Organizační úvod

Poznámka (Organizační úvod)
Nahrávky budou. (Z minulého roku anglicky, z letoška česky.)

1 Úvod

Definice 1.1 (Strojově čitelný soubor)

Strojově čitelný soubor je vlastnost konkrétního souboru, ne formátu (jelikož do formátu můžu nacpat data v jiném formátu).

Strojová čitelnost se špatně definuje.

Definice 1.2 (Binární soubor)

Binární soubor je takový, kde je struktura popsána na úrovni bitů (bit po bitu). Není čitelný textovými editory.

TODO!!!

2 RDF

Definice 2.1 (RDF – resource description framework)

RDF je formát popisu grafu, kde se každé tvrzení (tedy trojice) má tvar „subjekt predikát objekt“, tj. „kdo co s-čím“. Vše se identifikuje pomocí IRI odkazující na definici (nebo v případě některých objektů (často Stringů/čísel/datumů) – literálem).

Poznámka
Uri budeme často zkracovat (takové zkrácení se zapisuje jako např. `@prefix dcterms: https://…`).
Obecné zkratky lze najít na `prefix.cc`.

Definice 2.2 (Literál)

Literál má dvě části – text odpovídající formátu a uri na ?XML schéma toho typu. Nebo je tvaru `"text"@jazyk`.

Například
Nejčastější predikát je `rdf:type` – „je typu“.

Definice 2.3 (Blank node)

Existují i nepojmenované uzly.

Definice 2.4 (RDF serializace)

(Jak zapsat RDF do textu.)

- RDF 1.1 N-Triples = každá trojice se zapíše jako `<uri> <uri> <uri> .  # comment`.
- RDF 1.1 Turtle = použijí se prefixy, středníky na shodný subjekt a čárku na shodný subjekt i predikát + se používají relativní IRI (base se definuje pomocí `@base IRI`, implicitní je URL dokumentu) + multiline stringy a odescapované znaky + `rdf: type` má zkratku `a` + blank nody se píší pomocí hranatých závorek + běžné literály nemusí mít typ.
- RDF 1.1 N-Quads = místo trojice se kóduje i pojmenování grafu.
- RDF Trig = Turtle + pojmenované grafy (jsou reprezentovány jako bloky).

Definice 2.5 (Reifikace)

Pokud chci něco říct o naší trojici, můžu to udělat tak, že si definuji (zase pomocí trojic) objekt, který jako subjekt bude mít subjekt, atd. a navíc bude mít doplňující informace. Tato metoda se nazývá reifikace.

Definice 2.6 (Pojmenovaný graf, dataset)

Vztahy lze seskupit do tzv. pojmenovaného grafu.

Pojmenované grafy + defaultní graf se nazývá dataset.

Definice 2.7 (RDFS)

Nadstavba RDF, které umožňuje definovat třídy a dědičnost. `rdfs:Class`, `rdfs:subClassOf`, `rdf:Property`, `rdfs:range`, `rdfs:domain`, `rdfs:subPropertyOf`.

Oproti OOP není třeba definovat třídy, lze definovat property jako takové.

Také umožňuje label, comment, seeAlso: `rdfs:label`, `rdfs:comment`, `rdfs:seeAlso`, `rdfs:isDefinedBy?`.

Definice 2.8 (rdf:List a jiné kolekce)

Ve specifikaci RDF je přímo definován spojový seznam (`rdf:List` + anonymní prvky + `rdf:nil`).

`rdf:_i`, kde i je libovolné číslo jsou predikáty náležení do kolekce (`rdf:TODO`).

Definice 2.9 (Open World Assumption (OWA))

Tvrzení může být pravdivé, i když to nevíme. (Tj. máme i odpověď nevím.)

TODO!!!

Definice 2.10 (Otevřená data 5 hvězdičkova klasifikace dat)

První hvězdička je za uvedenou licenci, druhá je za strojovou čitelnost, třetí je za otevřený formát, čtvrtá za URI odkazy, pátá za připojení do systému LOD.

3 SPARQL

Definice 3.1 (SPARQL)

SPARQL je dotazovací jazyk nad daty v RDF. SPARQL endpoint je HTTP služba pro dotazování v SPARQL na daných open datech.

Poznámka
Doporučovaný user formulář je yasgui.

