

Datenmodellierung 2: Datenbanken

Vorlesung *Einführung in die Digital Humanities*
MSc Digital Humanities | Wintersemester 2019/20

Prof. Dr. Christof Schöch



Einstieg

Semesterüberblick

- 29.10.: Digital Humanities im Überblick
- 05.11.: Digitalisierung: Text und Bild
- 12.11.: Grundbegriffe des Programmierens
- 19.11.: Datenmodellierung 1: Modellierung
- **26.11.: Datenmodellierung 2: Datenbanken**
- 03.12.: Datenmodellierung 3: Text, Markup, XML
- 10.12.: Digitale Edition
- 17.12.: Geschichte der Digital Humanities
- 21.12.-5.1.: *Weihnachtspause*
- 07.01.: Informationsvisualisierung
- 14.01.: Natural Language Processing
- 21.01.: Quantitative Analyse 1: Stilometrie, Topic Modeling
- 28.01.: Quantitative Analyse 2: Superv. Machine Learning
- 04.02.: Open Humanities
- 11.02.: Klausurtermin

Sitzungsüberblick

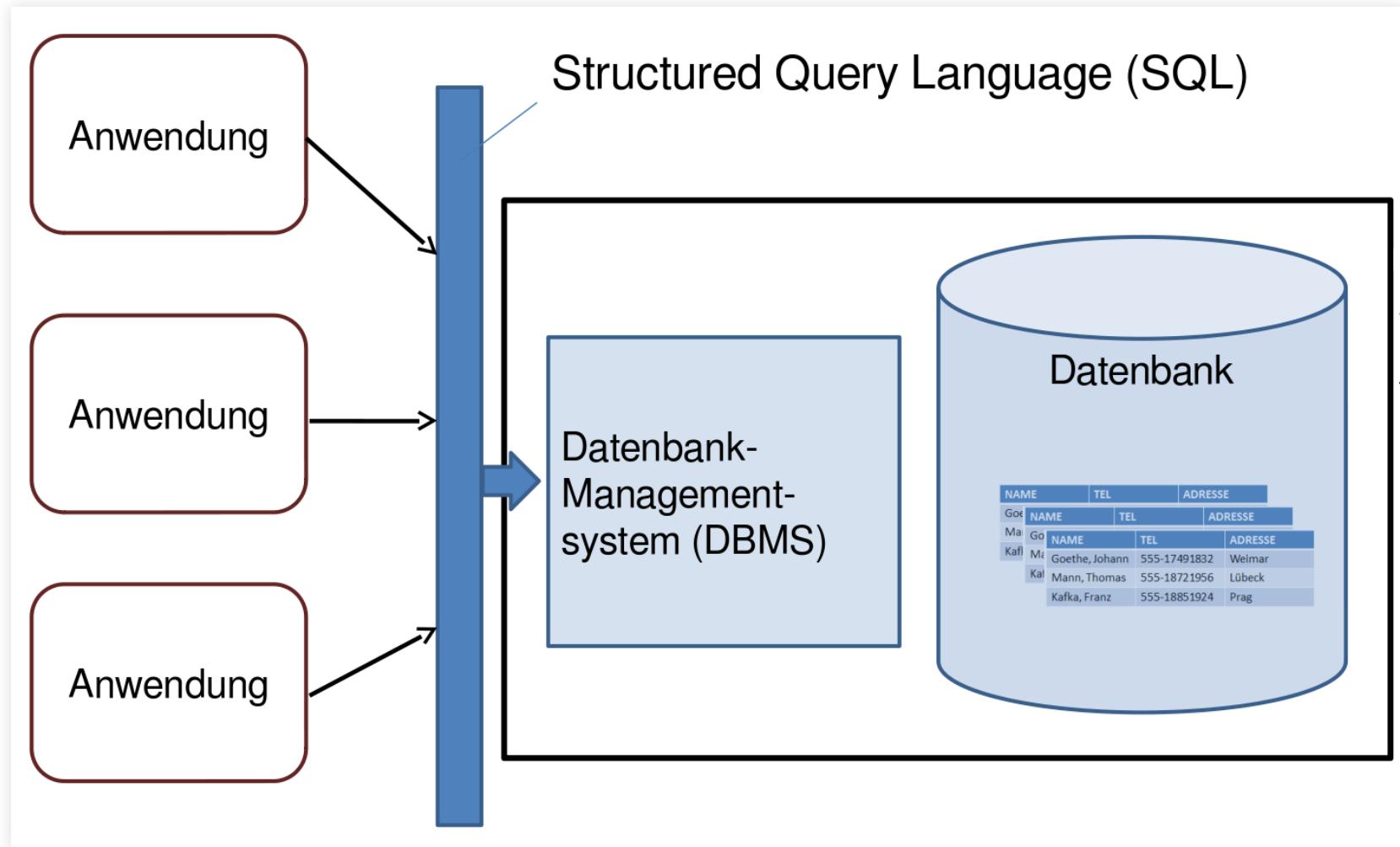
1. Datenbanken, wie und wozu?
2. Domäne: der zu modellierende Gegenstandsbereich
3. Konzeptuelles Modell: Entity-Relationship-Modell
4. Logisches Modell: Relational Database / Relationale Algebra
5. Implementierung: Structured Query Language

1. Datenbanken, wie und wozu?

(A) Datenbanken...

- eignen sich zur Organisation stark strukturierter Datenbestände
- können mit umfangreichen, detaillierten Datenbeständen befüllt werden
- erlauben präzise Suchabfragen auf diesen Beständen
- => erfüllen Grundbedürfnisse geisteswiss. Forschung nach Organisation, Speicherung und Abfragen von Informationen

