenie zawartości obiektu do Repozytorium cyfrowego |
| Warunki początkowe: | Brak. |
| Warunki końcowe: | Obiekt w Repozytorium z przywróconymi plikami. |
| Wykorzystywane zasoby | Biblioteka taśmowa aby odczytać migrowane paczki archiwalne.  Macierz pomocnicza aby przechować paczki archiwalne po odczytaniu z kasety. |
| Uwagi: | Brak. |

Poniżej zamieszczono diagram procesu.



## Pozostałe procesy biznesowe (P7)

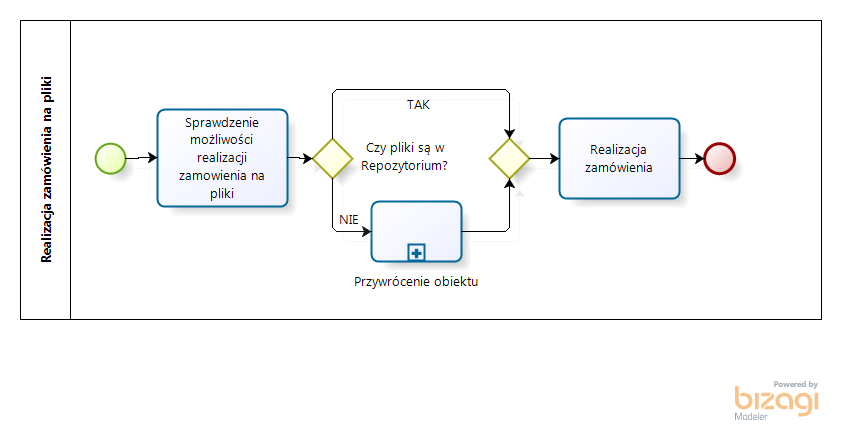
Poniżej została zamieszczona lista pozostałych procesów biznesowych, które będą analizowane w ramach tego projektu:

1. Usunięcie plików ze zarchiwizowanego obiektu, który został przywrócony.

Po przywróceniu plików do obiektu oraz po upłynięciu czasu karencji powinno nastąpić ponowne usunięcie plików ze zarchiwizowanego obiektu

1. Realizacja zamówień użytkowników na pliki z obiektów.

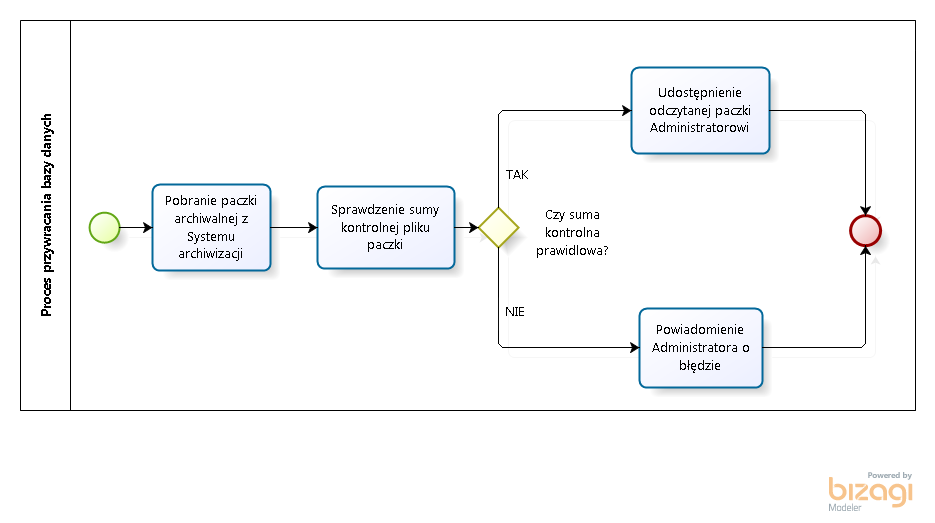
Użytkownicy powinni móc składać zamówienia na pliki z obiektów. W przypadku, kiedy zamówione pliki (lub ich część) nie są dostępne w Repozytorium, powinno następować przywrócenie obiektu, a następnie skompletowanie wszystkich zamówionych plików oraz udostępnienie zamówienia użytkownikowi. Poniżej przedstawiono diagram procesu realizacji zamówienia.



## Proces przywrócenia bazy danych (P8)

Proces archiwizacji powinien obejmować obiekty z Repozytorium oraz okresowo wykonywany zrzut bazy metadanych obiektów.

W przypadku archiwizacji bazy metadanych obiektów proces przywrócenia ich do Repozytorium nie jest przedmiotem niniejszego projektu, ale z uwagi na to, że pojawił się podczas etapu analizy procesowej, został on wstępnie zaprojektowany. Poniżej zamieszczony został koncepcyjny diagram tego procesu.



# Architektura

## Ogólne podsumowanie założeń architektonicznych systemu

System archiwizacji składać się będzie z węzła Archiwizator oraz wielu węzłów Archiwum. Na każdym z węzłów wykorzystywany będzie system kolejkowy służący do wymiany komunikatów sterujących oraz serwer bazy danych przechowujący metadane potrzebne do realizacji zaimplementowanych zadań i procesów.

Głównym założeniem koncepcji jest to, że mikrousługi (tzw. workery) realizujące procesy biznesowe będą wykonywać zadania powierzone im przez właściwego dla tego węzła Zarządcę (na poniższym diagramie architektury: Zarządca Archiwizator i Zarządca Archiwum).

Komunikacja między węzłem Archiwizatora a Archiwum będzie odbywała się z wykorzystaniem systemu kolejkowego Apache ActiveMQ Artemis, który jest następcą systemu Apache HornetQ zastosowanego w Repozytorium Cyfrowym.

Komunikacja między Repozytorium Cyfrowym a Systemem archiwizacji będzie odbywała się z wykorzystaniem interfejsu zaimplementowanego zgodnie z wzorcem architektonicznym REST oraz w niektórych przypadkach poprzez użycie wybranych artefaktów z Repozytorium Cyfrowego i bezpośredni dostęp do zasobów/zdarzeń (jak w przykładzie mikrousługi „MU 3” na poniższym diagramie). Do Systemu archiwizacji artefakty te zostaną dołączone w plikach pom.xml jako dependency lub poprzez wykorzystanie fragmentu kodu pochodzącego z systemu Repozytorium Cyfrowe. W przypadku użycia artefaktów z Repozytorium Cyfrowego należy mieć na uwadze, że w przypadku ich zmodyfikowania w ramach pozostałych, poza Systemem archiwizacji, prac programistycznych (w szczególności prac utrzymaniowych) może wystąpić potrzeba wykonania modyfikacji również po stronie komponentów Systemu archiwizacji wykorzystujących zmienione artefakty. Dlatego w przypadku wykonywania modyfikacji artefaktów będzie istniała potrzeba poinformowania Wykonawcy o planowaniu i wykonaniu takich prac.

Usługi pomocnicze za pomocą artefaktów z Repozytorium Cyfrowego będą łączyły się z bazą danych Repozytorium (np. w celu wyszukiwania obiektów do zarchiwizowania) oraz HCP (w celu wykonania redukcji obiektu).

Wybrane mikrousługi Archiwizatora za pomocą artefaktów z Repozytorium Cyfrowego będą łączyły się z bazą danych Repozytorium (np. w celu zmiany statusu przywracanego obiektu) oraz HCP (np. w celu przywrócenia zawartości obiektu).

Zarządca Archiwizatora za pomocą artefaktów z Repozytorium Cyfrowego będzie łączył się z bazą danych Repozytorium (np. w celu zmiany statusu obiektu), HCP (w celu oszacowania ilości miejsca potrzebnego do zbudowania paczki archiwalnej z obiektem) oraz system kolejkowym Repozytorium (np. w celu poinformowania o możliwości wznowienia procesów po przywróceniu zawartości obiektu).

W celu realizacji procesu archiwizacji Archiwizator będzie się komunikował z działającymi obecnie serwisami Repozytorium Cyfrowego m.in. z DocumentManagerem, JournalServicem, AcademicaMessaging.

Koncepcyjny diagram architektury został zaprezentowany poniżej.



Mikrousługi będą komunikowały się z Zarządcą w danym węźle z wykorzystaniem systemu kolejkowego Apache Artemis. Mikrousługa sterująca Biblioteką taśmową („MU LTO” na powyższym diagramie) w celu zebrania danych potrzebnych do realizacji operacji zapisu i odczytu kaset magnetycznych jako jedyna będzie łączyła się również do bazy danych danego węzła Archiwum.

Oprócz komunikacji oznaczonej na powyższym diagramie, występować będzie również komunikacja między Zarządcą Archiwizatora a bazami danych instancji Archiwów. Wynika ona z funkcjonalności wykonywania zrzutów baz danych w celu przygotowania paczki archiwalnej ze zrzutem metadanych.

Zadaniem przedstawionych na powyższym diagramie usług pomocniczych będzie wykonywanie następujących zadań:

1. Monitorowanie obiektów do zarchiwizowania i inicjowanie archiwizacji
2. Uruchamianie procesu archiwizacji metadanych/bazy danych
3. Monitorowanie obiektów do przywrócenia i inicjowanie przywrócenia
4. Usuwanie plików z przywróconego lub niezredukowanego obiektu

Wybrane usługi pomocnicze będą korzystały z funkcjonalności Systemu archiwizacji, które to funkcjonalności będą udostępnione poprzez programistyczny interfejs API Systemu archiwizacji.

Mikrousługi będą posiadały konfigurowalny zakres zadań, które mogą wykonywać. To oznacza, że jedna mikrousługa może podczas swojego działania wykonywać różne typy zadań, o ile tak zostanie skonfigurowana podczas jej uruchomienia. W ramach konfiguracji podczas uruchomienia mikrousługi będzie można wybrać, które typy zadań będą wykonywane. To rozwiązanie zapewni większą elastyczność i wydajność mikrousług na wybranym odcinku realizowanego procesu biznesowego. W tym rozwiązaniu, aby wybrane zadania nie czekały za długo na realizację, nadawane będą im priorytety uwzględniające typ zadania oraz czas oczekiwania.

Metadane z wykonanych zadań będą przechowywane w bazie danych MySQL, co zapewni ich bezpieczeństwo na wypadek awarii oraz transakcyjność zapisu metadanych odebranych od mikrousługi wykonującej konkretne zadanie.

Każdy worker (za wyjątkiem workera sterującego LTO) będzie podczas wykonywania zadania okresowo wysyłał do Zarządcy identyfikatory zadań nad którymi pracuje. Zarządca, w przypadku nieotrzymania przez skonfigurowany czas sygnału, że zadanie jest wciąż wykonywane, będzie uznawał zadanie za niewykonane i będzie kierował zadanie do ponownego wykonania.

W przypadku, w którym mikrousługa będąc w trakcie wykonywania zadania nieoczekiwanie zakończy swoje działanie, w momencie ponownego uruchomienia tej usługi wszystkie zlecone jej zadania uznaje się za zakończone niepowodzeniem, co skutkuje ponowieniem zlecenia wykonania tych zadań.

## Przyjęty sposób replikacji

Przyjęto, że za zapewnienie replikacji zapisania paczki w kilku bibliotekach taśmowych będzie odpowiadał komponent o nazwie Archiwizator. To Archiwizator będzie miał informacje o istnienia kilku instancji Archiwum oraz będzie weryfikował, czy paczka została skutecznie zapisana we wszystkich instancjach Archiwum.

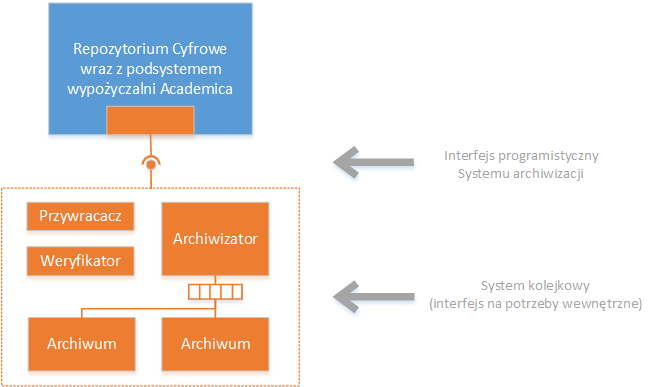
Archiwizator będzie pełnił następujące podstawowe zadania w zakresie replikacji pojedynczej paczki (niezależnie czy zawierającej obiekt czy zrzut bazy metadanych):

1. Przekazanie paczki do wszystkich instancji Archiwum.
2. W przypadku braku możliwości przekazania paczki do wybranej instancji Archiwum (np. z uwagi na niedostępność instancji), Archiwizator będzie ponawiać próbę przekazania paczki do tej instancji Archiwum.
3. Weryfikacja, czy paczka została skutecznie zapisana paczki wszystkich instancjach Archiwum. Dopiero skutecznie zapisanie obiektu we wszystkich instancjach Archiwum jest kryterium zakończenia procesu archiwizacji każdego obiektu.
4. Jeżeli paczka została skutecznie zarchiwizowana, to zainicjowanie czynności, które powinny zostać wykonane po zakończeniu zarchiwizowania danego typu paczki (paczka z obiektem, paczka ze zrzutem bazy metadanych).

## Specyfikacja interfejsu programistycznego (API)

API zostanie udostępnione jako zasób wewnętrzny, dostępny w sieci BN. API zostanie zaimplementowane zgodnie z wzorcem architektonicznym REST w warstwie modułu Archiwizatora, jako warstwa abstrakcji udostępniająca zunifikowany interfejs do Systemu Archiwizacji, zgodnie z poniższym diagramem. Dostęp do API zostanie zabezpieczony tokenem wygasającym. W ramach realizacji projektu zaimplementowany zostanie klient API, zapewniający dostęp do API z poziomu systemu Repozytorium.

Szczegółowy opis interfejsu API znajduje się w rozdziale Interfejs API Systemu archiwizacji.



## Środowisko

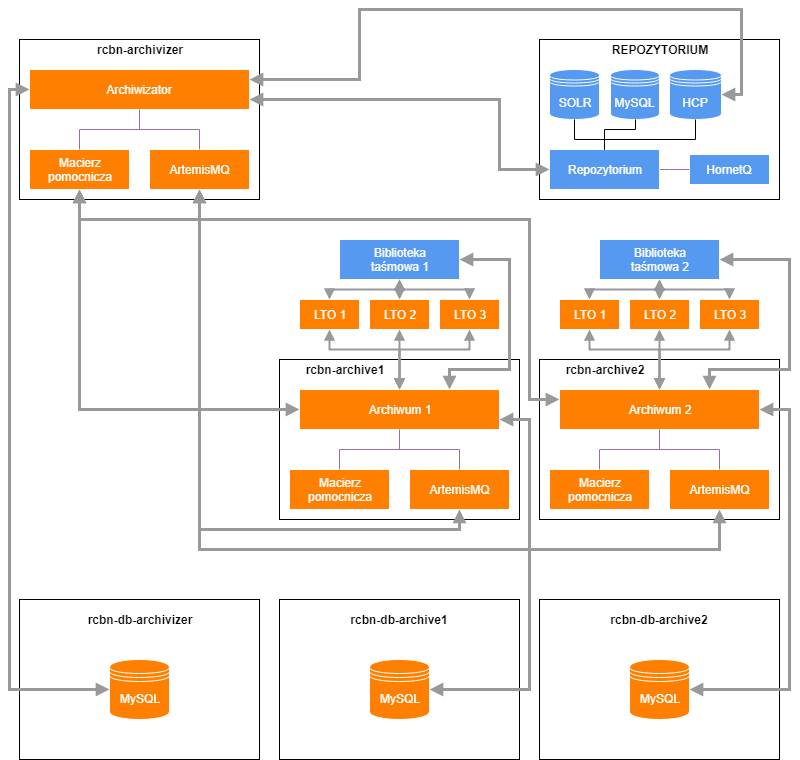
### Środowisko testowe

Do dyspozycji oddane jest jedno środowisko testowe przeznaczone na System Archiwizacji. Poniższy diagram przedstawia schemat tego środowiska.



### Środowisko produkcyjne

Poniższy diagram przedstawia schemat środowiska produkcyjnego przeznaczonego na System Archiwizacji:



## Protokoły komunikacyjne

Poniżej wymienione zostały protokoły komunikacyjne, wykorzystywane do wymiany danych pomiędzy elementami systemu.

