Dobrý den

Proč musí Shapefile zemřít?!

Jáchym Čepický[1]

[1]OOSGeo.cz http://osgeo.cz

GIVS 2015

Proč musí Shapefile zemřít?!

Jáchym Čepický[1]

[1]OOSGeo.cz http://osgeo.cz

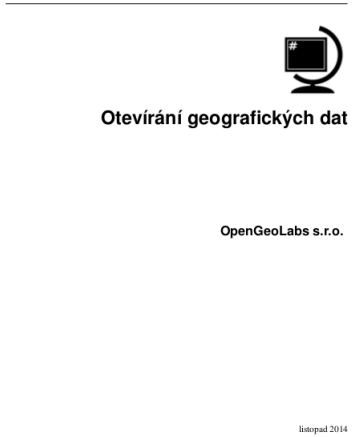
GIVS 2015

GIVS

K bulvárnímu názvu této prezentace mě vedla zkušenost, kterou jsme udělali při zpracování publikace ”Otevírání geografických dat”, zadané IPR Praha.

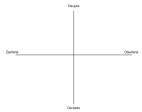
Mezi cíle studie patřilo jednak navržení a zhodnocení možností publikace prostorových dat, kde jsme se drželi linky, jak ji nastavila iniciativa INSPIRE, ale také zhodnocení formátů prostorových dat.

Otevírání geografických dat – Případová studie



http://opengeolabs.cz/publikace

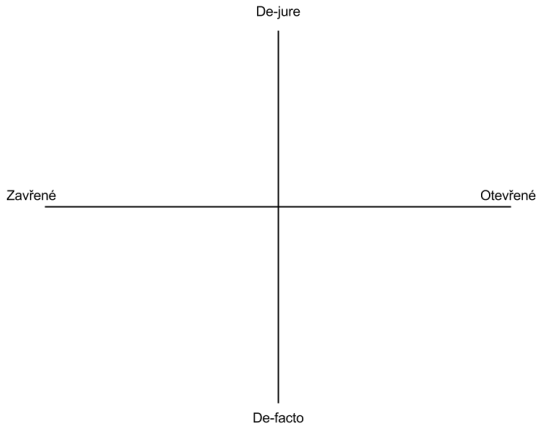
GIVS



Třídění standardů

Formáty prostorových dat jsou definovány standardy. Dělíme podle způsobu jejich vzniku a dalšího života na na otevřené a uzavřené, podle autority, která je prosazuje na de-facto a de-jure.

Otevřené standardy vznikají v otevřeném procesu, veřejnou diskusí, které se může účastnit teoreticky kdokoliv. Jsou dokumentované a žádná legislativní nebo technická překážka nebrání jejich dalšímu používání.



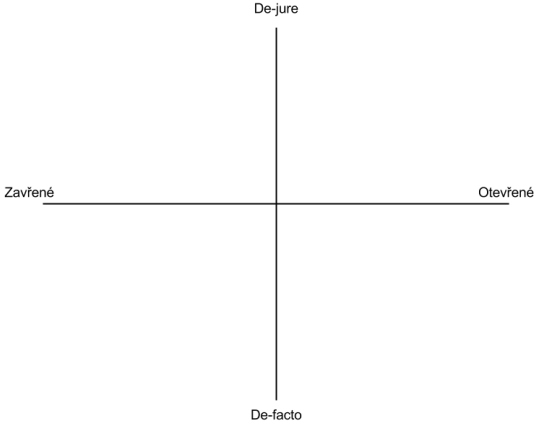
GIVS

Třídění standardů

Uzavřené standardy vznikají v rámci organizace bez veřejné diskuse, jejich praktická implementace je značně omezena legislativními a/nebo technickými překážkami.

De-jure standardy jsou vyžadovány legislativním procesem, kdežto de-facto standardy jsou tlačeny silou tržního podílu.

Teoreticky z pohledu otevírání dat je ideální používat otevřené de-jure standardy. Jako příklad můžeme vzít dokumenty publikované konsorciem OGC. Jako příklad otevřeného, de-facto standardu můžeme vzít formát ESRI Shapefile. Jedná se o zdokumentovaný formát, jehož implementace technicky ani legislativně nic nebrání. Je široce rozšířený ve všech softwarech, používaný jako výměnný formát a zdá se, že bez Shapefilu nemůžeme žít.



Co je špatného na ESRI Shapefile?

Ve zmíněné případové studii jsme se snažili preferovat otevřené de-jure standardy a spíše nedoporučovat jiné skupiny standardů. Samozřejmě jsme se ze strany IPR setkali s otázkou "co je špatného na esri shapefile?", každý to zná, používá, softwary to podporují bez ohledu na to, jsou-li open source nebo proprietární. Shapefile vypadá jako ideální formát.

Naše praktické zkušenosti ale hovoří jinak a pokusím se vás nyní přesvědčit o pohnout k tomu, abyste se pokusili překročit stín tohoto formátu a začali uvažovat o nějakém modernějším.

ESRI Shapefile

Co je špatného na ESRI Shapefile?