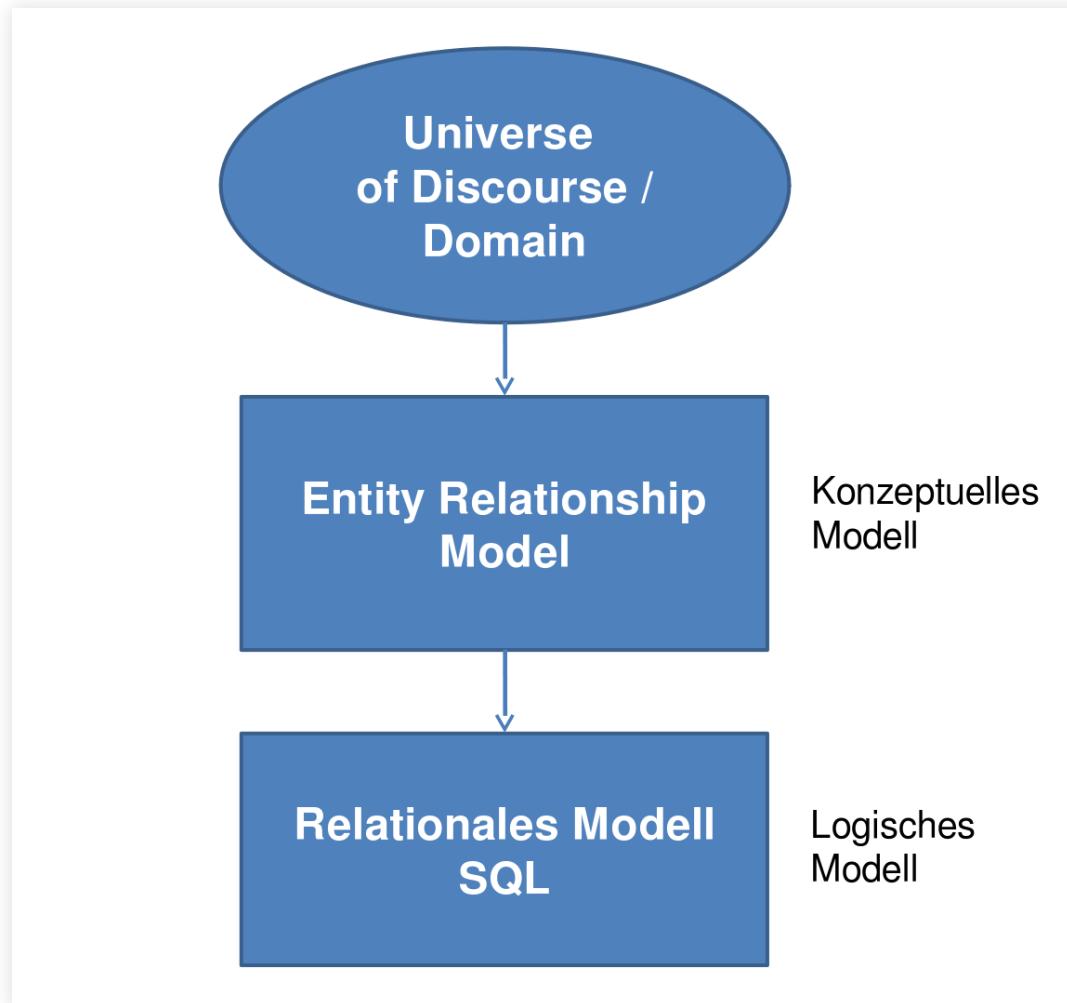
Datenbanksystem



Begriffe

- Datenbanksystem: das Gesamtsystem
- Datenbank-Management-System (DBMS): die Infrastruktur für die Datenbank
- Datenbank: enthält die Daten in strukturierter Form
- Datenbestand: die Datensätze, die vorhanden sind
- Anwendungen: greifen über Schnittstellen des DBMS auf die Daten zu

Zwei Aspekte



Prototypischer Ablauf

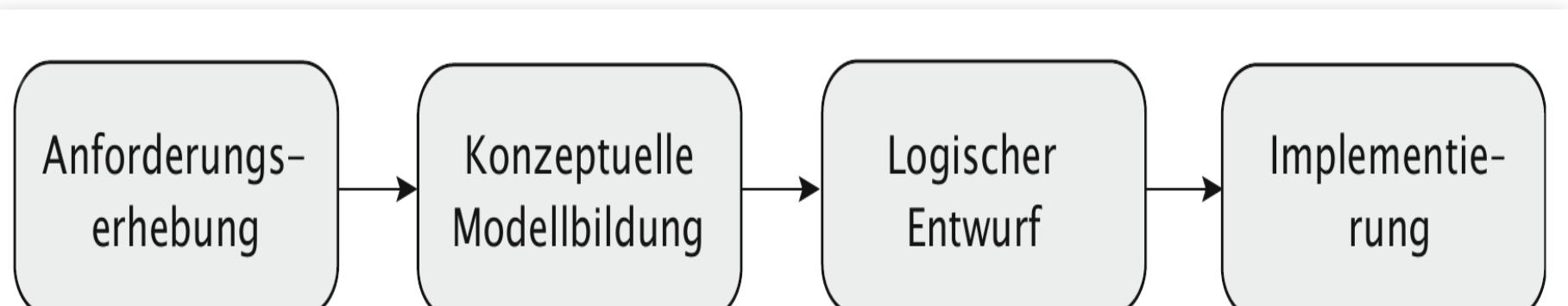


Abb. 21 Entwicklung einer Datenbankanwendung

(Quelle: Harald Klinke 2017)

2. Domäne und Anforderungen

Domäne: Bibliothek

Nutzungsszenario

Eine Geisteswissenschaftlerin möchte ihre Bibliothek verwalten. Die Bibliothek umfasst zahlreiche Texte vom Mittelalter bis zur Gegenwart. Die Datenbank soll es ermöglichen, schnell zu überprüfen, welche Autoren und welche Titel vorhanden sind. Die Geisteswissenschaftlerin möchte zudem sortieren können, und zwar nach dem Geburtsdatum bzw. Sterbedatum der Autoren, aber auch nach dem Namen der Autoren. Auf jeden Fall soll für jedes Buch die ISBN verzeichnet werden.

Anforderungsanalyse

- Autoren und Bücher unterscheiden
- Titel, Geburtsdatum, Sterbedatum, ISBN vorhalten
- Sortierbarkeit ermöglichen

Naiver Ansatz: Liste

- Marx, Karl; Das kommunistische Manifest ; 1818; 1883; 1242829340229
- Herder, J. ; Bildung der Menschheit ; 1744; 1803; 1534932829103
- Smith, J. ; An Inquiry into the Nature...; 1744; 1803; 1534932829103
- Marx, Karl ; Das Kapital ; 1818; 1883; 1231288828783
- Rousseau, J.-J.; Du contrat social ; 1712; 1778; 1665229181734

Verbesserung: Tabelle

Autor	Titel	Geb.	Tod	ISBN
Marx, Karl	Das kommunistische Manifest	1818	1883	1242829340229
Herder, J.	Bildung der Menschheit	1744	1803	1534932829103
Smith, J.	An Inquiry into the Nature...	1744	1803	1534932829103
Marx, Karl	Das Kapital	1818	1883	1231288828783
Rousseau, J.-J.	Du contrat social	1712	1778	1665229181734

So weit, so gut...

- Entitäten:
 - Bibliothek
 - Buch, Titel, ISBN
 - Autor, Name, Geburtsdatum, Sterbedatum
- Sortierbar, suchbar, filterbar
- Nachteile
 - Redundanzen
 - keine expliziten Beziehungen
 - keine Mehrfachbearbeitung
 - keine Schnittstelle
 - (und: wenig performant)

