

Workshop zu
Datenqualität

City Lab Berlin

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Workshop zu Datenqualität

City Lab Berlin

Datum (TBD)

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Einführung

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

• Theorie

- Was ist Daten-Qualität?
- Relevanz von Daten-Qualität

• Anwendung

- Excel
- CSV
- Andere Formate
- Formatentscheidung
- Doppelcheckliste

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

1. Theorie

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Was ist Datenqualität?

Datenqualität definieren

- *FAIR-Prinzip*

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

Datenqualität definieren

• **Findable:** *Wo finde ich die Daten?*

- Datensatz registriert oder indiziert
- Dauerhafte Identifikationsnummer
- Ausführliche Beschreibung

Quelle: Wilkinson, M. D. et al. (2016)

Datenqualität definieren

- **Accessible:** *Wie kann ich die Daten zugreifen?*
 - Klar und kostenlosen Zugriffsprotokoll
 - *Metadaten* (Beschreibung der Daten) und *Daten* (Inhalt) sind abrufbar
 - Metadaten sind **immer** verfügbar, auch wenn die Daten nicht mehr abrufbar sind

Quelle: Wilkinson, M. D. et al. (2016)

Datenqualität definieren

- **Interoperable:** *Wie kann ich die Daten verwenden?*

- Formale und zugängliche Sprache
- Verweise auf andere Daten
- Ermöglichte Integration mit anderen Datensätzen

Datenqualität definieren

- **Reusable:** *Wie kann ich die Daten replizieren?*
 - Detaillierter Herkunft
 - Klare und zugängliche Datennutzungslizenz

Quelle: Wilkinson, M. D. et al. (2016)

Datenqualität definieren

• 5-Sterne-Modell

• Kaskadierend Modell (von 1 bis 5 Sternen)

Das 5-Sterne Modell ist wie folgt definiert:

- ★ Stellen Sie Daten im Web unter einer offenen Lizenz bereit, das Datenformat für die Bereitstellung ist Ihnen überlassen.
- ★★ Stellen Sie Daten in einem strukturierten Format bereit.
- ★★★ Verwenden Sie offene, nicht proprietäre Formate.
- ★★★★ Verwenden Sie URIs, um Dinge zu bezeichnen, damit Daten verlinkt werden können.
- ★★★★★ Verlinken Sie Ihre Daten mit anderen Daten, um Kontexte herzustellen.



Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Datenqualität definieren

Daten im Web (ex.: PDF)



Daten in strukturiertem Format (ex.: XLS)



Daten in strukturiertem, nicht proprietärem
Format (ex.: CSV)

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Datenqualität definieren



Verwendung von eindeutigen URLs (ex.: RDF)



Verlinkung der eigenen Daten mit anderen
Daten (ex.: LOD)

Datenqualität definieren

- Liste von Merkmalen der Datenqualität:

- Aktualität
- Fehlerfreiheit
- Genauigkeit
- Konformität
- Konsistenz
- Transparenz & Vertrauenswürdigkeit
- Verlässlichkeit
- Verständlichkeit
- Vollständigkeit
- Zugänglichkeit & Verfügbarkeit

Datenqualität definieren

- Liste von Merkmalen der Datenqualität:

- Aktualität
- Fehlerfreiheit
- Genauigkeit
- **Konformität**
- Konsistenz
- Transparenz & Vertrauenswürdigkeit
- Verlässlichkeit
- **Verständlichkeit**
- Vollständigkeit
- Zugänglichkeit & Verfügbarkeit

Datenqualität definieren

- Bei diesem Workshop, werden wir vor allem auf **Konformität & Verständlichkeit** fokussieren
- Die anderen Merkmale sind natürlich auch wichtig, sind aber sehr von einzelnen Datensätzen/Kontexten abhängig

Datenqualität definieren

• Diskussion

- Die Datenqualität kann anhand verschiedener Standards definiert werden
- Metadaten und Daten müssen beide optimisiert sein
- Es gibt wichtige Prinzipien und Leitlinien... aber manchmal sind die nicht spezifisch genug!

