



Datenmodellierung 2: Datenbanken

Vorlesung *Einführung in die Digital Humanities*
MSc Digital Humanities | Wintersemester 2018/19

Prof. Dr. Christof Schöch



Einstieg

Semesterüberblick

- 23.10.: Digital Humanities im Überblick
- 30.10.: Digitalisierung: Text und Bild
- 06.11.: Grundbegriffe des Programmierens
- 13.11.: Datenmodellierung 1: Modellierung
- **20.11.: Datenmodellierung 2: Datenbanken**
- 27.11.: Datenmodellierung 3: Text, Markup, XML
- 30.11.: Digitale Edition (*Vorlesung statt Übung*)
- 04.12.: *Übung statt Vorlesung*
- 11.12.: Geschichte der Digital Humanities
- 18.12.: Informationsvisualisierung
- 22.12.-6.1.: *Weihnachtspause*
- 08.01.: Natural Language Processing
- 15.01.: Quantitative Analyse 1: Stilometrie, Topic Modeling
- 22.01.: Quantitative Analyse 2: Superv. Machine Learning
- 29.01.: Open Humanities
- 05.02.: Klausurtermin

Sitzungsüberblick

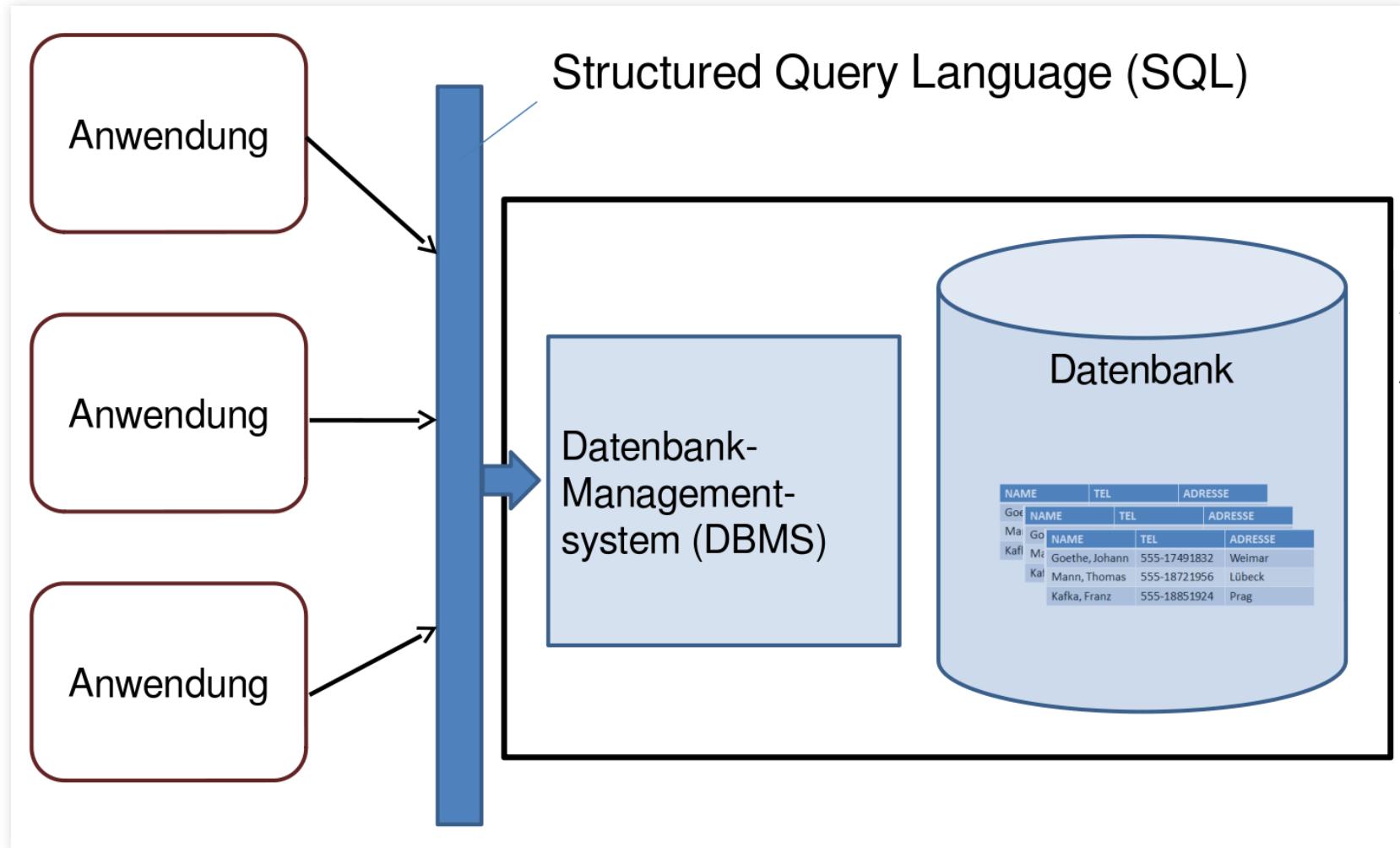
1. Datenbanken, wie und wozu?
2. Domäne: der zu modellierende Gegenstandsbereich
3. Konzeptuelles Modell: Entity-Relationship-Modell
4. Logisches Modell: Relational Database / Relationale Algebra
5. Implementierung: Structured Query Language

1. Datenbanken, wie und wozu?

(A) Datenbanken...