GIVS

nejrozšířenější
nejpodporovanější
good enough

Abych začal pozitivně, pokusím se shrnout to, co je na formátu ESRI Shapefile dobré:

- Jedná se o pravděpodobně nejrozšířenější formát pro vektorová data
- jeho licence nebrání implementaci v software třetích stran, také proto je podporován prakticky vším, co umí otevřít data se souřadnicemi x, y
- pro většinu případů je prostě ”dostatečně dobrý”

ESRI Shapefile – to dobré

- nejrozšířenější
- nejpodporovanější
- good enough

A nyní proč si myslíme, že se jedná o opravdu špatný formát a proč byste měli začít uvažovat o jeho náhradě.

ESRI Shapefile – to špatné

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

Zde je plný seznam vlastností formátu ESRI Shapefile, které považujeme za problematické, zkusíme je v rychlosti rozebrat:

ESRI Shapefile – to špatné

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

Jedná se o více souborový formát. Kdybychom se mohli spolehnout na to, že nám budou stačit vždy 3 soubory, tedy SHP, DBF a SHX, dalo by se s tím ještě žít, ale různé softwary si vytváří své vlastní metasoubory, do kterých ukládají informace o projekci, prostorové a databázové indexy nebo dokonce kartografii. Protože se tak děje mimo specifikaci, jsou tyto formáty uzavřené de-facto standardy, není možné je použít v dalších softwarech.

Vícesouborový formát také způsobuje problémy při komunikaci prostřednictvím webových služeb - shapefile musíte zabalit do jednoho archivu a doufat, že se váš klient shodne se serverem na kompresním algoritmu. Popis takového hybridu pomocí mimetype specifikace je prakticky nemožný.

ESRI Shapefile – to špatné

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

GIVS

Deset znaků v názvu atributu se může zdást jako dostatečné množstí ALE není. Atribut parcelní_číslo má 14 znaků v názvu, a to zdaleka není nejextrémnější případ.
Softwary se často chovají tak, že při konverzi z jiného formátu (např. z databáze PostGIS) název atributu natvrdo zkrátí. Shapefile je sice vytvořen, ale plná kompatibila formátů opravdu není zajištěna.

ESRI Shapefile – to špatné

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

GIVS

Dostane-li se vám do rukou ESRI Shapefile, potřebovali byste ho s celým rodkomenem jenom proto, abyste se mohli pokusit odhadnout znakovou sadu, ve které jsou data uložena. Automaticky to určit prakticky nejde, což značeně komplikuje život programátorům. I zkušený datový analytik se tímto problémem ale musí zabývat, musí se podívat na data a uhádnout, v jaké znakové sadě jsou uložena.

ESRI Shapefile – to špatné

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

GIVS

Limit velikosti databáze je omezen na dva gigabajty dat. To mohlo být dost před ještě deseti lety, dnes ale běžně pracujeme s daleko většími vektorovými soubory.

ESRI Shapefile – to špatné

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

GIVS

Formát nebosahuje topologické vztahy, což např. já považuji za nevýhodu. Softwary nic nenutí mít data topologicky validní a to často vede k velkým problémům. Ostatně datová sada obce publikované v registru RUIAN není topologicky validní a některé obce se musí začistit - topologie je problém, kterým se většina uživatelů nezabývá, ale měla by.

ESRI Shapefile – to špatné

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

GIVS

více-souborový
10 znaků na název atributu
znaková sada – neznámo
max. 2GB
bez topologie
jeden typ geometrie/soubor
komplikovanější vazby a stromová struktura není možné uložit

Omezení jednoho typu geometrie na soubor považuji za dnes již přežilé. Někdo by mohl říct, že to přece stačí, dokonce nám to umožní zmenšit počet stuňů uživatelské volnosti, což je vždycky dobré, nemožnost uložit parcely a jejich definiční body do jednoho souboru je ale špatně, abych jmenoval alespoň jeden případ užití.

ESRI Shapefile – to špatné

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

V databázi nelze popsat komplikovanší vazby mezi objekty, stromovou strukturu, relační vztahy atd. Vše se dohání až na úrovni aplikace, ale přenést tyto vazby mezi softwary je opět spíše problém.

ESRI Shapefile – to špatné

- více-souborový
- 10 znaků na název atributu
- znaková sada – neznámo
- max. 2GB
- bez topologie
- jeden typ geometrie/soubor
- komplikovanější vazby a stromová struktura není možné uložit

Nyní doufám, že se mi podařilo vás přesvědčit, že shapefile není dobrý formát.

ESRI Shapefile? Raději . . . ?

Než se pustíme do hledání po náhradě za Shapefile, musíme si položit otázku, k čemu se vlastně používá?
Doufám, že se mnou budete souhlasit, že pro vážnou práci, není shapefile dobrý formát. Ve větší organizaci se data nahrají do prostorové databáze a následně se s nimi pracuje dále.
Shapefile dnes využíváme především jako výměnný formát. A pro ten bychom se mohli pokusit najít náhradu.

Použití ESRI Shapefile

- výměnný formát
- úložiště dat

GIVS

Mohlo by to být GML? Ne. Je to sice otevřený de-jure standard, založený na XML. Můžeme do něj zapsat stromovou strukturu, ale je to veskrze upovídaný nepraktický formát. Parserování XML, obzvláště u větších datasetů může narazit na limity hardware.

