



WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS



Datenorientierte Systemanalyse

05/05/2014

Axel Polleres

Sie finden diese Folien unter: <http://polleres.net/>

Datenorientierte Systemanalyse/ Datenanalyse

- **Organisation**
- Die Lehrveranstaltung im Schnelldurchlauf
- Hands-on:
 - Zu unserem Server verbinden.
 - Unix/Linux basics
 - Anlegen einer einfachen Web-Seite.
 - Anlegen von Tabellen in Ihrer Datenbank in PostgreSQL

Organisation

Was heißt das für Sie?

- Lernen mit einer relationalen Datenbank umzugehen (PostgreSQL)
 - Tabellen anlegen
 - Tabellen richtig strukturieren
 - Abfragen
- Daten erfassen und verwalten
 - Web-User-Interface erstellen
 - Einfache Reports in HTML generieren
- Statistische Analyse mit R
- Andere Datenformate importieren und nutzen
 - CSV, JSON, Graph-basierte Daten, RDF)
- Andere Abfragesprachen als SQL:
 - SPARQL
- Ein eigenes kleines Projekt realisieren

1849 - Datenorientierte Systemanalyse

W/ 1849 - Datenorientierte System... x BACH - Wirtschaftsuniversität ... x +

https://learn.wu.ac.at/vwz/13w/1849#lvbeschreibung1

Google

Contents to top

Relationale Datenbanken, Graph- und Semantische Datenbanken, sowie andere Datenquellen

Auswertungen mit SQL sowie Skriptsprachen und Statistikpaketen

Learning outcomes to top

Verstehen der Konzepte der datenorientierten Systemanalyse und Fähigkeit zum Einsatz in praktischen Aufgaben bei der Erstellung von Applikationen zur dynamischen Extraktion von Datenbankinhalten und deren Aufbereitung mit einfachen statistischen Auswertungen

Teaching/learning method(s) to top

Vortrag und laufende Mitarbeit in Form von Stundenwiederholungen und Mitarbeit beim Lösen von praktischen Aufgaben

Assessment to top

Laufende Mitarbeit

Mehrere kurze schriftliche Prüfungen

Präsentation einer selbst erstellten Anwendung zu ausgewählten Themen

Organisation

Was heißt das für Sie?

- Typischer Ablauf einer Einheit:
 - eine **Stundenwiederholung** - durch die Teilnehmer
 - neuer Stoff - von mir vorgetragen, **teilweise** durch Folien, meist aber durch Vorzeigen am Rechner... **Learning by doing!**
 - **eigene Notizen** - erwünscht!
 - **eigene Recherche** nach guten Tutorials im Web - erwünscht!
- kleine **Hausübungen** wöchentlich - Abgabe über Learn@WU, als Basis für die Stundenwiederholungen
- Jeder Block dauert üblicherweise 2h
- Am Schluss:
Präsentation eines **Mini-Projekts**

The screenshot shows a web browser window with the title '1849 - Datenorientierte Systemanalyse'. The address bar contains the URL 'https://learn.wu.ac.at/vwz/13w/1849#lvbeschreibung1'. The page content is organized into sections: 'Contents', 'Learning outcomes', 'Teaching/learning method(s)', and 'Assessment'. The 'Teaching/learning method(s)' section is circled in red. The 'Contents' section lists 'Relationale Datenbanken, Graph- und Semantische Datenbanken, sowie andere Datenquellen' and 'Auswertungen mit SQL sowie Skriptsprachen und Statistikpaketen'. The 'Learning outcomes' section describes understanding concepts and practical application. The 'Teaching/learning method(s)' section mentions 'Vortrag und laufende Mitarbeit in Form von Stundenwiederholungen und Mitarbeit beim Lösen von praktischen Aufgaben'. The 'Assessment' section lists 'Laufende Mitarbeit', 'Mehrere kurze schriftliche Prüfungen', and 'Präsentation einer selbst erstellten Anwendung zu ausgewählten Themen'.

Organisation

Was heißt das für Sie?