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Relevanz von Datenqualität

Garbage in, garbage out

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

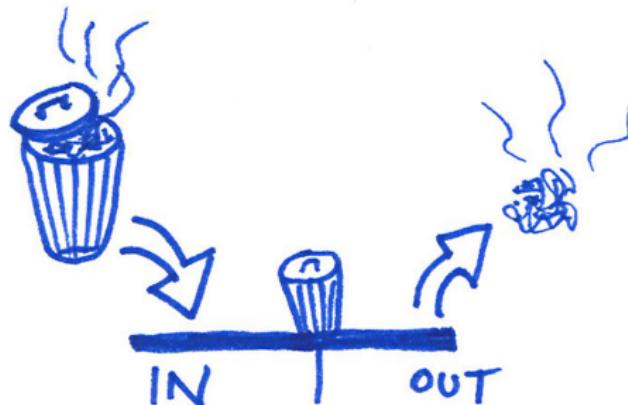
Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- “Wenn die Daten schlecht sind, ist die Analyse nutzlos”



Schritte der Datenqualität

- Drei Schritte, von denen die Datenqualität abhängt:
 - ① **Data Generating Process**
 - Wie werden die Daten gesammelt?
 - ② **Data Cleaning**
 - Wie sind die Datensatzdatei erstellt?
 - ③ **Data Publishing**
 - Wie ist der Datensatz veröffentlicht?
- Diese Schritte sind voneinander abhängig

Workshop zu Datenqualität

City Lab Berlin

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Quiz

Datensatz von Berlin Open Data

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- **Schulbausanierung Sommer 2020**

- [https://daten.berlin.de/datensaetze/
schulbausanierung-sommer-2020](https://daten.berlin.de/datensaetze/schulbausanierung-sommer-2020)

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel



Datensätze	Schulbausanierung Sommer 2020
Arbeitsmarkt	Berlin investiert in den kommenden Jahren 5,5 Milliarden Euro in die Sanierung und in den Bau von Schulgebäuden. Auch in den Sommerferien 2020 gehen an den Berliner Schulen die Bau- und Sanierungsmaßnahmen weiter.
Bildung	
Demographie	
Geographie und Stadtplanung	
Gesundheit	
Jugend	
Kunst und Kultur	
Öffentliche Verwaltung, Haushalt und Steuern	
Protokolle und Beschlüsse	
Sonstiges	

Informationen zum Datensatz

Lizenz:	Creative Commons Namensnennung
Kategorie:	Bildung
Geographische Abdeckung:	Berlin
Geographische Granularität:	Berlin
Zeitliche Granularität:	Keine
Veröffentlicht:	27.07.2020
Aktualisiert:	27.07.2020
Veröffentlichende Stelle:	Senatsverwaltung für Bildung, Jugend und Familie
E-Mail Kontakt:	karen.koenig AT senbjf.berlin.de
Website:	https://www.berlin.de/sen/bildung/service/daten/
Tags:	Bauen Sanieren Schulbaumaßnahmen ... (2 weitere)
Kommentare:	0

Fragen

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- Wie gut ist dieser Datensatz bezüglich der FAIR-Prinzipien?
 - **Findable:** *Wo finde ich die Daten?*
 - **Accessible:** *Wie kann ich die Daten zugreifen?*
 - **Interoperable:** *Wie kann ich die Daten verwenden?*
 - **Reusable:** *Wie kann ich die Daten replizieren?*
- Wie optimal ist das Datenformat?

Analyse

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- **Findable:** *Wo finde ich die Daten?*

✓ **Datensatz registriert oder indiziert**

✗ **Dauerhafte Identifikationsnummer**

≈ **Ausführliche Beschreibung**

Analyse

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- **Accessible:** *Wie kann ich die Daten zugreifen?*

✓ Klar und kostenlosen Zugriffsprotokoll

✓ Metadaten (Beschreibung der Daten) und Daten (Inhalt) sind abrufbar

Analyse

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- **Interoperable:** Wie kann ich die Daten verwenden?

✓ Formale und zugängliche Sprache

✗ Verweise auf andere Daten

≈ Ermöglichte Integration mit anderen Datensätzen

Analyse

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- **Reusable:** *Wie kann ich die Daten replizieren?*

✖ Detaillierter Herkunft

✓ Klare und zugängliche Datennutzungs-lizenz

Analyse

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

• Datenformat:

≈ Aktuell = **Excel**



✓ Optimal = **CSV**



Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

2. Anwendung

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Excel

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Was ist eine Excel-Datei?

