Dokumentacja API PBN w wersji v1.1

wersja dokumentacji: 1.0

20 września 2021

1 Wstęp

Dokument zawiera opis oraz przykłady użycia API PBN w wersji v1.1.

1.1 Struktura PBN

W systemie PBN można wyróżnić dwa moduły: Moduł Repozytoryjny oraz Profil Instytucji.

- Moduł Repozytoryjny: część systemu udostępniająca możliwość dodania publikacji do bazy PBN oraz ustanowienia powiązań pomiędzy encjami np. powiązanie autora z ORCID, powiązanie autora z POLON, powiązanie publikacji z ORCID itp. W ramach tego modułu istnieje Profil Autora, w którym prezentowany jest dorobek publikacyjny danej osoby. Większość danych dostępnych w tym module została wprowadzona ręcznie przez importerów publikacji oraz autorów.
- Profil Instytucji: część systemu gromadząca dane związane z instytucją. W ramach tego profilu osoby wyznaczone przez instytucje naukowe nadzorują proces gromadzenia dorobku naukowego. Dane zgromadzone w profilu przekazywane są do systemu SEDN.

Warto nadmienić, że aktualne wersje encji występujące w Module Repozytoryjnym i Profilu Instytucji mogą się różnić. W obecnym kształcie API celem importu danych jest zasilenie danych w Profilu Instytucji. Na podstawie aktualnych analiz oraz planowanych zmian walidacji w systemie, PBN w przyszłości będzie dążył do ujednolicenia zapisu publikacji w obu tych modułach.

Na potrzeby ewaluacji konieczne jest wykazanie osiągnięcia w ramach dyscypliny w danej instytucji. Obligatoryjny jest zatem zapis oświadczenia, tj. zapis publikacji w Profilu Instytucji. Poprawnie przesłane przez API dane publikacji, bez względu na czy była ona przedmiotem dodawania lub edycji, zostaną wiernie odwzorowane przy zapisie w Profilu Instytucji.

1.2 PBN a jakość danych

System PBN jest systemem, do którego dane wprowadzane są z tysięcy źródeł przez dużą liczbę użytkowników. W naturalny sposób dochodzi do pewnych różnic w podawanych danych, mimo iż dotyczącą one tego samego bytu. Wśród takich różnic mogą być znaki interpunkcyjne w tytułach publikacji bądź czasopism, pomylenie numerów ISSN oraz eISSN, niepełne dane, a także błędy ludzkie jak literówki.

1.2.1 Walidacja danych

Akceptowanie różnych danych dla tych samych bytów prowadziłoby do powstawawania duplikatów i niskiej jakości zawartości bibliograficznej. Dlatego system jest rygorystyczny pod względem jakości wprowadzanych danych oraz uniemożliwia generowanie duplikatów wynikających z przesyłanych braków. Taka strategia podwyższa jakość zawartości PBN, ale jednocześnie prowadzi do konfliktów, które są odczuwalne dla użytkowników systemu. Należy pamiętać, że sprzeczności pomiędzy danymi wprowadzanymi do PBN oraz danymi zapisanymi w bazie danych są naturalne i niemożliwe do uniknięcia, jednak wysokie wymagania na jakość danych owocują wyższą jakością zawartości bibliograficznej.

1.2.2 Duplikaty w systemie PBN

W systemie PBN mogą występować duplikaty rekordów. Sytuacja ta jest efektem zasilenia systemu danymi z innych systemów bibliograficznych. Obecnie realizowany jest proces automatycznej masowej deduplikacji rekordów. Jednak proces ten jest złożony, wymagający oraz musi być przeprowadzany ze szczególną ostrożnością, w związku z tym został rozłożony w czasie. W przyszłości należy się spodziewać polepszenia jakości danych w tym aspekcie.

19.08.2021 r. usunięto z bazy 172081 wydawnictw, które zostały uznane za duplikaty. W dalszej kolejności deduplikacja dotyczyć będzie czasopism, książek pod redakcją oraz konferencji.