1. http – wykorzystywany do implementacji mechanizmów API w modelu REST. Wybór protokołu jest uzasadniony tym, że HTTP jest de facto standardem w zakresie implementacji usług sieciowych, wspieranym przez większość technologii i języków programowania.
2. Apache ActiveMQ Artemis Core, AMQP, Stomp – wykorzystywane przez silnik kolejkowy Apache Apache ActiveMQ Artemis do wymiany komunikatów sterujących oraz ich replikacji. Silnik kolejkowy udostępnia cały zestaw protokołów, pozwalając klientowi wybrać dowolny z nich. Wykorzystywany sterownik programistyczny implementuje warstwę abstrakcji zawierającą obsługę wspieranych protokołów komunikacyjnych oraz wybór preferowanego z nich w sposób niezależny od kodu systemu. Pozwala to na uniezależnienie warstw logiki biznesowej od samego mechanizmu transportu komunikatów oraz otwiera możliwości przyszłej integracji systemu z poziomu praktycznie każdego modułu.
3. JDBC – wykorzystywany do komunikacji usług implementowanych w języku Java z bazą danych MySQL. JDBC jest uznanym standardem komunikacji aplikacji obiektowych uruchamianych za pomocą JVM (Java Virtual Machine) z relacyjnymi bazami danych. Sterowniki JDBC są najczęściej rozwijane przez producentów silników bazodanowych, co zapewnia utrzymanie ich kompatybilności wraz z rozwojem technologii.
4. RSync – wykorzystywany do synchronizacji i kopiowania plików pomiędzy macierzami pomocniczymi. Wybór protokołu Rsync podyktowany jest jego możliwościami automatycznego ponawiania transmisji pojedynczych bloków plików w przypadku zaistnienia błędów sieciowych. W kontekście wymiany dużych plików jest to cecha szczególnie wartościowa, ponieważ pozwala na oszczędność czasu i przepustowości łącz w przypadku konieczności ponowienia transmisji danych.
5. FiberChannel, SMB i inne protokoły sieciowo-plikowe – wykorzystywane do mapowania zasobów dyskowych na poszczególne maszyny wchodzące w skład środowisk zgodnie z możliwościami macierzy dyskowych znajdujących się w infrastrukturze sprzętowej Biblioteki Narodowej; przezroczyste z punktu widzenia kodu aplikacji.

## Integracja z biblioteką taśmową

System będzie obsługiwał procesy zapisu na taśmach i odczytu z taśm umieszczonych w zasobach biblioteki taśmowej za pomocą komend bashowych. Będą one wywoływane z poziomu kodu za pomocą dodatkowej warstwy abstrakcji, która zostanie zaimplementowana w ramach komponentu Archiwum. Integracja z biblioteką taśmową realizowana będzie z maszyny posiadającej bezpośredni dostęp sieciowy do biblioteki taśmowej oraz napędów (niezależnie od środowiska).

## Opis badań mechanizmu zapisu danych w systemie LTFS

### Partycja logiczna biblioteki taśmowej

Z poziomu webowej aplikacji do konfiguracji biblioteki taśmowej istnieje możliwość zdefiniowania kilku logicznych partycji biblioteki taśmowej. W ramach takiej definicji najważniejszymi parametrami są:

1. Liczba przypisanych napędów: na wyłączność lub w trybie „shared”
2. Liczba przypisanych kaset i ewentualnie maksymalnej liczby kaset
3. Liczby przypisanych slotów I/O (opcja, domyślnie wszystkie sloty I/O są wspólne dla wszystkich zdefiniowanych partycji)

W ramach biblioteki taśmowej wykorzystywanej w Bibliotece Narodowej ramię bądź ramiona robota są zasobem wspólnym dla wszystkich partycji logicznych biblioteki taśmowej. Nie ma możliwości przypisania ramienia robota do partycji logicznej biblioteki taśmowej.

### Numery kaset magnetycznych

Numer kasety magnetycznej (Volume Serial) składa się z 6 znaków oraz dwuznakowego oznaczenia klasy kasety. W Bibliotece Narodowej obecnie stosowany jest następujący schemat numerowania kaset:

1. Przedrostek „BN”
2. Właściwy numer składający się z 4 cyfr

Przykładami numerów kaset mogą być: BN0036, BN0037, BN0038.

Dodatkowo za numerem kasety znajduje się oznaczenie klasy danej kasety. W przypadku klasy LTO-5 oznaczenie to posiada wartość „L5”.

Warto podkreślić fakt, że do logicznej partycji biblioteki taśmowej za pomocą webowej aplikacji do konfiguracji biblioteki taśmowej przypisywane są kasety, a nie sloty.

### Sterowanie położeniem kaset

#### Numeracja slotów

Po podłączeniu partycji logicznej do systemu operacyjnego, sloty numerowane są z wykorzystaniem tzw. wirtualnych numerów, przykładowo 1, 2, 3, itd. Wirtualne numery slotów zapewniają użytkownikowi lub systemowi korzystającemu z logicznej partycji biblioteki taśmowej, że do momentu kiedy nie zostanie wydane z poziomu systemu operacyjnego polecenie zmiany lokalizacji kasety, numery te pozostają niezmienne.

Z poziomu aplikacji webowej do konfiguracji biblioteki taśmowej widoczne są fizyczne numery slotów, w których znajdują się kasety. Fizyczne adresy slotów określane są jako „Location”, i są to fizyczne współrzędne oznaczające lokalizacją kasety. Fizyczne adresy slotów składają się z następujących elementów: „Frame”, „Column”, „Row” oraz „Tier”. Jednak sterując położeniem kaset z poziomu systemu operacyjnego należy posługiwać się wirtualnymi numerami slotów.

#### Sterowanie położeniem kaset

Wkładanie oraz wyjmowanie kaset z napędu wykonuje się za pomocą polecenia mtx.

Przykładowe polecenie włożenia kasety ze slotu 1 do napędu zostało zaprezentowanie poniżej:

mtx -f /dev/sg3 load 1 0

Przykładowe polecenie wyjęcia kasety z napędu do slotu 1 zostało zaprezentowanie poniżej:

mtx -f /dev/sg3 unload 1 0

Jest również możliwość wyświetlenia aktualnego stanu partycji logicznej biblioteki taśmowej, co zostało zaprezentowanie poniżej (wraz z przykładową odpowiedzią):

mtx -f /dev/sg3 status | more

  Storage Changer /dev/sg3:1 Drives, 940 Slots ( 255 Import/Export )  
Data Transfer Element 0:Full (Storage Element 1 Loaded):VolumeTag = BN0033L5  
      Storage Element 1:Empty:VolumeTag=  
      Storage Element 2:Full :VolumeTag=BN0034L5  
      Storage Element 3:Full :VolumeTag=BN0035L5  
      Storage Element 4:Full :VolumeTag=BN0036L5  
      Storage Element 5:Full :VolumeTag=BN0037L5  
      Storage Element 6:Full :VolumeTag=BN0038L5  
      Storage Element 7:Full :VolumeTag=BN0039L5  
      Storage Element 8:Full :VolumeTag=BN0040L5  
      Storage Element 9:Full :VolumeTag=BN0041L5  
      Storage Element 10:Full :VolumeTag=BN0042L5

Należy zauważyć, że podczas operacji wkładania i wyjmowania kaset z napędu należy posługiwać się wirtualnymi numerami slotów.

Podczas analizy tego zagadnienia okazało się, że dzięki wykorzystaniu wirtualnych numerów slotów nie występuje potrzeba ręcznego sterowania położeniem kaset w fizycznych slotach w ramach biblioteki taśmowej. Przemieszczenie kasety wynika z potrzeby włożenia kasety do napędu oraz odłożenia kasety po wyjęciu jej z napędu. Sterowanie tymi dwoma operacjami odbywa się poprzez wykorzystanie komendy mtx wraz z podaniem wirtualnych numerów slotów.

### Przygotowanie kasety do pracy w systemie plików LTFS

Przed użyciem kasety z wykorzystywaniem systemu plików LTFS należy ją najpierw sformatować poleceniem mkltfs. Przykładowe polecenie sformatowania kasety zostało zaprezentowane poniżej:

mkltfs --device=/dev/IBMtape3

Sformatowanie kasety z wykorzystywaniem systemu plików LTFS powoduje podział kasety na następujące partycje:

1. Partycja na pliki indeksów, które zawierają informacje o zapisanych plikach na partycji danych.
2. Partycja na dane.

W przypadku kasety klasy LTO-5 posiadającej pojemność 1500GB, partycja na pliki indeksów posiada rozmiar 37,5GB, a partycja na dane 1425GB. Pomiędzy partycjami jest niewykorzystany obszar 37,5GB określany mianem ‘guard wrap’ zapewniający fizyczne odseparowanie tych obszarów na kasecie, aby zminimalizować interferencje podczas zapisywania danych.

### Podłączenie kasety do systemu operacyjnego jako zasób

Podłączenie kasety do systemu operacyjnego jako zasób wykonuje się za pomocą polecenia ltfs. Przykładowe polecenie zostało zaprezentowane poniżej:

ltfs -o devname=/dev/IBMtape0 /mnt/ltfs -o sync\_type=unmount

Podczas operacji podłączenia kasety do systemu operacyjnego jako zasób jest możliwość określenia trybu zapisywania na kasecie pliku indeksu (więcej informacji o indeksach znajduje się w rozdziale „Przygotowywanie kasety do pracy w systemie plików LTFS”). Są możliwe trzy następujące tryby automatycznego zapisywania pliku indeksu na partycji na pliku indeksów:

1. Co określony czas (domyślnie 5 min).
2. Po każdym zapisanym pliku.
3. Brak automatycznego zapisu pliku indeksu, poza zapisaniem indeksu podczas odmontowania kasety.

### Zapis kasety w systemie plików LTFS

W celu uzyskania wysokiej prędkości zapisu kasety, zapis plików na kasetę powinien być wykonywany w trybie ciągłym. To oznacza zapis w taki sposób, aby prędkość zapisu była zdeterminowana jedynie prędkością zapisu na kasetę, a nie przykładowo prędkością odczytu źródłowych plików lub też przestojami wynikającymi z krokowego odczytywania kolejnych segmentów danych a następnie ich zapisywania na kasecie (to znaczy odczyt segmentu, zapis segmentu, odczyt segmentu, zapis segmentu itp.).

Do zapisu danych na kasecie wykorzystującej system plików LTFS w trybie ciągłym służy narzędzie ltfscp. Dodatkowo narzędzie ltfscp zapewnia opcję batchfile, która umożliwia podanie pełnej listy plików, które mają zostać zapisane na kasecie.

Jednakże podczas przeprowadzonych testów okazało się, że w ramach przekazanej za pomocą opcji batchfile listy plików, polecenie ltfscp podczas kopiowania plików na kasetę nie uwzględnia podanej wymaganej kolejności plików.

Dlatego też zapis plików na kasetę musiał zostać zaimplementowany w taki sposób, aby była możliwość wymuszenia kolejności zapisywanych plików (przykładowo od największych do najmniejszych). Została zaimplementowana procedura kopiowania, która działa co do koncepcji analogicznie jak narzędzie ltfscp tzn. używa 2 wątków oraz 2 buforów pamięci tak, aby jednocześnie czytać dane kopiowanego pliku z macierzy pomocniczej i zapisywać go na kasecie zapewniając stały strumień danych podczas kopiowania.

W przypadku mechanizmu odczytu plików z kasety zaimplementowana procedura kopiowania plików sortuje pliki na podstawie odczytanego z indeksu ltfs atrybutu ltfs.startblock, co zapewnia, że pliki zostają odczytane z kasety w kolejności w jakiej zostały zapisane (brak potrzeby cofania kasety podczas odczytu kilku plików w różnej kolejności).

Podczas zapisu kasety może zostać wykorzystana, zapewniana przez napęd, sprzętowa kompresja danych. Obecnie sprzętowa kompresja jest włączona dla napędów wykorzystywanych w Bibliotece Narodowej. W przypadku napędów klasy LTO-5 wynosi ona do 2:1. Podczas przeprowadzonych testów okazało się jednak, że podczas zapisu plików graficznych kompresja ta w zasadzie nie występuje. Chociaż z drugiej strony nie można wykluczyć, że w przypadku niektórych danych kompresja ta wystąpi.

Dodatkowo podczas zapisu danych może wystąpić błąd zapisu w danym sektorze kasety, który powoduje pominięcie przez napęd danego sektora (lub sektorów) oraz kontynuowanie próby zapisu w kolejnym sektorze.

Z uwagi na powody przytoczone w poprzednich dwóch akapitach, przed zapisaniem danych na kasecie nie jest możliwe dokładne określenie ile danych zmieści się na kasecie.

Warto wspomnieć, że w przypadku kiedy podczas zapisu pliku na kasecie napęd wykrywa koniec kasety, zwracany jest błąd „no space left on device”, ale zapisany w części plik jest widoczny na kasecie i posiada taki rozmiar, jaki zmieścił się na kasecie.

### Przebieg procedury zapisu kasety

Procedura zapisania danych na nowej kasecie składa się z następujących kroków:

1. Włożenie kasety do napędu.
2. Sformatowanie kasety wykorzystując system plików LTFS.
3. Podłączenie (zamontowanie) kasety jako zasób w systemie operacyjnym.
4. Zapis kasety.
5. Odłączenie (odmontowanie) kasety.
6. Odłożenie taśmy do slotu.

## Aplikacja administracyjna Systemu archiwizacji

Aplikacja administracyjna będzie hostowana w ramach węzła Archiwizator. Technologią, w jakiej zostanie utworzona aplikacja będzie uwzględniała wykorzystanie frameworków AngularJS i JavaScript, przy założeniu, że końcowy pakiet aplikacyjny będzie udostępniany w formie wykonywalnego pliku .jar, opartego o bibliotekę spring-boot pozwalającą na zagnieżdżenie kontenera webowego w pakiecie aplikacyjnym. Źródłem danych, z którym zintegrowane zostanie środowisko serwera aplikacyjnego będzie nowy schemat w bazie danych MySQL węzła Archiwizator.

Aplikacja będzie gwarantować bezpieczeństwo dostępu do udostępnianych funkcjonalności poprzez mechanizmy autentykacji i autoryzacji oraz mapowanie ról aplikacyjnych (Drives Admin, Users Admin) do odpowiednich kont użytkowników aplikacji administracyjnej Systemu archiwizacji.

Celem gromadzenia kompletnych informacji audytowych i diagnostycznych wykorzystane zostaną techniki logowania zdarzeń z poziomu aplikacji, jak również komponentów serwerowych:

1. log aplikacyjny – diagnostyczny – użycie logback
2. log aplikacyjny – audytowy – użycie logback
3. log systemowy – diagnostyczny – użycie logback (plik archive-admin.log)

Rekomendowane jest zapewnienie w ramach zasobów systemu operacyjnego dedykowanego filesystemu, o wielkości odpowiadającej rozmiarowi wspomnianych powyżej plików, składowanych na nim z określoną retencją. Zalecane rozwiązanie zakłada utworzenie dedykowanych filesystemów zarówno dla logów, jak i binariów aplikacji.

Komunikacja aplikacji administracyjnej z Systemem archiwizacji będzie się odbywać za pośrednictwem REST API. Metody, udostępniane przez System archiwizacji pozwalają na pobranie informacji o napędach, ich statusie oraz wyzwolenie akcji odblokowania wybranego z nich.

# Wymagania funkcjonalne Systemu archiwizacji

Poniższe rozdziały odzwierciedlają przyjęty podział na poszczególne logiczne moduły projektowanego Systemu archiwizacji.

## Elementy wspólne

### Dziennik zdarzeń biznesowych

System będzie zapisywał zdarzenia biznesowe w dedykowanym do tego typu zdarzeń dzienniku zdarzeń. Lista tych zdarzeń zostanie zdefiniowana na etapie analizy wymagań danego komponentu logicznego Systemu archiwizacji.

Każda mikrousługa Systemu archiwizacji będzie posiadała własny dziennik zdarzeń.

Pliki logów będą rotowane.

### Parametry systemu

Poszczególne komponenty Systemu archiwizacji powinny być konfigurowalne za pomocą plików z konfiguracją.

Parametry, które są wspólne dla Systemu archiwizacji oraz Repozytorium powinny być przechowywane w Repozytorium.

Powinna być możliwość aktualizacji wybranych parametrów konfiguracji Systemu archiwizacji bez potrzeby ponownego uruchomienia korzystającego z nich komponentu/elementu systemu przy założeniu, że zaktualizowane wartości parametrów będą uwzględniane od momentu wykrycia przez system nowych wartości. Nie będzie implementowana logika biznesowa modyfikująca istniejące w systemie obiekty (np. istniejące w momencie wczytania nowych parametrów żądania odczytu paczki) w celu dostosowania ich do nowych wartości parametrów.

Wybrane parametry konfiguracji powinny być zmieniane tylko w przypadku wygaszenia i wyłączenia całego Systemu archiwizacji. Pod pojęciem wygaszenia systemu rozumiane jest zrealizowanie wszystkich otrzymanych żądań.

Wskazanie, które parametry są dynamiczne przeładowywane, których zmiana wymaga wyłączenia systemu, oraz których zmiana wymaga wygaszenia i wyłączenia systemu, znajduje się w ramach dokumentacji deweloperskiej (tzw. ReadMe).

### Powiadomienia

#### Powiadomienia do administratorów

W przypadku wystąpienia zdarzenia biznesowego o wysokim stopniu ważności WARNING, system wyśle powiadomienia e-mailowe na zdefiniowane adresy e-mailowe np. Administratorów (parametr w konfiguracji: notification.mail.recipients).

#### Powiadomienia do użytkowników

W przypadku przygotowania zamówienia na pliki (funkcjonalność planowana w ramach zmian w Repozytorium), system wyśle powiadomienie e-mailowe do użytkownika, który złożył to zamówienie.