- Tabelle, die mit Excel bearbeitet werden kann
- Komplexes Dateiformat
- Stärker Fokus auf Human-Readability

Zu vermeiden in Excel

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- Überflüssige Text-Formatierung
 - Effiziente Kategorisierung statt komplexes Farbsystem

Zu vermeiden in Excel

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel



Institution	Preise (von/bis)	ABO**	Schüler	Studenten	Auszubildende	Bundesfreiwilligendienstleistende	Rentner
Große Sprechbühnen							
Deutsches Theater / Kammerspiele	4,00-48,00 €	40%	9,00 €	9,00 €	9,00 €	9,00 €	Nein
Volksbühne	6,00-40,00 €	25%	50%	50% Studententag 2x1	50%	50%	Nein
Maxim Gorki Theater	10,00-30,00 €	50%	8,00 €	8,00 €	8,00 €	8,00 €	Nein
Berliner Ensemble	5,00-30,00 €	6,00-18,00 €	9,00 €	9,00 €	9,00 €	9,00 €	9,00 €
Hebbel am Ufer (HAU 1, 2, 3)	10,00-30,00 €	10er Karte Tanzcard	5,00-10,00 €	5,00-10,00 €	5,00-10,00 €	5,00-10,00 €	5,00-10,00 €
Schaubühne am Lehniner Platz	7,00-47,00 €	25% Tanzcard	9,00 €	9,00 €	9,00 €	9,00 €	Nein
Renaissance Theater Neue Theater Betriebs-GmbH	10,00-48,00 €	12,00-22,00 €	So-Do 6,00 €	So-Do 6,00 €	auf Anfrage	auf Anfrage	Nein
Kinder- und Jugendtheater							
Theater an der Parkaue	13,00-14,00 €	30%	9,00-10,00 € Kinder bis 12 Jahre 7 €	9,00-10,00 €	9,00-10,00 €	9,00-10,00 €	9,00-10,00 €
Grips Theater	6,00-20,00 €	Nein	3,00-12,00 €	3,00-12,00 €	3,00-12,00 €	3,00-12,00 €	Nein
Kleine und mittlere privatrechtlich organisierte Theater, Theater-Tanzgruppen (Konzeptförderung)							
Constanza Macras / Dorkypark GmbH	variiert	Tanzcard	variiert	variiert	variiert	variiert	variiert

Quelle: Open Data Portal, "Eintrittspreisregelungen öffentlich geförderter Berliner Kultureinrichtungen"

Zu vermeiden in Excel

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel



Art	Institution	Preise (von/bis)	ABO**	Schüler	Studenten
Große Sprechbühnen	Deutsches Theater / Kammerspiele	4,00-48,00 €	40%	9,00 €	9,00 €
Große Sprechbühnen	Volksbühne	6,00-40,00 €	25%	50%	50% Studententag 2x1
Große Sprechbühnen	Maxim Gorki Theater	10,00-30,00 €	50%	8,00 €	8,00 €
Große Sprechbühnen	Berliner Ensemble	5,00-30,00 €	6,00-18,00 €	9,00 €	9,00 €
Große Sprechbühnen	Hebbel am Ufer (HAU 1, 2, 3)	10,00-30,00 €	10er Karte Tanzcard	5,00-10,00 €	5,00-10,00 €
Große Sprechbühnen	Schaubühne am Lehniner Platz	7,00-47,00 €	25% Tanzcard	9,00 €	9,00 €
Große Sprechbühnen	Renaissance Theater Neue Theater Betriebs-GmbH	10,00-48,00 €	12,00-22,00 €	So-Do 6,00 €	So-Do 6,00 €

Quelle: Open Data Portal, "Eintrittspreisregelungen öffentlich geförderter Berliner Kultureinrichtungen"

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel



Zu vermeiden in Excel

• Aktive Formeln

- Daten sollten nicht versehentlich vom Benutzer geändert werden

A	B	C	D	E	F
1	Country	Population	Average Population	Greater than Average?	
2	China	1,389,618,778	435,810,199	TRUE	
3	India	1,311,559,204	435,810,199	TRUE	
4	USA	331,883,986	435,810,199	FALSE	
5	Indonesia	264,935,824	435,810,199	FALSE	
6	Pakistan	210,797,836	435,810,199	FALSE	
7	Brazil	210,301,591	435,810,199	FALSE	
8	Nigeria	208,679,114	435,810,199	FALSE	
9	Bangladesh	161,062,905	435,810,199	FALSE	
10	Russia	141,944,641	435,810,199	FALSE	
11	Mexico	127,318,112	435,810,199	FALSE	
12					
13					
14					

Auswertung

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- **Vorteil:** bessere Darstellungsmöglichkeiten
- **Nachteil:** kann schnell nicht-maschinenlesbar werden

Workshop zu Datenqualität

City Lab Berlin

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

CSV

Was ist ein CSV?

