Format SHP nie dla danych złożonych, nietypowych i obszernych

Co po „szejpie”?

Mimo swojej archaiczności Shapefile wciąż pozostaje najpopularniejszym formatem wymiany danych przestrzennych na świecie. Ale jego dni wydają się policzone. Co go zastąpi?

Jerzy Królikowski

Zapis danych wektorowych w pliku SHP został wprowadzony przez firmę Esri do jej oprogramowania ArcView jeszcze na początku lat 90., a więc w czasach, gdy GIS był na świecie mało znanym skrótem. Można przypuszczać, że to właśnie wczesne wypuszczenie formatu w połączeniu ze sporą popularnością produktów Esri zadecydowało o międzynarodowym sukcesie „szejpa”. Popularność tego rozwiązania pozostawała niezachwiana przez lata mimo ostrej konkurencji ze strony innych aplikacji GIS-owych – zarówno komercyjnych, jak i otwartych. Niewiele zmieniły tu nawet wysiłki standaryzacyjne Open Geospatial Consortium, które stworzyło tak popularne w geoinformatyce standardy, jak WMS czy WFS. Szybki postęp technologiczny w GIS-ie sprawia jednak, że wady i ograniczenia formatu SHP stają się dla użytkowników coraz bardziej dokuczliwe.

Akt oskarżenia

Pierwszą wadę SHP dostrzeże nawet laik – jedna warstwa zapisywana jest do kilku plików. Zgodnie ze specyfikacją formatu są ich przynajmniej trzy (rozszerzenia SHP, DBF, SHX), choć opcjonalnie listę tę można rozszerzyć nawet o 7 kolejnych. Nie dość, że praca z takim „bogactwem” jest niewygodna, to powszechne są sytuacje, że mniej doświadczony użytkownik na prośbę: „prześlij mi szejpa”, wysyła tylko jeden – kompletnie nieprzydatny – plik.

Kolejną wadą jest ograniczony do 2 GB rozmiar pliku, co przekłada się na nie więcej niż kilkadziesiąt milionów rekordów. Nikomu to nie przeszkadzało w latach 90., gdy królowały dyskietki, ale w erze big data stanowi już poważne ograniczenie.

Tych ograniczeń w „szejpie” jest zresztą znacznie więcej. No bo dlaczego obiekt nie może mieć więcej niż 255 atrybutów, a nazwa każdego z nich nie może przekraczać 10 znaków? Dlaczego dopuszczalne atrybuty to tylko: tekst (i to ograniczony do raptem 254 znaków), data (bez możliwości zapisu czasu) oraz liczba całkowita bądź ułamek (tylko do 13 znaków). A co np. z „dymkami” czy obrazkami?

Do wad SHP zaliczana jest także możliwość zapisu w jednym pliku tylko jednego rodzaju geometrii – punktu, multipunktu, linii lub poligonu. Pamiętać również należy, że standard ten służy wyłącznie do przechowywania danych wektorowych. Nie da się zatem zapisać w tym rozszerzeniu np. rastrów. Ponadto „szejp” nie pozwala opisywać relacji topologicznych. A możliwości zapisu tak popularnych obecnie danych 3D są w nim mocno ograniczone.

Do listy bolączek można jeszcze dodać problemy z obsługą znaków diakrytycznych i definicji układów współrzędnych. To ostatnie zadanie realizuje wprawdzie plik z rozszerzeniem PRJ, ale nie wszyscy wiedzą, że jest on tylko opcjonalnym elementem tego standardu. Podsumowując: SHP okazuje się szczególnie kiepskim formatem, gdy pracujemy na danych złożonych, nietypowych i obszernych.