- ...eignen sich zur Organisation stark strukturierter Datenbestände
- ...erfordern Aufwand beim Design und Befüllen
- ...erlauben präzise Suchabfragen auf diesen Beständen
- ... erfüllen Grundbedürfnisse geisteswiss. Forschung nach Organisation, Speicherung und Abfragen von Informationen

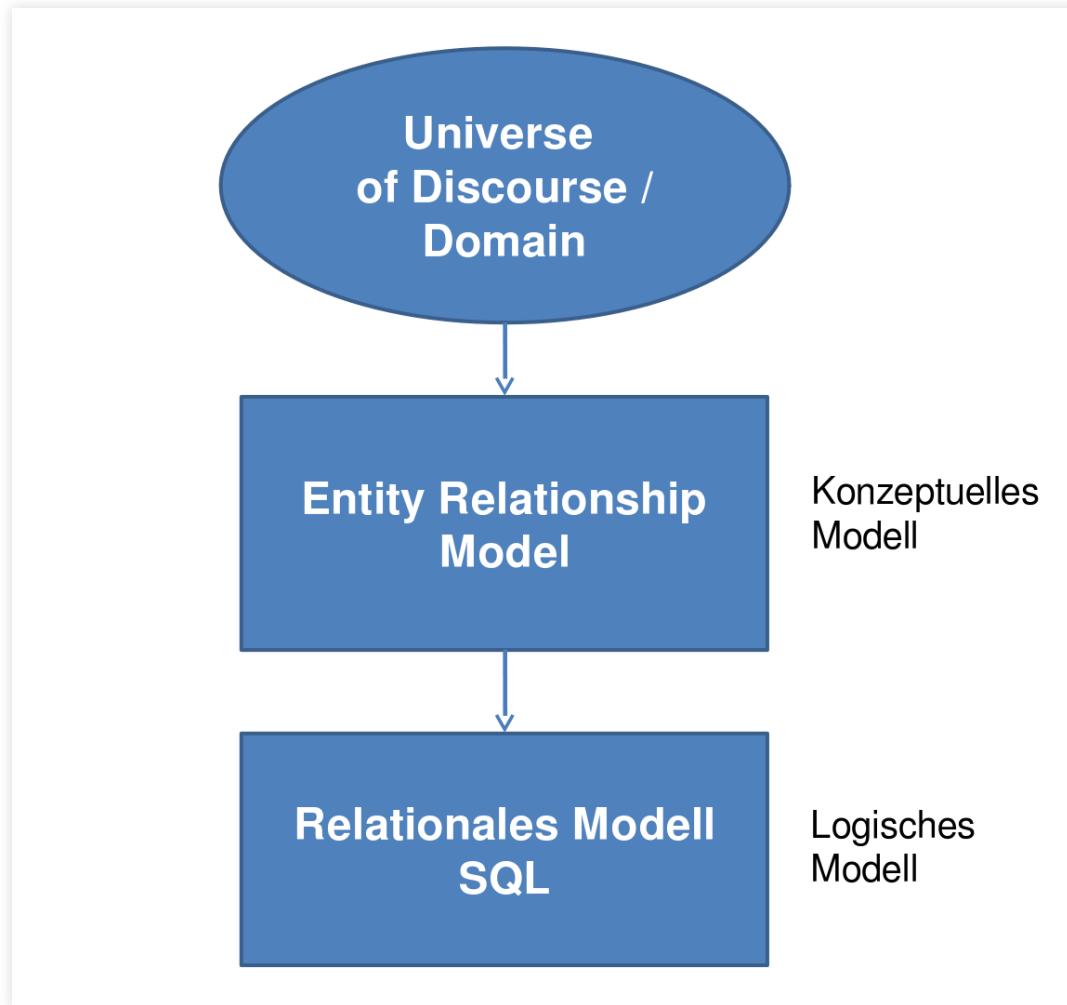
Datenbanksystem



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

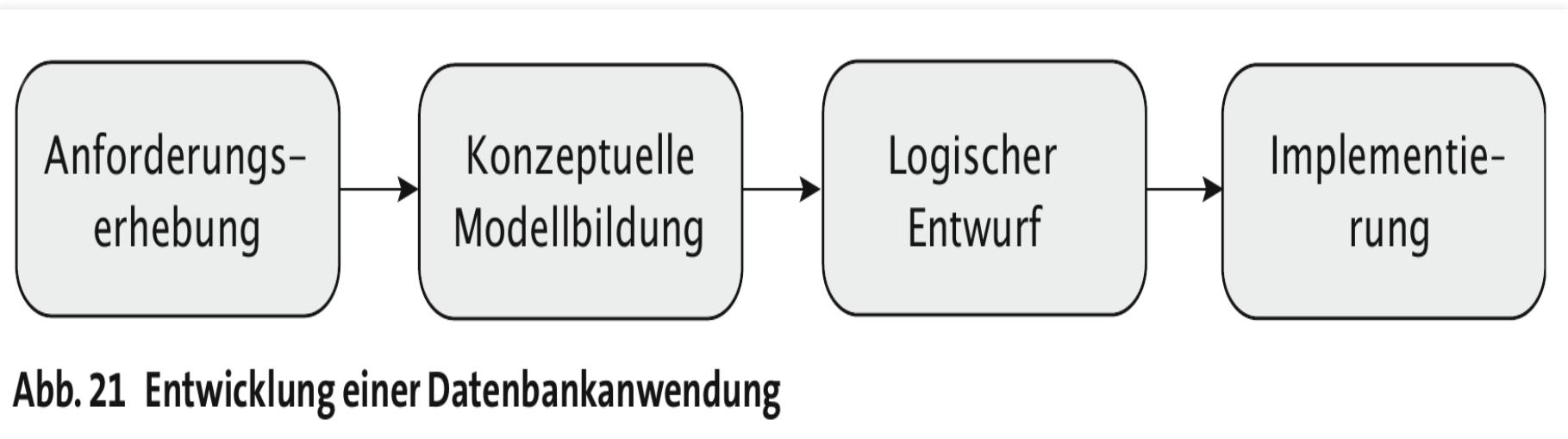


Abb. 21 Entwicklung einer Datenbankanwendung

(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

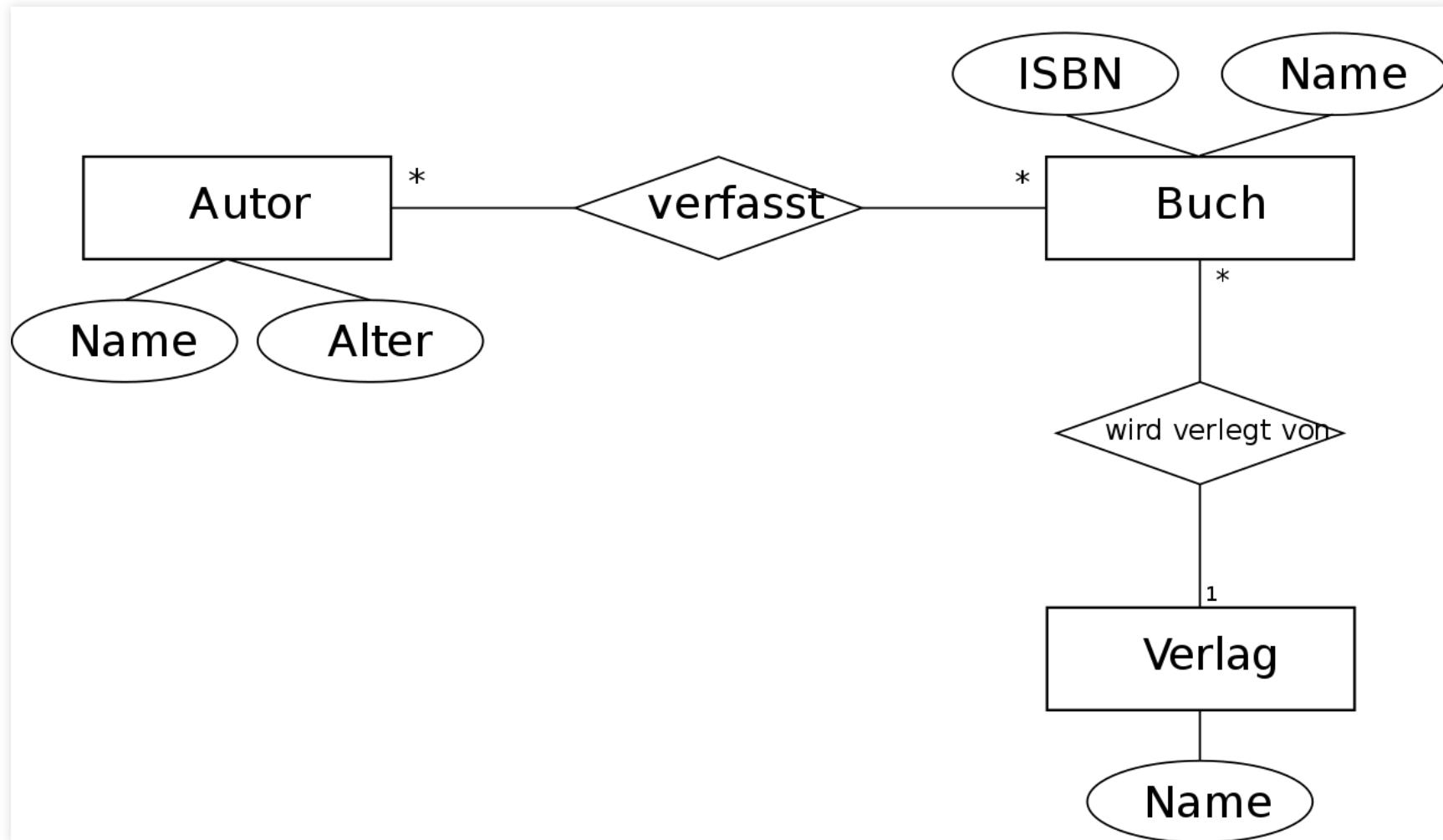
Vier Aufgaben des konzeptionellen Datenmodells

- Klassifizierung: Festlegung der Objekttypen / Entitäten
- Abstraktion: Bestimmung der relevanten Eigenschaften
- Beziehungen: Beschreibung der Zusammenhänge zwischen den Objekten
- Identifizierung: Festlegung von eindeutigen Namen / Eigenschaften

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)
- Beziehungen (inhaltlicher Zusammenhang zwischen Entitäten)
- Kardinalität von Beziehungen (mengenmäßige Beziehung)