Funguje tak, že se píší RDF trojice s `?nazevproměnné` v místě, kde chceme něco doplnit (a zjistit, co to je). To jsou tzv. datové vzory.

Výsledkem je pak tabulka řešení, kde je v každém řádku jeden match a v každém sloupci jedna proměnná, v políčkách je tam pak doplněno.

Do dotazu lze připsat `OPTIONAL` a výsledek pak bude matchovat, i když tato část bude chybět a v tabulce pak bude `NOT BOUND`. Také lze přidat `FILTER` pro podmínky s proměnnými.

Oproti SQL máme ještě RDF operátory: `bound, isIri, isBlank, isLiteral` a přístup k literálu: `str, language, typeOf`?

Taktéž fungují `/` jako v cestě k souboru, která se navíc zadává Regexem.

Jedním dotazem se můžeme ptát na více SPARQL endpointů, což uděláme pomocí příkazu `SERVICE`.

TODO!!!

4 Nejčastější slovníky RDF

Definice 4.1 (Dublin Core metadata)

Jeden z prvních slovníků, vznikl na popis knih (a dalších děl). Jsou to pojmy se zkratkou dcterms (Dublin Core Metadata Initiative).

Definice 4.2 (skos)

Konceptuální slovník. Důležité jsou např. `skos:prefLabel, skos:altLabel, skos:hiddenLabel`. Dále třeba `notation` a různé typy `skos:semanticRelation`.

Definice 4.3 (GoodRelations)

Slovník pro e-komerci.

Definice 4.4 (Schema.org)

Založen firmami Google, Microsofte, Yahoo a Yandex. Integruje existující slovníky. Určeno pro jednoduchou anotaci webových stránek, ne k dobré strukturalizaci.

Definice 4.5 (Wikidata)

Komunitní RDF data. Má k sobě také slovník. Běží na softwaru Wikibase.

TODO!!!

5 Hierarchické datové formáty

Definice 5.1 (Dokumentově orientované XML)

Dokument, do kterého se vloží značky (tj. bez značek je stále čitelný).

Definice 5.2 (Datově orientované XML)

To jsou pouze data se značkami (tj. bez značek je „nečitelný“).

Definice 5.3 (XML 1.0 a XML 1.1)

1.0 má list povolených znaků, 1.1 zakázaných. Aplikace ale zamrzly u 1.0.

TODO syntaxe XML

TODO!!!

TODO (Nebyl jsem)

6 Relační datový model

Definice 6.1 (Relační datový model)

V relačním modelu máme tabulku, která má řádky (záznamy) a sloupce (klíče). Pak máme primární klíč, jehož hodnoty určují jednoznačně každý záznam. A foreign klíč, tj. že se odkazujeme na cizí tabulku.

Definice 6.2 (SQL dump)

Vytvoří SQL příkaz, který vytvoří přesnou kopii dat.

Poznámka (Before CSV)
Delimiter-Separated Values (DSV): (delmiter = označerní kusu dat vs. separator = odděluje kusy dat (neoznačuje začátek + konec)), skoro jako CSV, jen jiné separátory a jiné kódování.

Tab-Separated Values (TSV): už má specifikaci, odděluje tabulátory.

Definice 6.3 (Comma-Separated Values (CSV))

Kódování defaultně UTF-8 (od 2014, předtím US-ASCII, jsou možná i jiná), oddělovačem je čárka, escape znakem je uvozovka (`""` je odescapovaná uvozovka, escapuje tak, že se věc uzavře do uvozovek),

Poznámka
Dále jsme si povídali o správných a špatných příkladech.

Definice 6.4 (URI Fragment Identifiers pro csv)

Když máme uri csv, můžeme na konec přidat `#col=rozsah`, `#row=rozsah` nebo `#cell=rozsah` rozsah může mít `-` a `*`.

Poznámka
Dále jsme se bavili o tom, jak popisovat schéma csv.

TODO!!!

7 Prostorové informace

Poznámka (Otázky, pro které potřebujeme prostorové informace)
Jak je to daleko? Kterým směrem to je? Co je nejbližší? Co největší, nejvyšší, atd.

Definice 7.1 (Prostorové informace)

Zabývá se jimi hlavně norma ISO/TC 211. Obor se nazývá Geoinformatika, Geomatika apod.

Geografická data = Geoprostorová data = Prostorová data – data a informace, která mají implicitní/explicitní lokaci oproti zemi.