3. Konzeptuelles Modell: Entity-Relationship-Modell

(A) Einführend

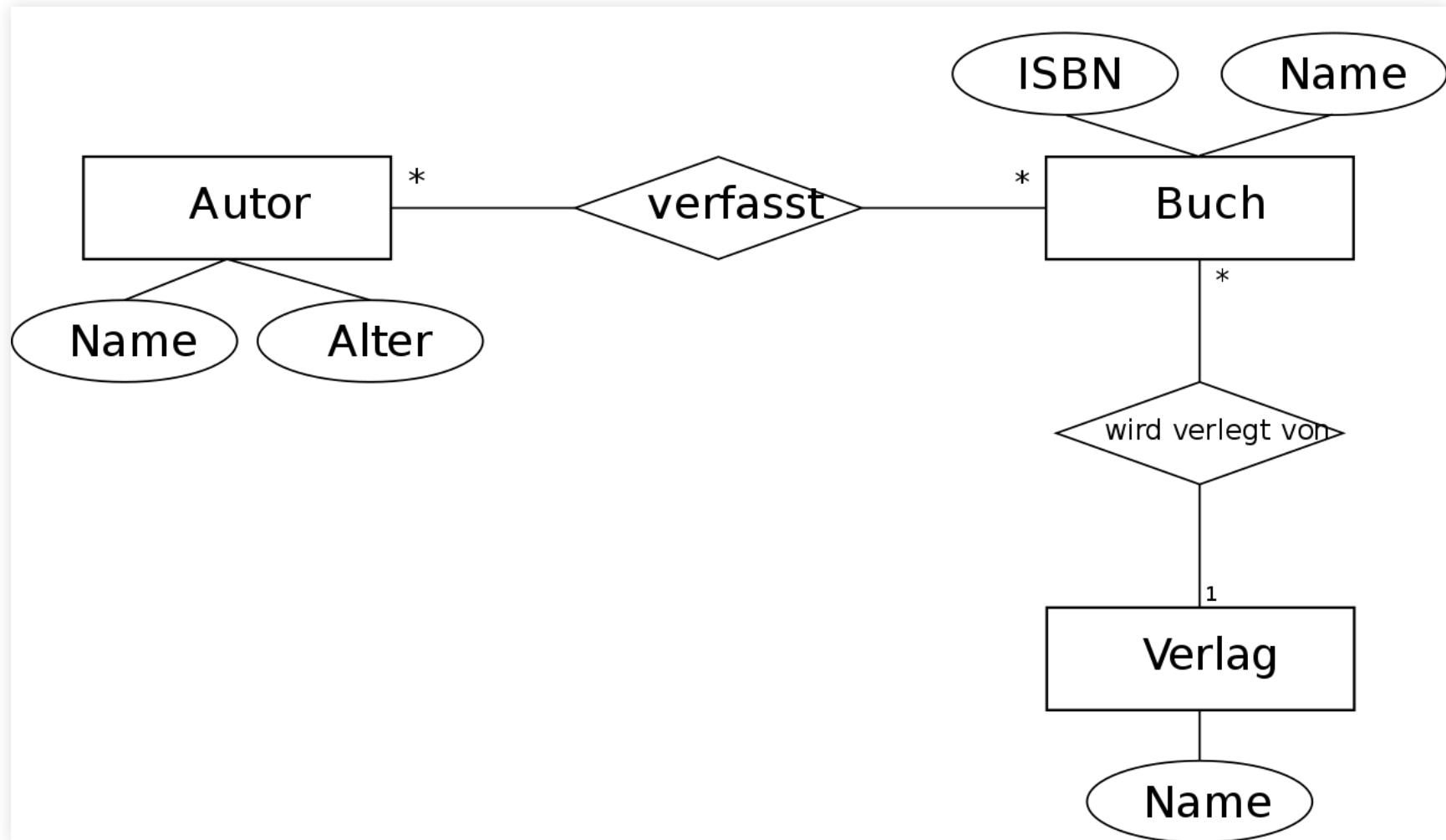
Vier Aufgaben des konzeptionellen Datenmodells

- Klassifizierung: Festlegung der Objekttypen (Entitäten)
- Abstraktion: Bestimmung der relevanten Eigenschaften (Attribute)
- Beziehungen: Beschreibung der Zusammenhänge zwischen den Objekten (Relationen)
- Identifizierung: Festlegung von eindeutigen Namen (Schlüssel)

Entity-Relationship-Modell

- konzeptuelles Modell: abstrakt
- erfüllt die vier genannten Aufgaben
- abstrakte Struktur der Daten (nicht die Daten selbst)
- bspw. in grafischer Notation festgehalten

Entity-Relationship-Diagramm



Elemente des ER-Modells

- Entitäten (Objekttypen)
- Attribute (Eigenschaft der Objekte)
- Werte (Ausprägung einer Eigenschaft)
- Relationen (Beziehungen zwischen Entitäten)
- Kardinalität von Beziehungen (mengenmäßige Beziehung)

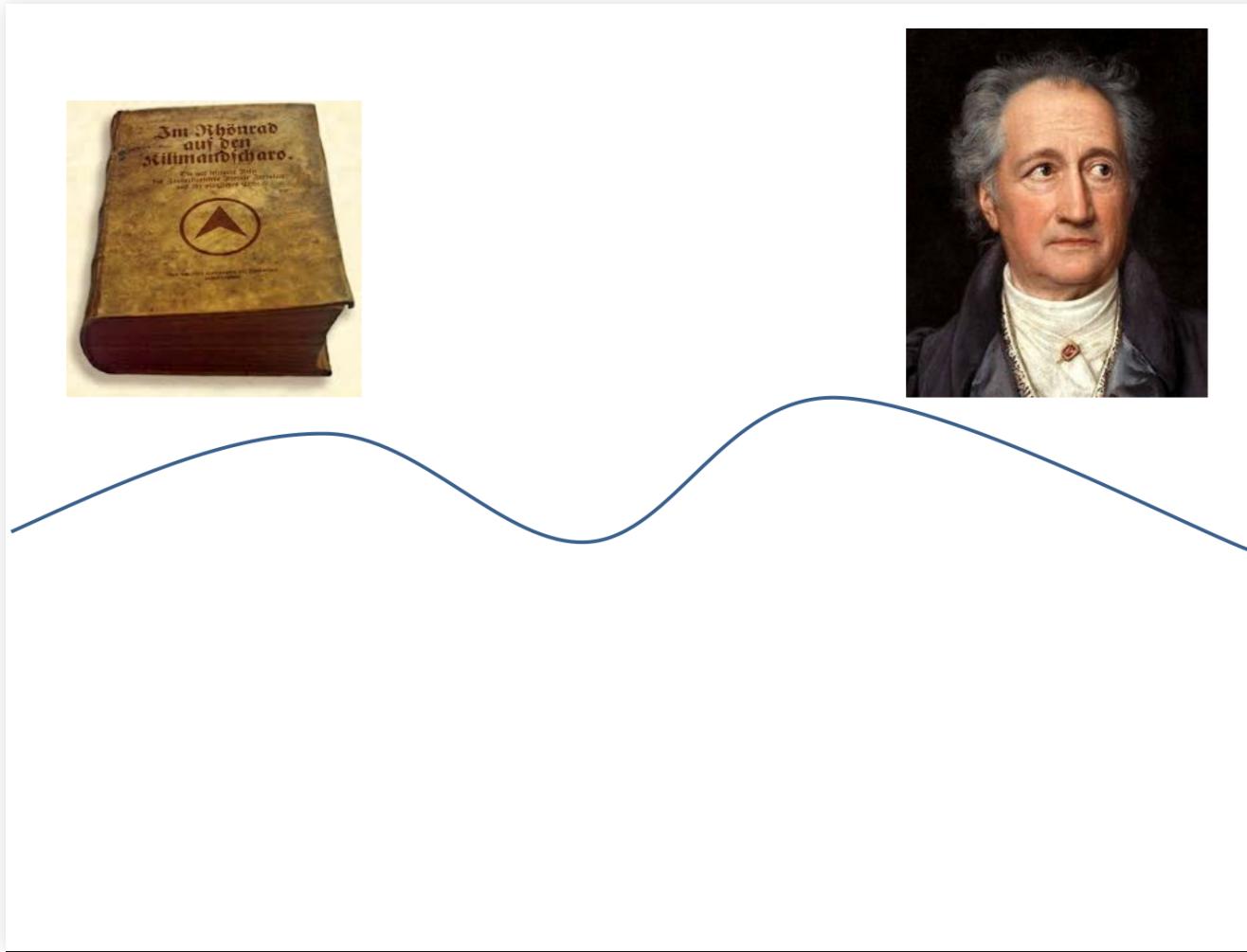
Kardinalität?



