OPEN REFINE – PŘEDZPRACOVÁNÍ DAT

11 LETNÍ ŠKOLA MATEMATICKÉ BIOLOGIE

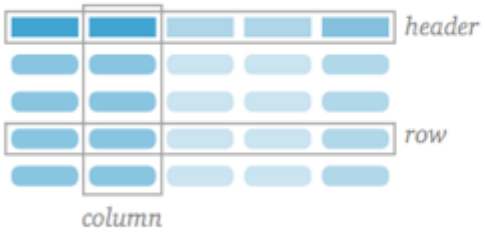
17.9.2015

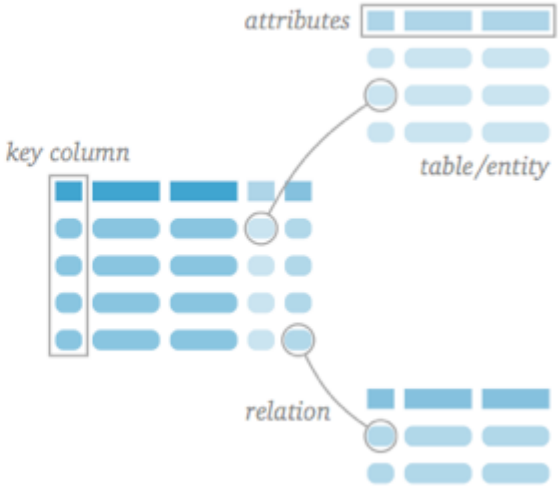
RNDr. Miroslav Kubásek, Ph.D.

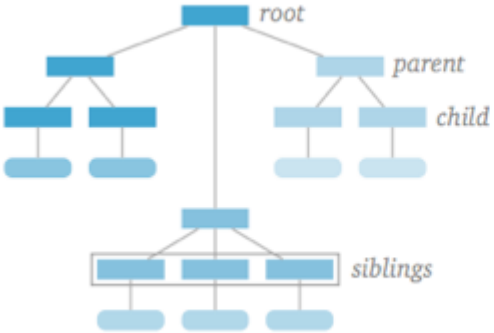
Program

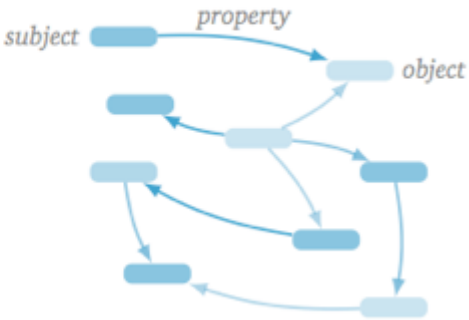
- ✓ Data
- ✓ Proč OpenRefine
- ✓ Instalace
- ✓ Základní orientace v programu
- ✓ Jazyk GREL
- ✓ Regulární výrazy
- ✓ Praktické příklady
  - ✓ Čištění dat
  - ✓ Získání dat z PDF souborů
  - ✓ Získání dat z webu
  - ✓ Doplnění dat pomocí API
  - ✓ Vizualizace

Jak modelovat data









Jak modelovat data - výhody

data model	(dis-)advantages	usage
tabular data	+ intuitive approach + very portable + technology agnostic – prone to redundancy and leading to inconsistencies – inefficient search and retrieval	import and export of data with a simple structure
relational model	+ handling of complex data + optimized queries + mature software market – binary format – schema dependent	management of complex data which require normalization
meta-markup	+ platform-independent + both human and machine readable – complicated implementation for complex data – verbosity	import and export of complex data
RDF	+ schemaless approach + discovery of new knowledge – loss of normalization – immature software market	making data available for linking

