Informatik Praktikum (Gr. Studienprojekt) 2014/2015

Betreuer: Dr. Michael Ley

Praktikant: Serge Oliver, Djomo Moungoue
Matrikelnummer: 1050291

Validierung und statistische Auswertung von dblp.xml
Softwaredokumentation

Meilensteine:	Erster Termin	→	Softwarespezifikation	→	Softwaredokumentation
Daten:	30.10.14		27.11.14		17.03.15

Inhaltsverzeichnis

I – Einführung

Die dblp computer science bibliography ist ein online Verweis für bibliographische Informationen über meistens Publikationen im Bereich Informatik. Binnen einem halbjährlichen Studienprojekt wurde ich aufgefordert eine Software zu entwickeln, welche die zugrunde liegende Datei der dblp *(dblp.xml)* validiert und statistisch auswertet. Die dblp.xml wurde am Mittwoch, dem 12. November 2014 von der dblp Webseite auf meinem Laptop heruntergeladen. Die war im Zeitpunkt des Downloads circa 1.4 GB groß. Die entwickelte Software speichert die generierten Daten je nach Art und Zweck in html, csv und log Dateien. Die html Dateien enthalten die Daten zur Darstellung von Säulendiagrammen und Box-Plot-Diagrammen auf die Website. Die csv und log Dateien bestehen jeweils aus den über die Häufigkeitsverteilung von Daten ermittelten Werten und einigen Ausreißern. Die Software lief circa zwei Minuten in der Entwicklungsumgebung.

II – Sofwarespezifikation

II.1 – Ausgewertete Statistiken (Darstellung als Säulendiagramm)

Zahl von

Autoren/Editoren Namen, die gleich viel Zeichen umfassen

- Titeln, die gleich viel Zeichen umfassen
- Titeln, die gleich viel Wörtern umfassen
- Querverweisen mit gleichem Umfang von Seiten
- Büchern mit gleich viel Querverweisen
- bearbeiteten Publikationen pro Monat
- bearbeiteten Publikationen pro Jahr
- Autoren/Editoren, die in demselben Jahr für das erste Mal publizierten
- generierten elektronischen Versionen von Publikationen pro Jahr
- Feldern pro Publikation

II.2 – Häufigkeitsverteilung von Daten (Darstellung als Box-Plot-Diagramm)

Häufigkeitsverteilung von

- Zeichen in Autoren/Editoren Namen
- Zeichen in Titeln
- Wörtern in Titeln
- Seiten in Querverweisen
- Querverweisen in Büchern
- bearbeiteten Publikationen pro Monat
- bearbeiteten Publikationen pro Jahr
- ersten Publikationsjahr von Autoren/Editoren
- generierten elektronischen Versionen von Publikationen pro Jahr
- Feldern pro Publikation