Typen: 1:1, 1:n, n:1, n:m

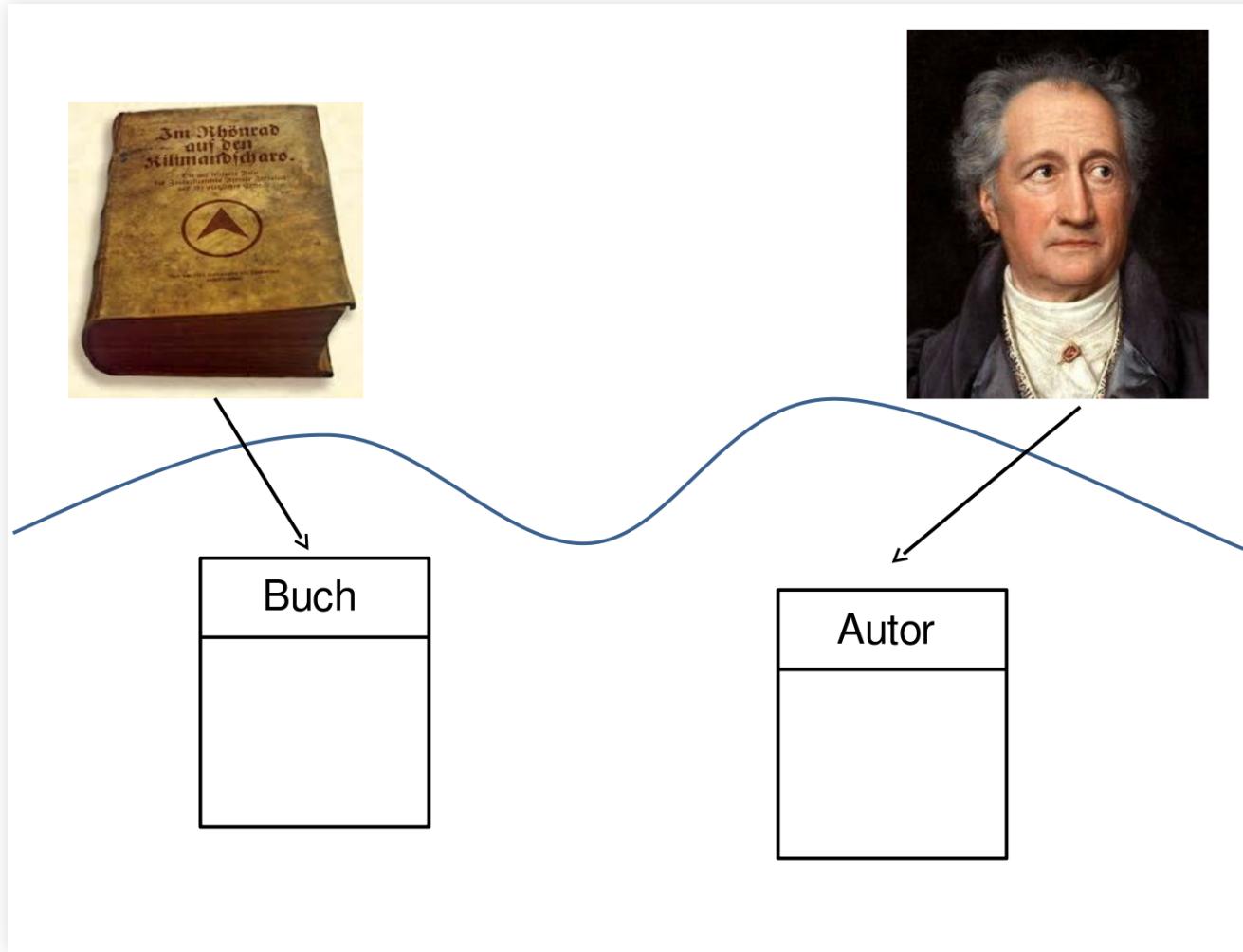
(B) Beispiel Bibliothek

Bibliothek: Bücher und Autoren



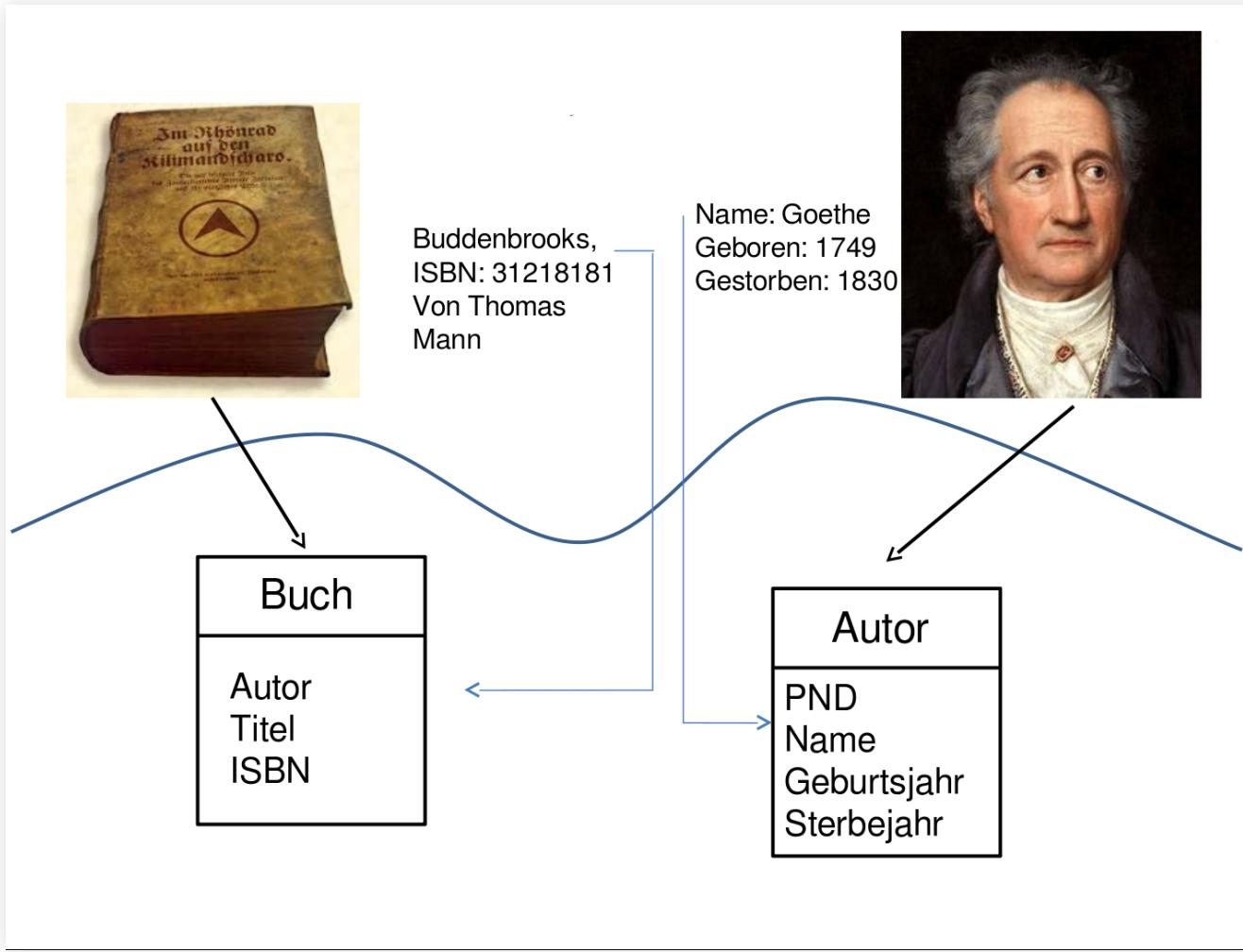
(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