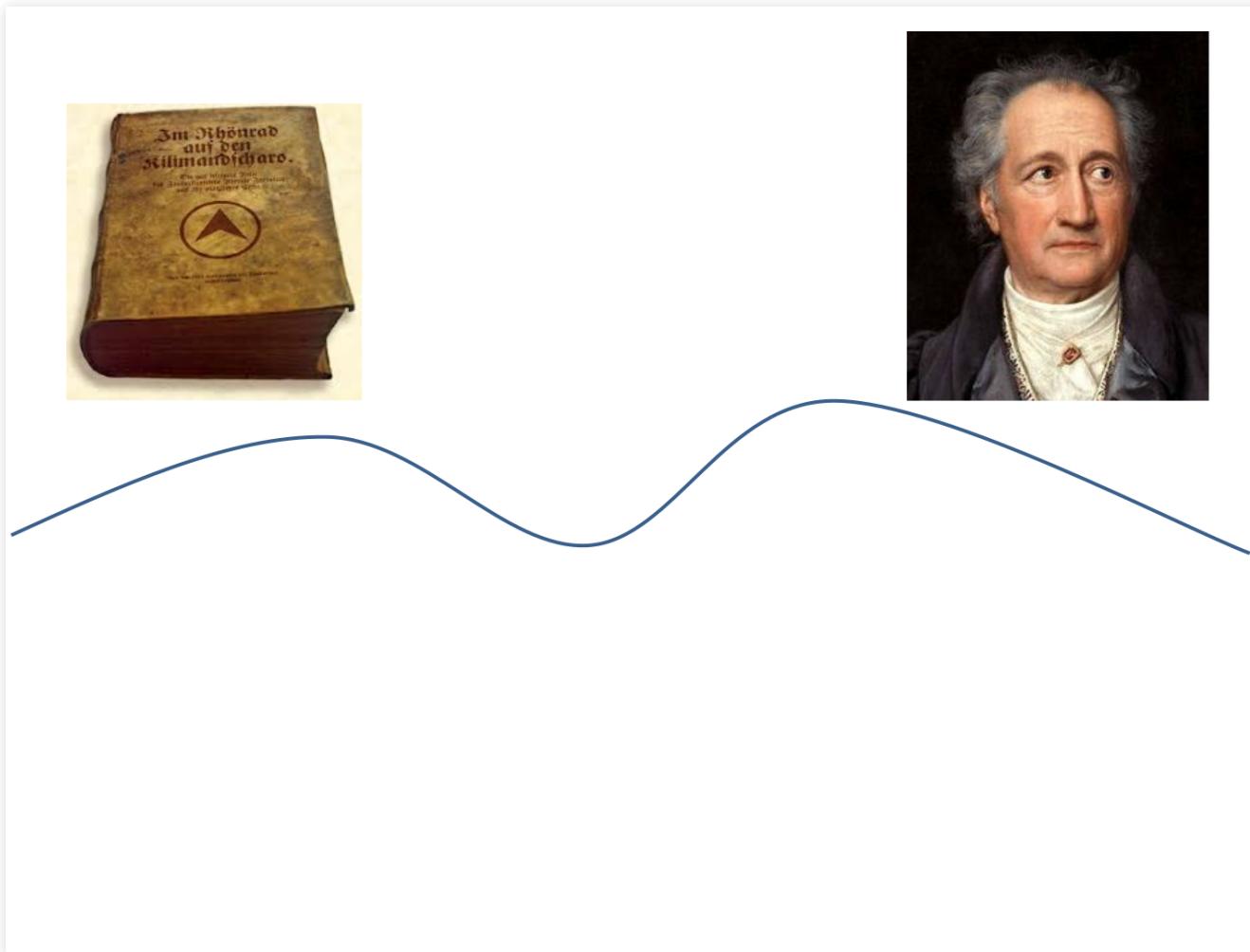
Kardinalität?