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- Text-Datei
- Einfaches Dateiformat
- Werte werden mit einem Trennzeichen abgetrennt
- Stärker Fokus auf Machine-Readability

Zu vermeiden in CSV

• Trennzeichen

- Verwendung von Trennzeichen an anderen Stellen (ex.: "10.5" ist besser als "10,5")



	A	B	C	D	E
1	id	altersgruppe	fallzahl	differenz	inzidenz
2	1	0-4	319	2	168,1
3	4	5-9	352	1	205
4	7	10-14	438	4	279,9
5	10	15-19	497	3	331,7
6	13	20-24	969	18	474,3
7	16	25-29	1235	14	430,1
8	19	30-39	2343	22	359,8
9	22	40-49	1701	5	351,3
10	25	50-59	1553	5	285,9

Quelle: Open Data Portal, "COVID-19 Fälle im Land Berlin, Verteilung nach Altersgruppen" (**verändert**)

Zu vermeiden in CSV

• Trennzeichen

- Verwendung von Trennzeichen an anderen Stellen (ex.: "10.5" ist besser als "10,5")



A	B	C	D	E
1	id	altersgruppe	fallzahl	differenz
2	1	0-4	319	2
3	4	5-9	352	1
4	7	10-14	438	4
5	10	15-19	497	3
6	13	20-24	969	18
7	16	25-29	1235	14
8	19	30-39	2343	22
9	22	40-49	1701	5
10	25	50-59	1553	5
				285.9

Quelle: Open Data Portal, "COVID-19 Fälle im Land Berlin, Verteilung nach Altersgruppen"

Zu vermeiden in CSV

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel



- Text-Formatierung

- Wird eh nicht angezeigt

	A	B	C	D
1	id	ssflag	pagenumberinocrdocument	authorfirstname
2	70526	2		0 Adler
3	70527	2		2 Bruno
4	70528	2		2 Felix
5	70529	2		2 Friedrich (Wolfgang)
6	70530	2		2 Georg
7	70531	2		2 Max
8	70532	2		2 Otto
9	70533	2		2 Viktor

Quelle: Open Data Portal, "Liste der verbotenen Bücher" (**verändert**)

Zu vermeiden in CSV

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- Text-Formatierung

- Wird eh nicht angezeigt



	A	B	C	D
1	id	ssflag	pagenumberinocrdocument	authorfirstname
2	70526	2		0 Adler
3	70527	2		2 Bruno
4	70528	2		2 Felix
5	70529	2		2 Friedrich (Wolfgang)
6	70530	2		2 Georg
7	70531	2		2 Max
8	70532	2		2 Otto
9	70533	2		2 Viktor

Quelle: Open Data Portal, "Liste der verbotenen Bücher"

Zu vermeiden in CSV

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel



	A	B	C	D	E	F
1	# (c) Der Bundeswahlleiter, Wiesbaden 2017					
2	#					
3	# Wahl zum 19. Deutschen Bundestag (24. September 2017)					
4	# Endgültige Ergebnisse der Erst- und Zweitstimmen sowie der Vorperiode nach					
5	#					
6	Nr	Gebiet	gehört zu	Wahlberechtigte		
7				Erststimmen	Zweitstimme	
8				Endgültig	Vorperiode	Endgültig
9	1 Flensburg – S		1	228471	226944	228471
10	2 Nordfriesland		1	186568	186177	186568
11	3 Steinburg – D		1	176636	176731	176636
12	4 Rendsburg-E		1	200831	198903	200831

Quelle: Der Bundeswahlleiter, "Wahl zum 19. Deutschen Bundestag"

Zu vermeiden in CSV

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- Überschriften (außer Spaltennamen), Copyright-Hinweise, usw.