GML
Geodatabáze
CSV
SQL Dump
KML
GeoJSON
SpatiaLite

Výměnný formát

- GML
- Geodatabáze
- CSV
- SQL Dump
- KML
- GeoJSON
- SpatiaLite

Nabízí se možnost jít podobnou cestou jako šel Shapefile - Geodatabáze je ale uzavřený de-facto standard, nemá pořádnou implementaci v softwarech třetích stran, je svázána pouze s produkty firmy ESRI, což příliš pro interoperabilitu nehovoří.

Výměnný formát

- GML
- Geodatabáze
- CSV
- SQL Dump
- KML
- GeoJSON
- SpatiaLite

GIVS

GML
Geodatabáze
CSV
SQL Dump
KML
GeoJSON
SpatiaLite

Jednoduchý formát CSV je opravdu příliš primitivní

Výměnný formát

- GML
- Geodatabáze
- CSV
- SQL Dump
- KML
- GeoJSON
- SpatiaLite

GIVS

Stejně jako dump databáze ... bohužel nelze vytvořit univerzální SQL dump, který bych mohli číst napříč databázovými systémy

GML
Geodatabáze
CSV
SQL Dump
KML
GeoJSON
SpatiaLite

Výměnný formát

- GML
- Geodatabáze
- CSV
- SQL Dump
- KML
- GeoJSON
- SpatiaLite

GML
Geodatabáze
CSV
SQL Dump
KML
GeoJSON
SpatiaLite

Chvíli to vypadalo, že vše převálcuje KML, ale mix geometrie s kartografií, omezení na jeden souřadnicový systém, těžkopádné XML, a tak dále způsobily, že se od KML zase v praxi poněkud ustupuje

Výměnný formát

- GML
- Geodatabáze
- CSV
- SQL Dump
- KML
- GeoJSON
- SpatiaLite

GIVS

GML
Geodatabáze
CSV
SQL Dump
KML
GeoJSON
SpatiaLite

Perspektivním formátem je GeoJSON. Tam kde potřebujeme ušetřit datovou linku a zároveň potřebujeme lidsky čitelný formát je GeoJSON nepřekonatelný. Formalisovat v něm ale datové struktury je velice obtížné.

Výměnný formát

- GML
- Geodatabáze
- CSV
- SQL Dump
- KML
- GeoJSON
- SpatiaLite

GML
Geodatabáze
CSV
SQL Dump
KML
GeoJSON
SpatiaLite

Slibnějším formátem je další komunitní standard (tedy otevřený de-facto standard) SpatialLite. Podobně jako PostGIS je prostorové rozšíření databáze Postgres, Oracle má svou prostorovou nadstavbu, je SpatiaLite prostorové rozšíření souborové SQL databáze SQLite. Slučuje tak výhody relačních databází s jednoduchou manipulací se soubory na úrovni operačního systému. Něco tomu ale chybí.

Výměnný formát

- GML
- Geodatabáze
- CSV
- SQL Dump
- KML
- GeoJSON
- SpatiaLite

GIVS

http://opengeospatial.org/standards/geopackage
Rastry i vektory
Založeno na SQLite
OGC Simple Features
Maximální velikost databáze je 140 TB
Data mohou mít různé typy geometrií
Podporováno GDAL (1.11), ArcGIS 10.2.1

GeoPackage

Formát OGC Geopackage se zatím do povědomí světové geoinformatiky neprosadil (kolik lidí ho zná?). Jedná se o otevřený de-jure standard, produkovaný konsorciem OGC částečně jako odpověď na potřeby uživatelů, částečně jako odpověď na proprietární Geodatabázi.
Geopackage umožňuje uložit libovolná vektorová data spolu s rastrovými (ty mohou být ve formě dlaždicové cache nebo velkých souborů, ve formátu GeoTIFF) do prostředí databáze SQLite.
Vektorová data jsou pak uložena v souladu se specifikací OGC Simple Features.
Maximální velikost databáze je 140 TB, což by pro dnešní praxi mělo několik let stačit.
A je již podporován ve většině softwarů.

http://opengeospatial.org/standards/geopackage

- Rastry i vektory
- Založeno na SQLite
- OGC Simple Features
- Maximální velikost databáze je 140 TB
- Data mohou mít různé typy geometrií
- Podporováno GDAL (1.11), ArcGIS 10.2.1

Takže abych se vrátil k titulku prezentace, zbavíme se někdy formátu ESRI Shapefile?

Mým cílem není zbavit se Shapefilu. Na množství usecasů je to dostatečný formát.

Nepoužívejme ho ale tam, kde není vhodný, používejte jiné formáty, zvažte, jestli pro distribuci vašich prostorových dat není Geopackage, jako otevřený progresivní formát vhodnější.

Dotazy?

jachym.cepicky@geosense.cz
http://geosense.cz
@jachymc

Dotazy?

jachym.cepicky@geosense.cz
http://geosense.cz
@jachymc

GIVS