Technische Universität Dresden
Fakultät Informatik

Wikidata Toolkit

A Java library for working with Wikidata

Markus Krötzsch

TU Dresden

Preface

These slides were used in a tutorial given at the Web Intelligence Summer School in St. Etienne in **August 2014**. You can find up-to-date documentation on Wikidata Toolkit on the Web: software will change, these slides will be outdated at some point.

To follow the hands-on session, students received a data file (Wikidata in mapdb format) offline before the tutorial.

If you want to reenact the hands-on, you can generate this binary file using the Java application "CreateDbExample" that is found inside the examples package in the code branch used in the tutorial.

WIKIDATA

Wikidata is a (Media)Wiki

- Wikidata runs on MediaWiki
  → Content organised in pages: text pages/data pages

- Each data page is about one **entity**
  - Two types of entity: **property** and **item**
  - Identified by opaque ids, such as Q42 or P31

- All data editing happens through form-based UI

Wikidata Toolkit

Wikidata Toolkit

- Java library for working with Wikidata
  https://www.mediawiki.org/wiki/Wikidata_Toolkit
- Goal: Support programming with Wikidata content
  - Provide access to **all** of the data
  - Facilitate **fast processing/analysis**
  - Support **arbitrary Wikibase** sites
- Supported by WMF Individual Engagement Grant

Project Status of Wikidata Toolkit

- Ongoing project (final report 15 Sept)
- Done:
  - Full implementation of Wikibase data model in Java
  - Processing MediaWiki dump files to extract data
  - Downloading current Wikimedia dumps
  - Export to other formats (RDF, JSON)
- Todo:
  - Support new JSON format
  - Local storage and query
  - API access (maybe)

Data Model

The Content of Wikidata

Douglas Adams (Q42) [edit]

English writer and humorist [edit]

Also known as: Douglas Noël Adams | Douglas Noel Adams | DNA | Bop Ad [edit]

date of birth	11 March 1952 ▸ 1 reference	[edit]

Wikipedia pages linked to this item (64 entries)

Language	Code	Linked page	
العربية	arwiki	دوغلاس آدمز	[edit]
مصرى	arzwiki	دوجلاس ادامز	[edit]
Boarisch	barwiki	Douglas Adams	[edit]
беларуская	be_x_oldwiki	Дуглас Адамз	[edit]

Terms and Languages

- Three kinds of terms: labels, descriptions, aliases
- Used for labelling and searching
  - Label-description pair globally unique (key)
- Term: string in a language (“monolingual text value”)
- Over 350 languages
  - Based on MediaWiki user language (UI language)
  - Different from Wikipedia languages (<300)

Site Links

Wikipedia pages linked to this item (64 entries)

Language	Code	Linked page	
العربية	arwiki	دوغلاس آدمز	[edit]
مصرى	arzwiki	دوجلاس ادامز	[edit]
Boarisch	barwiki	Douglas Adams	[edit]
беларуская	be_x_oldwiki	Дуглас Адамз	[edit]

- Links to other Wikimedia projects
- Used in all Wikipedias to create language links
- Site links as keys:
  - At most one link per project (functional)
  - At most one item per site link (inverse functional)

