

STANDARZY INTEROPERACYJNOŚCI I STANDARDY OTWARTOŚCI DANYCH

Podmioty realizujące zadania publiczne są zobowiązane do przestrzegania zasad wymiany informacji pomiędzy systemami teleinformatycznymi umożliwiającymi szybki i sprawny przepływ informacji oraz jej efektywne przetwarzanie. Zasady te zostały wprowadzone rozporządzeniem Rady Ministrów z dnia 12 kwietnia 2012 r. w sprawie Krajowych Ram Interoperacyjności, minimalnych wymagań dla rejestrów publicznych i wymiany informacji w postaci elektronicznej oraz minimalnych wymagań dla systemów teleinformatycznych (Dz. U. poz. 526). W rozporządzeniu określono cechy i standardy interoperacyjności, w tym formaty danych i standardy zapewniające dostęp do zasobów informacji udostępnianych za pomocą systemów teleinformatycznych używanych do realizacji zadań publicznych. Spośród ok. 50 formatów jedynie nieliczne posiadają cechy umożliwiające wykorzystywanie ich w systemie udostępniania danych publicznych, jakim jest portal DanePubliczne.gov.pl.

I. WYTYCZNE PRZYGOTOWYWANIA DANYCH

Kryteriami wyjściowymi, jakie powinny spełniać zasoby informacyjne są filary otwartości, o których mowa w Programie.

1. Wymagania dotyczące plików danych

Ze względu na specyfikę systemów udostępniania danych publicznych w celu ponownego wykorzystywania, zaleca się stosowanie następujących formatów plików danych (w kolejności od najbardziej preferowanych do najmniej):

dane ustrukturyzowane

- 1) XML, GML
- 2) CSV, SHP
- 3) XLSX, ODS
- 4) XLS

dane niestrukturyzowane

- 1) DOCX, ODT, RTF
- 2) DOC, TXT
- 3) PDF (dopuszczalny dla plików zawierających graficzne odwzorowania dokumentów)
- 4) JPG, TIF, PNG (dopuszczalny dla plików zawierających dokumenty graficzne)
- 5) archiwa skompresowane (dopuszczalny jedynie dla pakietów do pobrania)

2. Wymagania dotyczące organizacji i lokalizacji źródła danych

- 1) każdy zasób jest publikowany w odrębnym pliku,
- 2) wersjonowanie zasobów, zamiast nadpisywania (każdy nowy plik pod nowym adresem URL),

- 3) niezmienny adres URL każdego zasobu/pliku,
- 4) trwała dostępność każdego (już) opublikowanego zasobu danych publicznych,
- 5) tworzenie lokalnych repozytoriów danych publicznych lub wykorzystywanie systemów scentralizowanych, np. SSDIP, chroniących udostępnione zasoby i ich URL przed zmianami organizacyjnymi podmiotów,
- 6) udostępnianie zasobów w kilku standardach jednocześnie, np. obraz dokumentu w przeszukiwalnym pliku PDF, edytowalnym pliku ODF, plik(i) z danymi użytymi w dokumencie, w standardzie CSV.

3. Wymagania dotyczące standaryzacji i formatów plików danych

- 1) standard zapisu liczb w systemie dziesiętnym: minus „-” bez spacji dla liczb ujemnych, bez grupowania, przecinek „,” jako separator dziesiętny
- 2) standard zapisu daty: YYYY-MM-DD
- 3) standard zapisu czasu: hh:mm:ss
- 4) standard zapisu pliku CSV: kodowanie znaków UTF-8, średnik „;” jako separator pól, znak nowej linii (CRLF) jako separator rekordów
- 5) dopuszcza się pliki obrazowe PDF i graficzne (JPG, PNG, TIF lub SVG) wyłącznie jako wizualizację dokumentu stanowiącego informację publiczną (nie jako plik danych)
- 6) dopuszcza się przeszukiwalny format PDF, wykorzystywany wyłącznie jako zobrazowanie informacji publicznej (nie jako plik danych)
- 7) dopuszcza się plik w formacie XLS wyłącznie jako ustrukturyzowany zasób informacji publicznej (nie jako plik danych)
- 8) zaleca się stosowanie plików danych w formatach: CSV, XML, HTML, GML i Shapefile

4. Wymagania dotyczące interfejsu programistycznego (API)

- 1) interfejs nie może zawierać limitów lub innych zabezpieczeń, które by uniemożliwiły anonimowemu użytkownikowi pobranie w sposób automatyczny wszystkich informacji z zasobu informacyjnego
- 2) interfejs musi zwracać informacje w postaci JSON lub XML.