	A	B	C	D	E	F
1	Nr	Gebiet	gehört zu	Wahlberechtigte		
2				Erststimmen		Zweitstimme
3				Endgültig	Vorperiode	Endgültig
4	1	Flensburg - S		1	228471	226944
5	2	Nordfriesland		1	186568	186177
6	3	Steinburg - D		1	176636	176731
7	4	Rendsburg-E		1	200831	198903
8	5	Kiel		1	204650	205243
9	6	Plön - Neum		1	174937	174746
10	7	Pinneberg		1	238533	235610
11	8	Segeberg - S		1	247296	244240
12	9	Ostholstein -		1	181522	180022
						181522

Quelle: Der Bundeswahlleiter, "Wahl zum 19. Deutschen Bundestag"

Hinweise

• Zeichencodierung

- UTF-8 ist empfohlen
- Dokument muss in einer bestimmten Codierung gespeichert werden
- Codierungsprobleme beeinträchtigen die Lesbarkeit von Dokumenten
 - Ex.:

Adler, Alfred: Praxis und Theorie der Individualpsychologie. Mvºnchen			
Adler, Bruno: SV§mtliche Schriften.			
Adler, Felix: Der Moralunterricht der Kinder. Berlin: Dvºmmler 1894.			
Adler, Friedrich (Wolfgang): SV§mtliche Schriften.			
Adler, Georg: SV§mtliche Schriften.			
Adler, Max: SV§mtliche Schriften.			

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

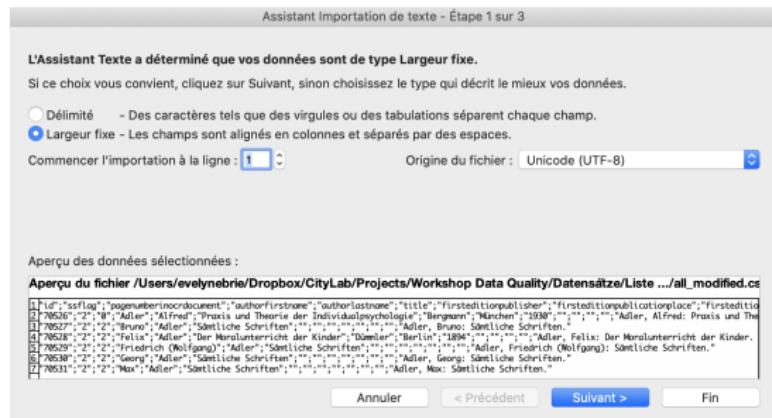
3. Beispiel

Hinweise

- Zeichencodierung ändern

- Fichier -> Importer (**Tori: we'll change this to German**)

- Origine du fichier: UTF-8



Auswertung

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

csv

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- **Vorteil:** am offensten
- **Nachteil:** sehr einfache Darstellungsmöglichkeiten

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Andere Formate

Daten Formate

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- **Nicht tabellarisch**

- JSON
- GeoJSON

- **Tabellarisch**

- SAV
- DTA
- RData

Daten Formate

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

• JSON

- “JavaScript Object Notation”
- Nicht tabellarisch, mehr flexibel

```
"Entity1" : {  
    "type" : "object",  
    "title" : "Entity1",  
    "properties" : {  
        "Attribute2" : {  
            "type" : "string",  
            "$ref" : "ReferencedDomains.json/properties/CharDomain"  
        },  
        "Attribute1" : {  
            "type" : "string",  
            "required" : true,  
            "maxLength" : 5,  
            "minLength" : 5  
        }  
    }  
}
```

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Daten Formate

• GeoJSON

- “Geographic JSON”
- Geografische Daten
- Basierend auf dem JSON-Format

```
1 {  
2   "type": "FeatureCollection",  
3   "features": [  
4     {  
5       "type": "Feature",  
6       "geometry": {  
7         "type": "Point",  
8         "coordinates": [-111.125, 33.375]  
9       },  
10      "properties": {  
11        "trackid": "AA-1234",  
12        "reported_dt": "12/31/2019 23:59:59"  
13      }  
14    },  
15    {  
16      "type": "Feature",  
17      "geometry": {  
18        "type": "Point",  
19        "coordinates": [-113.675, 35.875]  
20      },  
21      "properties": {  
22        "trackid": "AA-1234",  
23        "reported_dt": "12/31/2019 23:59:59"  
24      }  
25    }  
26  ]  
27 }
```

Daten Formate

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- SAV
 - SPSS Daten
- DTA
 - Python Daten
- RData
 - R Daten

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Formatentscheidung

Wann solltet man was nutzen?