Statements

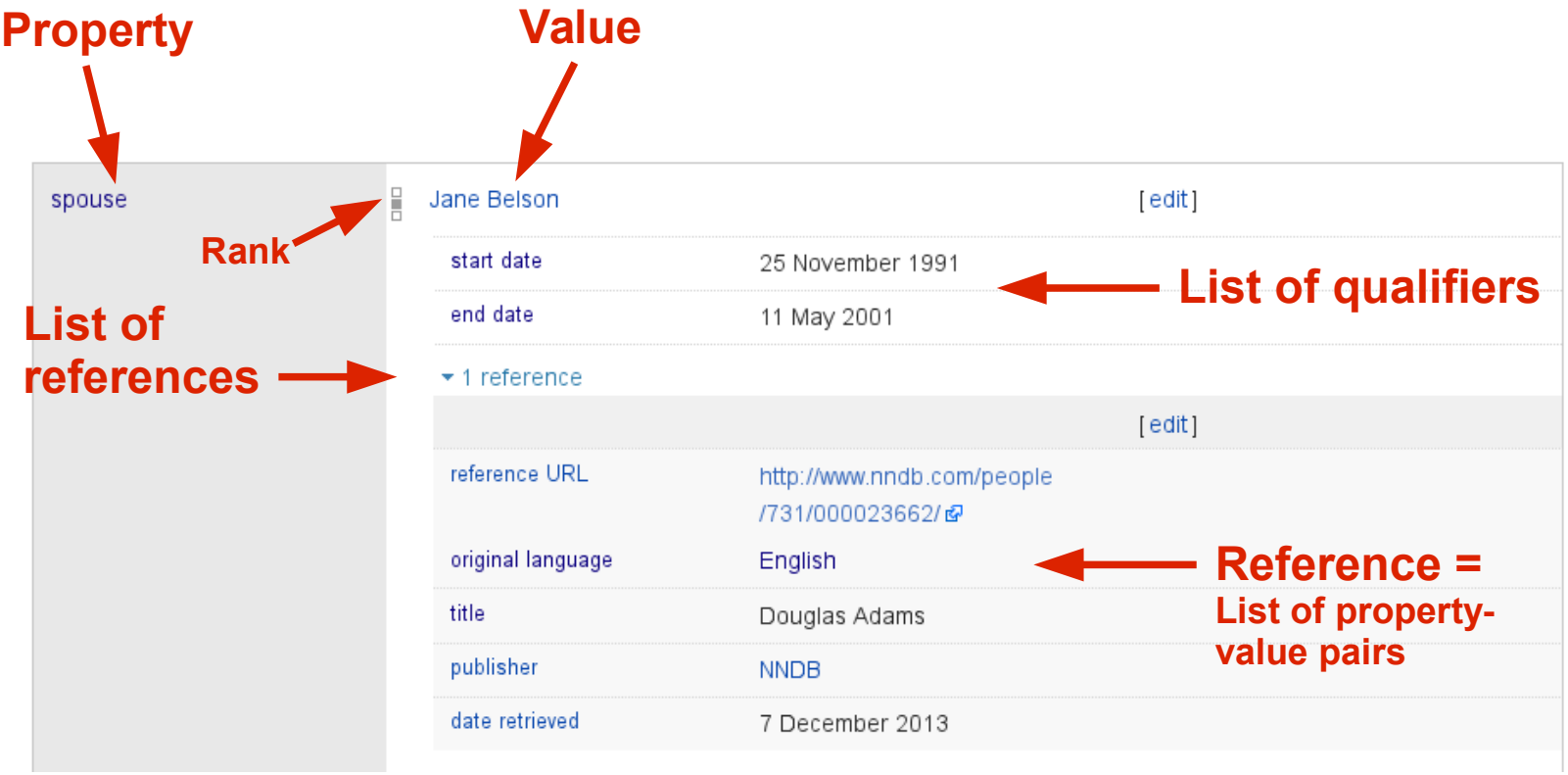
- The richest part of Wikidata's data

Statements

- The richest part of Wikidata's data

Statements

- The richest part of Wikidata's data



Statements

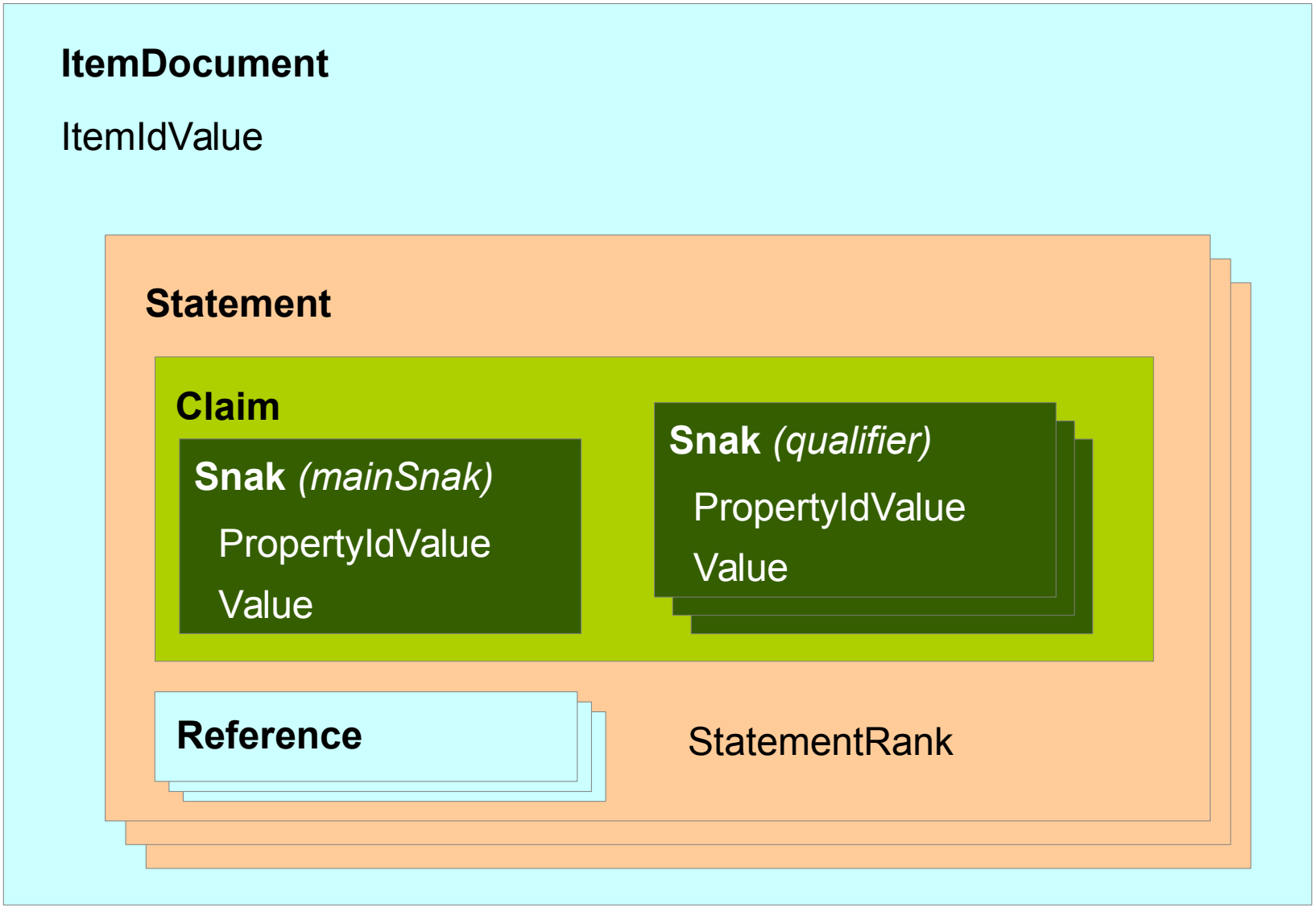
- The richest part of Wikidata's data
- Components of a statement:
  - Main property-value pair
  - List of qualifiers (property-value pairs)
  - List of references (each a list of property-value pairs)
  - Rank (preferred > normal > deprecated)
- Main property-value pair + qualifiers
  = claim (of the statement)

Property-value pairs and “Snaks”

- Properties have datatypes
  - Datatype fixed after creation
- Datatypes: Item, String, URL, CommonsMedia, Time, Globe Coordinates, Quantity
- Two special “values”:
  - *Some:* “there is a value” (that's all we can say)
  - *None:* “there is no value” (basic negative information)