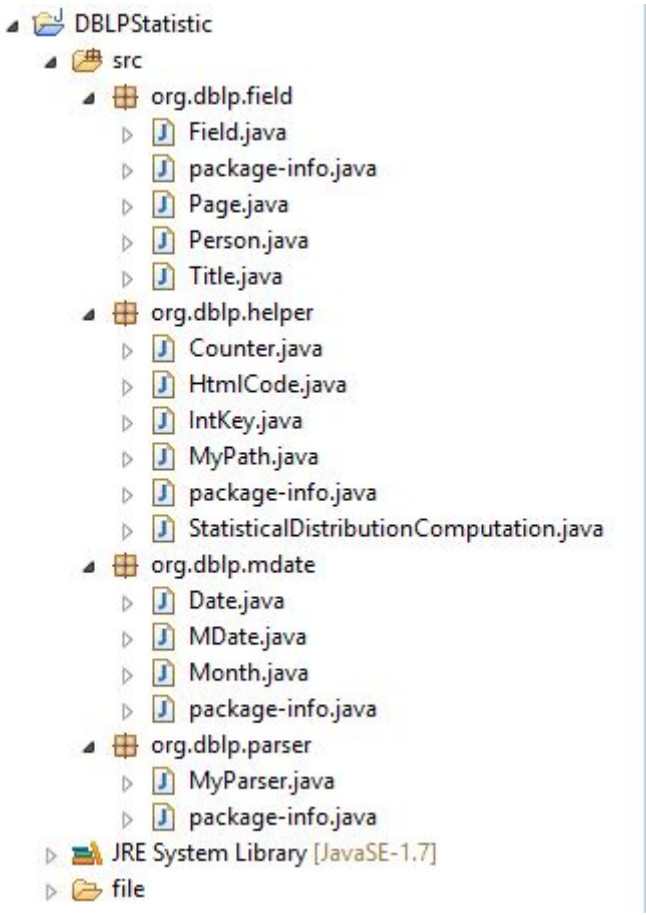
Liste von berechneten Maßen

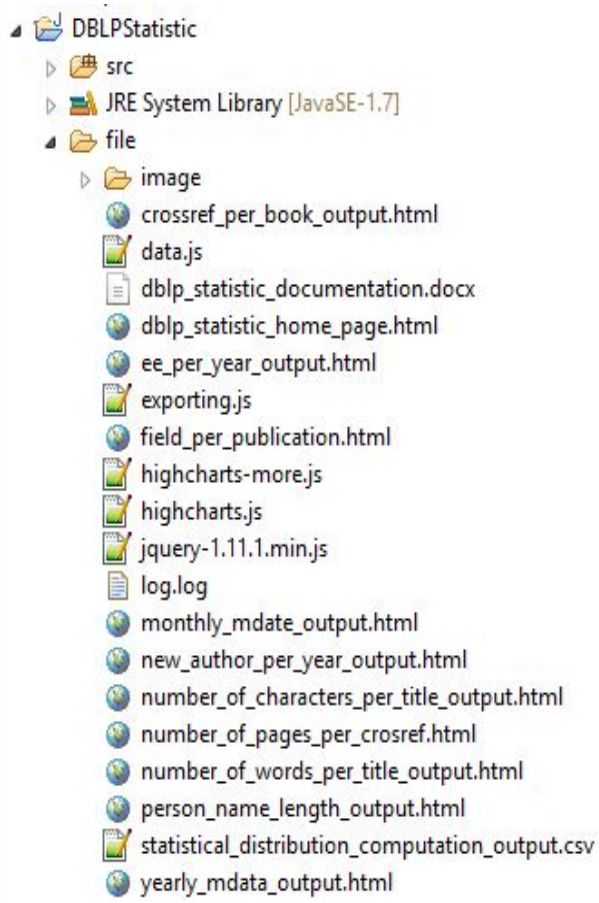
- Der obere Whisker
- Das obere Quartil
- Die Median
- Das untere Quartil
- Der untere Whisker
- Die Varianz
- Die Standard Abweichung
- Die Gesamtsumme

Ausführliche Liste in der Datei statistical_distribution_computatio_output.csv

III – Inhalt des Softwarepakets

III.1- Quellcode (136 KB) und Output-Dateien (3.44 MB)





IV – Entwicklungs- und Testumgebung

IDE	Eclipse SDK Version 3.8.0
VM Arguments	• ${-DensityExpensionLimit=2500000} • ${-Xms1024M} • ${-Xms1024M}
Externe Bibliotheken	• xerces.jar • jquery-1.11.1.min.js • highcharts.js • highcharts-more.js • data.js • exporting.js
Browser	Mozilla Firefox 36.0.1

Betriebssystem	Microsoft Windows 8

V – Veranschaulichung von Endergebnissen

V.1 – Startseite mit der Liste von Statistiken

DBLP Statistic - University of Trier

Statistics Summary

- Number of books which have the same number of cross references from 1 to 150
- Number of electronic versions made the same year
- Number of publication which have the same number of fields from 1 to 50
- Monthly modification frequence of publications
- Number of authors/editors which made their first publication in the same year
- Number of titles which have the same number of characters from 1 to 150
- Number of cross references which have the same number of pages from 1 to 100
- Number of titles which have the same number of words from 1 to 50
- Number of author/editror names which have the same number of characters
- Yearly modification frequence of publications

Please do not hesitate to contact us if you miss certain statistics.