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- Lieber CSVs bevorzugen
- Excel benutzen wenn...
 - es wichtig ist, dass die Tabellen etwas lesbarer für Menschen werden (vs. für andere Programme)
 - mehrere Blättern nötig sind

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Doppelcheckliste

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

5 Sachen zu überprüfen

① Nummern

② Daten

③ Standorte

④ Tippfehler

⑤ Identifikationsnummer

1. Nummern

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

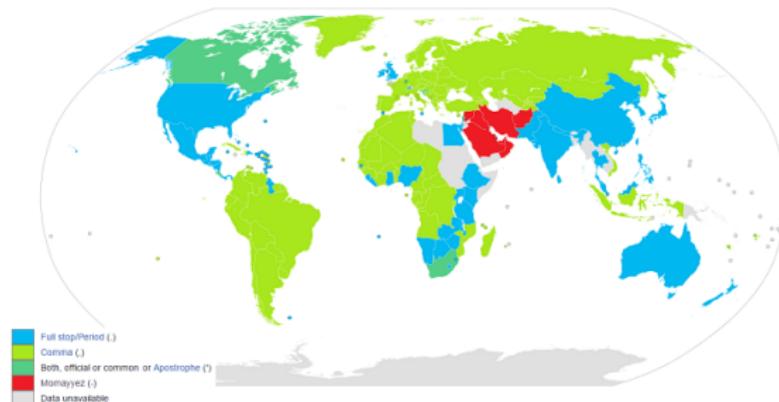
Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

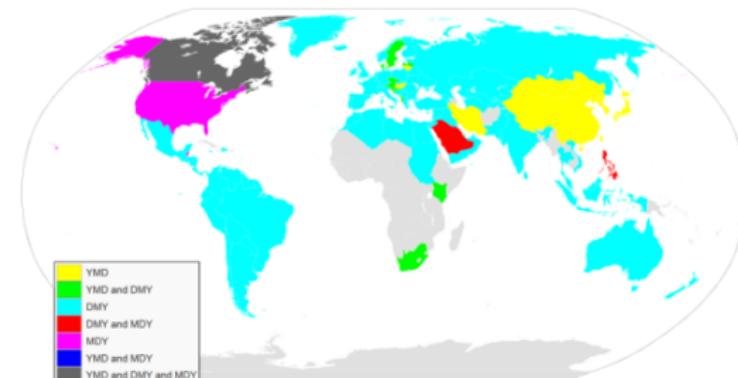
- Dezimaltrennzeichen (0.1 vs. 0,1)
 - Anders von Land zu Land
 - **Deutschland:** Komma System



2. Daten

- Reihenfolge der Tage, Monate und Jahre & Trennzeichen

- Anders von Land zu Land
- **Deutschland:** DD.MM.YY oder DD.MM.YYYY



3. Standorte

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

• Koordinaten

- Google-API (Adressen -> Koordinaten)
- Koordinatensystem angeben

• Adressen

- Konformität der Abkürzungen innerhalb des Dokumentes
 - Ex.: "Straße" vs. "Str."

4. Tippfehler

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- Spalten per Kategorie sortieren
 - Données -> Filtre avancé -> Extraction sans doublon

Buenos-Aires
Buss/Saar
Cassarate
Cassarete
Celje
Celle

Quelle: Open Data Portal, "Liste der verbannten Bücher"

- Rechtschreibkorrektur in Excel aktivieren

5. ID Nummer

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- Eine Identifikationsnummer pro Beobachtung
 - Meistens eine pro Reihe
- Wichtig um...
 - Änderungen zu verfolgen
 - Datensätze zusammenzuführen

Workshop zu Datenqualität

City Lab Berlin

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

Quiz

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel



Datensätze	Schulbausanierung Sommer 2020
Arbeitsmarkt	Berlin investiert in den kommenden Jahren 5,5 Milliarden Euro in die Sanierung und in den Bau von Schulgebäuden. Auch in den Sommerferien 2020 gehen an den Berliner Schulen die Bau- und Sanierungsmaßnahmen weiter.
Bildung	
Demographie	
Geographie und Stadtplanung	
Gesundheit	
Jugend	
Kunst und Kultur	
Öffentliche Verwaltung, Haushalt und Steuern	
Protokolle und Beschlüsse	
Sonstiges	