→ Can be used in all places where real values can

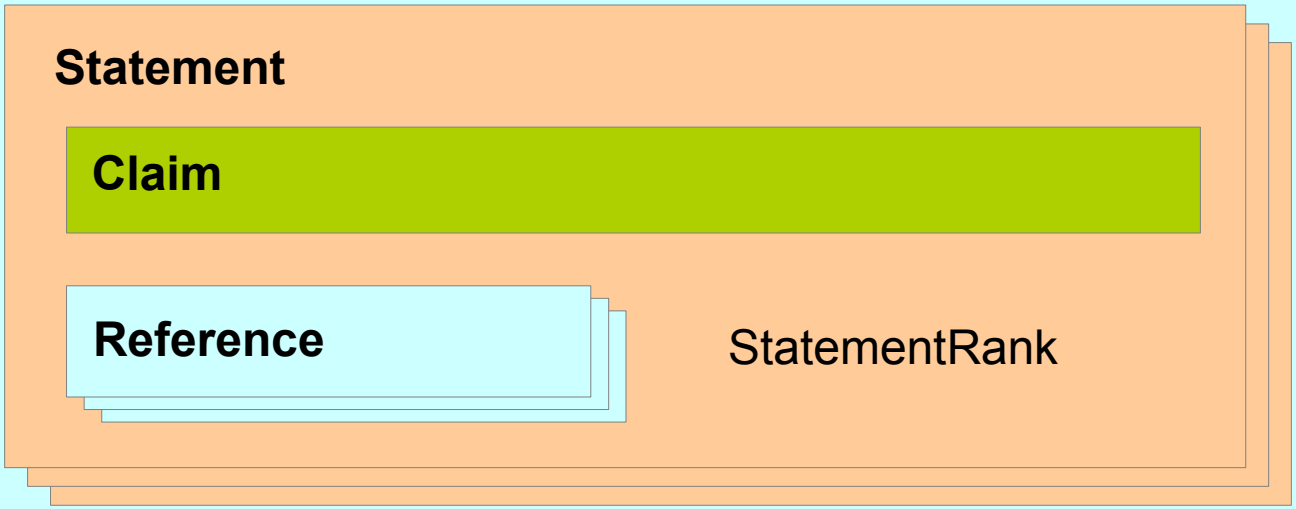
Statements


ItemDocument
ItemIdValue
Statement
Claim
Snak (mainSnak)
PropertyIdValue
Value
Snak (qualifier)
PropertyIdValue
Value
Reference
StatementRank