Geplante Themen:
Überblick
Datenbank-Modellierung
SQL als Abfragesprache
Web-Use-Interface erstellen
Andere Datenformate (CSV, JSON,...)
Graph-Daten (RDF&SPARQL)
Daten-Analyse mit R
evtl. Gastvortrag SAP HANA (in-memory Datenbanken)
Wrap-up & Q/A
Mini-Projekt Präsentationen
(1 slot „buffer“)

Day	Date	Room
Monday	05/05/14	D2.0.031 Workstation-Raum
Wednesday	05/07/14	D2.0.031 Workstation-Raum
Monday	05/12/14	D2.0.031 Workstation-Raum
Wednesday	05/14/14	D2.0.031 Workstation-Raum
Monday	05/19/14	D2.0.031 Workstation-Raum
Wednesday	05/21/14	D2.0.031 Workstation-Raum
Monday	05/26/14	D2.0.031 Workstation-Raum
Wednesday	05/28/14	D2.0.031 Workstation-Raum
Monday	06/02/14	D2.0.031 Workstation-Raum
Wednesday	06/04/14	D2.0.031 Workstation-Raum !
Wednesday	06/11/14	D2.0.031 Workstation-Raum
Monday	06/16/14	D2.0.031 Workstation-Raum

Organisation

Was heißt das für Sie?

- **Benotung:**
 - 1. Mitarbeit & Stundenwiederholungen (+/-)**
 - *z.B. eigene Recherche, selbständige Lösungsideen; auch Fragen, wenn man selbst nicht weitergekommen ist, sind ok!(aber nicht erst bei der Stundenwiederholung)*
 - 2. alle Hausübungen** abgegeben (+/-)
 - Stichproben
 - 3. Mini-Projekt am Schluss**
 - **Benotung (1-5)**
- **Endnote ergibt sich aus:**
 - **1.+2. (75%)**
 - **3. (25%)**
 - Aber :
 - keine positive Note ohne Abgabe von 3.

The screenshot shows a web browser window with the following content:

- Browser tabs: "1849 - Datenorientierte System...", "BACH - Wirtschaftsuniversität ..."
- Address bar: "https://learn.wu.ac.at/vwz/13w/1849#lvbeschreibung1"
- Page title: "1849 - Datenorientierte Systemanalyse"
- Section: **Contents** (with "to top" link)
 - Relationale Datenbanken, Graph- und Semantische Datenbanken, sowie andere Datenquellen
 - Auswertungen mit SQL sowie Skriptsprachen und Statistikpaketen
- Section: **Learning outcomes** (with "to top" link)
 - Verstehen der Konzepte der datenorientierten Systemanalyse und Fähigkeit zum Einsatz in praktischen Aufgaben bei der Erstellung von Applikationen zur dynamischen Extraktion von Datenbankinhalten und deren Aufbereitung mit einfachen statistischen Auswertungen
- Section: **Teaching/learning method(s)** (with "to top" link)
 - Vortrag und laufende Mitarbeit in Form von Stundenwiederholungen
 - und Mitarbeit beim Lösen von praktischen Aufgaben
- Section: **Assessment** (with "to top" link, circled in red)
 - Laufende Mitarbeit
 - Mehrere kurze schriftliche Prüfungen
 - Präsentation einer selbst erstellten Anwendung zu ausgewählten Themen

Fehlstunden

- **ACHTUNG:** Es besteht prinzipiell *Anwesenheitspflicht*: Sie können bei **max. 2 Einheiten** (egal welche und egal aus welchen Gründen) fehlen, aber wir behalten uns vor, Sie bei weiterem Fehlen negativ zu beurteilen!

Datenorientierte Systemanalyse

- Organisation
- **Die Lehrveranstaltung im Schnelldurchlauf**
- Hands-on:
 - Zu unserem Server verbinden.
 - Unix/Linux basics (login-shell aendern, ls, cd, mkdir, pico)
 - Anlegen einer einfachen Web-Seite (HTML)
 - Anlegen von Tabellen in Ihrer Datenbank in PostgreSQL

Relationale Datenbanken vs. Spreadsheets

- Haben Sie schon einmal mit einer Datenbank gearbeitet? ... Vermutlich ja, Spreadsheets sind in gewisser Weise auch „Datenbanken“, aber...