Informationen zum Datensatz

Lizenz:	Creative Commons Namensnennung
Kategorie:	Bildung
Geographische Abdeckung:	Berlin
Geographische Granularität:	Berlin
Zeitliche Granularität:	Keine
Veröffentlicht:	27.07.2020
Aktualisiert:	27.07.2020
Veröffentlichende Stelle:	Senatsverwaltung für Bildung, Jugend und Familie
E-Mail Kontakt:	karen.koenig AT senbjf.berlin.de
Website:	https://www.berlin.de/sen/bildung/service/daten/
Tags:	Bauen Sanieren Schulbaumaßnahmen ... (2 weitere)
Kommentare:	0

Fragen

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- In Excel, sortieren sie die Eingaben bei Finanzierung. Merken Sie etwas, was verbessert sein könnte?
- Merken Sie sich die Adressen in der "Adresse" Spalte. Was könnte besser gemacht werden?
- Wie kann man den Datensatz zu CSV konvertieren und speichern?

Analyse

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

• Finanzierung Spalte

- Konsistenz ist wichtig
- Auf Typos achten!

H	I	J
Kosten in Euro	Finanzierung	
1.175.000,1.254.000	baul. Unterhaltung	
1.262.000,00	BENE, Baubudget geplant	
415.000,00	Eigener Bezirkshaushalt	
1.920.000,00	Eigener Bezirkshaushalt	
31.100.000,00	Eigener Bezirkshaushalt	
1.066.000,102.000	Eigener Bezirkshaushalt	
3.775.000,00	Eigener Bezirkshaushalt	
190.215,606.831,962.553	Eigener Bezirkshaushalt; Kommunalinvestitionsprogramm	
519.929,244.198	Einsatz zweckgebundener Ein-nahmen für Infrastrukturmaßnahmen im Investitionsprogramm	
648.748,11.652.249.908	Investitionsprogramm	
98.250,00	Investitionsprogramm (Ausweichkosten)	
23.712.475,777	Investitionsprogramm	
720.000,00	Investitionsprogramm	
1.000.000,00	Investitionsprogramm	
1.500.000,800.000	Investitionsprogramm (Ausweichkosten)	
2.469.867,00	Investitionsprogramm SenStadtWohn	
280.000,00	Investitionsprogramm	
600.000,1.028.000	Investitionsprogramm	
110.000,00	Investitionsprogramm	
900.000,00	Investitionsprogramm; SIWA	
900.000,1.100.000	Investitionsprogramm (Ausweichkosten)	
800.000,00	SchulISP	
430.000,00	SchulISP	
1.500.000,00	Eigener Bezirkshaushalt	
590.000,00	Kommunalinvestitionsprogramm	
10.000.265.000,00	Eigener Bezirkshaushalt	
200.000,00	Städtebaulicher Datumsfehler	
330.000,200.000	SchulISP; Eigener Bezirkshaushalt	
1.200.000,00	BWA	
950.000,00	Kommunalinvestitionsprogramm	
1.942.681,00	SchulISP	
415.000,00	Eigener Bezirkshaushalt	
2.800.000,00	Geldanlagen im Denkmalschutz	
136.787	Baubudget gestart	
27.329	Baubudget gestart	
564.317	PV LAGEBO, Baubudget gestart	
400.000,00	Eigener Bezirkshaushalt	
775.000,00	Eigener Bezirkshaushalt	
400.000,00	Eigener Bezirkshaushalt	

Analyse

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

• Adresse Spalte

- “Straße” oder “Str.”?

Adresse
Ungarnstraße 75
Guineastrasse 17-18
Lützowstraße 83-85
Müllerstraße 158
Berolinastraße 8
Schöningstraße 17
Reinickendorfer Straße 60
Alt-Moabit 10
Bochumer Str. 8
Osloer Str. 23-26
Schwyzer Str. 6-8