Abbildung 1 *Startsete, die die Liste von Statistiken enthält*

V.2 –Darstellung von Säulendiagramm und Box-Plot-Diagramm

Säulendiagram

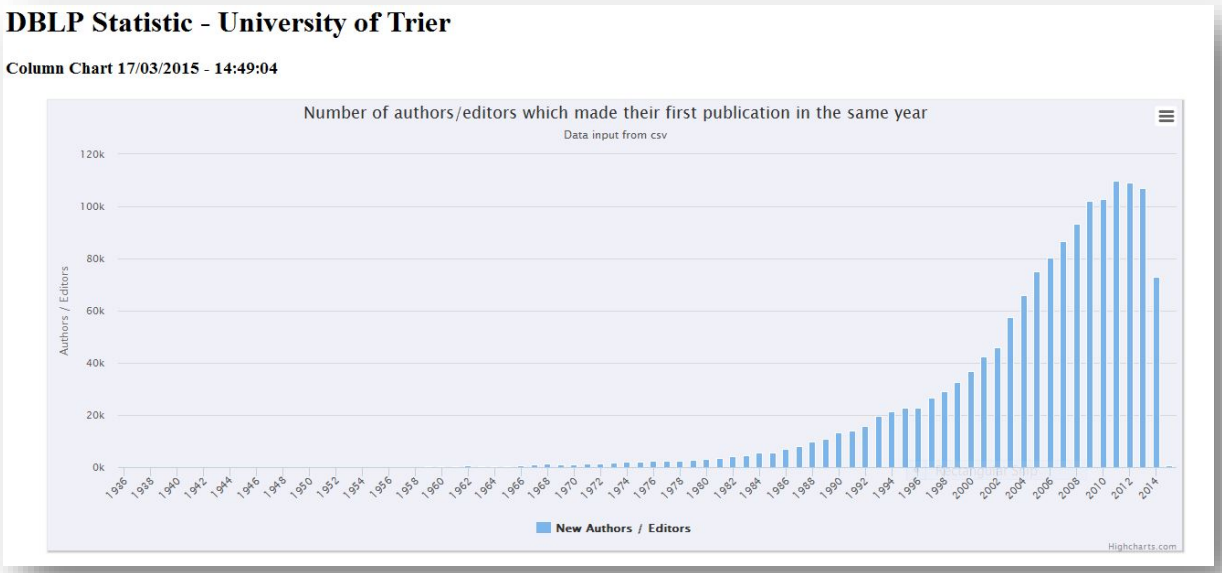


Abbildung 2 *Anzahl von Autoren/Editoren die publizierten für das erste Mal demselben Jahr*

