Załącznik nr 1

Standard udostępniania danych na portalu danepubliczne.gov.pl

Standardy interoperacyjności i standardy otwartości danych

Podmioty realizujące zadania publiczne są zobowiązane do przestrzegania zasad wymiany informacji pomiędzy systemami teleinformatycznymi umożliwiających 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[1]. 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. Zalecenia 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):

1) dane ustrukturyzowane:

a) JSON, XML, GML,

b) CSV, SHP,

c) XLSX, ODS,

d) XLS;

2) dane niestrukturyzowane:

a) DOCX, ODT, RTF,

b) DOC, TXT[2],

c) PDF (dla plików zawierających graficzne odwzorowania dokumentów),

d) JPG, TIF, PNG (dla plików zawierających dokumenty graficzne),

e) archiwa skompresowane (dla pakietów przygotowanych do pobrania jako całość).

2. Zalecenia dotyczące organizacji i lokalizacji źródeł danych

1) publikowanie każdego zasobu (jednostki informacji publicznej) w odrębnym pliku;

[1] Dz. U. z 2016 r. poz. 113.

[2] Plik tekstowy ustrukturyzowany z wykorzystaniem znaków pełniących rolę separatorów, np. tabulator, średnik, kreska pionowa (*pipe*).

2) stosowanie wersjonowania zasobów, zamiast zastępowania/nadpisywania (każdy nowy plik pod nowym adresem URL);

3) utrzymywanie niezmiennego adresu URL każdego zasobu (pliku);

4) zapewnienie trwałej dostępności każdego, już opublikowanego zasobu;

5) tworzenie lokalnych repozytoriów danych publicznych lub wykorzystywanie systemów scentralizowanych, np. SSDIP, w celu ochrony udostępnionych zasobów i ich URL przed zmianami organizacyjnymi podmiotów;

6) udostępnianie zasobów w kilku standardach jednocześnie, np. obraz dokumentu w przeszukiwalnym pliku PDF, edytowalnym pliku ODF, odrębne pliki z danymi użytymi w dokumencie, np. w standardzie CSV, lub też skompletowane w archiwum skompresowanym do pobrania.

3. Zalecenia 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) oprócz standardowego zapisu pliku CSV z przecinkiem jako separatorem, dopuszcza się dla zasobów utworzonych w aplikacjach lokalizowanych w Polsce: 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) zaleca się stosowanie plików danych w formatach: CSV, XML, GML i Shapefile.

4. Zalecenia dotyczące interfejsu programistycznego (API)

1) interfejs nie może zawierać limitów lub innych zabezpieczeń, które uniemożliwiałyby 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 (gdy ma znaczenie);

8) format przeznaczony do odczytu maszynowego.

*** <u>Udostępnienie danych w sieci Web (w dowolnym formacie) na warunkach otwartej licencji</u>**

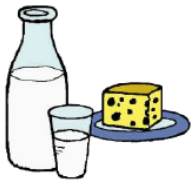
Przykład danych udostępnionych powszechnie w formacie PDF*

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 17/2016
05 maja 2016r.

NOTOWANIA Z OKRESU: 25.04-01.05.2016r.

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

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

1. Polska +UE.

Daty podane w tabelach oznaczają <u>ostatni dzień</u> analizowanego tygodnia (poniedziałek-niedziela).

1.1. Produkty mleczarskie płynne.

TOWAR	Zawartość tłuszczu	POLSKA		
		Cena [zł/100kg]		Tygodn. zmiana ceny [%]
		2016-05-01	2016-04-24	
Mleko spożywcze pasteryzowane	do 0,5%	--	--	--
	1,5-1,8%	nld	nld	-
	2%	164,02	164,17	-0,1
	3,2%	174,46	175,20	-0,4
	od 3,5%	259,42	261,09	-0,6
	Ogółem	**168,81**	**169,23**	**-0,2**
	do 0,5%	161,29	159,73	1,0

Objaśnienie
*Wykorzystano dane opublikowane 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. Dane pobrane w dniu 28.06.2016 r.
Źródło danych dotyczy każdego przykładu powołanego dalej.

** **Udostępnienie danych w formie ustrukturyzowanej (np. arkusz kalkulacyjny zamiast zeskanowanego obrazu tabeli)**

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

*** **Używanie formatów otwartych (np. CSV zamiast arkusza kalkulacyjnego)**

Przykład formatu CSV zobrazowany:

– **w tabeli arkusza kalkulacyjnego**

	A	B	C	D	E	F	G	H
1	Towar	Cena aktu	Cena tydz	Cena mies	Cena z po	Cena rok t	Cena 2 lata temu	
2	Mleko surowe	110,5		112,6	114,8	122,1	147,6	
3	Mleko w proszku odtłusz	700	703,8	707	712,3	777,6	1226,5	
4	Mleko w proszku pełne	807,8	852,6	879,3	967,5	1003,7	1436,6	
5	Masło w blokach	986,5	978,1	1014,9	1217,4	1196,4	1442,1	
6	Masło konfekcjonowane	1171	1194	1229,5	1352,4	1374,6	1597,9	
7	Ser Edamski	956,6	968,4	1007,3	1080,5	1106,4	1518	
8	Ser Gouda	939,3	963,3	1001,3	1055,3	1111,7	1462	
9								
10								

– **jako plik tekstowy**

```
Towar;Cena aktualna;Cena tydzień temu;Cena miesiąc temu;Cena z początku roku;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) analiza posiadanych praw do zasobu umożliwiających jego udostępnienie do ponownego wykorzystywania i analiza jego dostępności (zasób w dyspozycji podmiotu);
2) zamawianie danych u dostawców (wykonawców zamówień na opracowania, ekspertyzy, analizy, ankiety, itp.) na otwartych licencjach (o ile w ogóle są konieczne) i w formatach umożliwiających powtórne wykorzystywanie (o odpowiednio wysokim stopniu otwartości);
3) możliwość udostępniania w ramach domeny publicznej;
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, dobry poziom ustrukturyzowania danych, ograniczone wykorzystanie; lub np. w formacie XML – metoda zwykle trudniejsza do stosowania, lecz właściwa, gdy dane uzyskują tą drogą lepsze cechy użytkowe);
2) wybór lokalizacji źródła danych przeznaczonych do udostępniania:
   a) lokalny system teleinformatyczny, w tym lokalne repozytorium – opcja obarczona ryzykiem utraty dostępu do danych wskutek zmian organizacyjnych podmiotu,

b) 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) procedury obiegu dokumentów i danych.

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

1) publikowanie:

a) liczby odsłon,

b) liczby użytkowników obserwujących zasób,

c) liczby pobrań,

d) liczby abonentów zasobu,

e) opinii nt. zasobów;

2) tworzenie rankingów:

a) otwartości danych,

b) popularności zasobów,

c) jakości danych,

d) największej liczby zastosowań w ramach ponownego wykorzystywania.

III. Przygotowanie arkusza ustrukturyzowanych danych w formacie xls lub xlsx

1. Arkusz musi posiadać jeden wiersz nagłówka, 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, np. kolejnych danych adresu (adresat, ulica, miejscowość, itp.);

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 nagłówka, 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[3]. 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 opisania 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.

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

[3] Dz. U. poz. 361.