1.3 Identyfikacja encji

Jedną z podstawowych zasad przyjętych przy konstrukcji API jest zasada wyższości mniswId nad objectId. Zasada ta wykorzystywana jest przy wyszukiwaniu danych encji, która ma zostać powiązana w ramach importu. Przyjęto hierarchię priorytetów uwzględniającą ID:

- Domyślnie: mniswId > objectId
- Person: naturalId > orcidId > objectId
- Publication: objectId > doi i type

Gdzie oznaczenie mniswId > objectId określa, że system uważa mniswId za ważniejszy i w przypadku jego wystąpienia ignoruje pozostałe identyfikatory.

Ważne: w przypadku nie podania identyfikatora w ramach encji system PBN spróbuje ją automatycznie dodać do bazy danych co może prowadzić do utworzenia duplikatów.

1.4 Walidacja encji

Reguły dotyczące jakości danych przesyłanych do PBN oraz ich powiązań z danymi w bazie zostały przedstawione w dokumencie dotyczącym walidacji dostępnym w pomocy PBN. W związku z tym niniejszy dokument nie porusza szczegółów związanych z tymi zasadami.

1.5 Import encji

PBN udostępnia końcówki dotyczących importowania encji do systemu PBN. Końcówki te mają za zadanie import w **kontekście oświadczenia danej instytucji**. Dla książki, artykułu i rozdziału informacje dotyczące oświadczenia muszą być częścią żądania i są wymagane, dla książki pod redakcją opcjonalne, natomiast dla tomu pokonferencyjnego nie powinny być przesyłane. W przypadku, gdy encja nie istniała w bazie danych, tzn. identyfikacja na podstawie id zakończyła się niepowodzeniem, tworzony jest nowy byt w Repozytorium (wraz z powiązaniami) oraz w Profilu Instytucji. W przeciwnym wypadku encja jest poddana edycji całościowej wyłącznie w Profilu Instytucji, natomiast w Repozytorium zapisane są jedynie podstawowe metadane publikacji jak: tytuł, rok, tom, strony itp.

W ramach edycji w Repozytorium nie zostaną zapisane afiliacje, a także powiazania z innymi encjami, np. osobami bądź czasopismami. Reguła ta ma na celu ochronę przed nadpisaniem pracy człowieka przez błędy w działaniu skryptów, bądź podawaniu masowo nieprawidłowych lub wybrakowanych danych. Warto zwrócić uwagę, że zablokowana jest edycja, która dotyczy głównie publikacji, które zostały dodane przez użytkowników systemu ręcznie.

W PBN stosowana jest zasada, która uznaje, że publikacje wprowadzone przez osoby podlegają weryfikacji merytorycznej, która jest ceniona wyżej niż weryfikacja automatyczna wykonywana przez oprogramowanie. Zasada ta ma na celu ochronę pracy wykonanej przez użytkowników. Jednak w przyszłości, zgodnie z przeprowadzonymi analizami planowane zmiany w walidacji obiektów pozwolą na ujednolicenie zapisów, zgodnie z wcześniej przedstawionym opisem 1.1.

2 Uzyskanie uprawnień oraz sposób autentykacji

System PBN wykorzystuje metodę autentykacji trzystopniowej opartej o tokeny. W ramach uwierzytelnienia klient API musi przesłać identyfikator aplikacji oraz dwa unikatowe tokeny: aplikacji oraz użytkownika.

2.1 Pozyskanie tokenu aplikacji

Dodanie aplikacji do systemu - Pierwszym krokiem umożliwiającym dodanie aplikacji do systemu jest wystapienie z prośbą o nadanie tokenu aplikacji na wersji testowej PBN (PBN Alpha). Token nadawany jest na prośbę osoby z rolą Administratora Polon danej instytucji. Podczas rejestracji aplikacji system PBN generuje token aplikacji oraz identyfikator aplikacji.

2.2 Pozyskanie jednorazowego tokenu użytkownika

W celu uzyskania tokenu użytkownika wymagany jest etap pośredni, który przy pomocy logowania w systemie PBN uwierzytelni, że klient API ma odpowiednie uprawnienia. W tym celu należy wygenerować jednorazowy kod (One Time Token - OTT). Kod ten może zostać wykorzystany do uzyskania tokenu użytkownika, który pozwoli na dalsze korzystanie z API bez konieczności logowania.