The top screenshot shows LibreOffice Calc with the following data:

	A	B	C	D	E	F	G	H
	MatNr	Name	LVNr	LVTitel	Semester	LVLeiter	Note	
1								
2	812345	Maxeline Musterfrau	1849	Datenorientierte Systemanalyse	WS2013	A. Polleres		1
3	1212345	Tom Turob	1849	Datenorientierte Systemanalyse	WS2013	A. Polleres		3
4	1100815	Karl Karl	1849	Datenorientierte Systemanalyse	WS2013	A. Polleres		3
5	9947110	Max Mustermann	1849	Datenorientierte Systemanalyse	WS2013	A. Polleres		5
6	1211111	Donald Duck	1848	Datenorientierte Systemanalyse	WS2013	H. Mitlöchner		2
7	1222222	Daisy Duck	1848	Datenorientierte Systemanalyse	WS2013	H. Mitlöchner		3
8	1233333	Mickey Mouse	1848	Datenorientierte Systemanalyse	WS2013	H. Mitlöchner		4
9	1333333	Minney Mouse	1848	Datenorientierte Systemanalyse	WS2013	H. Mitlöchner		1
10	1211111	Donald Duck	1848	Datenorientierte Systemanalyse	WS2012	Mad Professor		5
11	1222222	Daisy Duck	1848	Datenorientierte Systemanalyse	WS2012	Mad Professor		5
12						Durchschnittsnote		3.2

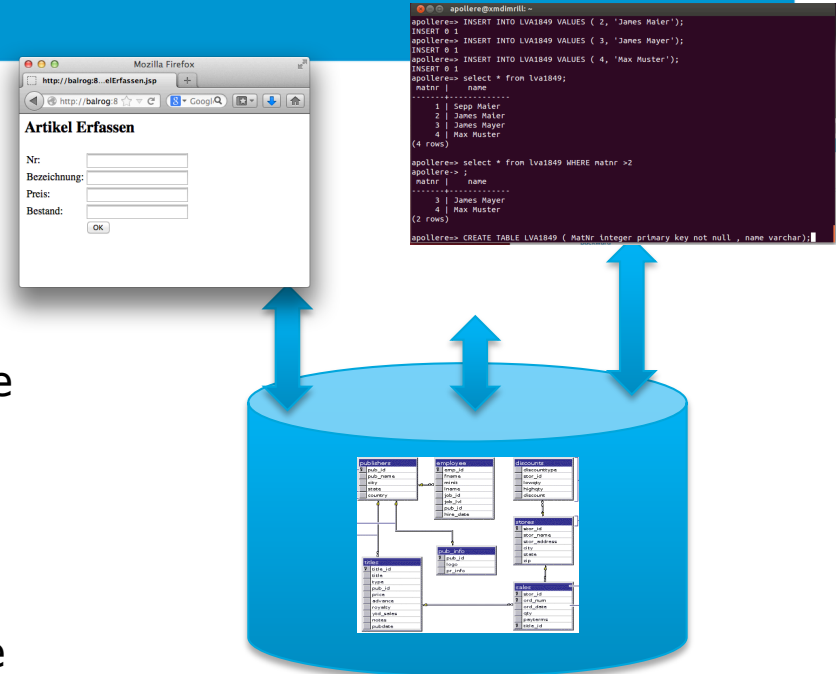
The bottom screenshot shows Google Docs with the same data and a formula bar containing `=MITTELWERT(G2:G11)`.

- Fehlende Information
- (unnötig) duplizierte Information
- Rohdaten und berechnete Daten (Abfragen) vermischt
- Skaliert nicht auf große Datenmengen
- Komplexe Anfragen sind schwierig zu formulieren
- Nicht-standardisierte Abfragesprachen, z.B. unterschiedl. Funktionsnamen Office vs. Excel, etc.



Relationale Datenbanken vs. Spreadsheets

- Redundanzfreie Speicherung von Daten, die miteinander in Beziehung stehen, in mehreren Tabellen
- Abfragen (queries, views) getrennt von Daten (tables)
- Skaliert auf große Datenmengen
- Mehrere Benutzer greifen über user-interface auf die daten zu (Concurrency/transaction support), z.b. über ein Web-Interface.
- Typischerweise unterschiedliche **Rollen** von Benutzern: Administratoren, „Power-user“, Endbenutzer, etc.
- Standardisierte Abfrage- und Datendefinitionssprache – SQL
- Einige Relationale Datenbanksysteme:



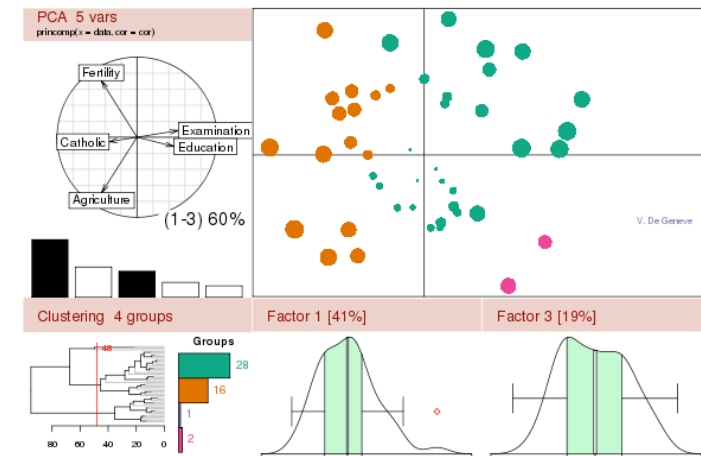
- ~~Fehlende Information~~
- ~~(unnötig) duplizierte Information~~
- ~~Rohdaten und berechnete Daten (Abfragen) vermischt~~
- ~~Skaliert nicht auf große Datenmengen~~
- ~~Komplexe Anfragen sind schwierig zu formulieren~~
- ~~Nicht-standardisierte Abfragesprachen, z.b. unterschiedl. Funktionsnamen LibreOffice vs. Excel, etc.~~
- ~~Typischerweise single user~~



- Wer hat schon mit relationalen Datenbanken gearbeitet?
- Wer kennt SQL?
- Wer kennt das Entity-Relationship Modell?
- Wer kennt Relationale Algebra?

Statistik als Analyse Tool

- Einfache Analysen (Balkendiagramme, etc.) sind wiederum in Spreadsheets möglich...
 - ... Ähnliche Probleme wie bei Datenbanken:
 - Unhandlich für grössere Datenmengen,
 - Keine Schnittstellen um reports zu generieren
 - Vorschau: wir werden uns hierzu mit dem Statistikpaket „R“ befassen
 - Skriptsprache
 - Austausch mit Daten aus relationalen Datenbanken
 - Report-Generierung
- <http://www.r-project.org/>
- Einstiegs-Tutorial: <http://mitloehner.net/lehre/rbasics/rbasics.html>
- Report-Generierung: <http://mitloehner.net/lehre/datsys/reports.html>



- Wer hat bereits Erfahrung mit Statistik-Tools?
- Mit „R“? Andere?
- Programmieren? Java? (zB. GZP)
- Skriptsprachen im Web (zur Report-Generierung)?
JSP? Ruby? etc.

Semantic Web Daten

- In den letzten Jahren, mehr und mehr Graph-Daten im Web als RDF verfügbar: Linked Data

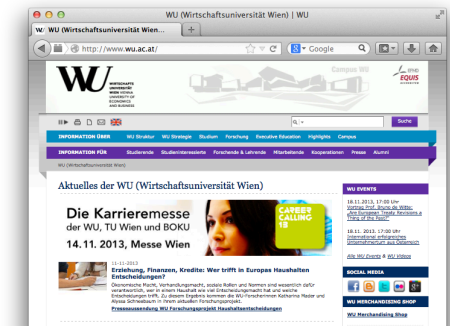
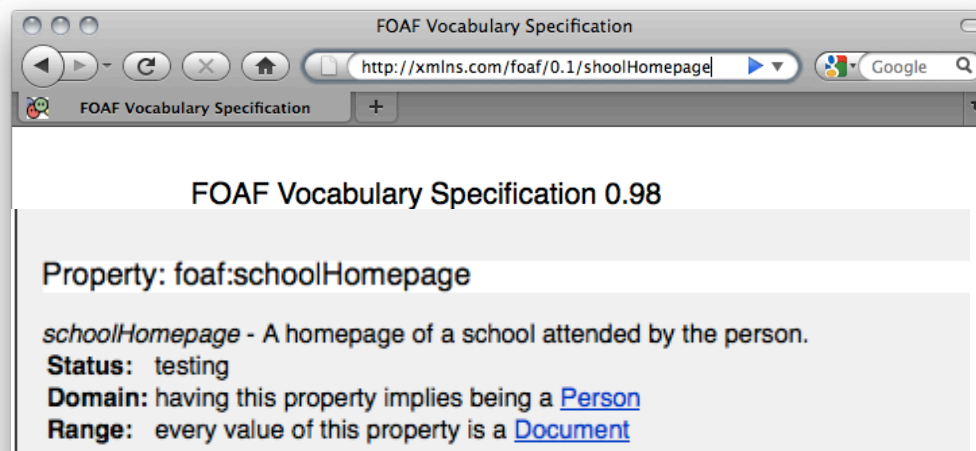
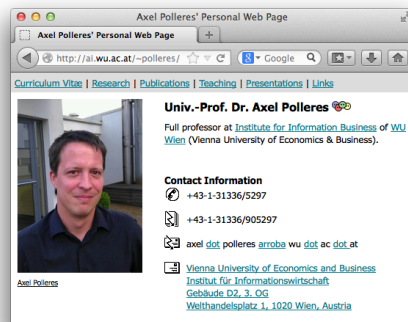
Linked Data Principles