## Archiwum

Główne realizowane zadania biznesowe – podproces zapisu danych w ramach archiwizacji:

1. Przyjęcie żądania zarchiwizowania paczki
2. Skopiowanie paczki do Archiwum
3. Wybranie paczek do zapisania na kasetę
4. Zapisanie kasety

Główne realizowane zadania biznesowe – podproces odczytu danych:

1. Przyjęcie żądania odczytu paczki/paczek
2. Wybranie kasety do odczytania
3. Odczyt paczki/paczek

Pozostałe główne realizowane zadania biznesowe:

1. Sterowanie dostępem do napędu
2. Obsługa dodawania taśm do puli taśm do wykorzystania

Wymagania biznesowe do modułu Archiwum:

1. Zapis paczek archiwalnych na kasecie powinien odbyć się podczas jednego pobytu w napędzie. Nie powinno być możliwości zapisywania kolejnych paczek archiwalnych podczas kolejnych pobytów kasety w napędzie.
2. W celu efektywnego wykorzystania szybkości zapisu na kasecie, zapis ten powinien odbyć się w trybie ciągłym. To oznacza zapis wybranego na podstawie skonfigurowanego rozmiaru kasety zestawu plików, bez weryfikacji w trakcie zapisu pozostałego miejsca, co spowodowałoby potrzebę przewinięcia kasety do początku w celu zamknięcia sesji kasety.
3. Nie powinno być możliwości kasowania poprawnie zapisanych paczek archiwalnych z kaset.
4. Archiwum powinno móc wykorzystywać wiele napędów w ramach udostępnionej jednej partycji biblioteki taśmowej.
5. Przekazywanie zadań do Archiwum powinno odbywać się wyłącznie przez udostępniony interfejs programistyczny. W ramach realizowanego projektu nie będzie udostępniony graficzny interfejs użytkownika umożliwiający monitorowanie i sterowanie pracą Archiwum.
6. Użycie interfejsu programistycznego do Archiwum wykorzystywanego przez Archiwizator w postaci systemu kolejkowego powinno wymagać uwierzytelniania.
7. Powinna być możliwość skonfigurowania pojemności kaset wykorzystywanych w danej instancji Archiwum. Powinien być mechanizm weryfikujący, czy otrzymana od Archiwizatora paczka nie jest większa niż skonfigurowana pojemność.
8. Powinna być możliwość skonfigurowania maksymalnego czasu oczekiwania paczki archiwalnej przyjętej do instancji Archiwum na zapisanie na kasecie. Po minięciu tego czasu, paczka powinna zostać przekazana do zapisania na kasecie pomimo niespełnienia kryterium stopnia zapełnienia kasety.

### Przyjęcie żądania zarchiwizowania paczki

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Skopiowanie paczki do Archiwum

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Przyjęcie żądania odczytu paczki/paczek archiwalnych

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Wybranie paczek do zapisania na kasetę

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Wybieranie kasety do odczytania

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Sterowanie dostępem do napędu

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Mechanizm pobierania i odkładania kaset

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Zapisanie kasety

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Odczyt paczek archiwalnych oraz weryfikacja sum kontrolnych

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Obsługa dodawania kaset do puli kaset do wykorzystania

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Wykrywanie powtarzalnego błędu zapisu kasety

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Mechanizm monitorujący błędy podczas odczytu kaset

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Mechanizmu ustawienia kasecie statusu PROBLEM

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

## Usługi pomocnicze

W niniejszym rozdziale zostały opisane usługi pomocnicze Systemu archiwizacji, które logicznie mają związek z zadaniami realizowanymi przez Archiwizator.

Mowa o mechanizmach, które powinny uruchamiać: proces archiwizacji obiektu, proces archiwizacji zrzutu bazy metadanych/danych, proces przywracania przywrócenia obiektów oraz proces usuwania plików z obiektu.

Przygotowanie tych mechanizmów jako osobne usługi pomocnicze umożliwia w przyszłości uwzględnienie innego sposobu uruchamiania w/w procesów, przykładowo w sposób ręczny.

Ilekroć w niniejszym rozdziale mowa o statusie obiektu, to jest to status obiektu w kontekście archiwizacji, którego opis znajduje się w rozdziale Zmiany encji obiekt.

### Monitorowanie obiektów do zarchiwizowania i inicjowanie archiwizacji

W ramach tego zadania Archiwizator powinien co określony skonfigurowany czas (parametr w konfiguracji: archivization.scanner.cronExpression) monitorować obiekty w Repozytorium w celu znalezienia obiektów spełniających kryteria archiwizacji.

Pełna lista kryteriów wyboru obiektów do archiwizacji została przedstawiona poniżej:

1. Obiekt posiada status NEW lub ARCHIVED.
2. Data ostatniej modyfikacji w kontekście archiwizacji jest późniejsza niż data ostatniej archiwizacji LUB data ostatniej archiwizacji jest pusta.
3. Od daty ostatniej modyfikacji w kontekście archiwizacji minął skonfigurowany okres karencji (parametr w konfiguracji: opisano w rozdziale Zmiany encji obiekt).
4. inWorkflow=false.

Powyższe warunki wybiorą zarówno obiekty do pierwszej archiwizacji jak i ponownej archiwizacji.

W ramach Repozytorium powinna zostać udostępniona metoda getDocumentIdsToArchive, która powinna zwracać kolekcję obiektów spełniających powyżej opisane kryteria.

Weryfikację okresu karencji od zakończenia przez obiekt przebiegu workflow planowano pierwotnie wykonać w oparciu o datę wykonania ostatniego kroku workflow wykorzystując encję „activity” w bazie danych. Jednakże z uwagi na dużą liczbę elementów w tej tabeli, wyszukiwanie danych mogłoby znacznie obciążać bazę danych Repozytorium. Dlatego weryfikacja okresu karencji od zakończenia workflow powinna zostać wykonana w taki sposób, aby każde wykonanie kroku w workflow aktualizowało datę ostatniej modyfikacji w kontekście archiwizacji. Wówczas wyszukiwanie obiektów do zarchiwizowania powinno zostać oparte o jeden wspólny okres karencji „od ostatniej modyfikacji w kontekście archiwizacji” oraz warunek inWorkflow=false. Dodanie oraz odpowiednia aktualizacja daty ostatniej modyfikacji w kontekście archiwizacji jest przedmiotem prac w ramach zmian w systemie Repozytorium.

Okres karencji od ostatniej modyfikacji w kontekście archiwizacji powinien być konfigurowalny (parametr w konfiguracji: opisano w rozdziale Zmiany encji obiekt).

Po znalezieniu obiektów spełniających kryteria archiwizacji, powinno nastąpić wywołanie przez interfejs Systemu archiwizacji żądania zarchiwizowania tego obiektu.

Podczas analizy kryteriów wyboru obiektów do archiwizacji pierwotnie rozważano rozdzielenie wykrywania obiektu do pierwszej archiwizacji od wykrywania obiektów do kolejnych archiwizacji. Analiza tego aspektu doprowadziła jednak do wniosku, że można zdefiniować jeden wspólny warunek wyszukiwania obiektów, dzięki czemu nie będzie potrzeby dwukrotnego monitorowania obiektów w poszukiwaniu kandydatów do archiwizacji, tym samym dwukrotnie obciążając bazę danych Repozytorium różnymi zapytaniami.

### Uruchamianie procesu archiwizacji metadanych/bazy danych

W ramach tego zadania Archiwizator powinien co określony skonfigurowany czas (parametr w konfiguracji: dbDump.cronExpression) uruchamiać proces archiwizacji metadanych poprzez wywołanie przez interfejs Systemu archiwizacji tego procesu.

### Monitorowanie obiektów do przywrócenia i inicjowanie przywrócenia

W ramach tego zadania Archiwizator powinien co określony skonfigurowany czas (parametr w konfiguracji: restore.scanner.cronExpression) monitorować obiekty w Repozytorium w celu znalezienia obiektów wymagających przywrócenia.

Obiekt powinien zostać zakwalifikowany do przywrócenia jeżeli spełniony jest chociaż jeden poniższy warunek:

1. Nastąpiła modyfikacja zredukowanego obiektu, co powoduje potrzebę przywrócenia oraz ponownej archiwizacji. Warunki zmodyfikowania zredukowanego obiektu są następujące:
   1. Obiekt posiada status REDUCED
   2. Obiekt jest zmodyfikowany: data ostatniej modyfikacji w kontekście archiwizacji jest późniejsza niż data ostatniej archiwizacji
2. Użytkownik złożył zamówienie na usunięte pliki ze zredukowanego obiektu.
3. Proces Repozytorium wymaga usuniętego pliku ze zredukowanego obiekt.

W ramach Repozytorium powinna zostać udostępniona metoda getDocumentIdsToRestore, która powinna zwracać kolekcję obiektów spełniających powyżej opisane kryteria. Dodatkowo metoda ta powinna zwracać informację, czy przywrócenie obiektu posiada wysoki priorytet.

Po znalezieniu obiektów spełniających kryteria przywrócenia, powinno nastąpić wywołanie przez interfejs Systemu archiwizacji żądania przywrócenia tego obiektu

### Usuwanie plików z przywróconego lub niezredukowanego obiektu

W ramach tego zadania powinno co określony skonfigurowany czas (parametr w konfiguracji: reduce.process.cronExpression) być wykonywane monitorowanie obiektów w Repozytorium w celu znalezienia obiektów spełniających kryteria redukcji obiektu oraz powinna być wykonywana próba redukcji tych obiektów.

Pełna lista warunków wyszukiwania obiektów do zredukowania została przedstawiona poniżej:

1. Obiekt posiada status ARCHIVED.
2. Obiekt jest nie jest zablokowany przez Repozytorium (UNLOCKED) oraz nie jest zmodyfikowany (data ostatniej archiwizacji jest późniejsza niż data ostatniej modyfikacji w kontekście archiwizacji).
3. Od daty ostatniego przywrócenia obiektu minął okres karencji lub data ostatniego przywrócenia jest pusta (parametr w konfiguracji: opisano w rozdziale Zmiany encji obiekt).
4. Od daty ostatniej próby redukcji obiektu minął okres karencji, lub data ostatniej próby redukcji jest pusta (parametr w konfiguracji: opisano w rozdziale Zmiany encji obiekt)..

W ramach Repozytorium powinna zostać udostępniona metoda getDocumentIdsToReduce, która powinna zwracać kolekcję obiektów spełniających powyżej opisane kryteria.

Po znalezieniu obiektów spełniających kryteria redukcji dla każdego znalezionego obiektu powinny zostać wykonane następujące kroki (dodatkowo poniżej zamieszczony został diagram tego procesu):

1. Dodatkowa weryfikacja możliwości zredukowania obiektu

W ramach tego kroku powinno nastąpić sprawdzenie, czy dla tego obiektu w Systemie archiwizacji znajduje się paczka archiwalna posiadająca znacznik czasowy późniejszy niż data ostatniej modyfikacji obiektu, czyli [data ostatniej archiwizacji wynikająca ze znacznika czasowego paczki archiwalnej] > [data ostatniej modyfikacji]:

* 1. Jeżeli w Systemie archiwizacji znajduje się paczka archiwalna posiadająca znacznik czasowy późniejszy niż data ostatniej modyfikacji obiektu, to obiekt może zostać zredukowany.
  2. Jeżeli w Systemie archiwizacji nie znajduje się paczka archiwalna posiadająca znacznik czasowy późniejszy niż data ostatniej modyfikacji obiektu, to:
     1. Obiekt nie zostanie zredukowany.
     2. Nastąpi ustawienie obiektowi statusu PROBLEM.
     3. Nastąpi zarejestrowanie zdarzenia w dzienniku oraz wysłanie powiadomienia do administratora.

Sprawdzenie dostępnej wersji paczki archiwalnej powinno zostać wykonane za pomocą API Systemu archiwizacji.

W ramach tego punktu zostanie dodane zdarzenie biznesowe wykrycia sytuacji skierowania obiektu do redukcji przy jednoczesnym braku zarchiwizowania obiektu w odpowiednio nowej wersji.

1. Zmiana statusu obiektu

W ramach tego kroku powinna nastąpić zmiana statusu obiektu na REDUCING.

1. Usunięcie plików z obiektu w Repozytorium

Przeanalizowane powinny zostać tylko te kontenty, które zawierają streamy spełniające kryteria (klucze stream’ów, mime-types plików) zadane w konfiguracji streamów archiwizowanych i usuwanych (parametr w konfiguracji: opisano w rozdziale Konfiguracja archiwizowanych i usuwanych kluczy stream’ów z obiektu).

Plik powinien zostać usunięty ze stream’a jeżeli spełnia następujące kryteria:

1. Kontent, do którego należy stream nie jest współdzielony pomiędzy różne obiekty; LUB jeżeli jest współdzielony, to wszystkie obiekty do których należy dany kontent są zredukowane:
   * 1. Jeżeli wszystkie pozostałe obiekty są zredukowane, to oznacza to, że analizowany obiekt jest ostatnim niezredukowanym obiektem, do którego należy analizowany kontent. W tej sytuacji streamy należące do analizowanego współdzielonego kontentu mogą i powinny zostać usunięte.
     2. Jeżeli chociaż jeden obiekt (z pozostałych obiektów) nie jest zredukowany, to streamy należące do analizowanego współdzielonego kontentu nie powinny zostać usunięte.
2. Stream posiada status ARCHIVED.
3. Plik posiada mime-type, które znajduje się na skonfigurowanej liście usuwanych mime-types plików – liście skonfigurowanej dla stream’ów o danym kluczu.

W przypadku wykonania usunięcia stream’a, powinien on otrzymać status DELETED.

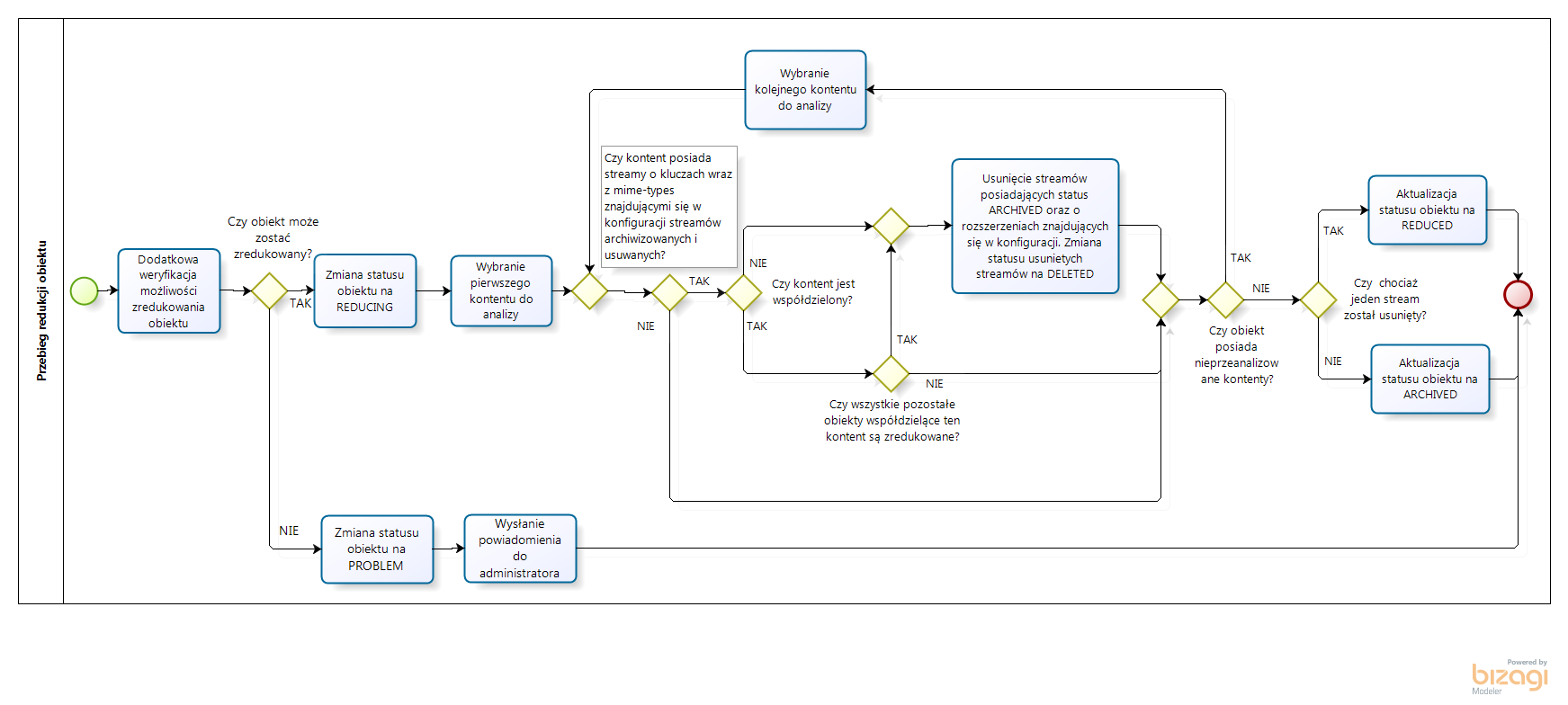
1. Aktualizacja statusu obiektu w Repozytorium

W ramach tego punktu powinno nastąpić zaktualizowanie statusu obiektu:

1. Jeżeli chociaż jeden stream w obiekcie jest usunięty – to zmiana statusu obiektu z REDUCING na REDUCED.
2. Jeżeli ani jeden stream w obiekcie nie jest usunięty – to zmiana statusu obiektu z REDUCING na ARCHIVED.