OTT wysyłane jest na zarejestrowany w systemie PBN adres zwrotny podany w ramach rejestracji aplikacji. Takie rozwiązanie pozwala na automatyzację procesu komunikacji z systemem PBN i ograniczenie niezbędnych interakcji do logowania.

2.3 Pozyskanie tokenu użytkownika

Token użytkownika należy pobrać, wysyłając żądanie do systemu PBN, które zawiera OTT, id aplikacji oraz token aplikacji. Token należy pobrać przez odpowiednie zapytanie, przykład został podany w instrukcji przedstawionej w pomocy PBN.

W wyniku procedury wystawiony zostanie token użytkownika, który pozwoli systemowi PBN zidentyfikować klienta API.

2.4 Dobre praktyki

- Długość życia wystawionych tokenów może być ograniczona. W zależności od przyjętych zasad bezpieczeństwa ich termin ważności może zostać skrócony i wymusić na użytkownikach ponowne wygenerowanie tokenu. Zasady te w systemie PBN mogą ulec zmianom w przyszłości. Sugerujemy, aby aplikacje wykonywały procedurę pozyskania tokenów w sposób automatyczny.

3 Scenariusze

W rozdziale przyjrzymy się bliżej scenariuszom korzystania z API. Postaramy się omówić je od strony biznesowych reguł działania systemu PBN.

3.1 Import artykułu

API PBN udostępnia dwie końcówki do dodawania publikacji:

1. `/publications` - służy do importowania pojedynczej publikacji;
2. `/publications/import` - służy do importowania listy publikacji.

W celu zaimportowania artykułu należy przygotować obiekt JSON zawierający zawartość przygotowaną do importu. System PBN sprawdza poprawność przesłanych danych i nie dopuści do importu w przypadku wykrycia błędów walidacji.

Istnieją dwa rodzaje artykułów akceptowanych przez system: wydane w czasopiśmie bądź związane z konferencją. W związku z tym wymagane jest podanie danych czasopisma albo konferencji. W przypadku braku tych danych zwrócony zostanie błąd.

3.1.1 Identyfikacja czasopisma po id

Użytkownik może podać jedynie objectId albo mniswId. System PBN wyszuka odpowiednie czasopismo i połączy je z importowanym artykułem. Jest to zalecana opcja.

3.1.2 Podanie danych czasopisma

Ten tryb uruchamiany jest kiedy klient nie poda wartości objectId bądź mniswId. W pierwszej kolejności system PBN spróbuje odnaleźć czasopismo o podanym numerze ISSN bądź eISSN. Jeżeli w bazie danych istnieje takie czasopismo zwrócony zostanie błąd walidacji wraz z podanym identyfikatorem czasopisma bedącym w konflikcie. Użytkownik, może wykorzystać ten identyfikator do wygenerowania żądania 3.1.1.

Ważne: w przypadku wyżej opisanego błędu zweryfikuj czy dane czasopisma w bazie PBN o podanym ID są zgodne z czasopismem przesłanym w JSONie.

Jeżeli system nie odnalazł w bazie danych czasopisma o podanych ISSN/eISSN oraz dane są poprawne system doda nowe czasopismo. Warto zwrócić uwagę, że walidacja wydawcy odbywa się w podobny sposób do walidacji czasopisma. System próbuje odnaleźć wydawcę o podanym Id albo utworzyć nowego wydawcę o podanej nazwie.

Walidacja danych czasopisma obejmuje:

1. tytuł (`title`);
2. numer ISSN lub eISSN (`issn` lub `eIssn`);
3. dano o wydawcy (`publisher` zagnieżdżone w `journal`);
4. strona internetowa czasopisma (`websiteLink`).

3.1.3 Identyfikacja serii konferencji i edycji konferencji po id

Tak samo jak ww. przypadku artykułu z czasopisma, użytkownik może podać w przesłanej konferencji identyfikatory: objectId oraz mniswId. Jeżeli zarówno seria jak i edycja konferencji zostały na podstawie tych identyfikatorów odnalezione w systemie PBN, weryfikowane jest czy ta edycja jest powiązana z tą serią. Jeżeli rezultat weryfikacji był pozytywny podana seria i edycja konferencji zostaną połączone z artykułem. W przeciwnym razie zostanie zwrócony błąd walidacji.