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

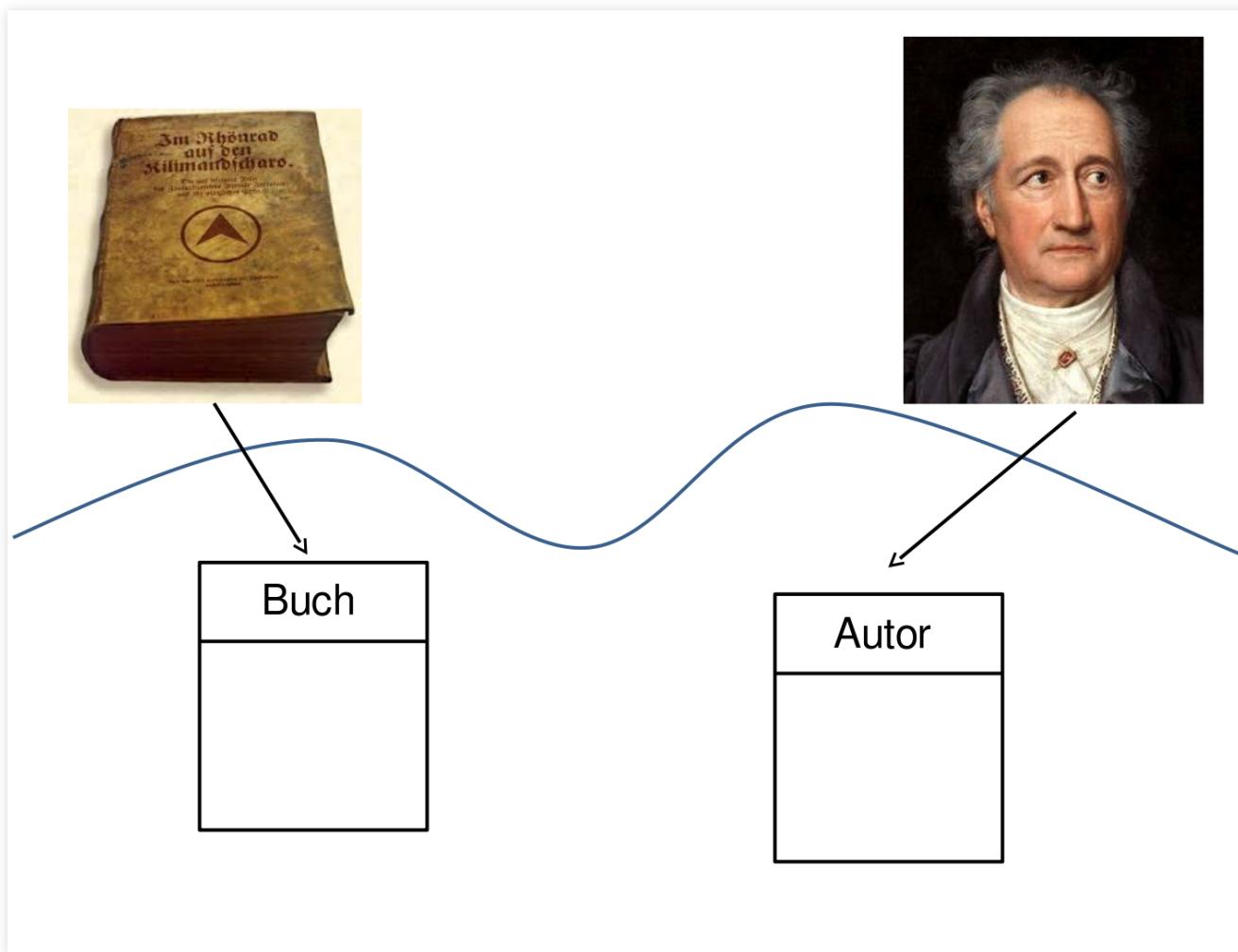
(B) Beispiel Bibliothek

Bibliothek: Bücher und Autoren



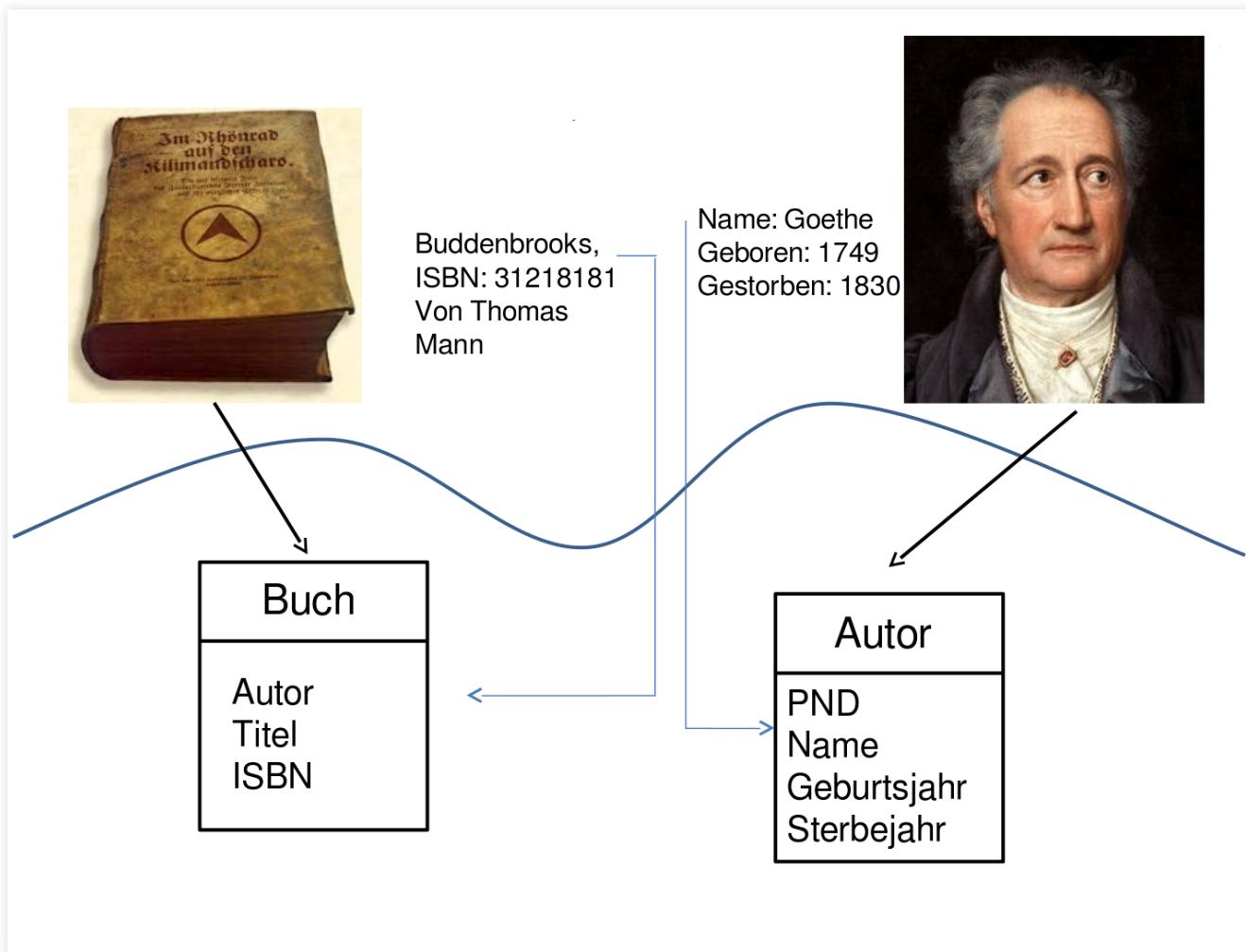