Dodatkowo po prawidłowym zakończeniu procesu redukcji powinno nastąpić zaktualizowanie w obiekcie daty ostatniej próby redukcji.

Poniżej zamieszczony został diagram tego procesu:



Konfiguracja usuwanych kluczy stream’ów z obiektu

Konfiguracja usuwanych kluczy stream’ów z obiektu została opisana w ramach rozdziału o nazwie Konfiguracja archiwizowanych i usuwanych kluczy stream’ów z obiektu.

## Archiwizator

Główne realizowane zadania biznesowe – właściwy proces archiwizacji:

1. Przyjęcie żądania zarchiwizowania obiektu
2. Przyjęcie żądania zarchiwizowania metadanych/bazy danych
3. Przygotowywanie paczki archiwalnej z obiektem
4. Przygotowywanie paczki archiwalnej ze zrzutem metadanych/bazy danych
5. Przekazanie żądania archiwizacji paczki archiwalnej do instancji Archiwum
6. Odebranie informacji o statusie zakończenia zadania archiwizacji paczki w instancji Archiwum wraz z podjęciem kolejnych kroków
7. Obsługa zakończenia procesu archiwizacji paczki z obiektem
8. Obsługa zakończenia procesu archiwizacji paczki z metadanymi

Główne realizowane zadania biznesowe – proces przywracania:

1. Przyjęcie żądania przywrócenia obiektu / udostepnienia paczki archiwalnej
2. Przyjęcie od instancji Archiwum żądania odczytu paczki archiwalnej
3. Odebranie informacji o statusie zadania odczytu paczki archiwalnej wraz z podjęciem kolejnych kroków
4. Monitorowanie statusu przywrócenia obiektu

Wymagania biznesowe do modułu Archiwizator:

1. Moduł powinien korzystać z dedykowanych usług realizujących elementarne dające się wydzielić czynności, przykładowo z usługi do pakowania/rozpakowania plików.
2. System powinien dbać o relację listy typów plików archiwizowanych i usuwanych z obiektu.
3. Czas oczekiwania na udostępnienie paczki autorowi żądania powinien być konfigurowalny. Po tym czasie podsystem powinien uznać, że nastąpiła sytuacja nietypowa i powinno to spowodować wysłanie powiadomienia do administratora. Nie powinno to spowodować przerwania realizacji żądania przywracania.

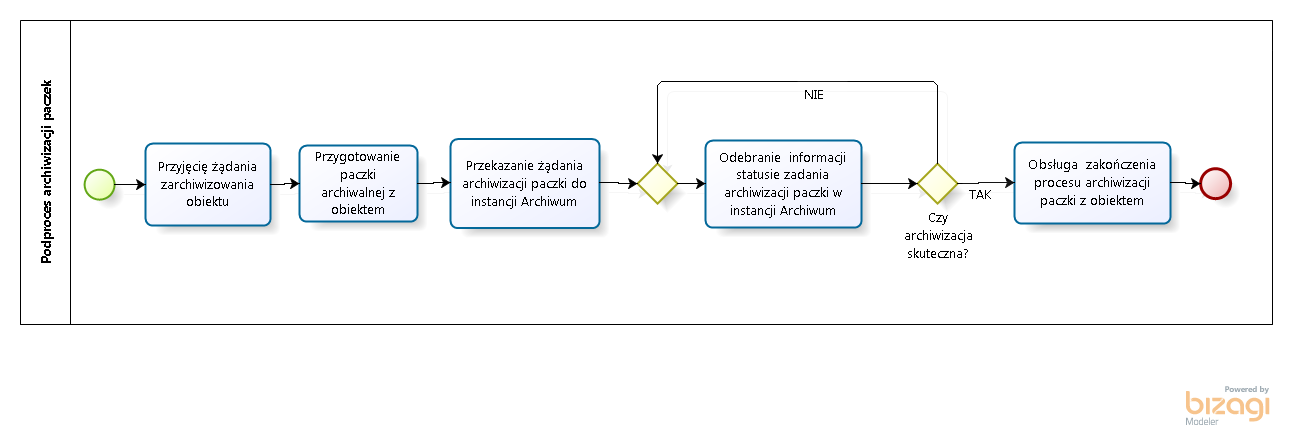
Ilekroć w niniejszym rozdziale mowa o statusie obiektu, to jest to status obiektu w kontekście archiwizacji, którego opis znajduje się w rozdziale Zmiany encji obiekt.

### Przyjęcie żądania zarchiwizowania obiektu

Po przyjęciu żądania archiwizacji obiektu oprócz uruchomienia procesu archiwizacji powinno nastąpić ustawienie obiektowi statusu ARCHIVING.

W przypadku, w którym system odbierze żądanie archiwizacji obiektu, który jest już archiwizowany – powinien zostać zwrócony odpowiedni komunikat lub kod odrzucenia żądania.

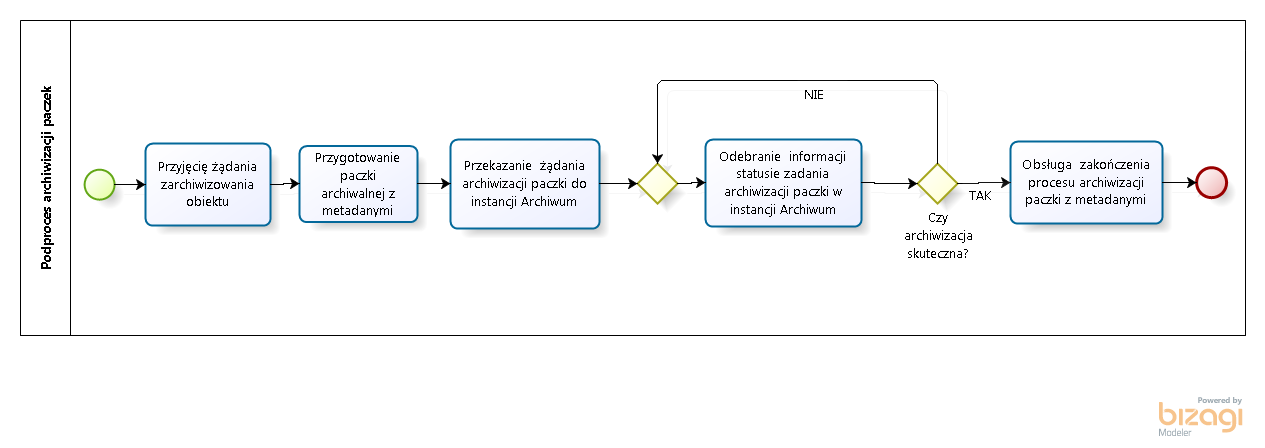
Poniżej zamieszczony został poglądowy diagram procesu archiwizacji obiektu. Diagram ten ma za zadanie przedstawić kolejne kroki procesu bez uwzględniania aspektu asynchroniczności poszczególnych kroków.



### Przyjęcie żądania zarchiwizowania metadanych/bazy danych

W ramach tego punktu powinno nastąpić rozpoczęcie procesu archiwizacji metadanych.

Poniżej zamieszczony został poglądowy diagram procesu archiwizacji metadanych. Diagram ten ma za zadanie przedstawić kolejne kroki procesu bez uwzględniania aspektu asynchroniczności poszczególnych kroków.



### Przygotowywanie paczki archiwalnej z obiektem

W ramach tego zadania powinna zostać przygotowana paczka archiwalna z danymi archiwizowanego obiektu, która powinna zawierać następujące elementy:

1. Metadane obiektu
2. Pozostałe informacje o obiekcie (np. dane z procesu workflow, segmentacja)
3. Strukturę obiektu – kontenty i ich struktura
4. Pliki stream’ów zgodnie z konfiguracją.

Aby wyeliminować sytuacje, kiedy proces przygotowania paczki archiwalnej zostanie rozpoczęty, ale w trakcie jego trwania miejsce na macierzy pomocniczej się wyczerpie, powinien istnieć prosty mechanizm rezerwacji miejsca dla rozpoczynanych zadań mogących wpłynąć na ilość wolnego miejsca na macierzy pomocniczej. Powinien także istnieć mechanizm umożliwiający ustawienie limitu miejsca zajętego (ang. quota) przez proces archiwizacji oraz proces przywracania/udostępniania – w systemie powinien być parametr definiujący, ile procent miejsca na macierzy pomocniczej przeznaczone jest dla procesu archiwizacji (parametr w konfiguracji: archivization.space.share); procent miejsca na macierzy pomocniczej przeznaczony dla procesu przywracania/udostępniania powinien wynikać z różnicy 100% - x%, gdzie x to parametr dla procesu archiwizacji. Ten mechanizm pozwoli uniknąć sytuacji zakleszczenia się obu procesów z uwagi na zużycie całego dostępnego miejsca na macierzach pomocniczych.

Zarządca Archiwizatora po kilku nieudanych próbach (parametr odświeżany dynamicznie o nazwie archivization.subsequent.allocation.fails.limit) sprawdzenia, czy według istniejących w systemie rezerwacji jest dostępna wystarczająca ilość miejsca na stworzenie paczki archiwalnej, powinien odkładać w czasie kolejne sprawdzenie do momentu wzrostu ilości wolnego miejsca (ilości miejsca obliczonej według istniejących w systemie rezerwacji).

Pod pojęciem momentu stworzenia paczki archiwalnej (szczególnie w kontekście znacznika czasowego w paczce archiwalnej) rozumie się moment rozpoczęcia pobierania danych z obiektu z Repozytorium do Systemu archiwizacji.

#### Zawartość paczki archiwalnej

Szczegółowa lista elementów obiektu do umieszczenia w paczce archiwalnej została przedstawiona poniżej:

1. Wersja Systemu archiwum – jako że w budowaniu baczki archiwalnej bierze udział tylko Archiwizator, to wystarczające jest uwzględnienie w tym miejscu wersji Archiwizatora.
2. Metadane obiektu (w dwóch wersjach, natywnym oraz przekonwertowane do DC – konwersja na DC obecnie dostępnymi gotowymi narzędziami w RCBN, natomiast w celu zapewnienia walidacji zawartości pliku METS, powinien powstać plik XSD, który umożliwi walidację w szczególności metadanych obiektu zapisanych w formacie natywnym Repozytorium jak i w formacie DC)
3. Metadane kontentów (w dwóch wersjach, natywnym oraz przekonwertowane do DC – konwersja na DC obecnie dostępnymi gotowymi narzędziami w RCBN, natomiast w celu zapewnienia walidacji zawartości pliku METS, powinien powstać plik XSD, który umożliwi walidację w szczególności metadanych obiektu zapisanych w formacie natywnym Repozytorium jak i w formacie DC)
4. Pozostałe techniczne atrybuty obiektu: data utworzenia obiektu, data ostatniej modyfikacji obiektu
5. Pozostałe biznesowe atrybuty obiektu: liczba dostępnych w BN fizycznych kopii, główny kontent, kontekst obiektu, informacja czy obiekt jest publiczny czy pod prawem autorskim
6. Historia workflow (ekran log obiektu w GUI):
   1. Nazwa kroku
   2. Data wykonania zadania
   3. Wykonujący (Obiekt przypisany do)
7. Historia skanowania: skanery na których były wykonane skany kontentów (graficznych)
8. Historia modyfikacji (ekran historia modyfikacji w GUI)
9. Lista (struktura) kontentów postaci zbiorczych + ich metadane jak w pkt 3 + lista dostępnych stream’ów + dodatkowe atrybuty (czy ukryty, stan obrotu dla plików graficznych)
10. Lista (struktura) kontentów w strukturze paginacyjnej (wraz z informacją o zachowanej kolejności oraz informacją o kontentach alternatywnych) + ich metadane + lista dostępnych stream’ów + szczegóły paginacyjne (obrót, rozdział, układ, renumeracja, strony, ciekawe, funkcja strony, opis strony) + dodatkowe atrybuty (informacja czy jest ukryty, stan obrotu dla plików graficznych)
11. Segmentacja w obiekcie
12. Pliki (streamy) + ich sumy kontrolne
13. Kategorie, kolekcje – nazwy kategorii i kolekcji przypisanych do obiektu
14. Odnośniki do innych obiektów – identyfikatory podobiektów

Aby było możliwe odtworzenie całego obiektu w bazie danych Repozytorium, jak również przywrócenie zawartości (plików) istniejącego obiektu w Repozytorium, potrzebna jest możliwość precyzyjnego zidentyfikowania, w które miejsce w obiekcie dany plik powinien zostać przywrócony. Dlatego w paczce archiwalnej każdy plik powinien być opisany identyfikatorem kontentu, do którego należy, oraz nazwą stream’u.

#### Konfiguracja archiwizowanych i usuwanych kluczy stream’ów z obiektu

Powinna być możliwość skonfigurowania typów (kluczy) stream’ów, podlegających archiwizowaniu, oraz w ramach tych kluczy stream’ów tzw. mime-types plików podlegających usuwaniu.

Z uwagi na to, że ta konfiguracja będzie wykorzystywana zarówno przez System archiwizacji, jak i przez Repozytorium, to konfiguracja ta powinna spełniać następujące wymagania:

1. Konfiguracja listy archiwizowanych kluczy streamów oraz usuwanych mime-types plików powinna być przechowywana w bazie danych Repozytorium (tabele: archivable\_stream, reducible\_stream\_type).
2. Powinna zostać przygotowana metoda do pobierania wartości tych konfiguracji z Repozytorium – z której to będzie korzystał System archiwizacji.
3. W Systemie archiwizacji ta konfiguracja będzie cache’owana przez czas zdefiniowany w parametrach Systemu archiwizacji (parametr w konfiguracji: archivization.config.refresh.delay).

Konfiguracja powinna umożliwiać podanie klucza stream’a (np. mainStream, download\_alto, text, download\_fullJPG) co było by równoznaczne z uwzględnieniem go w archiwizacji, oraz zdefiniowanie dla tego klucza, które tzw. mime-types plików powinny być usuwane. Przykład takiej konfiguracji został zaprezentowany poniżej:

1. Archiwizowany klucz=”mainStream”, usuwane mime-types:
   1. „image/tiff”
   2. „image/jpeg”
   3. „image/png”
2. Archiwizowany klucz=”download\_fullJPG”, usuwane mime-types:
   1. „image/jpeg”
3. Archiwizowany klucz=”download\_przyklad”, usuwane mime-types: (brak)

Punkt 3 w powyższym przykładzie należy rozumieć w taki sposób, że stream o kluczu „download\_przyklad” powinien być archiwizowany, ale nie powinno nastąpić usunięcie jakiegokolwiek pliku stream’a.

W związku z konfiguracją wykonywaną w opisany powyżej sposób wykluczającą na poziomie pliku konfiguracyjnego pomyłki polegającej na zdefiniowaniu usunięcia stream’u, który jednocześnie nie jest archiwizowany, nie ma konieczności wprowadzenia dodatkowego mechanizmu weryfikacji w plikach konfiguracyjnych relacji listy kluczy stream’ów archiwizowanych i usuwanych z obiektu.

Mechanizm usuwania stream’ów (plików) został szczegółowo opisany rozdziale dotyczącym usuwania plików z przywróconego/niezredukowanego obiektu (Usuwanie plików z przywróconego lub niezredukowanego obiektu). Zgodnie z opisem we wspomnianym rozdziale, podczas usuwania stream’ów (plików) powinna zostać zwrócona uwaga na to, aby nie usunąć stream’ów (plików), które nie są oznaczone jako zarchiwizowane.