Klassifikation



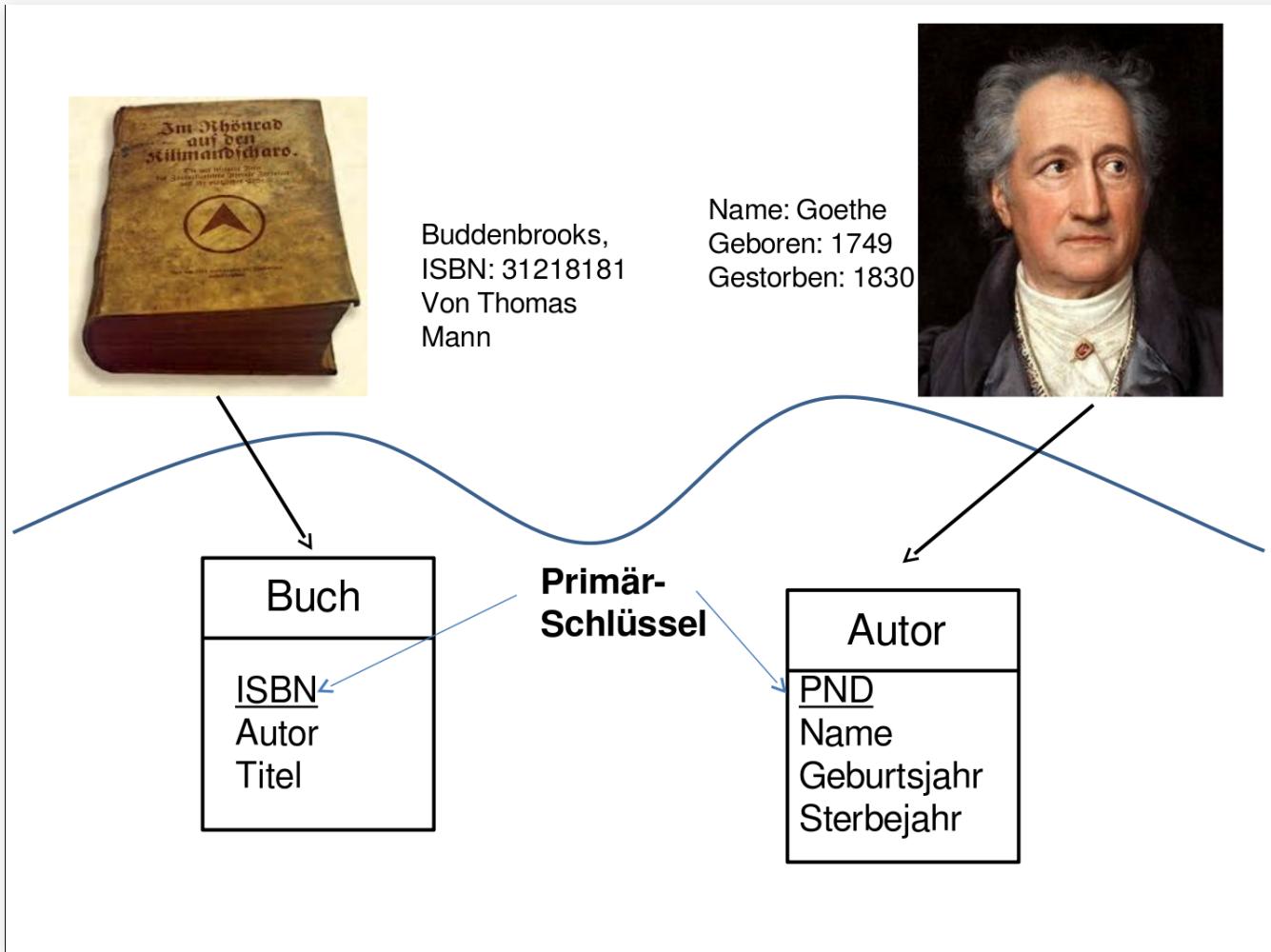
(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

Abstraktion



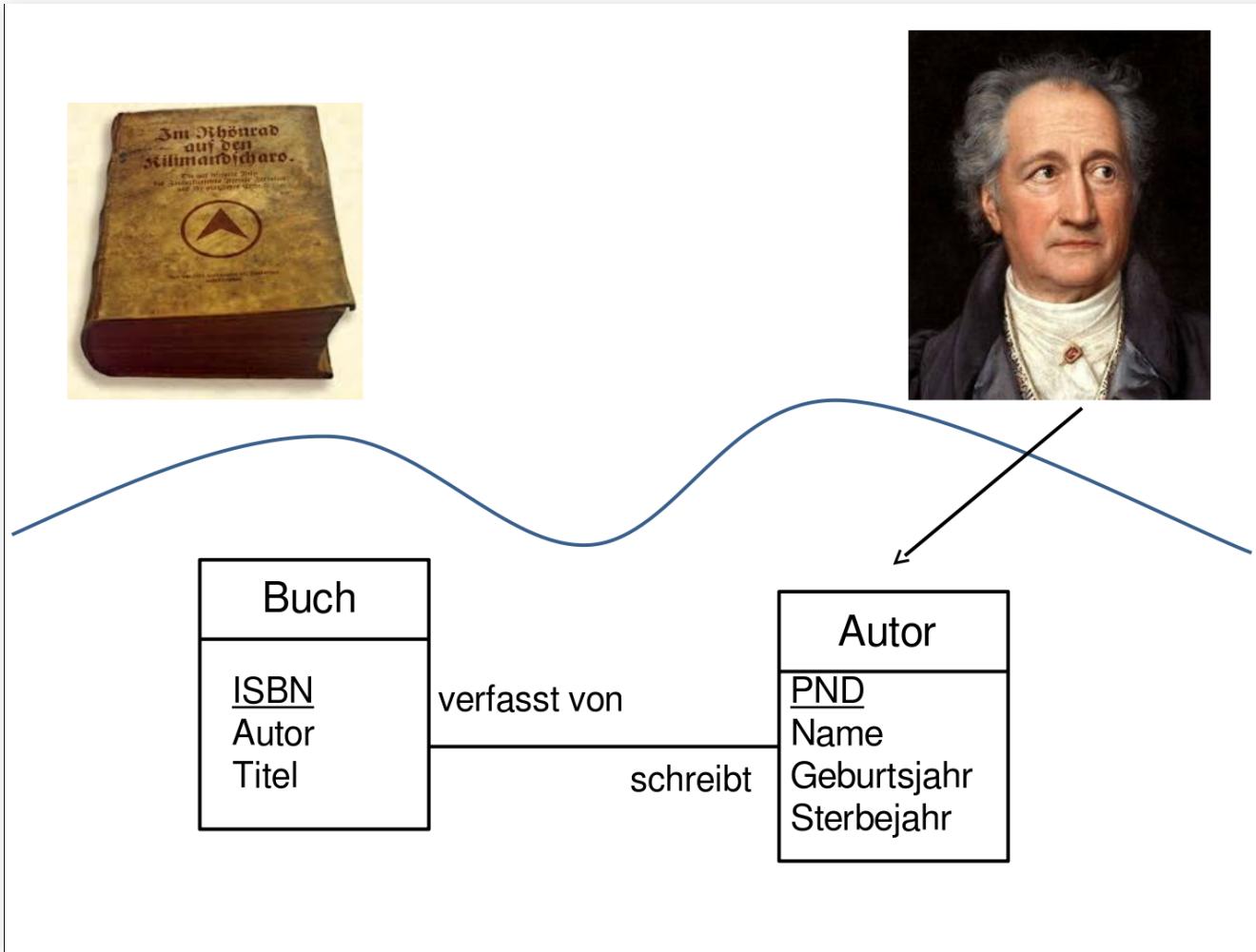
(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