Výběr správného nástroje

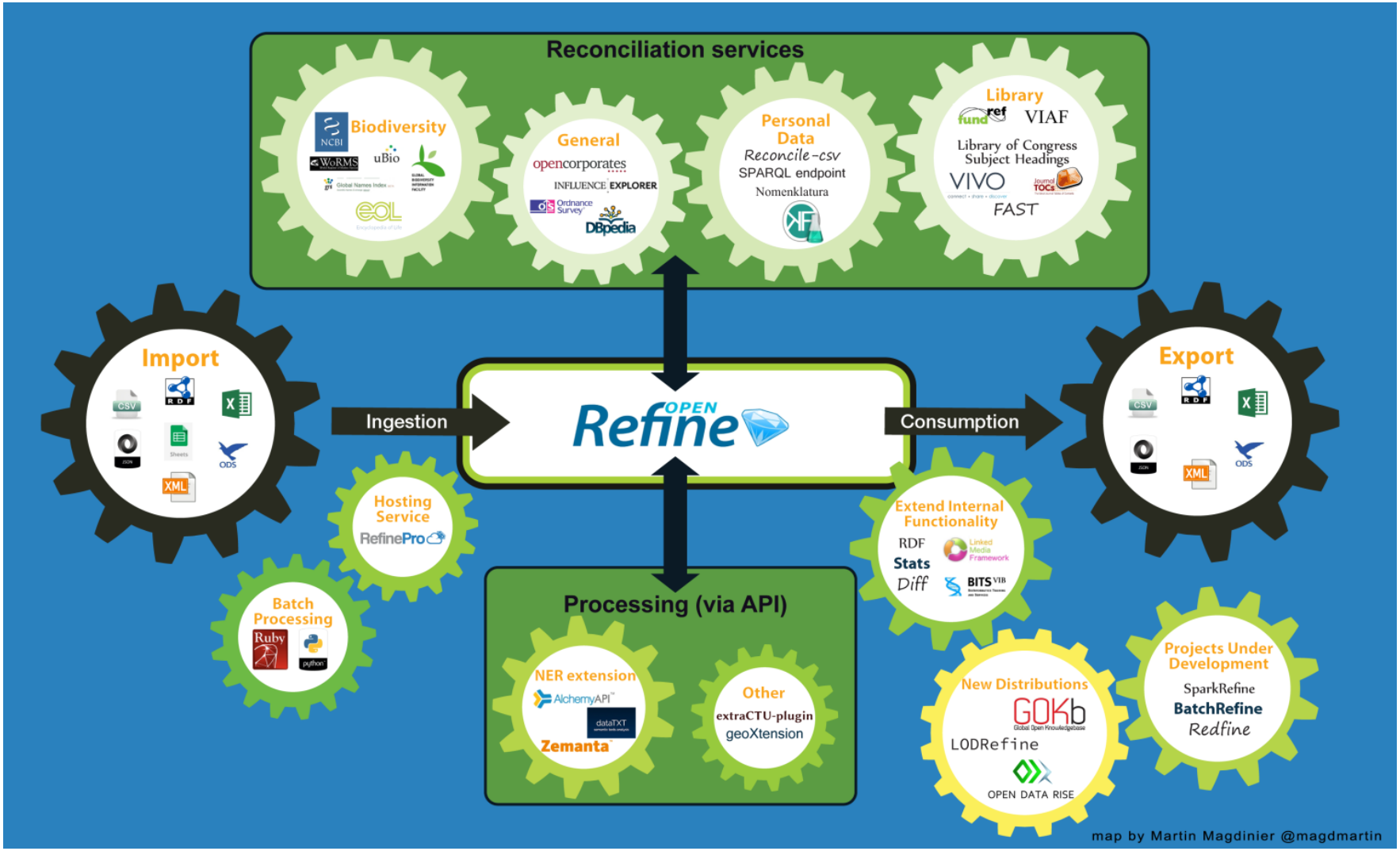
OpenRefine	Tabulkový editor (Excel)	Databáze
Je možné dávkově zpracovat jak řádky tak i sloupce. Umožňuje zkoumat a transformovat data. Není potřeba definovat schéma dat. Data jsou viditelná v každém kroku editace. Mnohem více interaktivní a vizuální nástroj.	V jeden okomžik lze editovat jen jednu buňku. Vhodné pro vkládání dat a provádění výpočtů. Není potřeba definovat schéma dat. Data jsou viditelná, omezený počet řádků. Vizualizace dat je pouze základní.	Je potřeba znát dotazovací jazyk (SQL) Pouze základní možnosti transformací. Je potřeba definovat schéma zpracovávaných dat. Data jsou skryta, dokud je dotazem nevyexportujete. Použití příkazového řádku pro spouštění dotazů.