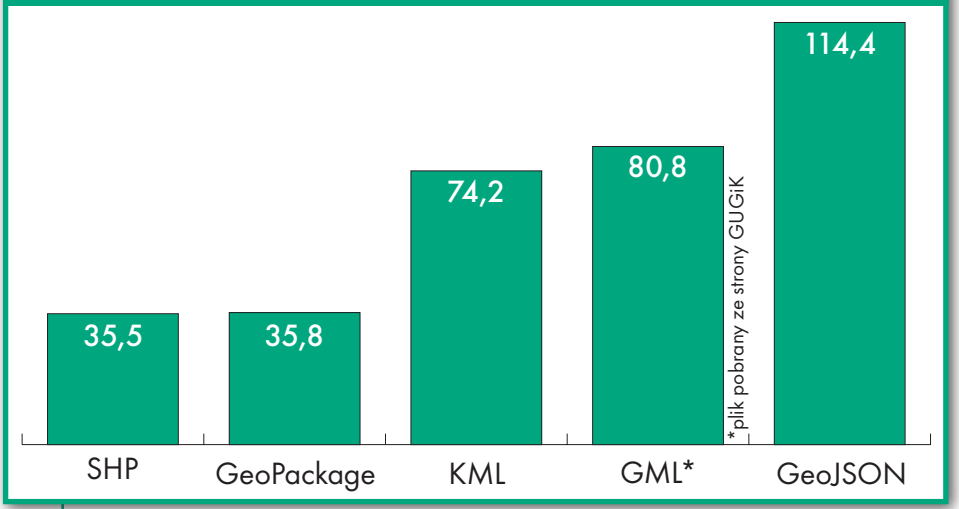
Oczywiście „szejp” ma także pewne zalety. Podstawowym plusem jest wspomniana już powszechność, dzięki której mamy niemal pewność, że gdy udostępnimy komuś dane w tym rozszerzeniu, to zostaną poprawnie odczytane w różnych aplikacjach, także darmowych. Specjaliści od geoinformatyki zwracają ponadto uwagę, że standard ten zapewnia relatywnie mały rozmiar pliku oraz wydajne wczytywanie.

Wszechstronna paczka

Skoro nie Shapefile, to w takim razie co? Dziś największe szanse na zastąpienie tego rozwiązania ma GeoPackage, w skrócie **GPKG**. Od 2014 roku jest on oficjalnym standardem OGC obsługiwanym przez coraz więcej GIS-owych aplikacji. Nie brak opinii, że rychły sukces GeoPackage przypieczętowała premiera popularnego otwartego programu QGIS w wersji 3.0. Jedną z fundamentalnych zmian wprowadzonych w tym pakiecie było bowiem ustanowienie GPKG domyślnym formatem zapisu danych (GEODETA 5/2018). Oczywiście standard ten jest obsługiwany również w wielu innych znanych GIS-owych aplikacjach, włącznie z programami firmy Esri.

Jakie są główne zalety GPKG? Mówiąc w skrócie: format ten jest pozbawiony kluczowych wad Shapefile’a. Przede wszystkim jeden zbiór danych jest tu jednym plikiem, a co więcej, możemy w nim przechowywać różne typy geometrii. Warto podkreślić, że GeoPa-

Porównanie rozmiaru pliku PRG z granicami powiatów zapisanego przy użyciu aplikacji QGIS 3.4 do różnych formatów [MB]



ckage pozwala zapisywać nie tylko dane wektorowe, ale również rastry (w tym ich „piramidy" ułatwiające przeglądanie warstwy w różnych skalach).

Na tym nie koniec możliwości tego formatu. Od wersji 3.8 „Zanzibar" oprogramowanie QGIS pozwala zapisywać w pliku GPKG również cały projekt, a także informacje o stylu wyświetlania warstwy, a więc np. rodzaje i kolory sygnatur czy grubość linii. Jest to możliwie dzięki temu, że format zaprojektowano tak, by dało się go rozbudowywać o dodatkowe rozszerzenia. Nie trzeba już chyba dodawać, że pliki GPKG są wolne od wspomnianych ograniczeń „szejpa" związanych z maksymalną objętością pliku czy długością nazw atrybutów.

Czy są jakieś wady tego formatu? Kluczową jest dziś jego młody wiek, co sprawia, że użytkownicy starszych wersji aplikacji GIS-owych będą mieli problemy z importem czy eksportem danych w tym rozszerzeniu. Wśród geoinformatyków można się także spotkać z opiniami, że standard ten kiepsko radzi sobie z rastrami, nie nadaje się ponadto do strumieniowej transmisji.