Box Plot-Diagramm

Abbildung 3 *Fünf-Nummer Verteile über die neuen Autoren/Editoren pro Jahr*

V.3 – UML- Paketdiagramm

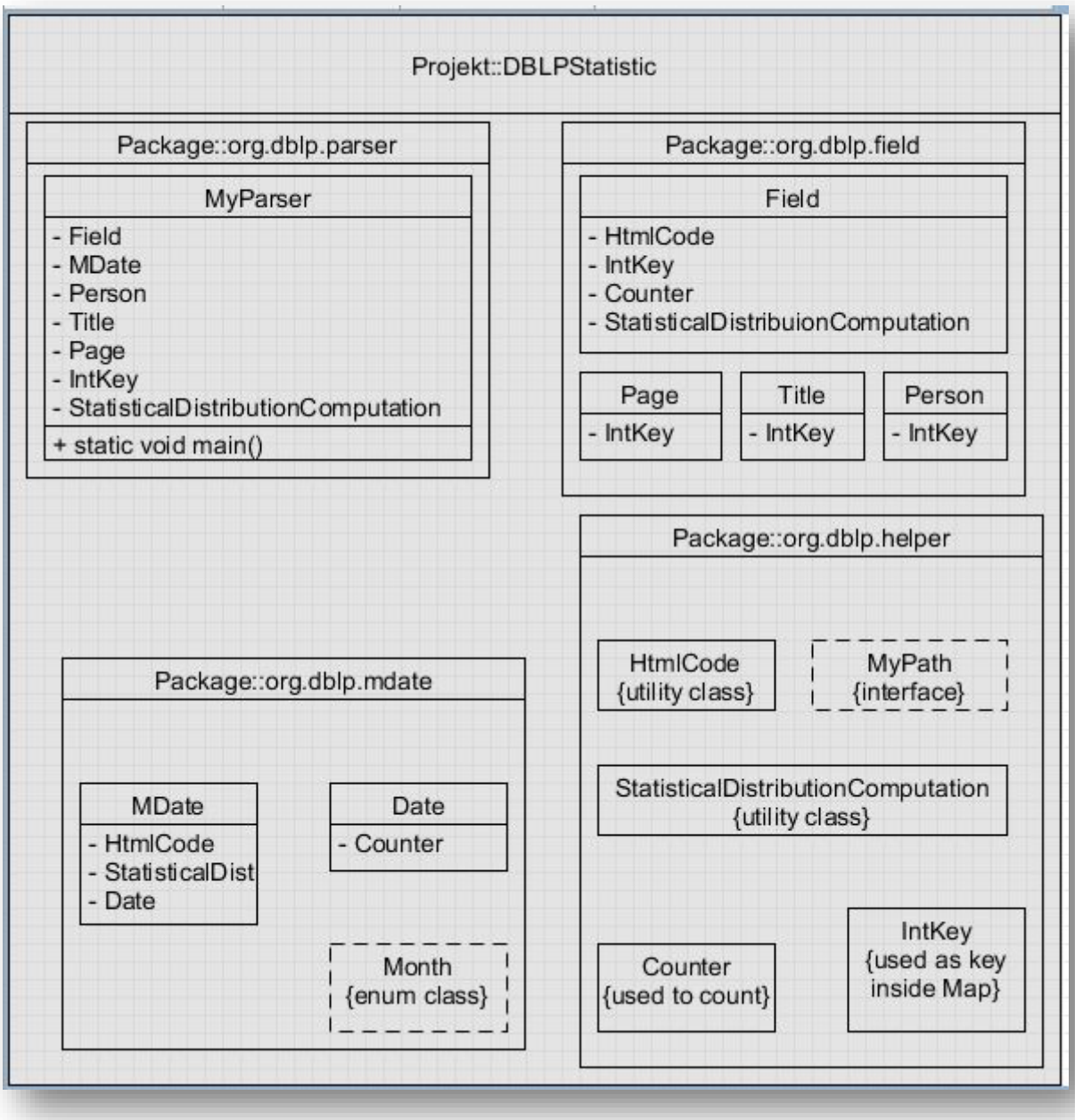


Abbildung 4 *Grobe Darstellung der Inhalt des Projektpaket*

V.4 – Liste von Ausreißern

```
Beispiel Auszug aus der Datei von Ausreißern (log.log)
[Length: 1] Title: . [Key: conf/cisis/Schatten09]
[Length: 1] Title: C [Key: books/bi/Mock1990]
[Length: 2] Title: 1. [Key: conf/aina/EvansR13]
[Length: 2] Title: A. [Key: conf/b/Groslambert07a]
[Length: 2] Title: ?. [Key: journals/pik/Potton06b]
[Extra White Space: 6] Title: The generalized bisymmetric solutions of the matrix equation A 1 X 1 B 1 + A 2 X 2 B 2 + + A  l  X  l  B  l  = C

[Page: (in the table of contents only)] [Key: conf/vldb/MahmoudR75]
[Page: 0-] [Key: books/crc/IIR2005/BertozziBM05]
[Page: 00:1-00:2] [Key: journals/lites/Burns14]
[Page: 1, 4-5] [Key: journals/ieeemm/Titsworth06]
[Page: 1-0] [Key: journals/cm/YuanZGCLBCS14]
[Page: 1-ix] [Key: journals/pvldb/JagadishZ13]
[Page: 1/2] [Key: journals/tsmc/MessingerRH91]
[Page: 10811084-] [Key: conf/igarss/SinghVKRM08]
[Page: 102-019] [Key: conf/gcc/NiuCZ06]
[Page: A1-A13] [Key: journals/jat/SommerS14]
[Page: A10] [Key: journals/bmcbi/AzizSR11]
[Page: ASMD1-ASMD6] [Key: journals/datascience/RumbleF12]
[Page: C-15-C-18] [Key: conf/compgeom/KedemY96]
[Page: I, 1-113] [Key: books/daglib/0020524]
[Page: I-XXXVIII, 1-774] [Key: books/daglib/0022095]
[Page: I-X] [Key: books/daglib/0016337]
[Page: O10] [Key: journals/bmcbi/MagarinosOCSDRCHNBRVA10]
[Page: P1.9] [Key: journals/combinatorics/CurrieS14]
[Page: P10] [Key: journals/bmcbi/DuncanPZ10]
[Page: S108-S112] [Key: journals/datascience/Peterson09]
```