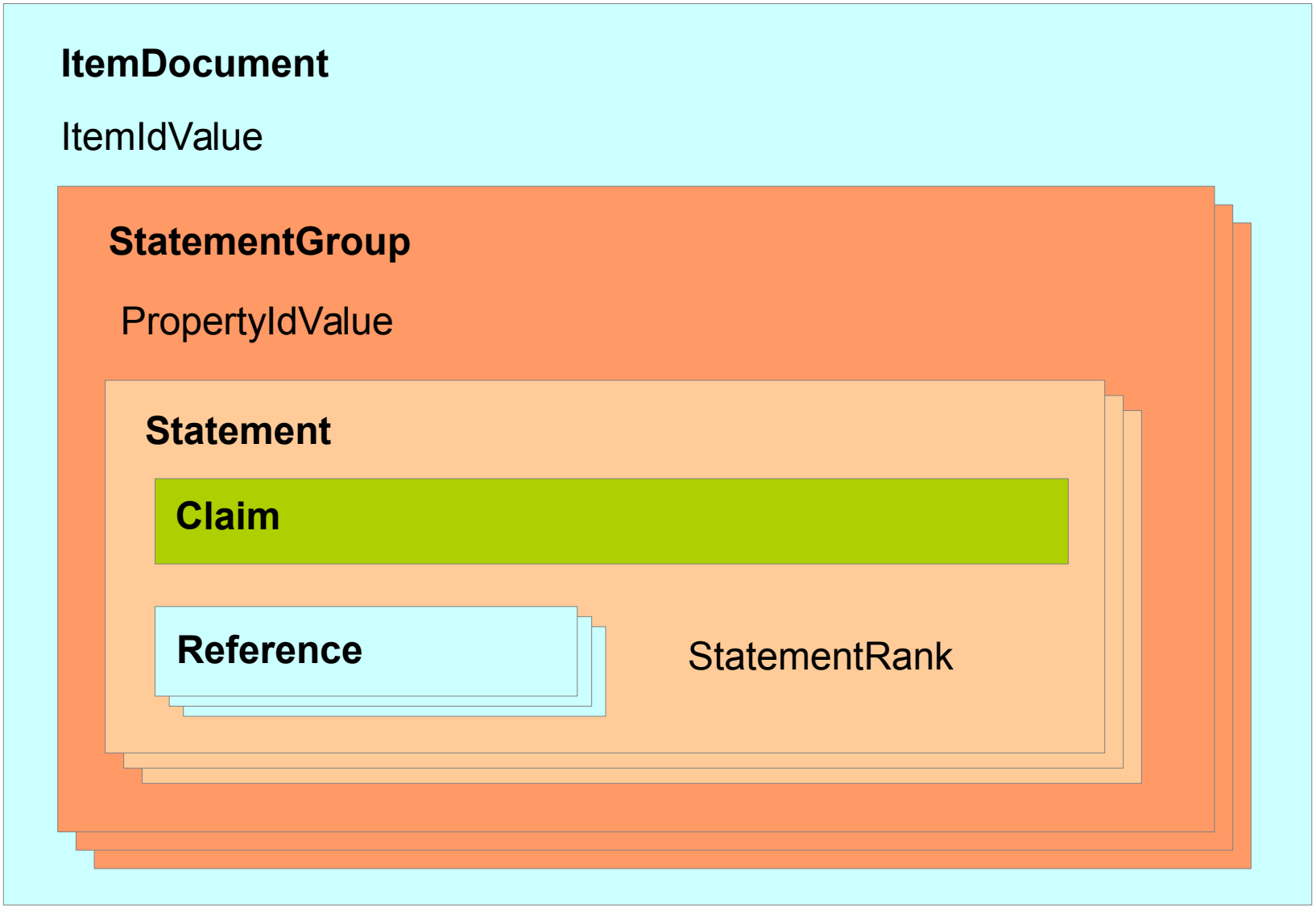
Statements

ItemDocument

ItemIdValue



Statements


ItemDocument
ItemIdValue
StatementGroup
PropertyIdValue
Statement
Claim
Reference
StatementRank

Qualifiers (data as of May 2014)

- Relatively rare: *136,187* statements use qualifiers
- Very diverse applications:
  - Temporal context (start date/end date)
  - Other context (e.g., *taxon author*, *asteroid taxonomy*)
  - N-ary relations (e.g., login for *web account on*)
  - Mixture (e.g., *character role* for *cast member*)
- Important: leaving away qualifiers may lead to a wrong claim (rather than just a "weaker" claim).

Classification (data as of May 2014)

- Properties *subclass of* (P279) and *instance of* (P31)
  - P31 is the most used property on Wikidata
- Often (but not always) used without qualifiers
- Interesting class hierarchy:
  - Entities used as classes: 41,868
  - Subclass of: 40,192 (without qualifiers)
  - Instance of: 6,169,821(without qualifiers)
- RDF/OWL file export at: http://tools.wmflabs.org/wikidata-exports/rdf/

Hands-on Session

What you will need for the hands-on session

- Hardware:
  - A few Gb of free disk space (at least 6Gb)
  - More RAM is better, but 1G should work.
- Software:
  - Java 7 (64bit is preferred)
  - Eclipse IDE with m2e (Maven) and git
- Data:
  - Wikidata binary file; decompressed

Setup instructions

For installing Eclipse properly, please follow

https://www.mediawiki.org/wiki/Wikidata_Toolkit/Eclipse_setup

Or search the Web for “Wikidata Toolkit Eclipse setup”.

Get Wikidata Toolkit from git

- Get the repository at

  https://github.com/Wikidata/Wikidata-Toolkit.git

- In Eclipse:
  - Select File->Import
  - Choose Maven->Check out Maven projects from SCM
  - Select "git" as the SCM (available after installing the connector)
  - Enter the repository URL
  - Finish

Get Wikidata Toolkit from git

- Switch to branch “db”
- How to do this in Eclipse:
  - Right-click on project explorer (on the left)
  - Team → Show in Repositories View
  - In Repositories View: Branches → Remote Tracking
  - Find branch “db”
  - Right-click on branch → Checkout
  - Confirm that you want to create a local branch too

This worked at the time of the tutorial, on 27 Aug 2014.
If you want to try this later on, you could find the according revision, or (better) look for more recent tutorials.

Prepare to run

- Build the project with Maven:
  - Right-click on module “parent”
  - Run as → Maven install
  - This downloads and installs all dependencies
- Copy the data files into the directory:

  eclipse-workspace/wdtk-parent/wdtk-example

Run

- Run tutorial example:
  - Open module “wdtk-examples”
  - Open package org.wikidata.wdtk.examples
  - Run “TutorialExampleDb.java” (as Java application)

Exercises

- Open the file TutorialDocumentProcessor.java
- The class documentation has some exercise descriptions

- Hints:
  - View pretty Javadoc output in Eclipse (tab “Javadoc”)
  - Use Ctrl+C and Ctrl+V on the project explorer to make copies of files (one for each exercise)

The tutorial continued by solving some of the exercises in the file, esp. the computation of the average life expectancy of people on Wikidata.
Code written at that time is found online.