3.1.4 Podanie danych serii konferencji i edycji konferencji

Pominięcie identyfikatorów objectId lub mniswId w ramach przesłanej konferencji interpretowane jest jako zamiar dodania nowej konferencji. Wobec tego dane tej konferencji zostają poddane walidacji. Seria i edycja podawane są w JSONie w oddzielnych węzłach.

Walidacja danych edycji konferencji obejmuje:

1. pełną nazwę konferencji (`fullName`);
2. skróconą nazwę (`shortName`);
3. stronę internetową z uwzględnieniem czy podany adres url ma właściwy format (`website`);
4. kraj (`country`);
5. miejscowość (`city`);
6. data początku (`startDate`);
7. data końca konferencji (`endDate`);

Walidacja danych serii konferencji obejmuje:

1. pełną nazwę konferencji (`fullName`);
2. skróconą nazwę (`shortName`);
3. stronę internetową z uwzględnieniem czy podany adres url ma właściwy format (`website`);
4. walidacja pod kątem unikalności ww. atrybutów względem stanu bazy PBN.

3.1.5 Opcjonalne wskazanie materiału pokonferencyjnego

Do artykułu można przypisać także tom pokonferencyjny. Jednak w tym przypadku możliwe jest to wyłącznie poprzez przesłanie objectId publikacji w `proceedings`.

3.1.6 Podanie danych sprawozdającego

System wymaga przekazania danych dotyczących oświadczenia co najmniej jednej osoby podanej w ramach publikacji jako autor. Dane identyfikacyjne podane w przesyłanej encji Person muszą być tożsame z danymi identyfikacyjnymi podanymi w ramach oświadczenia. Patrz: 1.4.

4 Synchronizacja danych

W celu zapobieganiu powstawania duplikatów oraz niedeterministycznej sekwencji edycji PBN synchronizuje żądania napływające do API. Konsekwencją tego zachowania może być odmowa pracy na obiektach, które są przetwarzane równolegle przez innych klientów. Dla przykładu rozważmy sytuację, w której pierwszy klient chce dodać publikację A, równolegle drugi klient dodaje publikację B o doi identycznym z doi A. W zależności od kolejności otrzymania żądań możliwe są trzy efekty:

1. Żądanie z A zostanie zarejestrowane jako pierwsze i na czas jego wykonania zablokuje wszystkie inne żądania zmieniające stan obiektów z doi z A. Żądanie z B zostanie odrzucone z kodem błędu HTTP 423 Locked. W efekcie do PBNu zapisana zostanie wyłącznie publikacja A.
2. Żądanie z B zostanie zarejestrowane jako pierwsze i na czas jego wykonania zablokuje wszystkie inne żądania zmieniające stan obiektów z doi z B. Żądanie z A zostanie odrzucone z kodem błędu HTTP 423 Locked. W efekcie do PBNu zapisana zostanie wyłącznie publikacja B.
3. Jeżeli żądanie, które zostanie zarejestrowane jako pierwsze będzie niepoprawne w dalszych krokach walidacji, w PBNie nie zostanie zapisana żadna publikacja.

Sytuacja konfliktu synchronizacji występuje niezwykle rzadko. W przypadku jej wystąpienia zalecamy powtórzyć żądanie po odczekaniu krótkiego czasu.

5 Dobre praktyki

W rozdziale postaramy się przedstawić ogólne zalecenia dotyczące korzystania z API PBN.

5.1 Korzystaj z trybu batch

W przypadku większej liczby encji staraj się korzystać z końcówek umożliwiających wysyłkę list encji, zamiast wysyłać serię zapytań na końcówki przyjmujące pojedyńcze encje. Pozwoli to ograniczyć obciążenie związane z narzutem komunikacyjnym a jednocześnie zwiększyć wydajność procesu. Należy się też spodziewać, że wraz z rozwojem API PBN, końcówki te będą coraz lepiej zoptymalizowane, jeszcze bardziej zwiększając wydajność względem końcówek obłsugujących pojedyncze zapytania. Jeżeli liczba danych jest zbyt duża (np. setki tysięcy encji) podziel dane na mniejsze paczki.