(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

Klassifikation



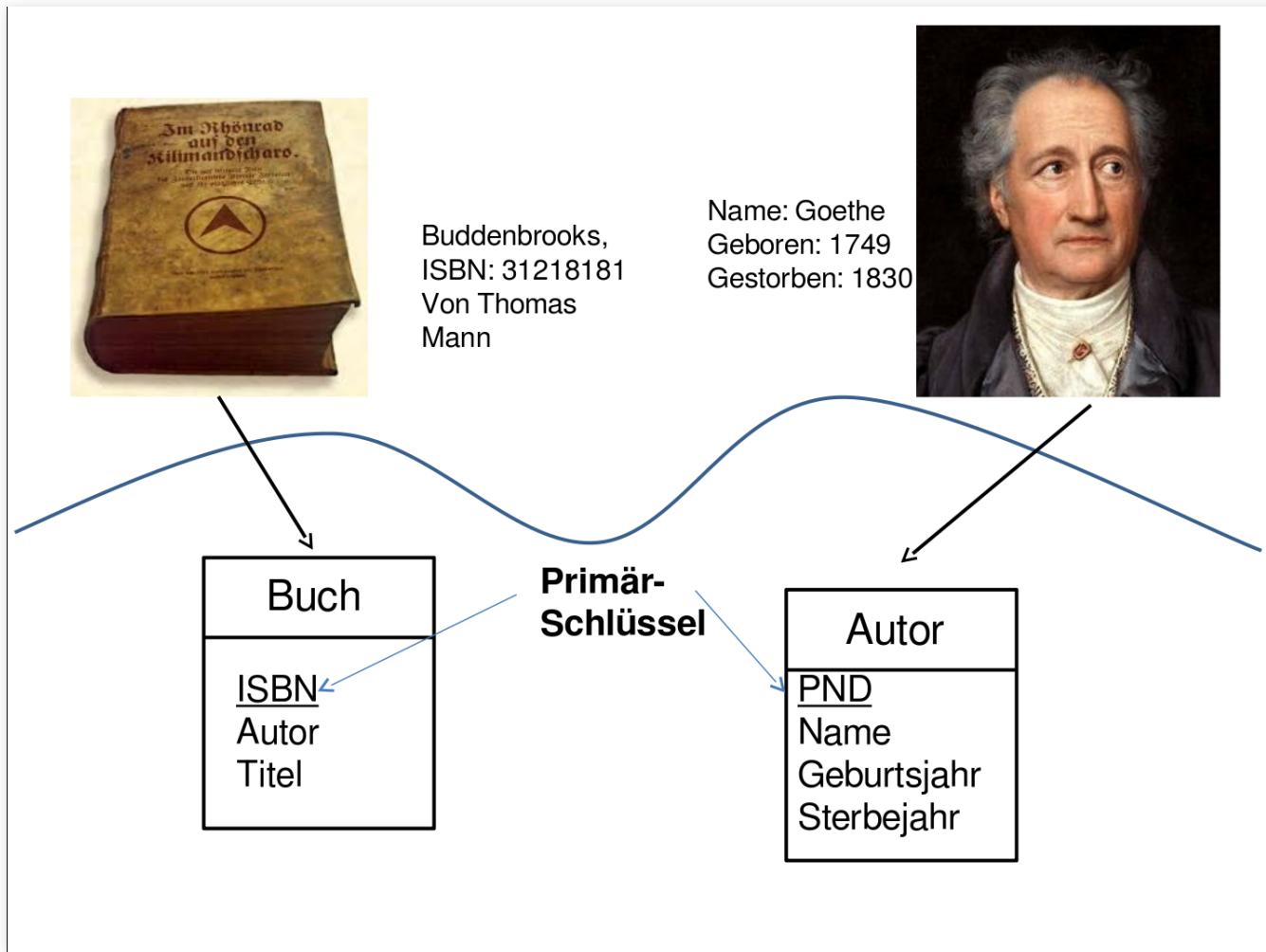
(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