Conclusions

Conclusions

- Wikidata is …
  - … fascinating and unpredictable
  - … full of unexplored potential
  - … only at its beginning
- Wikidata Toolkit gives you full access to **all** of Wikidata
  - For creating your own excerpts of the data
  - For aggregation and anaylisis
  - For high-speed, random, offline data access
- WDTK works with any other Wikibase installation

Further reading

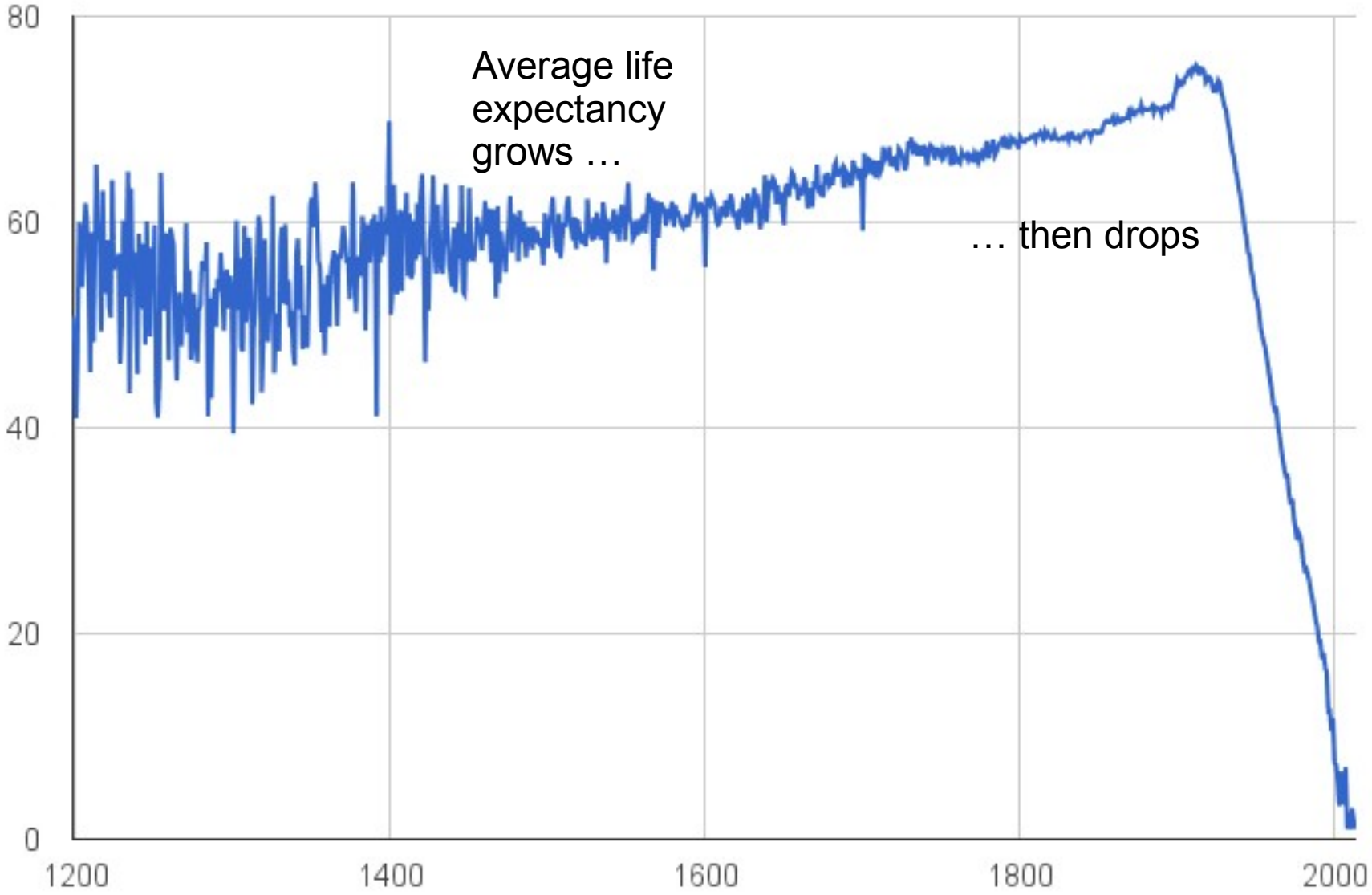
- Denny Vrandecic, Markus Krötzsch.
  Wikidata: A Free Collaborative Knowledge Base. CACM 2014. To appear
  → *general first introduction to Wikidata*

- Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, Denny Vrandečić.
  Introducing Wikidata to the Linked Data Web. 2014.
  → *introduction of the Wikidata RDF export and data model*

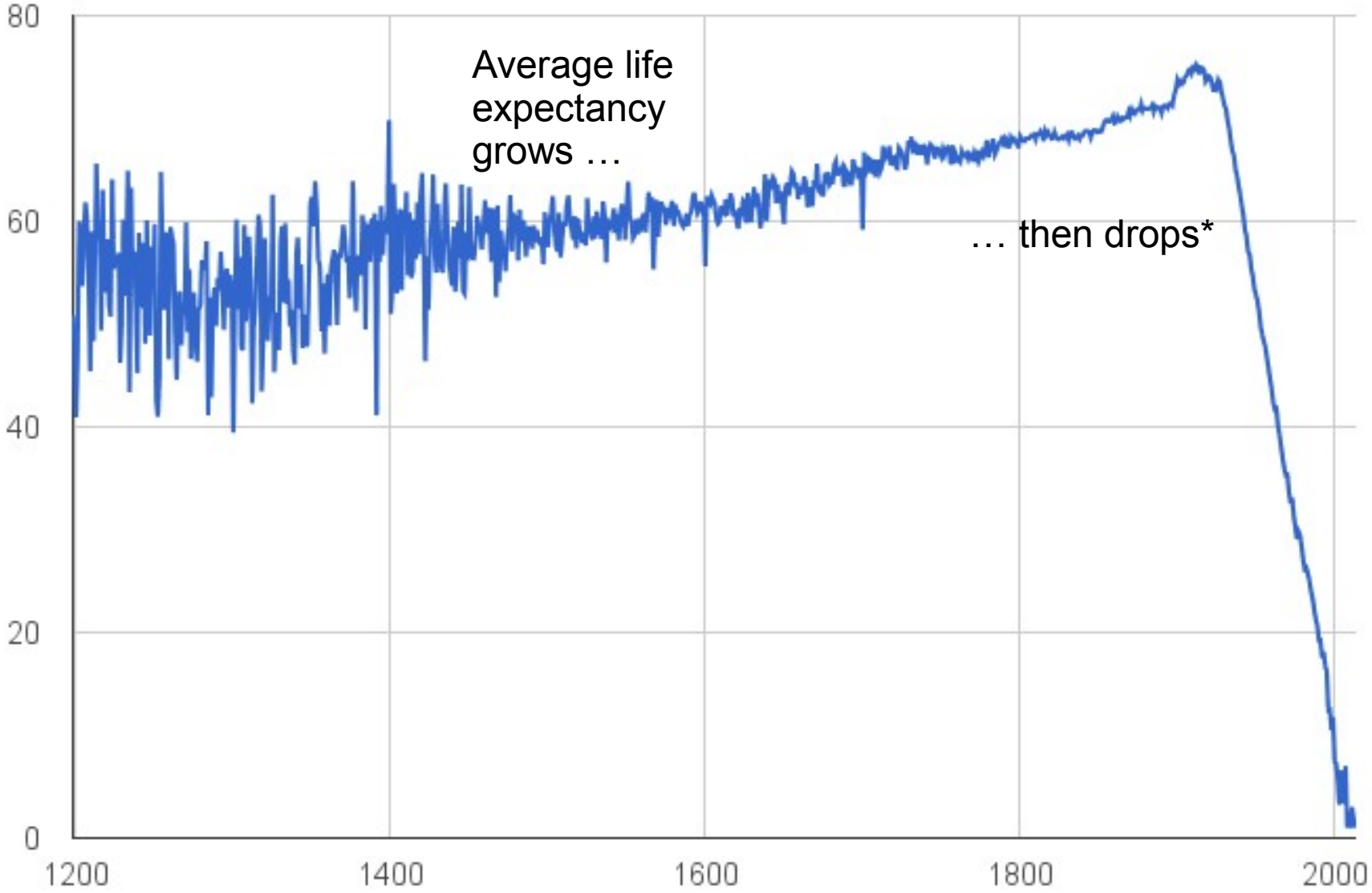
Outcome of Exercise 6

(Life expectancy by year of birth)

We will all die!



We will all die!



*) obviously, this must be so if we take the life expectancy of people dead already