5. Wymagania dotyczące jakości danych

- 1) aktualność
- 2) kompletność
- 3) poprawność formalna (kontrola danych, reguły poprawności)
- 4) wiarygodność
- 5) jednorodność (te same typy danych są zapisywane w tym samym standardzie formalnym, np. data, waluta, liczby)
- 6) brak redundancji (nadmiarowości/powtórzeń)
- 7) naturalny język danych (gdzie ma znaczenie)
- 8) format przeznaczony do odczytu maszynowego.

Przykłady:

Przykład danych udostępnionych powszechnie w formacie pdf na stronie <http://www.minrol.gov.pl/pol/Rynki-rolne/Zintegrowany-System-Rolniczej-Informacji-Rynkowej/Biuletyny-Informacyjne/Rynek-mleka/RYNEK-MLEKA-notowania-za-okres-02.05.2016-08.05.2016-r>

MINISTERSTWO ROLNICTWA I ROZWOJU WSI

ZINTEGROWANY SYSTEM ROLNICZEJ INFORMACJI RYNKOWEJ

(podstawa prawna: ustawa o rolniczych badaniach rynkowych z dnia 30 marca 2001r.)

RYNEK MLEKA

Nr 18/2016
12 maja 2016r.



NOTOWANIA Z OKRESU: 02-08.05.2016r.

Badanie cen sprzedaży prowadzone jest na reprezentatywnej próbie. Średnie ceny liczone są jako średnie ważone.

I. CENY SPRZEDAŻY (NETTO) PRODUKTÓW MLECZARSKICH w zł/100kg.

I. Polska +UE.

Daty podane w tabelach oznaczają ostatni dzień analizowanego tygodnia (poniedziałek-niedziela).

1.1. Produkty mleczarskie płynne.

TOWAR	Zawartość tłuszczu	POLSKA		Tygodn. zmiana ceny [%]
		Cena [zł/100kg]		
		2016-05-08	2016-05-01	

Przykład danych udostępnionych w formie ustrukturyzowanej arkusza kalkulacyjnego

Zmiana cen wybranych produktów mleczarskich (w zł/100kg)

w skali tygodnia, miesiąca, początku roku, roku i dwóch lat.

Skup - marzec 2016

Towar	Cena					
	aktualna	tydzień temu	miesiąc temu	początek roku	rok temu	2 lata temu
Mleko surowe	110,5		112,6	114,8	122,1	147,6
Mleko w proszku odtłuszczone	700,0	703,8	707,0	712,3	777,6	1226,5
Mleko w proszku pełne	807,8	852,6	879,3	967,5	1003,7	1436,6
Masło w blokach	986,5	978,1	1014,9	1217,4	1196,4	1442,1

Przykład formatu CSV zobrazowany:

- w tabeli arkusza kalkulacyjnego

Towar	Cena aktualna	Cena tydzień temu	Cena miesiąc temu	Cena z początku roku	Cena rok temu	Cena 2 lata temu
Mleko surowe	110.5	112.6	114.8	122.1	147.6	
Mleko w proszku odtłuszczone	700.0	703.8	707.0	712.3	777.6	1226.5
Mleko w proszku pełne	807.8	852.6	879.3	967.5	1003.7	1436.6
Masło w blokach	986.5	978.1	1014.9	1217.4	1196.4	1442.1
Masło konfekcjonowane	1171.0	1194.0	1229.5	1352.4	1374.6	1597.9
Ser Edamski	956.6	968.4	1007.3	1080.5	1106.4	1518.0
Ser Gouda	939.3	963.3	1001.3	1055.3	1111.7	1462.0

- jako plik tekstowy:

Towar,Cena aktualna,Cena tydzień temu,Cena miesiąc temu,Cena z początku roku,Cena rok temu,Cena

Mleko surowe,110.5,,112.6,114.8,122.1,147.6

Mleko w proszku odtłuszczone,700.0,703.8,707.0,712.3,777.6,1226.5

Mleko w proszku pełne,807.8,852.6,879.3,967.5,1003.7,1436.6

Masło w blokach,986.5,978.1,1014.9,1217.4,1196.4,1442.1

Masło konfekcjonowane,1171.0,1194.0,1229.5,1352.4,1374.6,1597.9

Ser Edamski,956.6,968.4,1007.3,1080.5,1106.4,1518.0

Ser Gouda,939.3,963.3,1001.3,1055.3,1111.7,1462.0

II. PROCEDURY PUBLIKACJI ZASOBÓW W REPOZYTORIUM

1. Etap wyboru zasobu informacyjnego

- 1) posiadanie praw właścicielskich do zasobu i jego dostępność (zasób w dyspozycji podmiotu),
- 2) zamawianie danych u dostawców (wykonawców zamówień na opracowania, ekspertyzy, analizy, ankiety itp.) na licencjach i w formatach umożliwiających powtórne wykorzystywanie (o odpowiednio wysokim stopniu otwartości),
- 3) możliwość udostępniania w ramach otwartej licencji,
- 4) ocena merytoryczna potencjału do powtórnego wykorzystywania (identyfikacja popytu na dane).

2. Etap analizy formatu danych i dostosowania do zasad otwartości i jakości

- 1) ocena dostępnych formatów danych,
- 2) ocena możliwości modyfikacji zasobów do formatu o wyższym stopniu otwartości, np. poprzez strukturyzowanie danych,
- 3) podział złożonych zasobów na zasoby o prostej strukturze, np. podział skoroszytów na pojedyncze arkusze,
- 4) posiadanie możliwości technicznych i organizacyjnych do niezbędnej przebudowy struktury i modyfikacji zawartości plików (np. usuwanie zakłóceń/agregacji struktury arkusza, usuwanie formuł oraz odniesień do danych poza źródłem, łączenie treści podzielonych między polami itp.) oraz poprawy jakości danych (czyszczenie) w celu zwiększenia użyteczności zasobów przeznaczonych do udostępniania.

3. Etap publikowania danych (formaty, miejsce publikacji, opis i metadane)

- 1) opracowanie/przygotowanie danych w możliwie najwyższym stopniu otwartości (postulatywnie: w formacie CSV – metoda najprostsza i najtańsza, ograniczone wykorzystanie; w formacie XML – metoda najbardziej złożona i najdroższa, najlepsze cechy użytkowe),
- 2) wybór lokalizacji źródła danych przeznaczonych do udostępniania:
 - lokalny system teleinformatyczny, w tym lokalne repozytorium – opcja obarczona ryzykiem utraty dostępu do danych wskutek zmian organizacyjnych podmiotu,
 - repozytorium centralne – opcja o wysokim stopniu pewności dostępu i bezpieczeństwa danych,

- 3) sporządzenie opisu zasobu, precyzyjnie charakteryzującego zawartość oraz istotne cechy wyróżniające i wartości brzegowe (np. data/okres, obszar, sektor gospodarczy itp.),
- 4) dobór słów kluczowych (postulowane rozwiązanie: wybór ze słownika).

4. Etap nadzoru nad aktualizacją zasobu, trwałości i dostępności danych

Opracowanie wewnętrznych zasad i procedur udostępniania danych, w szczególności określenie:

- 1) źródeł danych,
- 2) tymczasowej i docelowej alokacji zasobów,
- 3) zasad opracowania zasobów, formatów plików danych, sposobu sporządzania opisów i stosowania słów kluczowych,
- 4) osób upoważnionych do przetwarzania i publikowania zasobów oraz pełniących nadzór nad aktualnością danych,
- 5) osób odpowiedzialnych za udostępnianie,
- 6) ustalenie procedury obiegu dokumentów i danych.