5.2 Korzystaj z identyfikatorów

Jeżeli to możliwe staraj się wysyłać dane o encjach zawierające identyfikatory (`objectId`, `mniswId`, `naturalId` itp.). Korzystanie z identyfikatorów daje gwarancję, że obiekt zostanie powiązany jednoznacznie z istniejącą już encją w systemie. Przed wysłaniem danych do PBN skorzystaj z końcówek wyszukiwarek `/search`.

1. Przygotuj listę obiektów dla każdego typu encji: Person, Journal, Publisher itp. na podstawie danych, które zamierzasz zaimportować w danym kroku iteracji.
2. Skorzystaj z wyszukiwarek `/search/*` do zmapowania listy obiektów z Twojego systemu na objectId. Użyj cache'u do przechowywania mapowania encja → objectId.
3. Na podstawie mapperów przygotuj finalny format JSONa wysyłanego do PBN, wykorzystując jedynie identyfikatory wszędzie tam, gdzie to możliwe.
4. Postaraj się przechowywać w pamięci pobrane mappery aż do końca całej procedury importu (wskazane współdzielenie cache pomiędzy iteracji procedury batchowej).

5. Przed odpytaniem `/search/*` upewnij się, że Twoje mappery nie posiadają już mapowania dla tej encji.

Po wykonaniu procedury importu skasuj mappery. Dane w PBN są dynamiczne i mogą się zmieniać w dłuższych okresach, z tego względu nie zalecamy przechowywania mapperów w cache'u przez czas dłuższy niż kilkanaście godzin.

6 Częste pytania i przykłady

1. **Pytanie:** system PBN przy odpytaniu `/search/*` zwrócił wiele czasopism o tym samym ISSN. Które objectId wybrać?

   Odpowiedź: sytuacja ta jest wynikiem duplikacji rekordów w bazie. Jeżeli wyszukiwane czasopismo pochodzi z ministerialnego wykazu czasopism to zalecamy wybranie encji zawierającej mniswId. Wyszukiwanie domyślnie jest ograniczone tylko do czasopism z mniswId (parametr `mnisw:true`). Jeżeli czasopismo nie pochodzi z ministerialnego wykazu czasopism PBN, zwraca wszystkie wersje wprowadzone przez użytkowników. Klient powinien wybrać encję, która najlepiej odpowiada tej, którą chce powiązać z publikacją. Warto podkreślić, że wraz z postępem w procesie deduplikacji liczba zduplikowanych rekordów, także względem ISSN ulegnie zdecydowanej redukcji (patrz: 1.2.2).

2. **P:** przy odpytaniu `/search/*` zwrócił wiele czasopism o tym samym ISSN, ale żadna z wersji nie pasuje do tej, którą chcę sprawozdać. W jaki sposób wprowadzić swoją wersję czasopisma?

   O: przez wzgląd na zaostrzoną walidację, która ma na celu polepszanie jakości danych dodawanie nowych czasopism o tym samym ISSN zostało zablokowane. W przypadku braku poprawnej encji w bazie należy wyedytować ją przez interfejs graficzny udostępniony na stronie `https://pbn.nauka.gov.pl`.

3. **P:** publikacje zostały poprawnie przesłane przez API, ale w na stronie `https://pbn.nauka.gov.pl` nie zgadzają się dane autorów publikacji. Co jest przyczyną tej sytuacji?

   O: obecnie w niektórych przypadkach system PBN nie zapisuje pełnej wersji publikacji w Module Repozytoryjnym. Celem przesyłania publikacji przez API jest zapis publikacji w Profilu Instytucji. Pełna wersja publikacji powinnna zostać zapisana na tym profilu, natomiast w module repozytoryjnym zapisywane są dane, które nie są w konflikcie z obecną wersją w obszarze powiązań pomiędzy encjami. W przyszłości PBN będzie dążył do unifikacji zapisu (patrz: 1.1).

7 Ważne linki

1. Swagger API PBN v1
2. Walidacja encji
3. Autentykacja klienta
4. Centrum pomocy