Vlastnosti OpenRefine

„Spreadsheet on steroids"

✓ Pracujete s kopií dat
✓ Data jsou umístěna a zpracovávána na Vašem počítači, na žádný cizí server se nanahrávají (bezpečnost dat)
✓ Pracujete ve webovém prohlížeči (doporučuje se Google Chrome), díky tomu můžete pracovat na vzdáleném stroji
✓ OpenRefine udržuje historii provedených akcí (můžete se vrátit libovolně zpět)
✓ Provedené akce lze exportovat a znovu aplikovat
✓ Základní práce: filtering, faceting, cluster, split, join, fill, transpose
✓ Projekt lze vyexportovat a poté importovat na jiném počítači
✓ OpenRefine vhodný pro velké datové sady (nad 100 tis. záznamů), zvládá až milióny záznamů, záleží na velikosti paměti a výkonu počítače
✓ Import z formátů: TSV, CSF, *SV, Excel (.xls .xlsx), JSON, XML, RDF as XML, Google Data documents. Zvládá import přímo z webu, z zkomrimovaných souborů (více souborů)

OpenRefine ecosystem



Instalace

- ✓ http://openrefine.org/download.html
- ✓ spusťte soubor **refine.bat**
- ✓ přejdete v prohlížeči na adresu **http://127.0.0.1:3333/**

Všechny použité datové soubory použité v tomto textu včetně instalačních souborů jsou k dispozici na adrese

https://github.com/MiroslavKubasek/OpenRefine-tutorial

Nastavení OpenRefine

Under http://127.0.0.1:3333/preferences you can define the number of facet choices

Allocate more memory to Refine :
Windows : open openrefine.l4j.ini file, find the line that starts with -Xmx and override the default allocated memory of 1024M with for example 2048 M
Mac : close Refine, hold control and click on its icon, selecting Show package contents from the pop-up menu. Open the info. plist file from the Contents folder. Navigate to the Java settings and edit the value of VMOptions. Look for the part that starts with -Xmx and change its default value of 1024 M to the desired amount of memory
Linux: instead of starting OpenRefine with ./refine as you usually would do, just type in ./refine -m 2048M

OpenRefine praktické kódy

Součet: forEach(value.split(','),v,v.toNumber()).sum()

Průměr: with(value.split(','), a, forEach(a, v, v.toNumber()).sum() / a.length()).replace("NaN","").toNumber()

Získání titulku z webu:
value.substring(indexOf(value, "<title>")+7, indexOf(value, "</title>"))
value.parseHtml().select("title")[0].htmlText()