#### Struktura danych w kontenerze METS

Dane i informacje o strukturze obiektu powinny zostać umieszczone w kontenerze METS (Metadata Encoding & Transmission Standard), którego struktura została opisane w poniższej tabeli:

| Sekcja | Skrót | Elementy | Przykład |
| --- | --- | --- | --- |
| METS (korzeń) | mets:mets | OBJID=<Id obiektu (w Repozytorium)> LABEL=<Nazwa obiektu (Tytuł i Iteracja)> ARCHIVISER\_VERSION= <Wersja Archiwizatora> OWNER="Biblioteka Narodowa" | <mets:mets xsi:schemaLocation=" … " OBJID="2648414" LABEL="Bromberger Zeitung R:1894 Nr 057" ARCHIVISER\_VERSION="8.234" OWNER="Biblioteka Narodowa"> |
| METS header | metsHdr | CREATEDATE=<Data utworzenia pliku METS> ARCHIVINGDATE=<Data rozpoczęcia zadania kopiowania plików z Repozytorium do Systemu archiwizacji> | <mets:metsHdr CREATEDATE="2012-09-27T08:38:19Z" ARCHIVINGDATE="2012-09-27T08:41:47Z"> </mets:metsHdr> |
| Descriptive Metadata | dmdSec | 1. Metadane obiektu: a) natywnie jak w Repozytorium b) w standardzie DC 2. Metadane kontentów jak w Repozytorium a) postaci zbiorcze - poza podstawowymi metadanymi kontentów: - czy kontent ukryty - stan obrotu dla plików graficznych b) kontenty w strukturze paginacyjnej - poza podstawowymi metadanymi kontentów: - szczegóły paginacyjne (obrót, rozdział, układ, renumeracja, strony, ciekawe, funkcja strony, opis strony) - dodatkowe atrybuty (czy ukryty, stan obrotu dla plików graficznych) | <mets:dmdSec ID="REPO\_OBJECT"> … </mets:dmdSec> <mets:dmdSec ID="DC\_OBJECT"> … </mets:dmdSec> <mets:dmdSec ID="REPO\_CONTENT\_<Id contentu 1>"> … </mets:dmdSec> <mets:dmdSec ID="REPO\_CONTENT\_<Id contentu 2>"> … </mets:dmdSec> <mets:dmdSec ID="REPO\_CONTENT\_<Id contentu n>"> … </mets:dmdSec> |
| Administrative Metadata | amdSec | 1. Jako **techMD**: a) Pozostałe techniczne atrybuty obiektu: data utworzenia, data ostatniej modyfikacji b) Pozostałe biznesowe atrybuty obiektu: główny kontent, kontekst obiektu c) Segmentacja w obiekcie wraz z metadanymi artykułów d) Kategorie, kolekcje – nazwy kategorii i kolekcji przypisanych do obiektu  e) Odnośniki do innych obiektów – identyfikatory podobiektów  2. Jako **rightsMD**: a) Liczba dostępnych w BN fizycznych kopii b) czy obiekt jest objęty prawami autorskimi  3. Jako **sourceMD**: *Sekcja nie zostanie wykorzystana*  4. Jako **digiprovMD**: a) Historia workflow – log obiektu: - Nazwa kroku - Data wykonania zadania - Wykonujący (Obiekt przypisany do) b) Historia skanowania: skanery na których były wykonane skany kontentów (graficznych) c) Historia modyfikacji (ekran historia modyfikacjiw GUI) |  |
| File Section | fileSec | 1. Lista stremów (plików) wraz z sumami kontrolnymi 2. Użycie następujących grup plików: - streamy źródłowe (z kluczem mainSteam) - streamy pochodne (klucz inny niż mainStream) - pozostałe (serializowane dane obiektu z HCP)  Konwencja nazywania identyfikatorów plików: streamID\_streamName | <fileSec>  <fileGrp ID="MAINSTREAMS"> … </fileGrp>  <fileGrp ID="DERIVEDSTREAMS"> … </fileGrp>  <fileGrp ID="OTHER"> … </fileGrp> </fileSec> |
| Structural Map | structMap | 1. Kontenty w strukturze paginacyjnej: struktura kontentów + lista streamów - Elementy <div> na pierwszym poziomie określają contenty główne. - Jeśli dana strona ma contenty alternatywne używamy zagłębionego <div> numerując je kolejno przy pomocy Order zaczynając od 1. - Metadane kontentów linkowane są za pomocą atrybutu DMDID (oczywiście właściwe sekcje metadanych muszą być stworzone w ramach tego samego dokumentu w <dmdSec>).  2. Postaci zbiorcze: struktura kontentów + lista streamów | Przykład contentu w strukturze paginacyjnej:  <structMap LABEL="Struktura paginacyjna" TYPE="logical">  <!--content główny dla strony 243--> <mets:div ORDER="243" LABEL="Strona nr 243" TYPE="page" ID="DIV00243" DMDID="CID45654456meta">  <!--pliki contentu głównego-->  <mets:fptr FILEID="CID45654456\_mainStream"/>  <mets:fptr FILEID="CID45654456\_downloadALTO"/> <!--content alternatywny strony 243 nr 1--> <mets:div ORDER="1" ID="DIV243\_1" TYPE="alternative page" DMDID="CID45654457meta">   <mets:fptr FILEID="CID45654457\_mainStream"/>   <mets:fptr FILEID="CID45654457\_downloadALTO"/>  </mets:div> <!--content alternatywny strony 243 nr 2--> <mets:div ORDER="2" ID="DIV0243\_2" TYPE="alternative page" DMDID="CID45654458meta">   <mets:fptr FILEID="CID45654458\_mainStream"/>   <mets:fptr FILEID="CID45654458\_downloadALTO"/>  </mets:div> </mets:div> </structMap>  Przykład contentu zbiorczego:  <structMap LABEL="Pliki pozapaginacyjne " TYPE="logical">  <mets:div ORDER="1" LABEL="bulk1" TYPE="document" ID="bulk1" DMDID="CID45654490meta">   <mets:fptr FILEID="CID45654490\_mainStream"/> <mets:fptr FILEID="CID45654490\_text"/>   <mets:fptr FILEID="CID45654490\_download\_ALTO"/>  </mets:div> </structMap>  Stringi do wpisywania w wartość TYPE powinny być konfigurowalne.  „nazwa typu dla structMap kontentów w strukturze paginacyjnej”=”logical” „nazwa typu węzła strony głównej” = „page” „nazwa typu dla structMap kontentów postaci zbiorczej” = „alternative page” „nazwa typu dla węzła kontentów postaci zbiorczej” =”logical” „nazwa typu dla kontentu postaci zbiorczej” =”document” |
| Structural Links | structLink | *Sekcja nie zostanie wykorzystana* |  |
| Behavioral | behaviorSec | *Sekcja nie zostanie wykorzystana* |  |

Daty w paczce METS powinny być zapisane w standardzie ISO8601.

#### Struktura folderów i nazewnictwo plików w paczce archiwalnej

Wszystkie pliki i podfoldery w paczce archiwalnej powinny znajdować się w folderze, którego nazwą powinien być identyfikator obiektu w Repozytorium. Następnie w tym folderze powinien znaleźć się plik z kontenerem METS oraz podfoldery, których nazwą powinny być poszczególne klucze stream’ów (np. mainStream, download\_alto, text, download\_fullJPG, download\_lowJPG). W tych podfolderach powinny zostać umieszczone stream’y (pliki) archiwizowanych kontentów należących do obiektu.

Nazewnictwo plików stream’ów zostało przedstawione poniżej:

<idetyfikator obiektu z Repozytorium>separator<nazwa typu streamu>separator<kolejność w obiekcie>separator<Identyfikator kontentu z repozytorium>separator<identyfikator streamu>.rozszerzenie pliku

Pod pojęciem kolejności w obiekcie należy rozumieć systemową (w rozumieniu wewnętrznych identyfikatorów kolejności) w Repozytorium kolejność kontentów w strukturze paginacyjnej.

Numer będący kolejnością w obiekcie powinien być uzupełniany zerami do 4 cyfr, np. 0001, 0002, ..., 0010, 0011... .

W przypadku kontentów postaci zbiorczych kolejność należy rozumieć jako systemową (w rozumieniu wewnętrznych identyfikatorów kolejności) kolejność w Repozytorium – o ile taka jest dostępna w Repozytorium. Jeżeli taka kolejność w Repozytorium nie jest dostępna, to w nazwach stream’ów postaci zbiorczych zamiast kolejności pobranej z Repozytorium powinna zostać użyta na stałe liczba zero („0”).

Separator zostanie wybrany przez Wykonawcę na etapie implementacji. Może to być przykładowo znak podkreślenia dolnego „\_” lub myślnik „-”.

#### Postać paczki archiwalnej

Paczka archiwalna zawierająca obiekt powinna zostać przygotowana jako pojedynczy plik TAR stworzony bez zastosowania kompresji. Brak zastosowania kompresji argumentuje się tym, że największą część paczki archiwalnej pod kątem zajmowanego rozmiaru będą stanowiły pliki w formacie tiff, który to format już jest skompresowany. Dodatkowo napęd w ramach biblioteki taśmowej zapewnia funkcjonalność kompresji sprzętowej, dlatego wykonanie na etapie tworzenia paczki archiwalnej dodatkowej próby kompresji nie jest zasadne (a może w odległej przyszłości spowodować trudności z odpakowaniem paczki z uwagi na trudności w dostępie do narzędzi służących do odwrócenia procesu kompresji).

Jedynym istotnym argumentem za podziałem paczki archiwalnej na wiele plików były by problemy sieciowe, których skutkiem w przypadku przerwania transmisji była by potrzeba retransmisji całego dużego pliku paczki archiwalnej. Jednak użycie narzędzia rsync zapewnia obsługę tego typu problemów. Natomiast podział paczki archiwalnej na wiele plików powodowałby potrzebę dużo bardziej skomplikowanych algorytmów doboru paczek do zapisania na kasecie, jak również dużo bardziej skomplikowanego mechanizmu przekazującego paczkę archiwalną pomiędzy węzłami Systemu archiwizacji.

#### Nazwa pliku paczki archiwalnej

Nazwa paczki archiwalnej powinna składać się z następujących elementów:

1. Identyfikator obiektu z Repozytorium
2. Znacznik czasowy z momentu stworzenia paczki archiwalnej
3. Identyfikator klienta Systemu archiwizacji, który dla Repozytorium będzie cyfrą „1”
4. Stały tekst: „Owner\_BibliotekaNarodowa”

Znacznik czasowy powinien być w formacie „Unix time”, który oznacza liczbę sekund, które upłynęły od 1 stycznia 1970 wykorzystując strefę czasową UTC.

Przykład nazwy pliku został zaprezentowany poniżej:

Id\_201671#Time\_1475139982#Source\_1# Owner\_BibliotekaNarodowa.TAR

### Przygotowywanie paczki archiwalnej ze zrzutem metadanych/bazy danych

W ramach tego zadania powinna zostać przygotowana paczka archiwalna metadanych, która powinna zawierać elementy opisane w niniejszym rozdziale.

Aby wyeliminować sytuacje, kiedy proces przygotowania paczki archiwalnej zostanie rozpoczęty, ale w trakcie jego trwania miejsce na macierzy pomocniczej się wyczerpie, powinien istnieć prosty mechanizm rezerwacji miejsca dla rozpoczynanych zadań mogących wpłynąć na ilość wolnego miejsca na macierzy pomocniczej.

#### Zawartość paczki archiwalnej

Szczegółowa lista elementów do umieszczenia w paczce archiwalnej została przedstawiona poniżej:

1. Numer wersji Systemu archiwizacji (zawierający numer wersji poszczególnych instancji Archiwum oraz wersję Archiwizatora), np. w pliku XML lub tekstowym
2. Plik z bazą Repozytorium
3. Pliki z bazami Systemu archiwizacji: osobne pliki dla poszczególnych baz węzłów Systemu archiwizacji

Należy pamiętać, że pliki ze zrzutami Systemu archiwizacji będą zawierały zadania/żądania do wykonania, których stan będzie aktualny na moment przygotowania paczki archiwalnej. Dlatego ewentualne przywrócenie tych baz do Systemu archiwizacji w szczególności spowoduje utratę spójności danych w Systemie archiwizacji.

#### Format zrzutu baz danych

Formatem zrzutu baz danych powinien być tzw. DUMP (zrzut) zawierający strukturę i zawartość bazy danych, który powinien być w momencie wykonywania zrzutu od razu kompresowany z użyciem GZIP (polecenie mysqldump .... | gzip > dump.sql.gz). Takie podejście umożliwia zaoszczędzenie miejsca na macierzy pomocniczej już na tym etapie, a nie dopiero po skompresowaniu całej paczki archiwalnej. Zastosowanie kompresji w stosunku do zrzutów baz danych jest zasadne, ponieważ zrzut typu DUMP posiada postać tekstową, a pliki tekstowe zazwyczaj poddają się przynajmniej częściowej kompresji.

#### Struktura folderów i nazewnictwo plików w paczce archiwalnej

Wszystkie pliki powinny znajdować się w jednym katalogu, którego nazwą będzie data utworzenia katalogu w formacie YYYYMMDD, np. 20160930.

Nazewnictwo plików ze zrzutami baz danych powinno być następujące:

MySQLDump<separator><identyfikator węzła>.SQL

Separator zostanie wybrany przez Wykonawcę na etapie implementacji. Może to być przykładowo znak podkreślenia dolnego „\_”.

#### Postać paczki archiwalnej

Paczka archiwalna zawierająca zrzut metadanych powinna zostać przygotowana jako pojedynczy plik TAR.

Jedynym istotnym argumentem za podziałem paczki archiwalnej na wiele plików były by problemy sieciowe, których skutkiem w przypadku przerwania transmisji była by potrzeba retransmisji całego dużego pliku paczki archiwalnej. Jednak użycie narzędzia rsync zapewnia obsługę tego typu problemów. Natomiast podział paczki archiwalnej na wiele plików powodowałby potrzebę dużo bardziej skomplikowanych algorytmów doboru paczek do zapisania na kasecie, jak również dużo bardziej skomplikowanego mechanizmu przekazującego paczkę archiwalną pomiędzy węzłami Systemu archiwizacji.

W pierwotnej koncepcji założono, że w pierwszym kroku powinien nastąpić zrzut baz danych w postaci plików DUMP, a następnie powinna nastąpić kompresja tych plików z użyciem GZIP. Jednakże analiza tego aspektu doprowadziła do wniosku, że można wykonać zrzut bazy danych w postaci plików DUMP kompresując go już podczas wykonywania zrzutu za pomocą narzędzia GZIP. Dzięki temu wygenerowany plik zrzutu zostaje zapisany na dysku już w postaci skompresowanej, więc miejsce na macierzy pomocniczej wykorzystywanej przez Archiwizator może być lepiej wykorzystane.

#### Nazwa pliku paczki archiwalnej

Nazwa paczki archiwalnej powinna składać się z następujących elementów:

1. Stały tekst: „MetadataDump”
2. Znacznik czasowy z momentu stworzenia paczki archiwalnej
3. Identyfikator klienta Systemu archiwizacji, który dla Repozytorium będzie cyfrą „1”
4. Stały tekst: „Owner\_BibliotekaNarodowa”

Znacznik czasowy powinien być w formacie „Unix time”, który oznacza liczbę sekund, które upłynęły od 1 stycznia 1970 wykorzystując strefę czasową UTC.

Przykład nazwy pliku został zaprezentowany poniżej:

Id\_MetadataDump#Time\_1475139982#Source\_1#Owner\_BibliotekaNarodowa.TAR

### Przekazanie żądania archiwizacji paczki do instancji Archiwum

W ramach tego zadania Archiwizator powinien przekazać żądanie archiwizacji paczki do wszystkich powiązanych z nim instancji Archiwum. Przekazane żądanie może dotyczyć obiektu lub też zrzutu metadanych.

Archiwizator przekazując paczkę archiwalną powinien przekazać informacje, czy paczka powinna podlegać replikacji. Informacja ta powinna być przekazana na podstawie następujących danych:

1. Paczka archiwalna z obiektem – domyślnie każda paczka powinna być replikowana, chyba że metadana obiektu o nazwie „replicate” posiada wartość „false”, to paczka nie powinna zostać zreplikowana w ramach instancji Archiwum. Założono, że metadana „replicate” będzie typu „string”, dlatego też sprawdzenie jej wartości powinno odbyć się poprzez porównanie tekstów.
2. Paczka archiwalna ze zrzutem metadanych – zgodnie z parametrem zdefiniowanym w konfiguracji Archiwizatora (parametr w konfiguracji: metadataDump.replication).

Archiwizator powinien przechowywać informacje, do których instancji Archiwum przekazał żądanie archiwizacji, aby być następnie w stanie określić, czy archiwizacja danej paczki zakończyła się sukcesem i tym samym obsłużyć zdarzenie zakończenia procesu archiwizacji paczki archiwalnej z obiektem oraz z metadanymi.

Zastosowanie systemu kolejkowego do zapewnienia komunikacji między Archiwizatorem i Archiwum wyklucza potrzebę zapewnienia mechanizmu ponowienia wysłania przez Archiwizator żądania w przypadku niedostępności instancji Archiwum.

### Odebranie informacji o statusie zakończenia zadania archiwizacji paczki w instancji Archiwum wraz z podjęciem kolejnych kroków

W ramach tego zadania po odebraniu od instancji Archiwum informacji o statusie zleconego żądania archiwizacji paczki w danej instancji Archiwum, Archiwizator w zależności od statusu żądania powinien:

1. Jeżeli żądania archiwizacji zakończyło się powodzeniem

Archiwizator powinien zweryfikować, czy paczka jest już zarchiwizowana we wszystkich instancjach Archiwum powiązanych z Archiwizatorem w momencie przekazania paczki do instancji Archiwum. Jeżeli paczka jest skutecznie zarchiwizowana, to Archiwizator powinien przejść do czynności zakończeniowych procesu archiwizacji, których przebieg zależy od tego czy:

* 1. Paczka zawiera obiekt – wówczas wykonanie czynności opisanych w rozdziale “Obsługa zakończenia procesu archiwizacji paczki z obiektem”.
  2. Paczka zawiera metadane – wówczas wykonanie czynności opisanych w rozdziale „Obsługa zakończenia procesu archiwizacji paczki z metadanymi”.

1. Jeżeli żądanie archiwizacji zakończyło się niepowodzeniem

W tym przypadku Archiwizator powinien w zależności od otrzymanej przyczyny niepowodzenia podjąć dalsze kroki opisane w poniższych podpunktach. Zakłada się następujące przyczyny niepowodzenia zadania archiwizacji:

* 1. Plik jest większy niż obsługiwane kasety w tej instancji

W tym przypadku powinno nastąpić zalogowanie błędu wraz z powiadomieniem administratora.