Abbildung 5 *Auszug aus der log Datei*

Manche Publikationen haben eine Groß Zahl von Felder wegen zu viele *cite* Felder

Zum Beispiel: Book key: books/aw/AbiteboulHV95 enthält 751 Felder (circa 740 *cite* Felder)

Ausführliche Liste in der Datei *log.log*

V.5 – Baumstruktur von dblp.xml

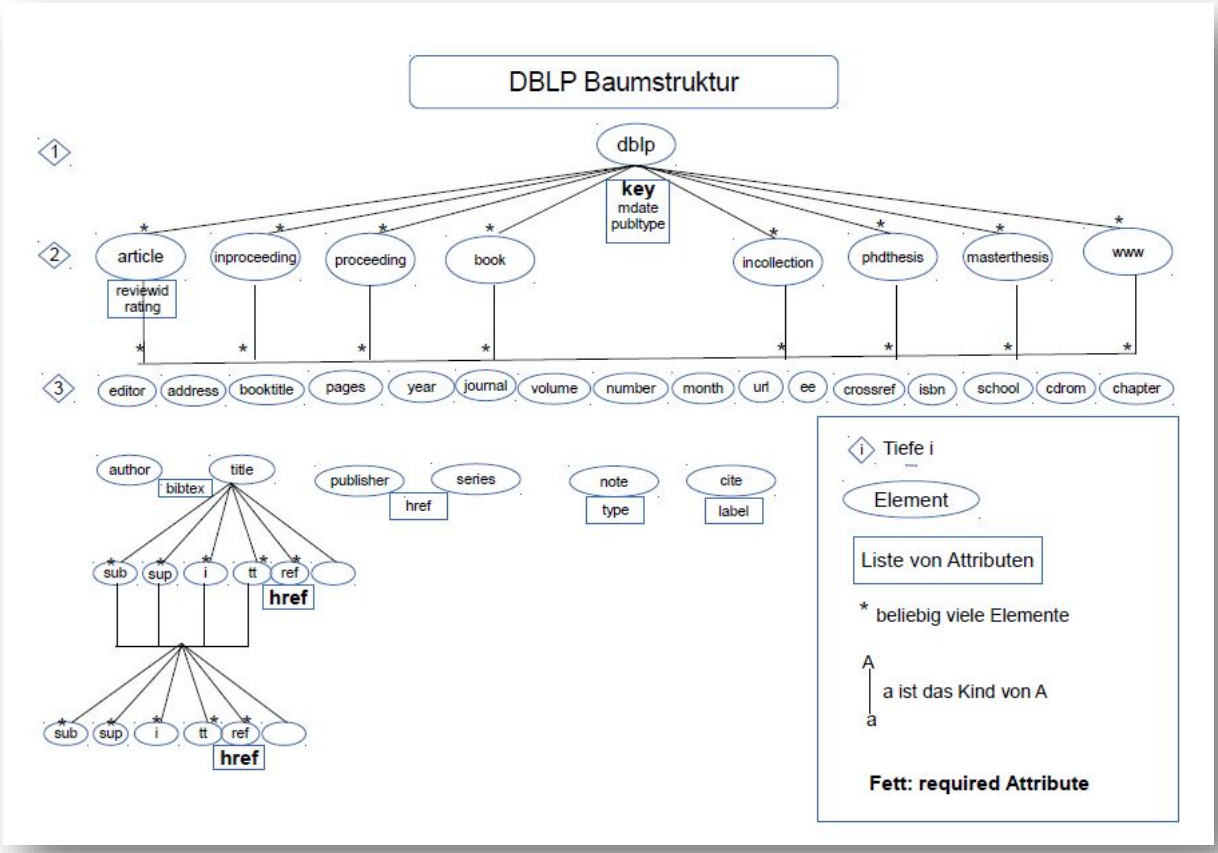


Abbildung 6 *Baumstruktur der dblp.xml*

VI – Schluss

Die Software ist laut die durchgeführten Tests zuverlässig und hat eine Laufzeit von circa zwei Minuten auf meinem Laptop. Alle Diagramme bis auf das Box-Plot-Diagramm der Statistik über Felder in Publikationen lassen sich gut darstellen. Ich habe die Ursache bislang nicht herausfinden können.