1. Everything gets a URI (papers, people, talks, organizations, topics...)
2. These URIs are linked via RDF describing relations
3. Relations are URIs again (e.g. :name)
4. **When I dereference the URIs, I should find more information about them, defining them.**

polleres.net#me

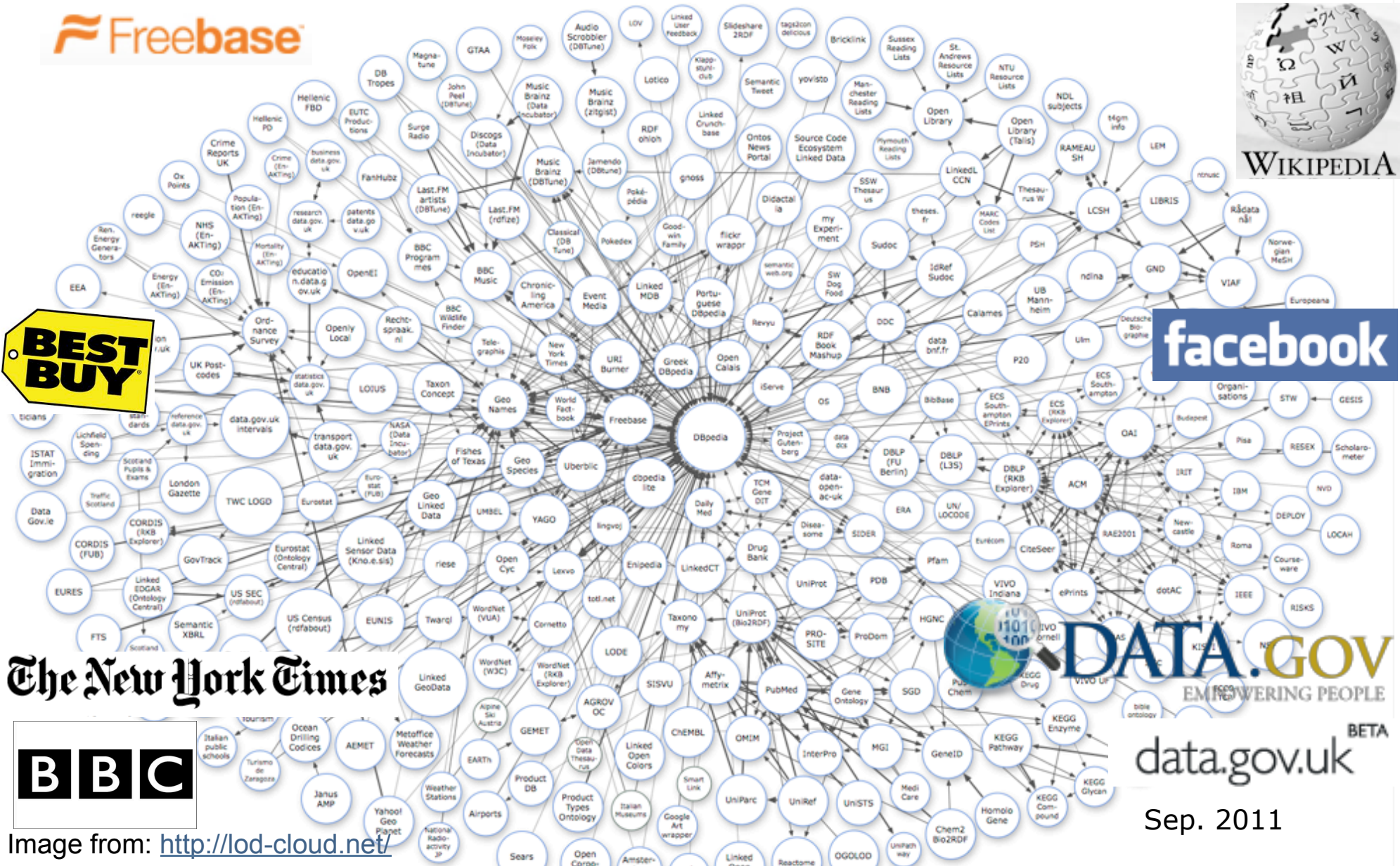
xmlns.com/foaf/0.1/workplaceHomepage

wu.ac.at



Linked Data on the Web: Adoption

Freebase



The New York Times



DATA.GOV
EMPOWERING PEOPLE
data.gov.uk
BETA

Sep. 2011

Image from: <http://lod-cloud.net/>

A standard Query language for Web Data: SPARQL

SPARQL = *SQL look-and-feel for Linked Data*



“Latest news on NYT about technology companies with a revenue greater than 10B EUR”

```
SELECT * WHERE
{ ?C rdf:type NYT:Org .
  ?C dbpedia:revenue ?R .
  ?C NYT:latestArticle ?A .
  FILTER( ?R > 10000000000 ) }
```



- Wer hat schon von RDF gehört?
- Wer hat schon von SPARQL gehört?
- ... Ein kleines Beispiel:
 - „österreichische Universitäten und Ihre derzeitigen Rektoren“
 - http://dbpedia.org/sparql/?query=SELECT+*+WHERE+{%0D%0A+%3FU+rdf%3Atype+%3Chttp%3A%2F%2Fdbpedia.org%2Fontology%2FUniversity%3E+%3B%0D%0A++++%3Chttp%3A%2F%2Fdbpedia.org%2Fontology%2Fcountry%3E+%3Chttp%3A%2F%2Fdbpedia.org%2Fresource%2FAustria%3E+%3B%0D%0A++++%3Chttp%3A%2F%2Fdbpedia.org%2Fproperty%2Frector%3E+%3FR++++%0D%0A%7D