* 1. Plik jest większy niż całkowite wolne miejsce na macierzy pomocniczej

W tym przypadku powinno nastąpić zalogowanie błędu wraz z powiadomieniem administratora

W przypadku wystąpienia jednego z wyżej wymienionych problemów, żądanie archiwizacji jest kontynuowane w pozostałych instancjach Archiwum, przy czym całe żądanie archiwizacji kończy się niepowodzeniem oraz ustawieniem w obiekcie statusu PROBLEM, dzięki czemu wyklucza się wybieranie tego obiektu do ponownej archiwizacji dając jednocześnie administratorowi możliwość i czas na weryfikację przyczyny problemu. Przyjęto, że w ramach tego projektu nie zostanie przygotowane narzędzie do wyjścia z tego statusu, dopuszczając tym samym, że wyjście z tego statusu będzie wymagało wykonania modyfikacji statusu obiektu z poziomu bazy danych.

### Obsługa zakończenia procesu archiwizacji paczki z obiektem

Po skutecznej archiwizacji paczki z obiektem Archiwizator powinien odnotować ten fakt w swojej bazie informacji, a następnie powinien przystąpić do następujących kroków:

1. Aktualizacja daty ostatniej archiwizacji

W ramach tego kroku system powinien zmodyfikować w obiekcie datę ostatniej archiwizacji – wypełniając ją datą rozpoczęcia zadania pobrania plików i przygotowania paczki archiwalnej.

1. Aktualizacja statusów stream’ów

W ramach tego kroku system powinien zmodyfikować status tych stream’ów, które zostały zarchiwizowane – status powinien zostać zmieniony na ARCHIVED. Chodzi o właściwą obsługę sytuacji, kiedy podczas trwania archiwizacji w kontencie zostanie dodany nowy stream – wówczas ten stream nie może zostać oznaczony jako ARCHIVED ponieważ nie został zarchiwizowany.

1. Aktualizacja statusu obiektu w Repozytorium

W ramach tego punktu powinno nastąpić zaktualizowanie statusu obiektu z ARCHIVING na ARCHIVED.

Analiza możliwości oraz usunięcie plików z obiektu będzie realizowane przez osobną usługę, która będzie usuwała pliki zarówno z obiektów niedawno zarchiwizowanych, jak i z obiektów przywróconych po upłynięciu okresu karencji od przywrócenia.

### Obsługa zakończenia procesu archiwizacji paczki z metadanymi

Po skutecznej archiwizacji paczki ze zrzutem metadanych Archiwizator powinien odnotować ten fakt w swojej bazie informacji oraz usunąć paczkę archiwalną ze swojej macierzy pomocniczej.

### Przyjęcie żądania przywrócenia obiektu / udostępnienia paczki archiwalnej

W ramach tego zadania Archiwizator powinien obsłużyć otrzymane żądanie poprzez realizację następujących kroków:

1. Weryfikację, czy obiekt lub zrzut metadanych o podanej wersji w żądaniu jest zarchiwizowany w ramach Systemu archiwizacji. Jeżeli nie jest, to żądanie powinno zostać odrzucone wraz z zalogowaniem zdarzenia w dzienniku zdarzeń oraz poinformowaniem administratora.
2. Ustawienie obiektowi statusu RESTORING w Repozytorium – tylko jeżeli jest to proces przywracania.
3. Decyzję, z której instancji Archiwum przywrócić dany obiekt / udostępnić paczkę archiwalną.

Kryterium wyboru instancji Archiwum powinien być współczynnik obciążenia każdej instancji Archiwum. Przyjęto, że na obciążenie instancji Archiwum wpływa bezpośrednio sumaryczny rozmiar plików, które oczekują na archiwizację (zapis) oraz odczyt. Na podstawie tych rozmiarów oraz skonfigurowanych prędkości odczytu i zapisu w danej instancji Archiwum można obliczyć całkowity czas oczekujących zapisów oraz odczytów. Z drugiej strony każda instancja Archiwum może dysponować różną liczbą napędów, dlatego w ramach obliczeń należy wziąć pod uwagę liczbę tych napędów, dzieląc sumaryczny czas operacji do wykonania właśnie przez liczbę napędów.

Przyjęto, że skonfigurowane prędkości odczytu i zapisu będą ręcznie dobierane przez operatora Systemu archiwizacji.

Współczynnik obciążenia instancji Archiwum powinien być liczony w następujący sposób:

Dane wejściowe:

* 1. Ln : Liczba napędów
  2. PpO: Przyjęta prędkość odczytu (parametr w konfiguracji: drive.average.read.speed)
  3. PpZ: Przyjęta prędkość zapisu (parametr w konfiguracji: drive.average.write.speed)
  4. RozO : Sumaryczny rozmiar plików oczekujących na odczytanie
  5. RozZ : Sumaryczny rozmiar plików oczekujących na zapisanie

Obliczenia:

[Odczyty potrwają]: [RozO] / [PpO]

[Zapisy potrwają]: [RozZ] / [PpZ]

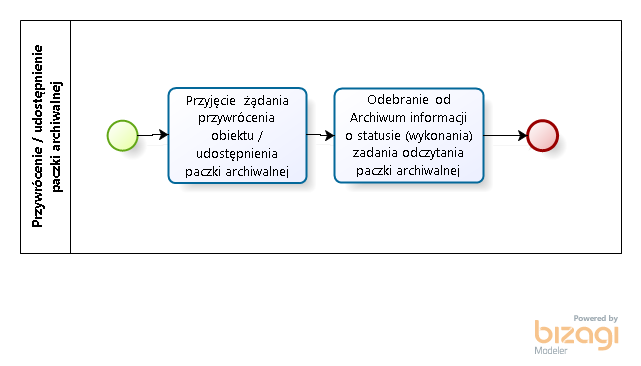
W powyższych obliczeniach należy zwrócić uwagę na jednostki, tak aby finalnie uzyskać całkowity czas oczekujących odczytów oraz całkowity czas oczekujących zapisów.

Wynik:

Współczynnik obciążenia: ([Odczyty potrwają]+[Zapisy potrwają])/[Ln]

1. Przekazanie wybranej w poprzednim kroku instancji Archiwum zadania odczytania paczki archiwalnej z obiektem / ze zrzutem metadanych. Instancja Archiwum powinna potwierdzić przyjęcie zadania odczytania. Jeżeli tego nie zrobi w skonfigurowanym czasie (parametr w konfiguracji: file.retrieval.acceptanceTimeout), to Archiwizator przekaże zadanie do jednej z pozostałych instancji Archiwum – decydując o wyborze instancji wykorzystując wartość współczynnika obciążenia. Dodatkowo to zdarzenie powinno zostać zapisane w dzienniku zdarzeń wraz z wysłaniem powiadomienia do administratora.

Poniżej zamieszczony został poglądowy diagram procesu przywrócenia obiektu / udostepnienia paczki archiwalnej. Diagram ten ma za zadanie przedstawić kolejne kroki procesu odwzorowane w niniejszym dokumencie jako rozdziały, bez uwzględniania aspektu asynchroniczności poszczególnych kroków.

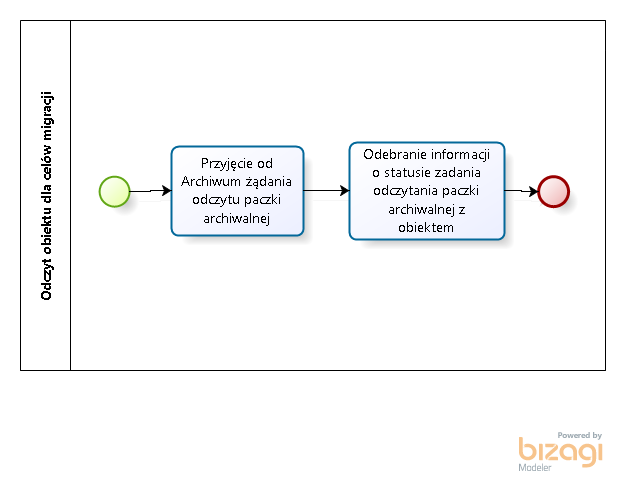


### Przyjęcie od instancji Archiwum żądania odczytu paczki archiwalnej

W przypadku wystąpienia procesu migracji kasety w ramach instancji Archiwum, może wystąpić potrzeba wykorzystania kopii paczki archiwalnej z innej instancji Archiwum. Wówczas instancja Archiwum powinna wysłać do Archiwizatora żądanie odczytu paczki archiwalnej z innej instancji Archiwum, a Archiwizator powinien przyjąć takie żądanie oraz wykonać następujące kroki:

1. Weryfikacja, czy paczka archiwalna w podanej w żądaniu wersji jest zarchiwizowana oraz czy jest aktualnie dostępna w innej instancji Archiwum (ponieważ może zajść sytuacja, w której wszystkie pozostałe kopie nie są dostępne / również są w trakcie migracji).
2. Decyzja, z której instancji Archiwum przywrócić dany obiekt. Decyzja ta powinna zostać podjęcia na podstawie współczynników obciążenia instancji Archiwów opisanych w rozdziale dotyczącym przyjęcie żądania przywrócenia obiektu / udostępnienia paczki archiwalnej.
3. Zlecenie odczytania paczki archiwalnej z obiektem do wybranej instancji Archiwum.

Poniżej zamieszczony został poglądowy diagram procesu odczytu obiektu dla celów migracji kasety wykonywanego w jednej z instancji Archiwum. Diagram ten ma za zadanie przedstawić kolejne kroki procesu odwzorowane w niniejszym dokumencie jako rozdziały, bez uwzględniania aspektu asynchroniczności poszczególnych kroków.



### Odebranie informacji o statusie zakończenia zadania odczytu paczki archiwalnej wraz z podjęciem kolejnych kroków

W ramach tego zadania po odebraniu od instancji Archiwum informacji o statusie zleconego zadania odczytu paczki archiwalnej z danej instancji Archiwum, Archiwizator w zależności od statusu zadania powinien:

1. Jeżeli zadanie odczytu zakończyło się powodzeniem i zadanie odczytu było zainicjowane w ramach procesu przywrócenia obiektu do Repozytorium:
   1. Rozpakowanie paczki archiwalnej

W ramach tego kroku powinno nastąpić rozpakowanie pliku paczki archiwalnej.

* 1. Zainicjowanie przywrócenia obiektu do Repozytorium

W ramach tego kroku Archiwizator powinien wykonać procedurę przywrócenia obiektu do Repozytorium wykorzystując funkcjonalność Przywracacza.

1. Jeżeli zadanie odczytu zakończyło się powodzeniem i zadanie odczytu było zainicjowane w ramach procesu udostępnienia paczki archiwalnej, to w ramach tego kroku Archiwizator powinien przegrać paczkę archiwalną na zasób sieciowy wskazany w wywołaniu żądania.
2. Jeżeli zadanie odczytu zakończyło się powodzeniem i zadanie odczytu było zainicjowane w ramach procesu migracji – Archiwizator powinien przekazać plik paczki archiwalnej do instancji Archiwum, która była autorem zadania odczytu.
3. Jeżeli zadanie odczytu zakończyło się niepowodzeniem

W tym przypadku Archiwizator powinien w zależności od otrzymanej przyczyny niepowodzenia podjąć dalsze kroki opisane w poniższym podpunkcie. Zakłada się następującą przyczynę niepowodzenia zadania odczytu paczki archiwalnej – brak poprawnej kopii paczki o podanym identyfikatorze. W tym przypadku powinno nastąpić zalogowanie zdarzenia w dzienniku zdarzeń wraz z powiadomieniem administratora oraz zlecenie odczytu paczki do innej instancji Archiwum.

### Monitorowanie statusu przywrócenia obiektu

W ramach tego zadania Archiwizator powinien dla oczekujących żądań przywrócenia oraz udostępnienia obiektu monitorować okres obsługi przez System archiwizacji tego żądania. W przypadku upłynięcia skonfigurowanego czasu powinno nastąpić zalogowanie zdarzenia w dzienniku zdarzeń oraz wysłanie powiadomienia do administratora (parametr w konfiguracji: expected.retrieval.time). Nie powinno to jednak spowodować przerwania realizacji tego żądania.

## Interfejs API Systemu archiwizacji

### Specyfikacja interfejsu API

Specyfikacja interfejsu API udostępnionego przez System archiwizacji zgodnego ze wzorcem architektonicznym REST jest dostępna w postaci odrębnego pliku, stanowiącego załącznik nr 1 do Projektu technicznego.

Z uwagi na to, że część operacji uruchamianych przez wywołanie interfejsu może trwać dłuższy czas, wynik operacji (np. przywrócenie obiektu) nie będzie mogło zostać wykonane synchronicznie. Dlatego zakłada się, że autor żądania będzie okresowo odpytywał System archiwizacji, używając udostępnionego interfejsu REST, czy zlecone zadanie jest już ukończone:

1. W przypadku ukończenia zadania, zostanie zwrócony:
   1. Status ukończenia.
   2. Ewentualne pozostałe dane będące wynikiem zadania.
2. W przypadku, kiedy zadanie nie będzie ukończone, zostanie zwrócona informacja, po jakim czasie autor żądania powinien ponownie sprawdzić status tego zadania. Czas ten będzie konfigurowalny (parametr w konfiguracji: task.status.checkInterval). W przyszłości powinna być możliwość dodania funkcjonalności obliczania tego czasu.

Przyjęto, że w przypadku wywołania, które skutkuje przysłaniem danych przez webservice, dane te powinny zostać przekazane w formacie JSON.

### Uwierzytelnianie

Uwierzytelnianie zostanie zrealizowane poprzez czasowy token wystawiany dla użytkownika. Token będzie wystawiany na skonfigurowany w parametrach czas ważności (parametr w konfiguracji: api.token.lifespan). Wywołanie udostępnionego interfejsu REST z wykorzystaniem tokena, pod warunkiem jego ważności w momencie wysłania żądania, będzie przedłużało jego ważność o skonfigurowany w parametrach czas (parametr w konfiguracji: api.token.lifespan.prolongation). Token po upłynięciu czasu ważności będzie wygaszany.

W ramach Systemu archiwizacji powinna istnieć baza użytkowników/klientów Systemu archiwizacji, uprawnionych do korzystania z udostępnionego interfejsu REST. Każdy użytkownik powinien posiadać nazwę, hasło oraz identyfikator klienta Systemu archiwizacji. Wykorzystując polecenia w konsoli systemu operacyjnego (z rodziny Linux) powinna być możliwość:

1. Wyświetlania listy użytkowników.
2. Dodawania użytkowników.
3. Edycji użytkowników – ponownego ustawienia hasła.
4. Oznaczania użytkowników jako nieaktywnych.

### Zapisywanie zdarzeń wywołania interfejsu

Wywołanie metod z interfejsu powinno powodować zapisywanie zdarzeń w dzienniku zdarzeń.

W ramach zdarzenia w dzienniku powinny być zapisywane następujące informacje:

1. Data i godzina UTC
2. Identyfikator hosta
3. Wywołana metoda, np. GET, PUT
4. Poziom zdarzenia
5. Pełna ścieżka/adres URL wywołania
6. Status operacji, np. 200, 403.
7. Zawartość nagłówka http X-NL-AUTH-TOKEN żądania
8. Zawartość body http żądania, w przypadku wywołania /api/token/ POST „Wygenerowanie i pobranie tokena dla wskazanego użytkownika” gdzie w body jest podany login i hasło, hasło nie powinno być zalogowane
9. Zawartość body http odpowiedzi.

W przypadku wystąpienia statusu 403 Forbidden, zdarzenie to powinno posiadać poziom WARNING. W pozostałych przypadkach zdarzenie powinno otrzymać poziom INFO.

Jeżeli brak jednej z informacji np. parametrów wywołania URL, to zawartość tego segmentu w zdarzeniu należy pozostawić pustą uwzględniając jednak wszystkie separatory pipe „|” przed i po tym segmencie.

Format i kolejność zdarzeń powinna być zgodna z następującą specyfikacją:

<Data i godzina UTC>|<Identyfikator hosta>|<Metoda>|<Poziom zdarzenia>|<Status operacji>|<Pełna ścieżka/adres URL wywołania>|<Zawartość nagłówka HTTP>|<Zawartość body HTTP>

W przypadku, w którym jeden z powyższych segmentów będzie wielolinijkowy (np. zawartość body HTTP), dopuszcza się, że jedno zdarzenie będzie w pliku zdarzeń zapisane w więcej niż jednej linijce.

Wysyłanie powiadomień do administratora powinno występować tylko dla zdarzeń z poziomem WARNING. Powinien zostać zastosowany identyczny mechanizm wysyłania powiadomień, co w przypadku pozostałych modułów Systemu archiwizacji.

## Przywracacz

Funkcjonalność Przywracacza ma za zadanie zrealizowanie zadania przywrócenia plików do obiektu w Repozytorium.

Funkcjonalność ta powinna być wykorzystywana przez Archiwizator w procesie przywracania obiektu do Repozytorium.

Przywracacz nie jest osobnym komponentem, a jedynie funkcjonalnością w ramach Archiwizatora. Dlatego nie występuje potrzeba przekazywania, w rozumieniu przykładowo interfejsu API, paczki archiwalnej z obiektem z Archiwizatora do Przywracacza.