A co z tak ważną kwestią, jak objętość plików? To oczywiście zależy od tego, jakiego zbioru użyjemy do porównania. W krótkim redakcyjnym teście sięgnęliśmy po dane o granicach powiatów z PRG, a wyniki zaprezentowaliśmy na wykresie. Jak widać, różnica między standardami SHP i GPKG jest nieznaczna, ale gdy uwzględnić inne formaty, rozbieżności robią się spore. Oczywiście podkreślamy, że wynik testu może okazać się całkiem odmienny, gdy sięgniemy po inne dane.

Kolejka pretendentów

Lista formatów, które mogą zastąpić SHP, jest znacznie dłuższa i można by na niej umieścić nawet kilkadziesiąt pozycji. W Polsce uprzywilejowane miejsce zajmuje **GML** (*Geography Markup Language*), który zgodnie z przepisami wykonawczymi do *Prawa geodezyjnego i kartograficznego* jest u nas podstawowym formatem wymiany danych przestrzennych. Podobnie jak GeoPackage oferuje szerokie możliwości zapisu różnego typu danych i jest pozbawiony kluczowych wad Shapefile'a. W praktyce jest jednak wciąż rzadko stosowany, głównie z uwagi na swoją złożoność, która niekiedy prowadzi do problemów z odczytem bardziej skomplikowanych baz. Po wymowną ilustrację tego problemu odsyłamy do wpisu pt. „GML madness" na blogu „Geo tricks and tips". Jak wylicza jego autor, specyfikacja standardu GML pozwala zapisać kwadrat na… 25 sposobów! Zresztą nie ma co sięgać po przykłady tak daleko, skoro i w geodezji widać znaczne opory z wdrażaniem tego formatu.

Sporą popularnością na całym świecie od lat cieszy się **KML** (*Keyhole Markup Language*). To przede wszystkim zasługa tego, że jest on podstawowym formatem popularnej i darmowej aplikacji Google Earth. Nie bez znaczenia jest także możliwość łatwego zapisu i odczytu różnych typów danych – nie tylko wektorowych (wraz z informacją o stylu wyświetlania), ale także rastrowych czy modeli 3D. Kto miał do czynienia z tym formatem, wie jednak, że jego pełną i bezproblemową obsługę gwarantuje w zasadzie tylko sam Google Earth. Pozostałe aplikacje miewają zaś trudności z wyświetlaniem bardziej złożonych danych. Wśród wad KML wymienia się także brak wsparcia dla układów współrzędnych innych niż WGS-84 czy relatywnie duże rozmiary plików.

Coraz więcej zastosowań ma także format **GeoJSON**. Do jego zalet zalicza się m.in. prostotę, obsługę nawet złożonych danych wektorowych czy możliwość strumieniowej transmisji. Na liście wad znajdziemy natomiast duży rozmiar pliku oraz obsługę jedynie układu WGS-84. Ze względu na swoje cechy format ten jest stosowany przede wszystkim w różnego rodzaju aplikacjach internetowych.

Jako następcę SHP niekiedy wymienia się również format Esri Geodatabase (**GDB**), czyli tzw. geobazy, które – podobnie jak GeoPackage – są pozbawione kluczowych wad Shapefile'a. Jest to jednak relatywnie rzadko stosowane rozwiązanie, głównie z uwagi na jego zamkniętą specyfikację. W rezultacie posługują się nim przede wszystkim użytkownicy oprogramowania firmy Esri, choć – przeglądając rządowe witryny – ostatnio coraz częściej trafiamy na przykłady publikacji otwartych danych w tym formacie.

Który z wymienionych formatów zastąpi zatem Shapefile'a? Jako że każdy z nich ma swoje wady i zalety, nie można wykluczyć, że nie będzie jednego zwycięzcy. Dysponenci danych już teraz powinni jednak porzucić format SHP, a przynajmniej zacząć udostępniać swoje zasoby również w innym standardzie. Odejście poczciwego „szejpa" w przeszłość wydaje się bowiem bardziej niż pewne.

Przy pisaniu artykułu korzystałem ze strony „Shapefile must die" (switchfromshapefile.org)