Identifizierung



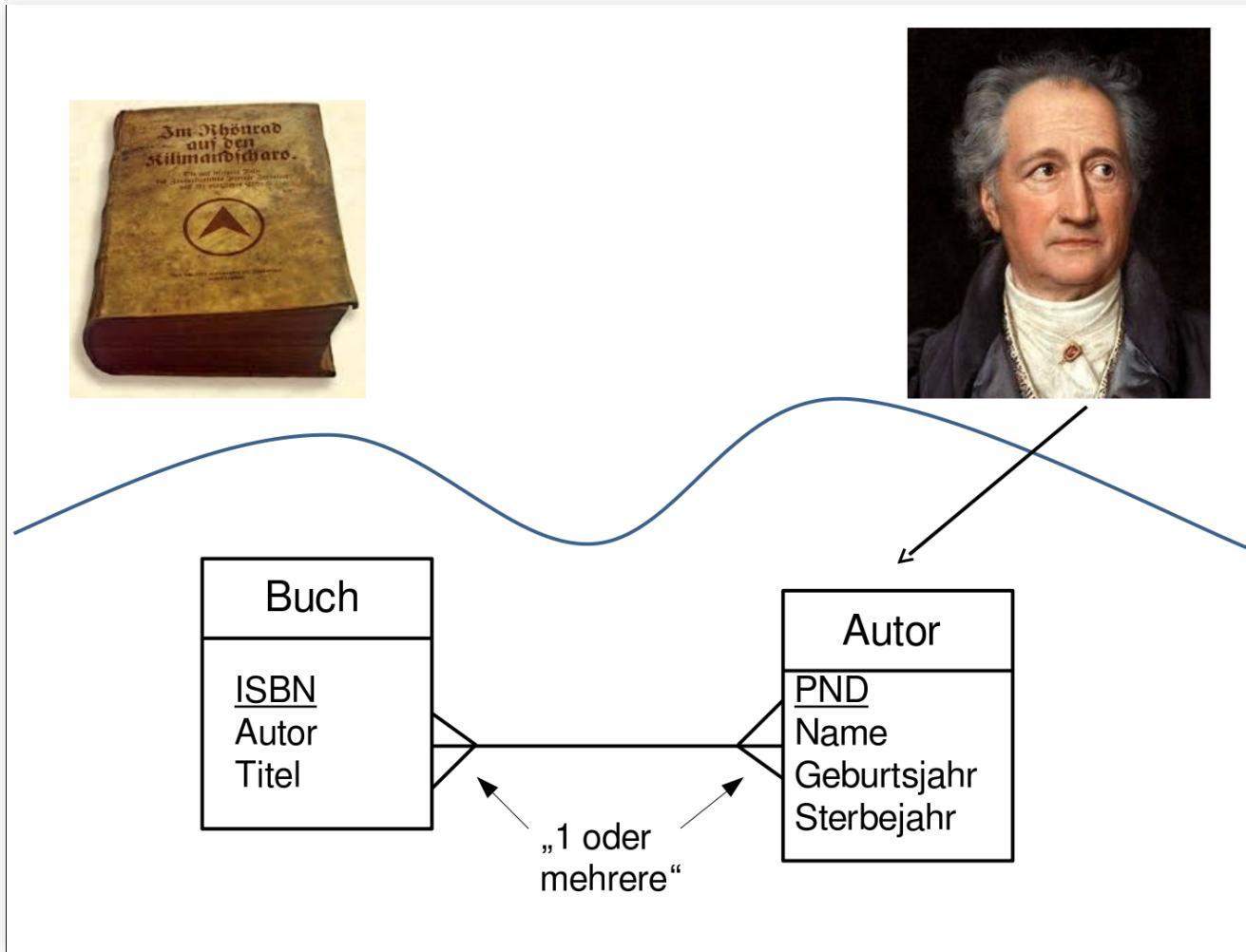
(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

Beziehungen



(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

Kardinalität



(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

Logisches Datenmodell: Relational Database Model (RDM)

(A) Grundideen

Grundideen des RDM

- ER-Modell: Entitäten, Attribute, Relationen
- Jede Klasse von Entität (gleiche Attribute) bekommt eine separate Tabelle
- Jeder Eintrag bekommt einen Identifier ("key")
- Relationen zwischen Entitäten laufen über die "keys"

Datenbankschema

- Legt die Struktur der Datenbank fest
- welche Tabellen ("Relationen") gibt es?
- Für jede Tabelle: Welche Spalten gibt es?
- Für jede Spalte: Welcher Datentyp ist erlaubt? (str, int, bool)

(B) Normalisierung

Normalisierung: Definition

Unter Normalisierung eines relationalen Datenschemas (Tabellenstruktur) versteht man die Aufteilung von Attributen (Tabellenspalten) in mehrere Relationen (Tabellen) gemäß den Normalisierungsregeln (s. u.), so dass eine Form entsteht, die keine Redundanzen mehr enthält.

(Quelle: Wikipedia, Art. "Normalisierung (Datenbank)",
[https://de.wikipedia.org/wiki/Normalisierung_\(Datenbank\)](https://de.wikipedia.org/wiki/Normalisierung_(Datenbank)))

Was sind "Normalformen"?