Abstraktion



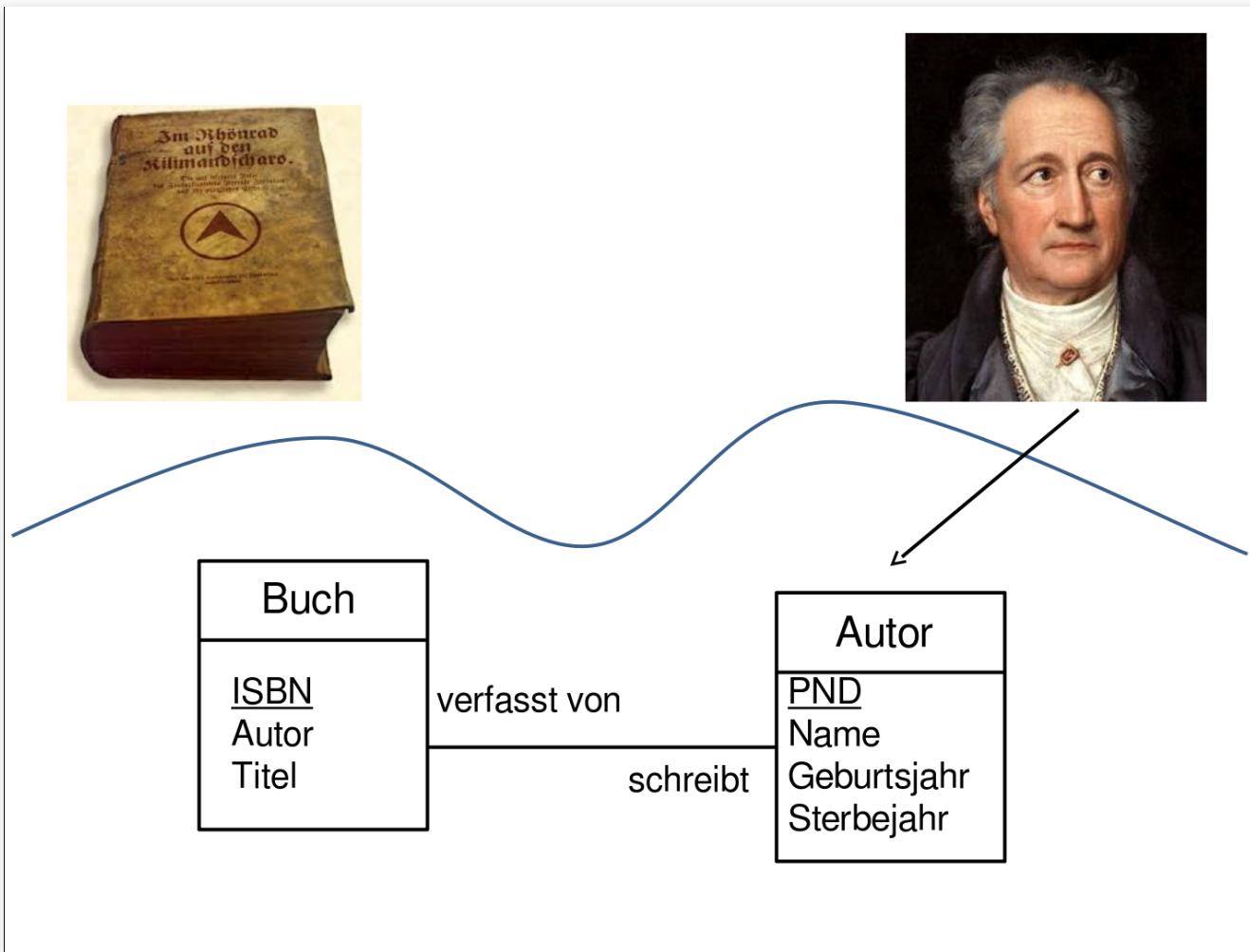
(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