K souřadnicím se většinou používá tzv. referenční elipsoid, který máme „umístění“ kolem země a projektujeme na něj. Má chybu cca půl metru? Ale existuje i mnoho dalších možností.

Implicitní = souřadnice, vzdálenosti směry; Explicitní = pojmenování, adresy, geografická jména.

Definice 7.2 (Typy)

Points, Multipoints, Lines (Linestrings), Multilines, Polygons, Multipolygons, Surface.

Dále lze používat i křivky jako kružnice, ale spíš se to nedělá (kružnice se rozseká a udělá se z ní Polygon).

Definice 7.3 (Points, Multipoints)

Points – určují „bodový“ objekt. Mutipoints jsou například zastávky – jednomu objektu odpovídá více „diskrétních“ bodů.

Definice 7.4 (Lines, Multilines, Polygons, Multipolygons)

Jako Points, jen s liniovými/mnohoúhelníkovými objekty.

Definice 7.5 (Well-Known Text (WKT))

OGC standart (nebo v placeném ISU 19125?), téměř všechny knihovny předpokládají WGS-84 (referenční elipsoid, souřadnice ve stupních). Je tvaru `TYP(souradnice1x souradnice1y souradnic`

Definice 7.6 (Geometry Markup Language (GLM))

XML, které kromě typů obsahuje i odkaz na souřadnicový systém, dimenzi, atd.

Definice 7.7 (Další formáty)

GeoJSON, Shapefile, GeoPackage, CSV, GeoSPARQL, …

Definice 7.8 (Feature)

(Časky geografický popis objektu?, nepoužívá se.)

Objekt, který může mít nějaký geografický popis (např. dům). Featura má atributy (které mluví o negeografických vlastnostech) a geometrii (geografická data).

Poznámka
Geografické objekty lze kreslit třeba v geojsonu… (Lze tam i převádět z geoJSONu.)

Definice 7.9 (Prostorové relace)

Topologické: Within, Touches, Crosses, Overlaps.

Směrové: Left, Right.

Vzdálenostní: Closer Further.

A ještě všechno může být určené v čase.

Definice 7.10 (Prostorové operace)

Buffer (nafoukne objekt na polygon všech bodů, které jsou blíže než nějaká určená vzdálenost od nějakého bodu objektu). Union, Difference, Intersection, Clip (odříznutí), Distance, Convex Hull.

Definice 7.11 (Geografické informační softwary a knihovny pro zpracování prostorových informací)

QGIS (zdarma), PostGIS (rozšíření pro PostgreSQL), ESRI ArcGIS (velký komerční projekt).

TODO?

8 Key-values formáty

Definice 8.1 (Properties file)

Pouze v Javě. Používá pouze Latin 1. Říká se mu také hash table.

Má i XML variantu.

Definice 8.2 (INI file)

Originálně MS-DOS, ve windows postupně nahrazován registry (ale i ve Windows 10 stále jsou INI). Bez standardu. Lze se na něj dívat jako na dvojrozměrný hashtable.

Definice 8.3 (TOML)

Podobný jako INI. Vznikl v roce 2021. Kódování je unicodové.

Nabízí i to, že knihovny pro dané jazyky umí z tohoto souboru číst i různé typy.

Syntax: komentář se značí #, hodnoty se ukládají jako klíč = hodnota, kde klíč může být buď bare, nebo odeskapovaný v uvozovkách (když má speciální znaky), nebo s tečkou, viz dále. Hodnota může být string, číslo (i v hexadecimálním, oktálním, binárním zápise, i ∞, i NaN), datetime, pole, slovníky, … Mimo to můžeme mít ještě označení tabulky `[nazev tabulky]`, nebo `[[nazev pole tabulek]]`.

Tabulky fungují jako slovník a buď se dají vyplňovat tak, že se označí název tabulky a pak se píšou klíč = hodnota, nebo se můžou napsat i jako `nazevtabulky = {klic = hodnota, …}`.

Definice 8.4 (YAML)

Nadmnožina jsonu. Přidává např. začátek `---` a konec dokumentu `...`. key-value (syntaxí: `key: value`) se zde nazývá mapping, *map*, dictionary nebo hash. Záleží v něm na odřádkování.

Poznámka (Ještě existují)
HOCON, JSON5 (/ JSON for Humans), Strict YAML