Przed rozpoczęciem przywracania paczka obiektu powinna zostać rozpakowania. Aby wyeliminować sytuacje, kiedy proces rozpakowania zostanie rozpoczęty, ale w trakcie jego trwania miejsce na macierzy pomocniczej się wyczerpie, powinien istnieć prosty mechanizm rezerwacji miejsca dla rozpoczynanych zadań mogących wpłynąć na ilość wolnego miejsca na macierzy pomocniczej.

### Przywrócenie plików do obiektu w Repozytorium

Przywrócenie zawartości obiektu powinno polegać na przywróceniu do Repozytorium usuniętych plików stream’ów pochodzących z paczki archiwalnej z uwzględnieniem faktu, że kontent, a więc również stream’y, może należeć do wielu obiektów.

Przed przystąpieniem do przywrócenia najpierw należy sprawdzić, jaki status posiada obiekt:

1. Jeżeli posiada RESTORING, to można przystąpić do wykonania przywrócenia zawartości obiektu.
2. Jeżeli posiada inny niż RESTORING (np. ARCHIVED), to przywrócenie zawartości obiektu nie powinno zostać wykonane.

Przywrócenie zawartości obiektu powinno składać się z następujących kroków:

1. Identyfikacja stream’ów znajdujących się w paczce archiwalnej

W ramach tego kroku na podstawie informacji w paczce METS należy zidentyfikować stream’y dostępne w paczce archiwalnej.

Lista ta powinna zostać zbudowana na podstawie jednego z następujących sposobów, wybranego na etapie implementacji:

* 1. Lista wynika z ostatniej sekcji nazwy plików stremaów w paczce archiwalnej.
  2. Lista wynika z danych w sekcji METS o nazwie fileSec, w której zapisane są relacje nazw plików do identyfikatorów stream’ów i nazw stremów – konwencja nazywania identyfikatorów plików jest streamID\_streamName.

Unikalnym identyfikatorem stream’ów jest streamID.

Jeżeli chodzi o pliki stream’ów, to znajdują się one w paczce archiwalnej w podfolderach, których nazwą są poszczególne klucze stream’ów (np. mainStream, download\_alto, text, download\_fullJPG, download\_lowJPG).

Pliki stream’ów posiadają nazwy zgodnie z poniższym schematem:

<idetyfikator obiektu z Repozytorium>separator<nazwa typu streamu>separator<kolejność w obiekcie>separator<Identyfikator kontentu z repozytorium>separator<**identyfikator streamu**>.rozszerzenie pliku

1. Identyfikacja stream’ów wymagających przywrócenia

W ramach tego kroku należy przeanalizować w obiekcie stream’y wymagające przywrócenia.

Lista ta powinna zostać zbudowana na podstawie streamów znajdujących się obecnie w obiekcie i posiadających status DELETED. Takie kryteria wykluczą streamy nieusunięte oraz nowe.

Unikalnym identyfikatorem stream’ów jest streamID.

1. Sprawdzenie możliwości przywrócenia obiektu

W ramach tego kroku powinno nastąpić sprawdzenie, czy wszystkie stream’y wymagające przywrócenia (lista stworzona w ramach kroku 2) są dostępne w paczce archiwalnej (lista stworzona w ramach kroku 1).

* 1. Jeżeli w paczce archiwalnej znajdują się wszystkie wymagane streamy, można przejść do kolejnego kroku – próby przywrócenia streamów.
  2. Jeżeli w paczce archiwalnej brakuje chociaż jednego streama, należy wykonać następujące czynności:
     1. Ustawienie w Repozytorium statusu obiektu na PROBLEM.
     2. Zalogowanie zdarzenia w dzienniku zdarzeń wraz z poinformowaniem administratora.
     3. Zakończenie niepowodzeniem procesu przywrócenia obiektu.
  3. Jeżeli obiekt nie zawiera streamów do przywrócenia (przykładowo w międzyczasie użytkownik usunął z obiektu kontenty zawierające zredukowane streamy), to etap przywracania streamów powinien zostać pominięty i można przejść do kroku aktualizacji statusu przywracanego obiektu.

1. Przywrócenie stream’ów

W ramach tego kroku należy wykonać operacje przywrócenia streamów polegającą na zapisaniu w HCP pliku pod tym samym identyfikatorem jaki znajduje się zarówno w HCP jak i w nazwie pliku w paczce archiwalnej.

Po przywróceniu stream’a jego status powinien zostać zmieniony z DELETED na ARCHIVED.

W przypadku wystąpienia problemu z przywróceniem pliku (np. chwilowa niedostępność HCP), powinno nastąpić zapisanie zdarzenia w dzienniku zdarzeń. Następnie po zdefiniowanym w konfiguracji czasie (parametr w konfiguracji: restoration.waitForHcpDelay) powinna nastąpić ponowna próba przywrócenia pliku:

* 1. Jeżeli operacja się powiedzie, należy kontynuować przywracanie stream’ów.
  2. Jeżeli ponownie wystąpi błąd, należy wykonać następujące czynności:
     1. Ustawienie w Repozytorium statusu obiektu na PROBLEM.
     2. Zalogowanie zdarzenia w dzienniku zdarzeń wraz z poinformowaniem administratora.
     3. Zakończenie niepowodzeniem procesu przywracania obiektu.

1. Aktualizacja statusu przywracanego obiektu

W ramach tego kroku należy zmienić status obiektu na ARCHIVED oraz zaktualizować datę ostatniego przywrócenia obiektu.

Dodatkowo należy zarejestrować w dzienniku zdarzenie zakończonego sukcesem przywrócenia obiektu.

Przywrócenie obiektu nie powinno powodować aktualizacji daty ostatniej modyfikacji obiektu w kontekście archiwizacji.

Dzięki wykorzystaniu podczas przywracania identyfikatorów streamów, nie występuje potrzeba uwzględniania i analizowania sytuacji, kiedy to contenty w Repozytorium są w innej kolejności.

W przypadku, w którym występuje współdzielenie kontentów, a więc streamy/pliki nie przynależą tylko do jednego obiektu, to podczas przywracania jednego obiektu może nastąpić przywrócenie części kontentów i streamów należących też do innego lub innych obiektów. Ustalono, że w przypadku wystąpienia takiej sytuacji nie jest zasadne analizowanie statusu pozostałych obiektów.

Podczas analizy zagadnienia przywracania zawartości obiektu wzięto pod uwagę fakt, iż przed rozpoczęciem przywrócenia zawartości obiektu należy zweryfikować, czy wszystkie wymagane w obiekcie streamy znajdują się w paczce archiwalnej. Ta weryfikacja jest wykonywana w powyżej opisanych krokach o numerach 1-3, i w przypadku wystąpienia braku chociażby jednego streamu, obiekt jest odpowiednio oznaczany a informacja o tym zdarzeniu jest zapisywana w dzienniku zdarzeń wraz z poinformowaniem administratora. Dodatkowo podczas analizy przypadków brzegowych okazało się, że może zdarzyć się sytuacja, w której w momencie przywrócenia zawartości obiektu, obiekt nie posiada streamów do przywrócenia. Dlatego w ramach opisanych powyżej kroków przebiegu przywrócenia uwzględniono taki przypadek.

## Weryfikator

Funkcjonalność Weryfikator działa w ramach Archiwum i ma za zadanie realizowanie zadania weryfikacji stanu kaset oraz sterowanie procesem migracji kaset.

Wyróżnia się dwa następujące rodzaje weryfikacji:

1. Pełna weryfikacja stanu kasety, polegająca na próbie odczytania wszystkich paczek z wybranej kasety wraz ze sprawdzeniem ich sum kontrolnych.
2. Podstawowa weryfikacja stanu kasety, polegająca na ocenie kondycji kasety.

### Statusy kaset w Archiwum

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Wybranie kasety i zlecenie jej odczytu

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Ocena kondycji kasety podczas odczytu danych

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Rozpoczęcie i przeprowadzenie migracji kasety

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

### Procedura usuwania uszkodzonych kaset z Archiwum

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

## Administracyjna aplikacji GUI

Zostanie stworzona aplikacja webowa umożliwiająca wyświetlanie informacji o Systemie archiwizacji wraz z możliwością wykonywania określonych czynności w Systemie archiwizacji.

Aplikacja webowa będzie hostowana w ramach węzła Archiwizator.

Dostęp do aplikacji będzie ograniczony do użytkowników, którzy zautoryzują się za pomocą nazwy użytkownika oraz hasła przechowywanych w ramach Systemu archiwizacji.

Będzie możliwość skonfigurowania, do których z następujących funkcjonalności może mieć dostęp dany użytkownik:

1. Uprawnienie zbiorcze do widoku podglądu statusu napędów wraz z możliwością wywoływania dostępnych na tym widoku akcji.

Aplikacja GUI będzie posiadała własny plik zdarzeń biznesowych obejmujący następujące zdarzenia:

1. Zalogowania się użytkownika do GUI
2. Wylogowanie się użytkownika z GUI
3. Skorzystanie przez użytkownika z funkcjonalności Zarządzanie listą napędów.

Szczegóły działania aplikacji GUI:

1. Użytkownik będzie mógł się zalogować do aplikacji podając nazwę użytkownika i hasło. W przypadku niepoprawnych danych autoryzacyjnych lub jeżeli użytkownik będzie zablokowany, pojawi się komunikat: „Podane dane są nieprawidłowe”.
2. Po zalogowaniu użytkownik będzie mógł wybrać w menu podstronę:
   1. Użytkownicy
   2. Zarządzanie listą napędów
   3. Moje ustawienia
3. Na podstronie „Użytkownicy” administrator będzie mógł dodawać i edytować użytkowników oraz konfigurować ich uprawnienia.
4. W ramach definicji danych użytkownika będzie możliwość podania następujących informacji:
   1. Sekcja „Dane użytkownika”:
      1. Nazwa użytkownika (pole obowiązkowe)
      2. Hasło (pole obowiązkowe)
      3. Czy zablokowany: Tak/Nie (domyślnie wybrane Nie)
   2. Sekcja „Uprawnienia”:
      1. Użytkownicy
      2. Zarządzanie listą napędów

Podstrona „Moje ustawienia” będzie dostępna dla wszystkich zalogowanych użytkowników.

Nie będzie możliwości usuwania użytkowników – uniemożliwienie logowania się użytkownika będzie możliwa poprzez zablokowanie użytkownika.

1. Na podstronie „Moje ustawienia” użytkownik będzie mógł zmienić swoje hasło. W tym celu będzie musiał wprowadzić swoje poprzednie hasło oraz 2 razy nowe hasło.
2. System (podczas edycji użytkownika oraz podczas zmiany swojego hasła) będzie wymagał wprowadzenia hasła spełniającego następujące kryteria:
   1. Długość: minimum 8 znaków
   2. Liczba dużych liter: minimum 1
   3. Liczba cyfr: minimum 1

### Widok listy napędów w instancji Archiwum

Widok listy napędów w danej instancji Archiwum będzie niezbędny, aby móc odblokować napęd.

Administrator będzie miał możliwość wykonania następujących operacji:

1. Wybranie instancji Archiwum.
2. Wyświetlenie listy napędów wraz z ich statusem (niezablokowany, zablokowany).
3. Dla zablokowanego napędu będzie możliwość jego odblokowania.

Ta funkcjonalność wymaga wykonania zmiany w interfejsie API polegającej na dodaniu zasobu (metody) w interfejsie REST Systemu archiwizacji umożliwiającej pobranie listy napędów oraz odblokowanie wybranego napędu.

W ramach tego punktu zostanie dodane zdarzenie biznesowe odblokowana napędu.

# Wymagania funkcjonalne zmian w Repozytorium

W niniejszym rozdziale zostały umieszczone wymagania dotyczące zmian wprowadzanych w Repozytorium.

## Zmiany encji obiekt

### Status obiektu w kontekście archiwizacji

W ramach realizowanych zmian w Repozytorium powinien zostać dodany nowy status w obiekcie, status mówiący o tym, w jakim stanie pod kątem archiwizacji jest obiekt. Status ten będzie aktualizowany przez System archiwizacji.

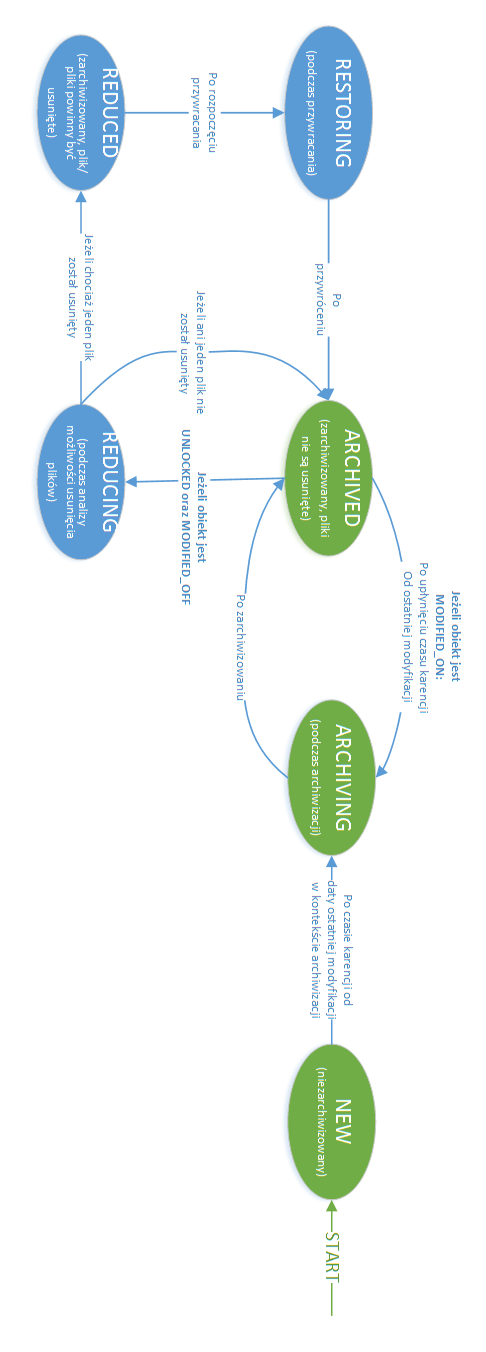
Diagram stanów i możliwości przejść między nimi został przedstawiony poniżej. Założenia do poniższego diagramu są następujące:

1. Po uruchomieniu Systemu archiwizacji, wszystkie obiekty będą posiadały status NEW. Po zmianie statusu NEW na inny – obiekt już nigdy go nie osiągnie.
2. Na niebiesko zaznaczono stany, w których obiekt jest zredukowany w Repozytorium, czyli chociaż jeden stream jest usunięty.
3. Na zielono zaznaczono stany, w których obiekt jest dostępny w Repozytorium, czyli ani jeden stream nie jest usunięty.
4. Pod pojęciem MODIFIED\_ON należy rozumieć sytuacje, w której „Data ostatniej modyfikacji w kontekście archiwizacji” jest późniejsza niż „Data ostatniej archiwizacji”
5. Po zakończeniu procesu próby usunięcia plików z obiektu status obiektu powinien zmienić się zgodnie z poniższym opisem:
   1. Jeżeli został usunięty chociażby 1 stream, to obiekt powinien przyjąć status REDUCED.
   2. Jeżeli nie został usunięty ani 1 stream, to obiekt powinien przyjąć status ARCHIVED.

Dodatkowo obiekt będzie mógł przyjąć status PROBLEM oznaczający, że podczas ostatniego procesu archiwizacji lub przywracania wystąpił problem z tym obiektem. Status ten będzie umożliwiał wykluczenie takiego obiektu podczas wybierania obiektów do ponownego procesowania.

W przypadku wystąpienia błędu powodującego potrzebę ustawienia kasecie statusu PROBLEM:

1. W dodatkowym polu zostanie zapisany status kasety sprzed zmiany tego statusu na PROBLEM.
2. W dodatkowym polu zostanie zapisany kod błędu, który wystąpił. Lista kodów błędów będzie zaimplementowana w aplikacji i nie będzie możliwości jej konfiguracji. W przypadku wystąpienia błędu nieposiadającego przypisanego kodu, zapisany zostanie kod błędu oznaczający UNDEFINED.



### Blokowanie obiektu (tzw. LOCK)

W ramach realizowanych zmian w Repozytorium powinien zostać dodany mechanizm blokowania obiektu. Chodzi o sytuacje, kiedy przykładowo proces OCR lub zamówień na pliki będzie za niedługo potrzebował danego obiektu, i chce zapobiec ewentualnej redukcji (usunięciu streamów) tego obiektu.

Status ten będzie aktualizowany przez Repozytorium, po stronie Systemu archiwizacji będzie sprawdzanie statusu zablokowania w kilku miejscach zdefiniowanych na diagramie statusów obiektu w miejscach, w których występuje sprawdzenie warunku LOCKED oraz UNCLOKED.

### Status streama w kontekście archiwizacji

W ramach realizowanych zmian w Repozytorium powinien zostać dodany nowy status w stream’ie mówiący o tym, w jakim stanie pod kątem archiwizacji jest stream. Status ten będzie aktualizowany przez System archiwizacji.