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns# >

```
SELECT * WHERE {  
  ?U rdf:type <http://dbpedia.org/ontology/University> ;  
    <http://dbpedia.org/ontology/country> <http://dbpedia.org/resource/Austria> ;  
    <http://dbpedia.org/property/rector> ?R .  
}
```

Datenorientierte Systemanalyse

- Organisation
- Die Lehrveranstaltung im Schnelldurchlauf
- **Hands-on:**
 - Zu unserem Server verbinden (`ssh`)
 - Unix/Linux basics (login-shell ändern, files anlegen, löschen, editieren: `ls`, `cd`, `mkdir`, `pico`, `rm`, `rmdir`; command-line history: CTRL-r)
 - Anlegen einer einfachen Web-Seite (`html`) im `www` Verzeichniss
 - Anlegen von Tabellen in Ihrer Datenbank in PostgreSQL (`psql`)

ssh – Secure Shell

- Öffnen Sie ein Terminalfenster
 - Verbinden Sie sich mit `ssh` mit unserem Server `balrog` verbinden

```
ssh -X -l username balrog.wu.ac.at
```

Im WU-Netz reicht:

```
ssh -X -l username balrog
```

- Ihr `username` und `password` sind Ihre Studierenden ID `hXYZXYZX` und WU-passwort.
- Sie müssen zuhause kein Linux installieren!
- Wenn Sie sich von zuhause verbinden wollen, gibt es `ssh-clients` auch für
 - Windows (<http://www.putty.org/>)
 - Mac OS X (`ssh` client built in in Terminal app)