Identifizierung



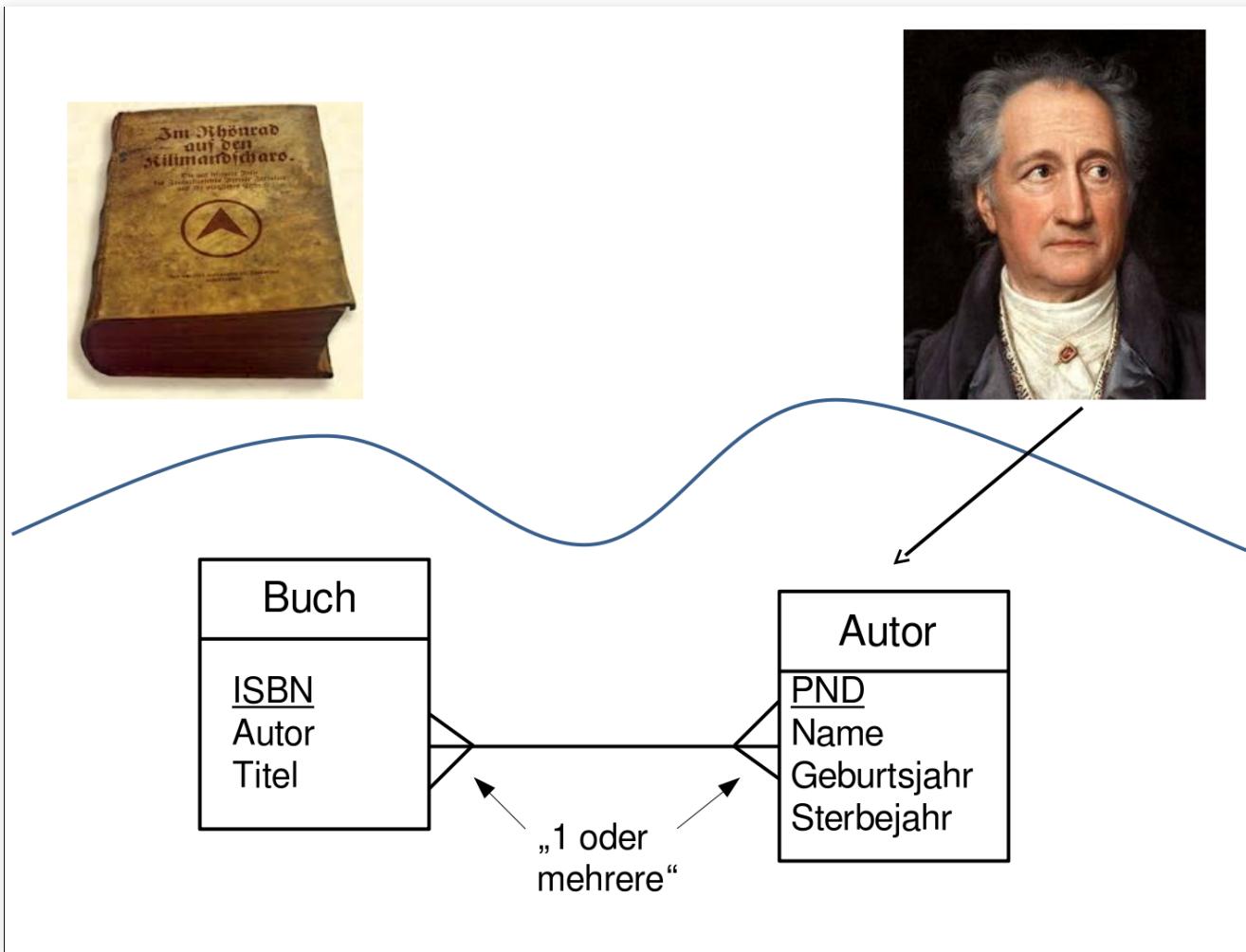
(Quelle für Beispiel und Darstellung: Fotis Jannidis, Würzburg. Bildquelle: Wiki Commons, [https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_\(Josef_Stieler\).jpg](https://commons.wikimedia.org/wiki/File:Johann_Wolfgang_von_Goethe_(Josef_Stieler).jpg)), gemeinfrei.

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

Was ist Normalisierung?

- Ziel ist die Reduktion von Redundanz
- "Normalformen" = Klassen von Qualitätskriterien
 - Erste Normalform: keine Wiederholungsgruppen; nur einfache Werte
 - Zweite Normalform: Attribute sind "voll funktional abhängig" vom (ganzen) Primärschlüssel
 - Dritte Normalform: Keine transitiven Abhängigkeiten: aus einem Nichtschlüsselattribut folgt ein anderes Nicht-Schlüsselattribut
- Normalisierung hat ihre Grenzen: es kann auch ineffizient werden

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

Atomisierung (= 1. NF)

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

Auftrennung (+Primärschlüssel) (=2. NF)