- Normalformen sind Klassen von Qualitätskriterien
- Sie werden nach und nach angewandt, um die Qualität der Datenbankstruktur zu verbessern
- Wir beschränken uns auf die erste, zweite und dritte Normalform
- Normalisierung hat ihre Grenzen: es kann auch ineffizient werden

Normalformen (NF)

- 1. NF: Jedes Attribut der Relation hat atomare Werte; es gibt keine Wiederholungsgruppen
- 1. NF: Attribute sind "voll funktional abhängig" vom (ganzen) Primärschlüssel
- 1. NF: Kein Nichtschlüsselattribut ist von einem anderen Nichtschlüsselattribut abhängig

Ausgangslage

ISBN	Titel	Autor	Geb.	Tod
1242829340229	Das kommunistische Manifest, Das Kapital	Marx, Karl	1818	1883
1534932829103	Bildung der Menschheit	Herder, J.	1744	1803
1534932829103	An Inquiry into the Nature...	Smith, J.	1744	1803
1665229181734	Du contrat social	Rousseau, J.-J.	1712	1778

Erste Normalform

ISBN	Titel	AutorVN	AutorNN	Geb.	Tod
1242829340229	Das kommunistische Manifest	Karl	Marx	1818	1883
1242829340229	Das Kapital	Karl	Marx	1818	1883
1534932829103	Bildung der Menschheit	J.	Herder	1744	1803
1534932829103	An Inquiry into the Nature...	J.	Smith	1744	1803
1665229181734	Du contrat social	J.-J.	Rousseau	1712	1778

(= Atomisierung, keine Wiederholungsgruppen)

Zweite Normalform

ISBN	Titel
1242829340229	Das kommunistische Manifest
1534932829103	Bildung der Menschheit
1534932829103	An Inquiry into the Nature...
1231288828783	Das Kapital
1665229181734	Du contrat social

GND	AutorVN	AutorNN	Geb.	Tod
19283746	Karl	Marx	1818	1883
98761234	J.	Herder	1744	1803
55652008	J.	Smith	1744	1803
11223344	J.-J.	Rousseau	1712	1778

(= Abhängigkeit der Attribute vom Primärschlüssel)

(a) Fremdschlüssel

ISBN	Titel	GND
1242829340229	Das kommunistische Manifest	19283746
1534932829103	Bildung der Menschheit	98761234
1534932829103	An Inquiry into the Nature...	55652008
1231288828783	Das Kapital	19283746
1665229181734	Du contrat social	11223344

GND	AutorVN	AutorNN	Geb.	Tod
19283746	Karl	Marx	1818	1883
98761234	J.	Herder	1744	1803
55652008	J.	Smith	1744	1803
11223344	J.-J.	Rousseau	1712	1778

(b) Assoziationsstabelle

ISBN	Titel
1242829340229	Das kommunistische Manifest
1534932829103	Bildung der Menschheit

... ...

GND	AutorVN	AutorNN	Geb.	Tod
19283746	J.	Herder	1744	1803
98761234	J.	Herder	1744	1803

... ...

GND	ISBN
19283746	1242829340229
98761234	1534932829103

... ...

Dritte Normalform

Verlags-ID	Name	PLZ	Ort
ABC123	De Gruyter	10785	Berlin
CDE456	Metzler	70182	Stuttgart
FGH789	transcript	33602	Bielefeld

Dritte Normalform

Verlags-ID	Name	PLZ
ABC123	De Gruyter	10785
CDE456	Metzler	70182
FGH789	transcript	33602

PLZ	Ort
10785	Berlin
70182	Stuttgart
33602	Bielefeld

(Indirekte Abhangigkeit aufgelost)

(C) Relationale Algebra

Edgar F. Codd (1923-2003)

Information Retrieval

A Relational Model of Data for Large Shared Data Banks

E. F. CODD
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on n -ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

Entwickelte das relationale Datenbankmodell

Codd, Edgar F. (1970). "A relational model of data for large shared data banks". Communications of the ACM 13/6.

(Porträt von Codd: siehe: https://en.wikipedia.org/wiki/Edgar_F._Codd#/media/File:Edgar_F_Codd.jpg)

Codd

"It was Codd's very great insight that a database could be thought of as a set of relations, that a relation in turn could be thought of as a set of propositions ..., and hence that all of the apparatus of formal logic could be directly applied to the problem of database access and related problems."

(Date, C. J. (2001). The Database Relational Model: A Retrospective Review and Analysis. Reading: Addison-Wesley.)

Formale Logik?

- Jede Tabelle ist ein Typ von Relation
- Jede Tabellenzeile enthält Aussagen
 - Marx, ist Autor von, Das Kapital
 - Marx, ist gestorben, 1883
- Formales Schließen:
 - Das Kapital, wurde verfasst vor, 1883
 - (Denn: Autoren verfassen nur zu Lebzeiten Werke)

4. Structured Query Language (SQL)

SQL

- SQL – Structured Query Language
- Standardsprache zur Erzeugung, Abfrage und Verwaltung von Datenbanken
- Keine 1:1 Umsetzung des relationalen Datenmodells, aber nahe dran
- Wird von allen relationalen Datenbanken unterstützt
- ANSI-Standard (aber es gibt Dialekte)

Drei Bereiche von SQL

- Datendefinition
 - Data Definition Language
 - bspw.: Tabelle erstellen:
 - CREATE TABLE Autoren ...
- Datenmanipulation
 - Data Manipulation Language
 - bspw.: Eintrag in einer Tabelle vornehmen:
 - INSERT INTO Autoren ...
- Datenabfrage
 - Data Query Language
 - bspw.: Suchabfrage formulieren
 - SELECT Name FROM Autoren ...

SQL in Python

- Library: sqlite3
- Dokumentation:
<https://docs.python.org/3.7/library/sqlite3.html>

Data Definition: CREATE

```
CREATE TABLE Buecher (
    ISBN INTEGER PRIMARY KEY,
    GND INTEGER,
    TITEL CHARACTER (50)
);
```

- Erstellt eine neue Tabelle "Buecher" mit ISBN, GND und Titel
- Weitere Befehle: Ändern (ALTER), Löschen (DROP) einer Tabelle

Data Manipulation: INSERT

```
INSERT INTO Buecher (ISBN, GND, TITEL)  
values (3211810002, 449382, "Faust")
```

- Fügt neue Datensätze in eine Tabelle ein

Data Query: Bausteine

- SELECT: welche Informationen sollen angezeigt werden?
- FROM: welcher Tabellen/Spalten sollen abgefragt werden?
- WHERE: welche Bedingungen werden formuliert?

Einfaches Beispiel

```
SELECT * FROM Buecher  
WHERE TITEL=="Faust"
```

- Zeige alle Spalten an,
- aus der Tabelle "Buecher";
- und zwar für diejenigen Einträge,
bei denen der Titel "Faust" lautet

Einfaches Beispiel

```
SELECT TITEL, JAHR FROM Buecher  
WHERE GND=="98765"
```

- Zeige die Spalten "Titel" und "Jahr" an;
- aus der Tabelle "Buecher";
- und zwar für den Eintrag,
der die GND 98765 hat.

Anwendungen

SQL in Python

- Library: sqlite3
- Dokumentation:
<https://docs.python.org/3.7/library/sqlite3.html>

Beispiel in LibreOffice Base

Screenshot of LibreOffice Base showing a database application for "figurenkonstellation.odb".

The interface includes:

- Left Sidebar:** Database navigation with tabs for Tables (selected), Queries, Forms, and Reports.
- Central Area:** Relation Design window showing three tables: Figuren, Szenen, and Anwesenheit, connected by relationships.
- Bottom Area:** Three data entry windows for Szenen, Figuren, and Anwesenheit.