Diagram stanów i możliwości przejść między nimi został przedstawiony poniżej. Założenia do poniższego diagramu są następujące:

1. Po uruchomieniu Systemu archiwizacji, wszystkie streamy będą posiadały status NEW. Nowy stream (w rozumieniu dodany do istniejącego contentu) również będzie otrzymywał status NEW.
2. Na niebiesko zaznaczono stany, w których stream w Repozytorium jest zredukowany.
3. Na zielono zaznaczono stany, w których stream w Repozytorium jest dostępny.



### Modyfikacja obiektu w kontekście archiwizacji

Przez modyfikację obiektu w kontekście archiwizacji rozumiane są następujące zmiany w obiekcie:

1. Dodanie nowego kontentu: postaci zbiorczej lub kontentu w strukturze paginacyjnej z przynajmniej jednym stream’em.
2. Dodanie lub podmiana (np. przez proces OCR) w przynajmniej jednym istniejącym kontencie stream’a znajdującego się na konfigurowalnej liście archiwizowanych stream’ów (konfiguracja brana pod uwagę podczas budowana paczki archiwalnej).

Modyfikacja obiektu w kontekście archiwizacji powinna powodować aktualizację daty ostatniej modyfikacji w kontekście archiwizacji.

Uwaga: przywrócenie plików przez System archiwizacji do zarchiwizowanego obiektu nie może powodować aktualizacji daty ostatniej modyfikacji.

Aby System archiwizacji mógł prawidłowo zakwalifikować obiekt do archiwizacji, powinien zostać sprawdzony warunek czy minął okres karencji od zakończenia przez obiekt workflow. Sprawdzanie daty zakończenia workflow w oparciu o datę wykonania ostatniego kroku workflow (encja „activity” w bazie danych Repozytorium) może powodować nadmierne obciążenie bazy danych Repozytorium, dlatego planowane jest inne rozwiązanie polegające na tym, aby po wykonaniu każdego kroku workflow aktualizowana była data ostatniej modyfikacji w kontekście archiwizacji. Dzięki czemu będzie możliwe sprawdzenie wspólnego okresu karencji od zakończenia workflow i ostatniej modyfikacji w kontekście archiwizacji, bez potrzeby odwoływania się do encji „activity” w bazie danych Repozytorium.

### Data ostatniej próby redukcji

Do obiektu w Repozytorium powinna zostać dodana data ostatniej próby redukcji. Data ta powinna być uaktualniana przez System archiwizacji za każdym razem:

1. Po udanym zakończeniu redukcji obiektu (czyli po przejściu obiektu ze statusu REDUCING w status REDUCED)
2. Jak również po zakończeniu redukcji obiektu bez usunięcia jakiegokolwiek streamu (czyli po przejściu obiektu ze statusu REDUCING w status ARCHIVED).

W przypadku ustawienia bardzo wysokiej wartości okresu karencji od ostatniej próby redukcji (np. 360 dni), może dojść do sytuacji, w której obiekt po próbie redukcji został zmodyfikowany i przeszedł ponowną archiwizację, ale nie jest redukowany z uwagi na to, że nie minął jeszcze okres karencji od ostatniej próby redukcji. Dlatego data ostatniej próby redukcji powinna być czyszczona w przypadku zmodyfikowania obiektu, jak również w przypadku pomyślnego zakończenia procesu przywracania obiektu.

### Konfiguracja okresów karencji

Powinna być możliwość skonfigurowania okresu karencji od ostatniej modyfikacji w kontekście archiwizacji, po którym powinna zostać wykonana ponowna archiwizacja, oraz okresu karencji od ostatniego przywrócenia, po którym powinna zostać wykonana próba zredukowania obiektu.

Z uwagi na to, że ta konfiguracja będzie wykorzystywana zarówno przez System archiwizacji, jak i przez Repozytorium, to konfiguracja ta powinna spełniać następujące wymagania:

1. Konfiguracja okresu karencji od przywrócenia i od ostatniej modyfikacji w kontekście archiwizacji powinna być przechowywana w bazie danych Repozytorium – tabela archival\_config:
   1. modification\_timeout – po jakim czasie od ostatniej modyfikacji obiektu w kontekście archiwizacji, usługa pomocnicza Systemu archiwizacji może przystąpić do archiwizacji obiektu (zarówno pierwszej archiwizacji, jak i kolejnej). Czas ten jest podawany w dniach jako liczba całkowita.
   2. restore\_timeout – po jakim czasie od przywrócenia obiektu do Repozytorium, usługa pomocnicza Systemu archiwizacji może przystąpić do redukcji obiektu. Zostanie przygotowana metoda do pobierania wartości tych konfiguracji z Repozytorium – z której to będzie korzystał System archiwizacji. Czas ten jest podawany w dniach jako liczba całkowita.
   3. reduction\_timeout – okresu karencji w dniach od ostatniej próby redukcji. Aktualizacja wartości tego parametru nie będzie wymagała zatrzymania systemu Repozytorium cyfrowe. Parametr ten będzie definiowany z dniach.
2. W Systemie archiwizacji ta konfiguracja będzie cache’owana przez czas zdefiniowany w parametrach Systemu archiwizacji (parametr w konfiguracji: archivization.config.refresh.delay).

# Monitorowanie usług Systemu archiwizacji

### Monitorowanie miejsca na macierzach pomocniczych przez Zarządcę

Zarządca każdego węzła powinien monitorować dostępne miejsce na przypisanej do tego węzła macierzy pomocniczej.

Monitorowanie powinno być wykonywane osobno dla miejsca przydzielonego do operacji odczytu, a osobno do operacji zapisu danych.

Stopień zajętego miejsca powinien być sprawdzany cyklicznie co zdefiniowany czas, np. co 5 minut.

W przypadku, w którym stopień zajętego miejsca dla danego typu operacji (zapis, odczyt) będzie wyższy niż skonfigurowany pierwszy próg, w dzienniku zdarzeń powinno zostać zarejestrowane zdarzenie o poziomie INFO.

W przypadku, w którym stopień zajętego miejsca dla danego typu operacji (zapis, odczyt) będzie wyższy niż skonfigurowany drugi próg, w dzienniku zdarzeń powinno zostać zarejestrowane zdarzenie o poziomie WARNING wraz z wysłaniem powiadomienia do administratora. Powiadomienia o przekroczeniu drugiego progu wykorzystania macierzy powinny być wysyłane z częstotliwością ustawioną w ramach parametrów systemu, parametr message.delay.threshold.second=10m (20s, 20h itd.), który powinien być odświeżany dynamicznie i być wyrażony w sekundach.

### Monitorowanie miejsca na dyskach twardych

Narzędzie monitorujące powinno weryfikować ilość ogólnego miejsca na widocznych dyskach twardych/woluminach systemu operacyjnego każdego z węzłów.

W przypadku, w którym stopień zajętego miejsca na dysku/woluminie będzie wyższy niż skonfigurowany pierwszy próg, w logu systemu monitorującego powinno zostać zarejestrowane zdarzenie wraz z wysłaniem powiadomienia do administratora.

W przypadku, w którym stopień zajętego miejsca dla danego typu operacji (zapis, odczyt) będzie wyższy niż skonfigurowany drugi próg, w logu systemu monitorującego powinno zostać zarejestrowane zdarzenie wraz z wysłaniem powiadomienia do administratora.

## Monitorowanie usług

Narzędzie monitorujące powinno weryfikować czy są uruchomione następujące usługi:

1. Miejsce na partycji głównej (z zainstalowanym systemem).
2. Miejsce na macierzy pomocniczej (na pliki).
3. Wszystkie usługi Systemu archiwizacji na danym węźle + połączenie do serwera MySQL danego węzła.
4. Działanie serwera MySQL oraz brokera Artemis MQ.

W przypadku wykrycia problemu narzędzie monitorujące powinno:

1. Zalogować problem do swojego dziennika zdarzeń.
2. Wysłać powiadomienie do administratora.
3. Zrestartować usługę/proces z którym wystąpił problem.

# Procedura wyłączenia i włączenia Systemu archiwizacji

System archiwizacji składa się z następujących elementów:

1. Podstawowe: usługi pomocnicze, workery, zarządcy.
2. Wspomagające: instancje systemu kolejkowego oraz baz danych

Dodatkowo uruchomiona jest usługa monitorująca działanie usług podstawowych i wspomagających Systemu archiwizacji.

W Systemie archiwizacji występują następujące węzły:

1. Archiwizator, który zawiera następujące elementy:
   1. Usługi pomocnicze
   2. Zarządca udostępniający interfejs API
   3. Workery
   4. Baza danych
   5. System kolejkowy
2. Archiwum, który zawiera następujące elementy:
   1. Zarządca
   2. Workery
   3. Baza danych
   4. System kolejkowy

Pod pojęciem workera rozumiana jest mikrousługa realizująca zadania zlecone przez Zarządcę.

## Wyłączenie Systemu archiwizacji

Wyłączenie Systemu archiwizacji polega na wyłączeniu podstawowych oraz wspomagających elementów Systemu archiwizacji.

Bezpieczne wyłączenie się podstawowych elementów Systemu archiwizacji oznacza dokończenie wykonywanych właśnie elementarnych zadań oraz niepodejmowanie wykonywania następnych. Pod pojęciem elementarnego zadania rozumie się przykładowo zadanie skopiowania pliku paczki archiwalnej z Archiwizatora do Archiwum. Różne elementarne zadania składają się na zadanie główne – przykładowo zadanie archiwizacji konkretnego obiektu. Dodatkowo wymagane jest zatrzymanie elementów wspomagających: wykorzystywanych instancji systemów kolejkowych oraz instancji baz danych.

Wyłączenie podstawowych elementów Systemu archiwizacji powinno być możliwe poprzez narzędzie monitorujące System archiwizacji.

Administrator powinien móc uruchomić funkcjonalność bezpiecznego wyłączenia się:

1. Wszystkich podstawowych elementów z wszystkich węzłów Systemu archiwizacji.
2. Wszystkich podstawowych elementów z wybranego węzła Systemu archiwizacji.
3. Usług pomocniczych w ramach węzła „Archiwizator”.

Kolejność wyłączenia Systemu archiwizacji jest następująca:

1. Wyłączenie następujących elementów podstawowych: usługi pomocnicze.
2. Wyłączenie następujących elementów podstawowych: zarządcy(ów), workera(ów).
3. Wyłączenie elementów wspomagających.
4. Wyłączenie usługi dodatkowej do monitorowania Systemu archiwizacji.

Po każdym powyższym kroku administrator powinien weryfikować, czy elementy przewidziane w tym kroku zakończyły swoje działanie.

Ewentualne wyłączenie systemu Repozytorium powinno nastąpić dopiero po wyłączeniu Systemu archiwizacji.

Ewentualne wyłączenie serwera pocztowego wykorzystywanego przez System archiwizacji do wysyłania powiadomień do administratorów powinno nastąpić dopiero po wyłączeniu Systemu archiwizacji.

Decyzja, które konkretnie usługi (wszystkie czy tylko wybrane) powinny zostać wyłączone należy do administratora: czy wyłączony powinien zostać cały System archiwizacji, wybrany węzeł lub wybrane usługi (np. tylko usługi pomocnicze w ramach Archiwizatora).

### Wyłączenie elementów podstawowych Systemu archiwizacji

Wyłączenie elementów wspomagających Systemu archiwizacji powinno być wykonywane poprzez zatrzymanie za pomocą narzędzia monitorującego System archiwizacji usług w następującej kolejności:

1. Usługi pomocnicze
2. Workery Archiwum
3. Zarządca Archiwum
4. Workery Archiwizatora
5. Zarządca Archiwizatora

### Przebieg wyłączenia elementów podstawowych Systemu archiwizacji

Przebieg wyłączenia się podstawowych elementów Systemu archiwizacji został opisany poniżej.

Usługi pomocnicze

Po uruchomieniu procedury wygaszenia usługi pomocniczej usługi te powinny zaprzestać realizacji swoich zadań. Jeżeli jakieś zadanie usługi pomocniczej jest już w trakcie obsługi, to powinno ono zostać dokończone.

Następnie usługa powinna zakończyć swoje działanie.

Zarządca Archiwizatora

Po uruchomieniu wyłączenia się usługi zarządcy Archiwizatora interfejs Systemu archiwizacji powinien zaprzestać przyjmowania i obsługi wszystkich nowych żądań – niezależnie od typu żądania, to znaczy czy jest to przykładowo żądanie przywrócenia obiektu, archiwizacji czy też pytanie o status wcześniejszego żądania. Jeżeli jakieś żądanie jest już w trakcie obsługi – powinno zostać dokończone.

Po wyłączeniu nasłuchiwania powinno nastąpić wyczyszczenie listy dostępowych tokenów.

Następnie usługa zarządcy Archiwizatora powinna:

1. Zaprzestać odbierania nowych wiadomości z instancji Archiwum.
2. Dokończyć wykonywanych właśnie elementarnych zadań/czynności.
3. Nie podejmować nowych zadań/czynności.
4. Samoczynnie zakończyć swoje działanie.

Zarządca Archiwum

Usługa zarządcy Archiwum powinna:

1. Zaprzestać odbierania nowych wiadomości z instancji Archiwizatora.
2. Dokończyć wykonywanych właśnie elementarnych zadań/czynności.
3. Nie podejmować nowych zadań/czynności.
4. Samoczynnie zakończyć swoje działanie.

Workery Archiwizatora oraz Archiwum

Usługi workerów powinny:

1. Zaprzestać odbierania nowych wiadomości od swojego Zarządcy.
2. Dokończyć wykonywanych właśnie elementarnych zadań.
3. Nie podejmować nowych zadań.
4. Samoczynnie zakończyć swoje działanie.

Narzędzie do dodawania nowych kaset

Narzędzie do dodawania nowych kaset do puli do wykorzystania – w ramach tego elementu należy zaimplementować mechanizm sprawdzenia, czy są dostępne usługi wspomagające wymagana przez aplikację (baza danych oraz system kolejkowy). W przypadku braku dostępności powinien zostać wyświetlony komunikat „Trwa wyłączanie Systemu archiwizacji, spróbuj ponownie później.”.

Narzędzie do zarządzania użytkownikami interfejsu API

Narzędzie do zarządzania użytkownikami interfejsu API – w ramach tego elementu należy zaimplementować mechanizm weryfikacji, czy jest dostępna baza danych, jeżeli nie to powinien zostać wyświetlony odpowiedni komunikat.

### Wyłączenie elementów wspomagających Systemu archiwizacji

Wyłączenie elementów wspomagających Systemu archiwizacji powinno polegać na zatrzymaniu za pomocą narzędzia monitorującego System archiwizacji właściwych usług wyłączanych węzłów: usług baz danych oraz systemów kolejkowych

### Wyłączenie usługi dodatkowej do monitorowania Systemu archiwizacji

Sposób wyłączenia narzędzia do monitorowania Systemu archiwizacji został opisany w ramach dokumentacji deweloperskiej (tzw. ReadMe).

## Włączenie Systemu archiwizacji

Włączenie Systemu archiwizacji polega na włączeniu usługi dodatkowej do monitorowania Systemu archiwizacji i za jej pomocą włączenie elementów podstawowych oraz wspomagających

W momencie uruchamiania Systemu archiwizacji system Repozytorium powinien być dostępny.

W momencie uruchamiania Systemu archiwizacji serwer pocztowy wykorzystywany przez System archiwizacji do wysyłania powiadomień do administratorów powinien być dostępny.

Kolejność włączenia Systemu archiwizacji jest następująca:

1. Włączenie usługi dodatkowej do monitorowania Systemu archiwizacji.
2. Włączenie wszystkich elementów wspomagających.
3. Włączenie następujących elementów podstawowych: zarządcy(ów), workera(ów).
4. Włączenie następujących elementów podstawowych: usługi pomocnicze.

Po każdym powyższym kroku administrator powinien weryfikować, czy elementy przewidziane w tym kroku rozpoczęły swoje działanie.

Decyzja, które konkretnie usługi (wszystkie czy tylko wybrane) powinny zostać włączone zależy od tego, które z usług zostały wyłączone, to znaczy czy wyłączony został cały System archiwizacji czy tylko wybrany węzeł lub wybrane usługi (np. tylko usługi pomocnicze).

### Włączenie usługi dodatkowej do monitorowania Systemu archiwizacji

Sposób uruchomienia narzędzia do monitorowania Systemu archiwizacji został opisany w ramach dokumentacji deweloperskiej (tzw. ReadMe).

### Włączenie elementów wspomagających

Włączenie elementów wspomagających Systemu archiwizacji powinno polegać na uruchomieniu za pomocą narzędzia monitorującego System archiwizacji właściwych usług włączanych węzłów: usług baz danych oraz systemów kolejkowych

### Włączenie elementów podstawowych

Włączenie elementów wspomagających Systemu archiwizacji powinno polegać na uruchomieniu za pomocą narzędzia monitorującego System archiwizacji usług w następującej kolejności:

1. Zarządca Archiwizatora
2. Workery Archiwizatora
3. Zarządca Archiwum
4. Workery Archiwum
5. Usługi pomocnicze

# Schemat i opis struktury bazy danych

## Baza danych Archiwizator

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę oświadczenia zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.

## Baza danych Archiwum

Zawartość rozdziału zostanie udostępniona po podpisaniu przez Wykonawcę oświadczenia zobowiązania do zachowania poufności i dostarczeniu oryginału tego dokumentu do Zamawiającego.