Data pro testování

https://github.com/MiroslavKubasek/OpenRefine-tutorial/tree/master/data

http://datahub.io/
http://www.data.gov/
https://data.cityofnewyork.us/

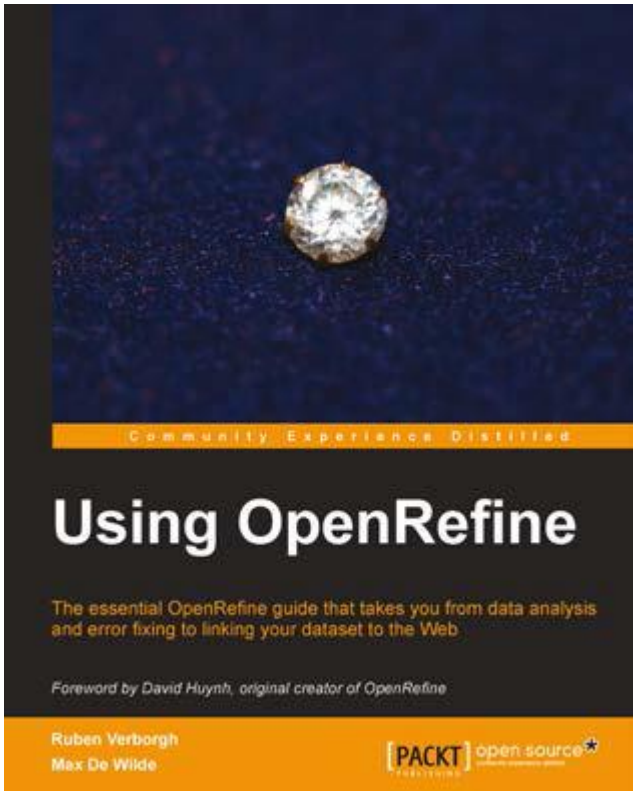
OpenRefine tutoriály

Getting Started with OpenRefine - http://thomaspadilla.org/dataprep/

Scraping multiple Pages using the Scraper Extension and Refine - http://schoolofdata.org/handbook/recipes/scraping-multiple-pages-with-refine-and-scraper

Cleaning Data with Refine - http://schoolofdata.org/handbook/recipes/cleaning-data-with-refine/

OpenRefine literatura



http://www.amazon.com/Using-OpenRefine-Ruben-Verborgh/dp/1783289082

http://casci.umd.edu/wp-content/uploads/2013/12/OpenRefine-tutorial-v1.5.pdf

Vizualizace dat

- ✓ **Gephi** – nástroj pro vizualizaci sítí - http://gephi.github.io/
- ✓ **Tagul** – tvorba word clouds - https://tagul.com/
- ✓ **Google Chart API** – online tvorba grafů - https://developers.google.com/chart/
- ✓ **D3** – Data Driven Documents, JavaScript - http://d3js.org/
- ✓ **Dimple** – nádstavba nad D3, jednodušší použití - http://dimplejs.org/

Vizualizace dat – příklad dimple

```
var data = {
 "rows" : [
  {
   "type" : " Fighting",
   "number" : 3
  },
  {
   "type" : " Electric",
   "number" : 7
  },
  {
   "type" : " Psychic",
   "number" : 9
  },
  {
   "type" : " Ghost",
   "number" : 10
  },
  {
   "type" : " Ice",
   "number" : 11
  },
  {
   "type" : " Poison",
   "number" : 11
  },
  {
   "type" : " Dragon",
   "number" : 12
  },
  {
   "type" : " Steel",
   "number" : 13
  },
  {
   "type" : " Fire",
   "number" : 14
  },
  {
   "type" : " Dark",
   "number" : 16
  },
  {
   "type" : " Ground",
   "number" : 17
  },
  {
   "type" : " Rock",
   "number" : 24
  },
  {
   "type" : " Normal",
   "number" : 29
  },
  {
   "type" : " Grass",
   "number" : 31
  },
  {
   "type" : " Bug",
   "number" : 45
  },
  {
   "type" : " Water",
   "number" : 45
  }
 ]
};
```

```
var order = true;
$("#button").click(function() {
 x._orderRules.pop();
 if(order) {
  x.addOrderRule("number", true);
 } else {
  x.addOrderRule("number", false);
 }
 myChart.ease = "bounce";
 myChart.draw(800, false);
 order = ! order;
});

var svg = dimple.newSvg("graph", "100%", 400);
var myChart = new dimple.chart(svg);
myChart.data = data.rows;
myChart.setBounds(20, 20, "90%", 330);
var x = myChart.addCategoryAxis("x", "type");
x.addOrderRule("number", false);

myChart.addMeasureAxis("y", "number");
myChart.addSeries(null, dimple.plot.bar);

myChart.draw(800);
```

http://output.jsbin.com/cevilihuyo/