Szenen - figurenkonstellation (Table View):

	Szenen-ID	Akt	Szene
0	1	1	1
1	1	1	2
2	1	1	3
3	1	1	4
4	2	1	
5	2	2	
6	2	2	3
7	2	2	4
8	2	2	5

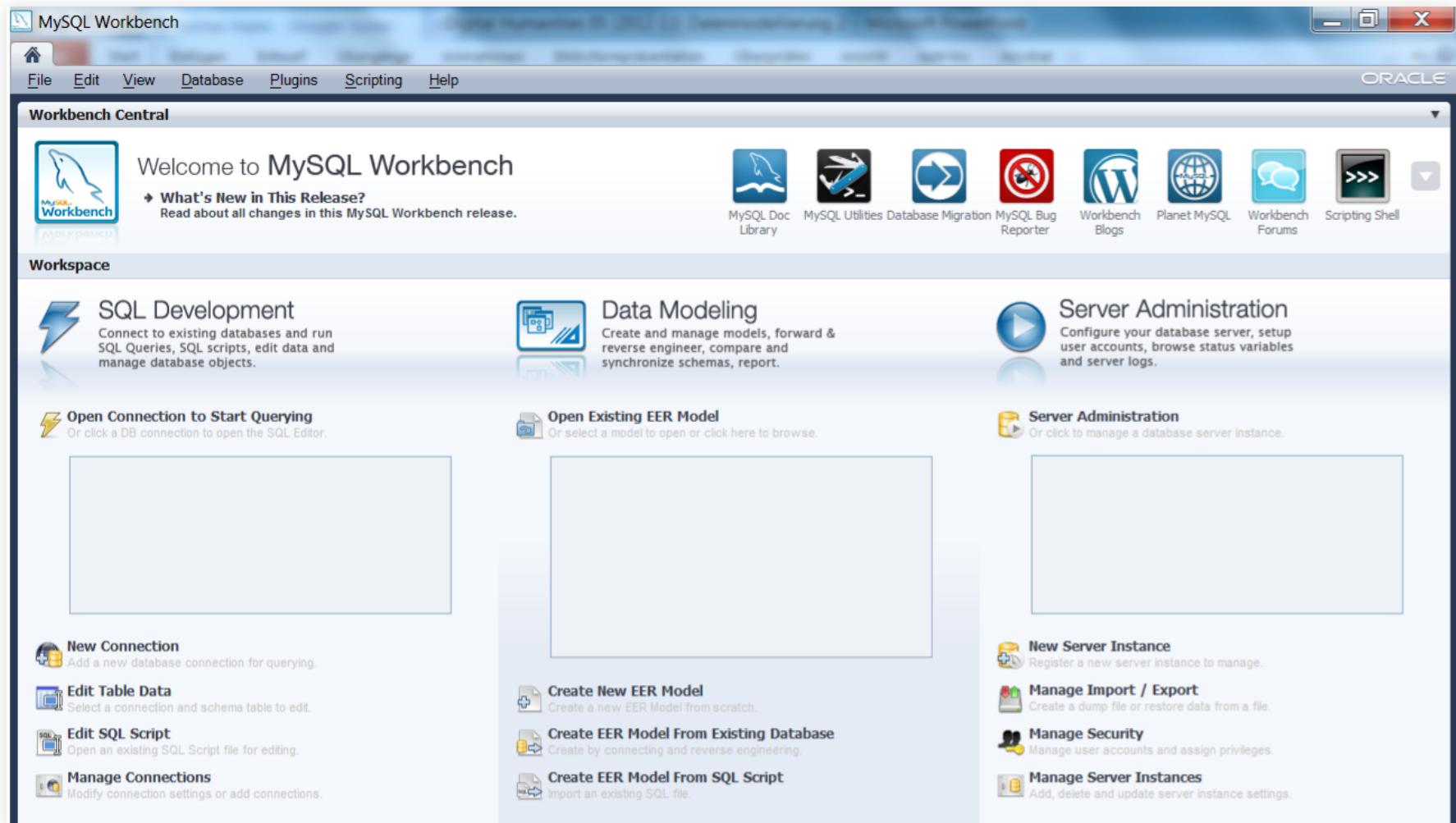
Figuren - figurenkonstellation (Table View):

	Figur-ID	Name	Geschlecht	Status
0	Andromache	weiblich	adelig	
1	Pyrrhus	maennlich	adelig	
2	Hermione	weiblich	adelig	
3	Pylade	maennlich	adelig	
4	Cléone	weiblich	nicht-adeli	
5	Céphise	weiblich	nicht-adeli	
6	Phoenix	weiblich	nicht-adeli	
7	Suite Ores	diverse	nicht-adeli	
8	Oreste	maennlich	adelig	

Anwesenheit - figurenkonstellation (Table View):

	ID	Szenen-ID	Figur-ID
0	0	0	8
1	0	0	3
2	1	1	1
3	1	1	8
4	1	1	6
5	2	1	1
6	2	2	6
7	3	1	1
8	3	0	0
9	3	3	5

MySQL Workbench



Wikidata SPARQL Endpoint

The screenshot shows the Wikidata Query Service interface. On the left, there is a sidebar with various icons: a blue info icon, a red cross icon, a green diamond icon, a blue folder icon, a grey circular arrow icon, a red trash bin icon, and a blue link icon. Below these is a large blue play button icon. At the top, there is a navigation bar with the text "Wikidata Query Service" and links for "Examples", "Help", "More tools", and "English".

The main area contains a SPARQL query:

```
1 SELECT ?country ?countryLabel ?capitalLabel
2 WHERE {
3     wd:Q458 wdt:P150 ?country.    # European Union
4     OPTIONAL{ ?country wdt:P36 ?capital. }
5     SERVICE wikibase:label { bd:serviceParam wikibase:language "de". }
6 }
```

Below the query results, there is a table with three columns: "country", "countryLabel", and "capitalLabel". The table contains the following data:

country	countryLabel	capitalLabel
wd:Q27	Irland	Dublin
wd:Q28	Ungarn	Budapest
wd:Q29	Spanien	Madrid
wd:Q31	Belgien	Brüssel
wd:Q32	Luxemburg	Luxemburg

At the bottom right of the table, there are buttons for "Code", "Download", and "Link". Above the table, a status bar indicates "28 results in 78498 ms".

Abschluss

Fragen?

Lektürehinweise

- Harald Klinke, "Datenbanken", in: *Digital Humanities: Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, 2017, S. 109-127.

Weitere Empfehlungen

- Stephen Ramsay, "Databases", in: *The Companion to Digital Humanities*, ed. by Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell, 2008. URL:
<http://www.digitalhumanities.org/companion/> (kostenfrei)
- Timo Hempel, "Normalisierung von Datenbanken", 2014. URL:
<https://www.tino-hempel.de/info/info/datenbank/normalisierung.htm> (gut erklärt)

Darüber hinaus

- René Steiner. *Grundkurs Relationale Datenbanken*. 6. Auflage. Braunschweig: Vieweg, 2006. [ER, SQL]

Nächste Sitzung

- 27.11.: Thema: Datenmodellierung 3: Markup, XML, TEI



Christof Schöch, 2018
<http://www.christof-schoech.de>

Lizenz: Creative Commons Attribution 4.0