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

(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) Assoziationstabelle

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

Autorschaftstabelle

GND	ISBN
19283746	1242829340229
98761234	1534932829103

(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 FROM Buecher  
WHERE GND=="98765"
```

- Zeige die Spalten "name" und "geburt" an;
- aus der Tabelle "autoren";
- und zwar für diejenigen Einträge,
bei denen "geburt" größer als 1700 ist.

Anwendungen

SQL in Python

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

Beispiel in LibreOffice Base

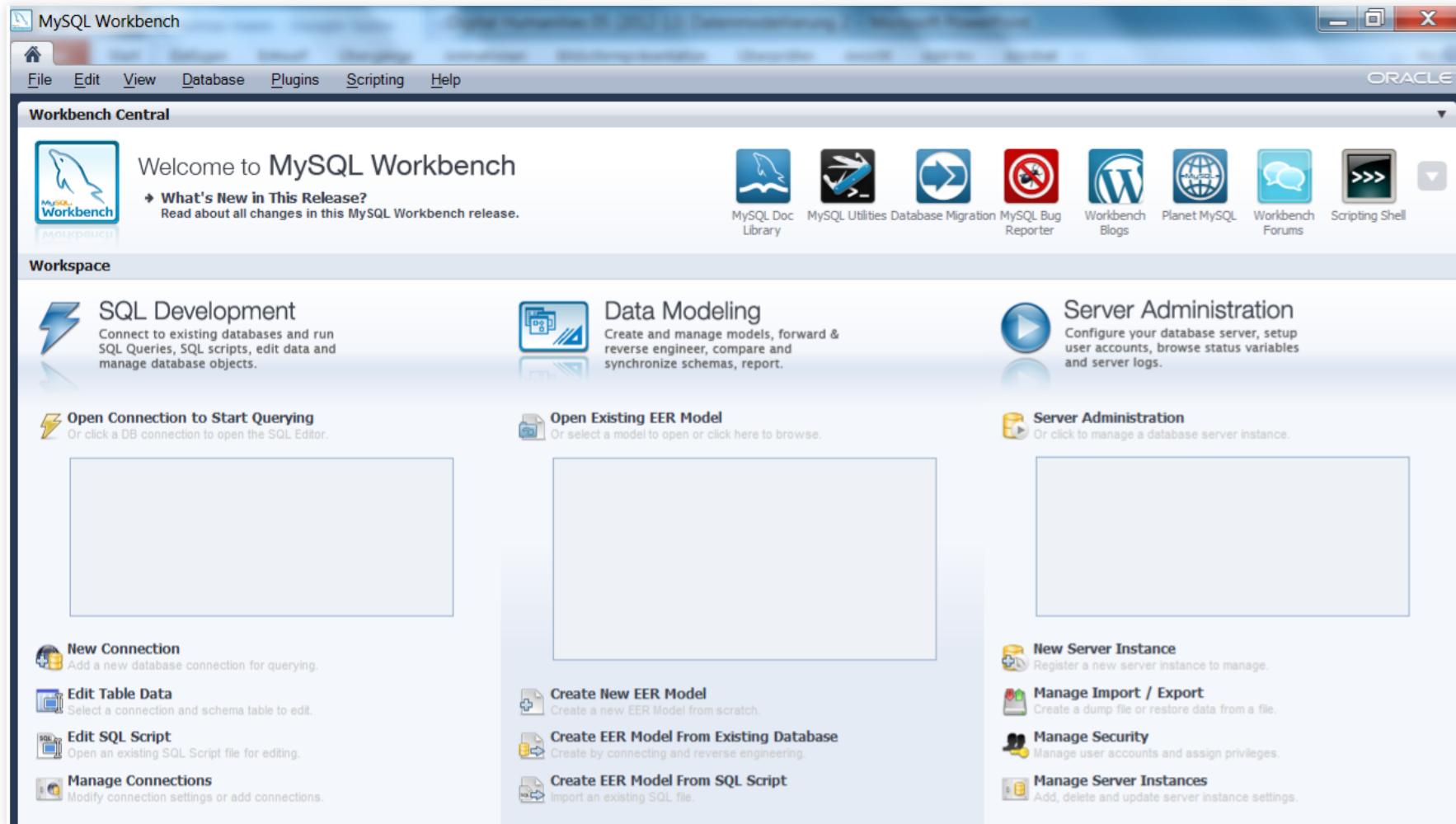
- Domäne / Anforderungen
 - Beziehungen zwischen Figuren in Theaterstücken
- Konzeptuelles Datenmodell (ER-Modell)
 - Entitäten: Figuren, Szenen
 - Relation: Anwesenheit von F in S
- Logisches Datenmodell (Tabellen)
 - Figuren (Name, Geschlecht, Status)
 - Szenen (Aktnummer, Szenennummer)
 - Relation (Szene, Figur)
- Implementierung
 - hier in Libre Office Base

Figurenkonstellation.odb - LibreOffice Base

The screenshot shows the LibreOffice Base interface with the following components:

- Database Sidebar:** Shows icons for Tables (selected), Queries, Forms, and Reports.
- Tasks:** A list of options including "Create Table in Design View...", "Use Wizard to Create Table...", and "Create View...".
- Tables:** A list of tables: Anwesenheit, Figuren, and Szenen.
- Relation Design Window:** Displays the database schema with three tables:
 - Figuren:** Contains fields Figur-ID, Name, Geschlecht, and Status. It has a one-to-many relationship with Anwesenheit (Figur-ID) and a many-to-one relationship with Szenen (Szenen-ID).
 - Szenen:** Contains fields Szenen-ID, Akt, and Szene. It has a many-to-one relationship with Anwesenheit (Szenen-ID) and a one-to-many relationship with Figuren (Szenen-ID).
 - Anwesenheit:** Contains fields ID, Szenen-ID, and Figur-ID. It connects the other two tables.
- Szenen Table View:** Shows data for the Szenen table with columns Szenen-ID, Akt, and Szene. Data rows include (1, 1, 1), (1, 1, 2), (1, 1, 3), (1, 1, 4), (2, 1, 1), (2, 2, 2), (2, 2, 3), (2, 2, 4), and (2, 2, 5).
- Figuren Table View:** Shows data for the Figuren table with columns Figur-ID, Name, Geschlecht, and Status. Data rows include (0, Andromache, weiblich, adelig), (1, Pyrrhus, maennlich, adelig), (2, Hermione, weiblich, adelig), (3, Pylade, maennlich, adelig), (4, Cléone, weiblich, nicht-adelig), (5, Céphise, weiblich, nicht-adelig), (6, Phoenix, weiblich, nicht-adelig), (7, Suite Ores, diverse, nicht-adelig), and (8, Oreste, maennlich, adelig).
- Anwesenheit Table View:** Shows data for the Anwesenheit table with columns ID, Szenen-ID, and Figur-ID. Data rows include (0, 0, 8), (1, 0, 3), (2, 1, 1), (3, 1, 8), (4, 1, 6), (5, 2, 1), (6, 2, 6), (7, 3, 1), (8, 3, 0), and (9, 3, 5).

MySQL Workbench



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