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

3. Beispiel

Giez den Kiez

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

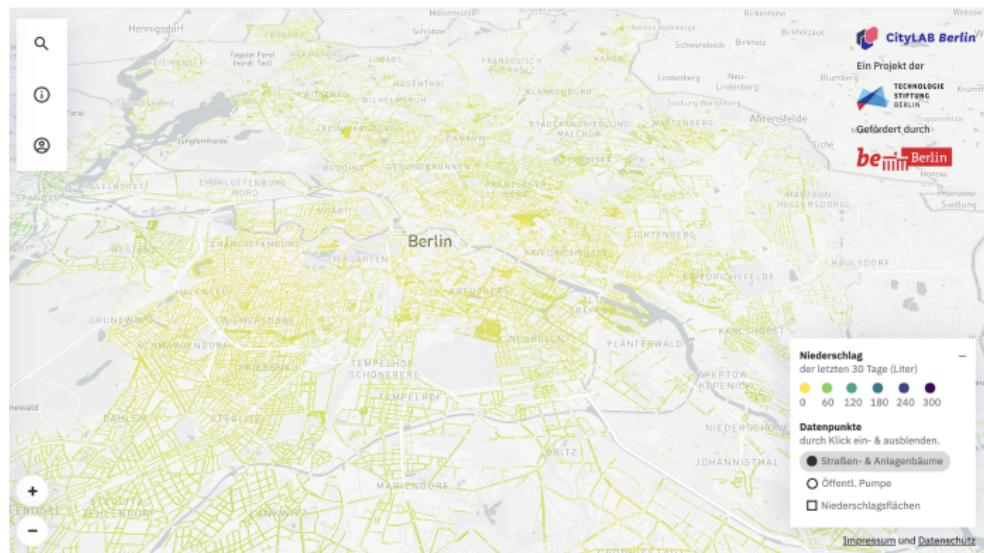
Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel



Giez den Kiez

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

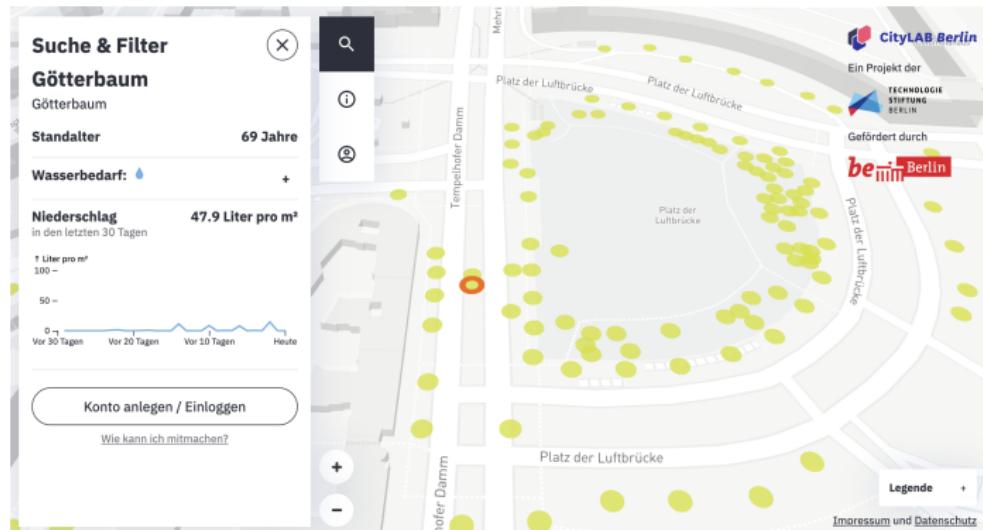
Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel



Fehlende ID Nummer

Einführung

1. Theorie

Was ist Datenqualität?

Relevanz von Datenqualität

Quiz

2. Anwendung

Excel

CSV

Andere Formate

Formatentscheidung

Doppelcheckliste

Quiz

3. Beispiel

- **Ziel:**

- Bewässerungsdaten von der Regierung in den Giez den Kiez Datensatz hinzufügen

- **Problem:**

- Kein gemeinsame Identifikationsnummer zwischen:

① Bewässerungsdaten von der Regierung

② Bewässerungsdaten von Giez den Kiez

Versuch #1

- **Methode:** ID selber von mehrere Spalten generieren

- Pflanzjahr + Standalter + Stammumfang + Stammdurchmesser + Baumhöhe + Kronendurchmesser + Sorte

```
bewaesserung$id <- NA  
bewaesserung$id <- paste0(bewaesserung$Pflanzjahr,  
                           bewaesserung$Standalter,  
                           bewaesserung$Stammumfang,  
                           bewaesserung$Stammdurchmesser,  
                           bewaesserung$Baumhöhe,  
                           bewaesserung$Kronendurchmesser,  
                           bewaesserung$Sorte.deutsch)
```

- **Problem:** Duplikate (i.e. Bäume mit identischen Charakteristiken im Datensatz)