5. Etap oceny wykorzystania i informacja zwrotna od użytkowników, analiza efektów udostępnienia danych

- 1) publikowanie ilości otwarć,
- 2) publikowanie ilości użytkowników śledzących zasób,
- 3) publikowanie ilości pobrań,
- 4) publikowanie abonentów zasobu,
- 5) tworzenie rankingów otwartości danych, popularności, jakości, największej ilości zastosowań w ramach wtórnego wykorzystywania itp. akcje promocji zasobów i dostawców danych,
- 6) tworzenie i publikowanie opinii nt. zasobów.

III. PRZYGOTOWANIE ARKUSZA USTRUKTURYZOWANYCH DANYCH W FORMACIE XLS LUB XLSX

1. Arkusz musi posiadać jeden wiersz nagłówek, który zawiera opisy wszystkich kolumn z danymi.

Niedopuszczalne jest:

- 1) scalanie komórek zarówno w ramach wiersza, jak i kolumny – zaburza to strukturę arkusza danych,
 - 2) umieszczanie treści dotyczących jednego rekordu w kilku wierszach,
 - 3) wprowadzanie dodatkowych tytułów, opisów, przypisów i komentarzy poza obszarem danych tabeli,
 - 4) używanie formuł agregujących dane, np. funkcji SUMA, lub funkcji filtrowania danych,
 - 5) używanie formuł pobierających dane spoza źródła danych, które będą niedostępne po przeniesieniu arkusza na inną platformę.
2. Każda kolumna tabeli, poza wierszem nagłówek, musi zawierać dane jednego rodzaju, tzn. liczbę, tekst lub ciąg cyfrowo-tekstowy o określonej strukturze i znaczeniu, np. numer telefonu, kod pocztowy.

3. Każda kolumna tabeli powinna zawierać dane dotyczące pojedynczej wielkości (elementarnej), np. numer kodu, a nie numer kodu i miejscowość; ułatwia to sortowanie i łączenie danych z wielu tabel.
4. Zapis w każdym polu w kolumnie musi być zgodny z założonym formatem i modelem – maską, właściwą dla danego typu zawartości.
5. Nie należy stosować formatowania tekstu i niektórych znaków interpunkcyjnych właściwych dla formatowania zaawansowanego, które w pewnych kontekstach mogą imitować operatory arytmetyczne lub zmieniać wartość liczby, np. łącznik (dywiz) może prowadzić do utraty danych generując w arkuszu wynikowym błąd formuły lub zmieniać wartość liczby na ujemną.
6. Jeżeli nie można powyższych warunków spełnić ze względu na złożoność danych zawartych w tabeli, należy tabelę podzielić w sposób, który umożliwi ich spełnienie.
7. Zaleca się opracowywanie danych w postaci pojedynczych arkuszy, a nie tzw. skoroszytów.

IV. STANDARD METADANYCH

Zestaw elementów metadanych zasobów udostępnionych na portalu danepubliczne.gov.pl określony został w rozporządzeniu Rady Ministrów z dnia 12 marca 2014 r. w sprawie Centralnego Repozytorium Informacji publicznej. Część metadanych jest wpisywanych do systemu automatycznie w momencie dodawania zasobu. Są to: identyfikator zasobu, urząd obsługujący dostawcę, data udostępnienia, data aktualizacji, format pliku oraz określany jest stopień otwartości zasobu.

Na dostawcy spoczywa obowiązek opisanie zasobu zestawem metadanych obejmujących:

- 1) informacje niestandardowe (dowolnie określane przez dostawcę):

nazwę zasobu z elementami charakterystycznymi, opis zawartości, słowa kluczowe określające zawartość, typ zasobu,

- 2) informacje standardowe (wybierane z określonych w systemie list)

URL zasobu, częstotliwość aktualizacji, kategorię, sposób prezentacji oraz warunki ponownego wykorzystywania, jeżeli są inne niż bez ograniczeń i bezpłatnie.

Szczególą rolę odgrywa właściwie dobrany i bezbłędny zestaw metadanych określanych przez dostawcę, gdyż są to informacje najczęściej wykorzystywane do precyzyjnego wyszukiwania zasobów.