- Die „shell“ (commandline) ist die Kommando-Zentrale von Linux, über die shell lässt sich alles steuern, ganz ohne Maus und `Klicken`:
 - Verzeichnisse auflisten: `ls`
 - Verzeichnis anzeigen, ändern: `pwd`, `cd`
 - Neues Verzeichnis anlegen: `mkdir Verzeichnisname`
 - (Text-)Datei anlegen, editieren: `pico Dateiname`
 - Datei löschen: `rm Dateiname`
 - Verzeichnis löschen: `rmdir Verzeichnisname`
 - Datei umbenennen/verschieben: `mv SOURCE TARGET`
 - Datei kopieren: `cp SOURCE TARGET`
 - „Hilfe“ zu einem Befehl: `man Befehl`
 - Ausgabe eines Befehls in eine Datei „umleiten“, `>`
z.B. `ls > meinVerzeichnisInhalt.txt`
 - Command-Line History: `history` (Suche in history: CTRL-r)
 - Die folgenden Dinge sollten Sie im Normalfall nicht brauchen:
 - login shell ändern: `chsh ...` Sollte: `/bin/bash` sein
 - Zugriffsrechte auf eine Datei ändern: `chmod`
 - Betriebssystem anzeigen: `uname`

Datenorientierte Systemanalyse

- Organisation
- Die Lehrveranstaltung im Schnelldurchlauf
- **Hands-on:**
 - Zu unserem Server verbinden (`ssh`)
 - Unix/Linux basics (login-shell ändern, files anlegen, löschen, editieren: `ls`, `cd`, `mkdir`, `pico`, `rm`, `rmdir`; command-line history: CTRL-r)
 - **Anlegen einer einfachen Web-Seite (`html`) im `www` Verzeichniss**
 - Anlegen von Tabellen in Ihrer Datenbank in PostgreSQL (`psql`)

Anlegen einer einfachen Webseite

- `(cd)`
- `(mkdir www)`
- `cd www` ... Ihr `www` Verzeichnis wird am Webserver automatisch unter <http://balrog.wu.ac.at/~username/> angezeigt

Damit etwas angezeigt wird, muss es eine Datei mit dem Namen `index.html` in diesem Verzeichnis geben, legen wir diese an:

- `pico index.html`

Kleines Linux Tutorial

- Tutorial für Linux
 - z.B. <http://mitloehner.net/lehre/linux/linux.pdf>
 - Zahlreiche andere gute Tutorials im Netz, z.B.
 - http://linuxcommand.org/learning_the_shell.php
- Tutorial für einfaches HTML, z.B.:
 - <http://de.selfhtml.org/>

PostgreSQL

- Postgres über die Commandline aufrufen:
 - `psql`
- Wichtige commands:
 - `\? ... help`
 - `\d ... List all tables`
 - `\q ... Quit (oder CTRL-D)`
- Ein Backup Ihrer Datenbank anlegen:
 - `pg_dump > backupfile.sql`
- Andere PostgreSQL Kommandos & Einführung in SQL:
<http://mitloehner.net/lehre/sql/all.html>
... und im PostgreSQL Manual:
<http://www.postgresql.org/docs/manuals/>
- Viele andere Quellen & Tutorials online!

Hausübung:

- 1) Machen Sie sich mit der Arbeitsumgebung (Linux) vertraut, lesen Sie `man` pages, oder Linux-Tutorials. Sie sollten mindestens einen neuen, nützlichen Befehl oder Parameter eines Befehls erklären können, den wir nicht besprochen hatten.
 - 2) Legen Sie eine einfache Webseite an (`index.html`) die auf eine andere Webseite verlinkt (`Anchortext`)
 - 3) Postgres: Lesen Sie das Postgres-Tutorial unter <http://mitloehner.net/lehre/sql/all.html> bis zum Kapitel „**Tabellen Löschen**“
 - 4) Legen Sie mindestens zwei verschiedene Tabellen mit mindestens je 3 Attributen an.
 - Befüllen Sie jede Tabelle mit mindestens 10 Tupeln
 - Ändern/Löschen Sie bestehende Datensätze
 - ... Dokumentieren Sie die verwendeten SQL-Komandos in einer Textdatei (Hinweis: Backup-Option im SQL Tutorial! `pg_dump`) ... Sie sollen alle Schritte am Rechner vorzeigen können!
- Punkte 1) & 2) sind bis Mittwoch zu lösen, für 3)+4) letzte mögliche Abgabe: Sonntag 11/05/2014, 20:00
 - Nächstes Mal: More on Postgres & SQL Tabellen richtig strukturieren, ER-Model, Abfragen