

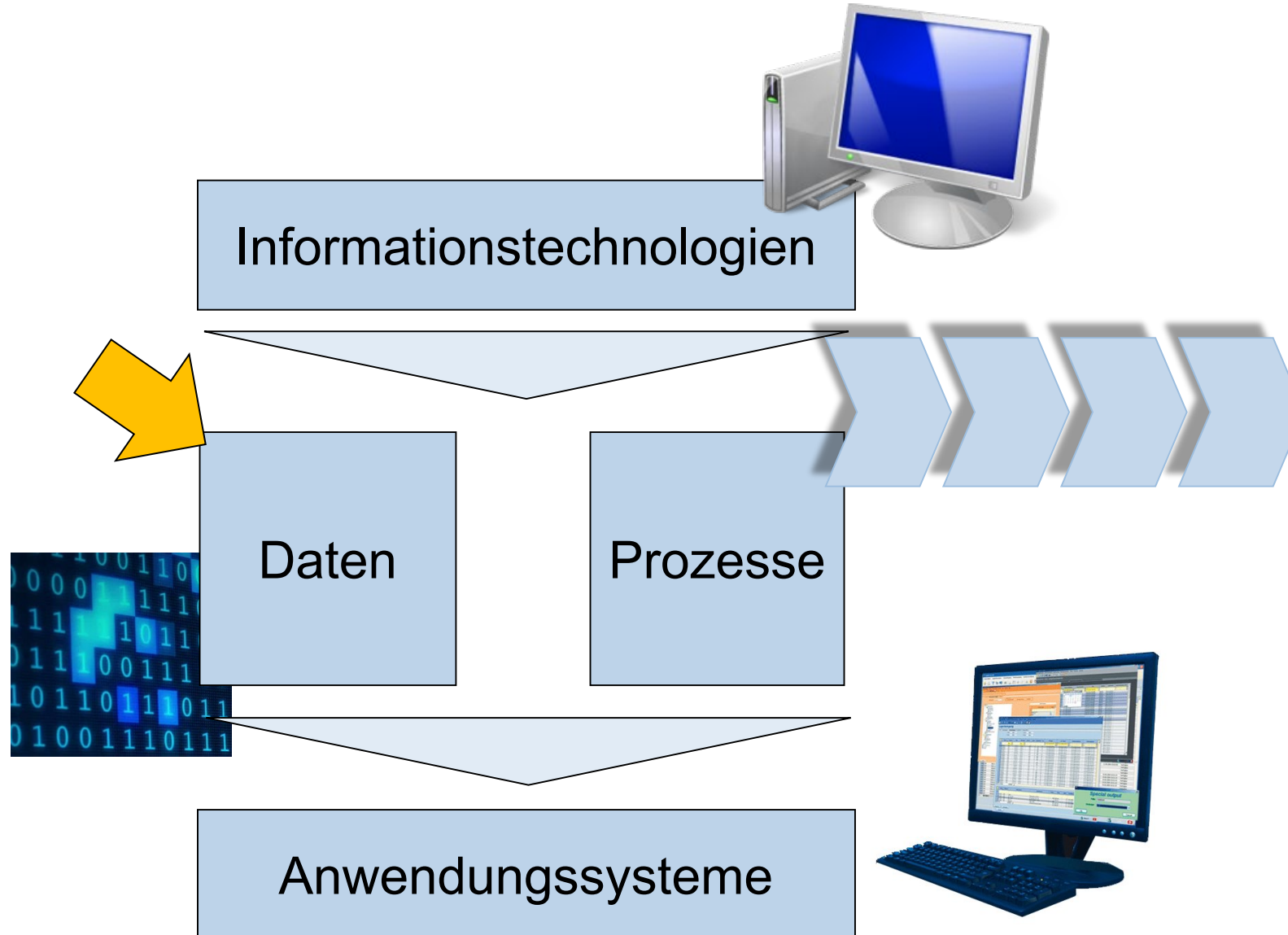
# Einführung in die Wirtschaftsinformatik

## Dateien und Datenbanksysteme: *Strukturierte und unstrukturierte Dateien*

**Prof. Dr. Thomas Myrach**  
**Universität Bern**  
**Institut für Wirtschaftsinformatik**  
**Abteilung Informationsmanagement**

Bern, 18. März 2020

# Logischer Aufbau



# Die «Informations-Kutsche»:

## Das Papierdokument

- Unser Umgang mit Informationen ist bis heute sehr stark geprägt von der Nutzung verschiedenartiger Dokumente.
- Diese werden durch ein bestimmtes Trägermedium charakterisiert, vor allem Papier.
- Inhalte erhalten durch die physische Bindung an das Trägermedium auch eine logische Einheit.
- Diese schlägt sich etwa nieder in Briefen, Dossiers, Akten, Büchern.



# Aufbewahrung von Aufzeichnungen

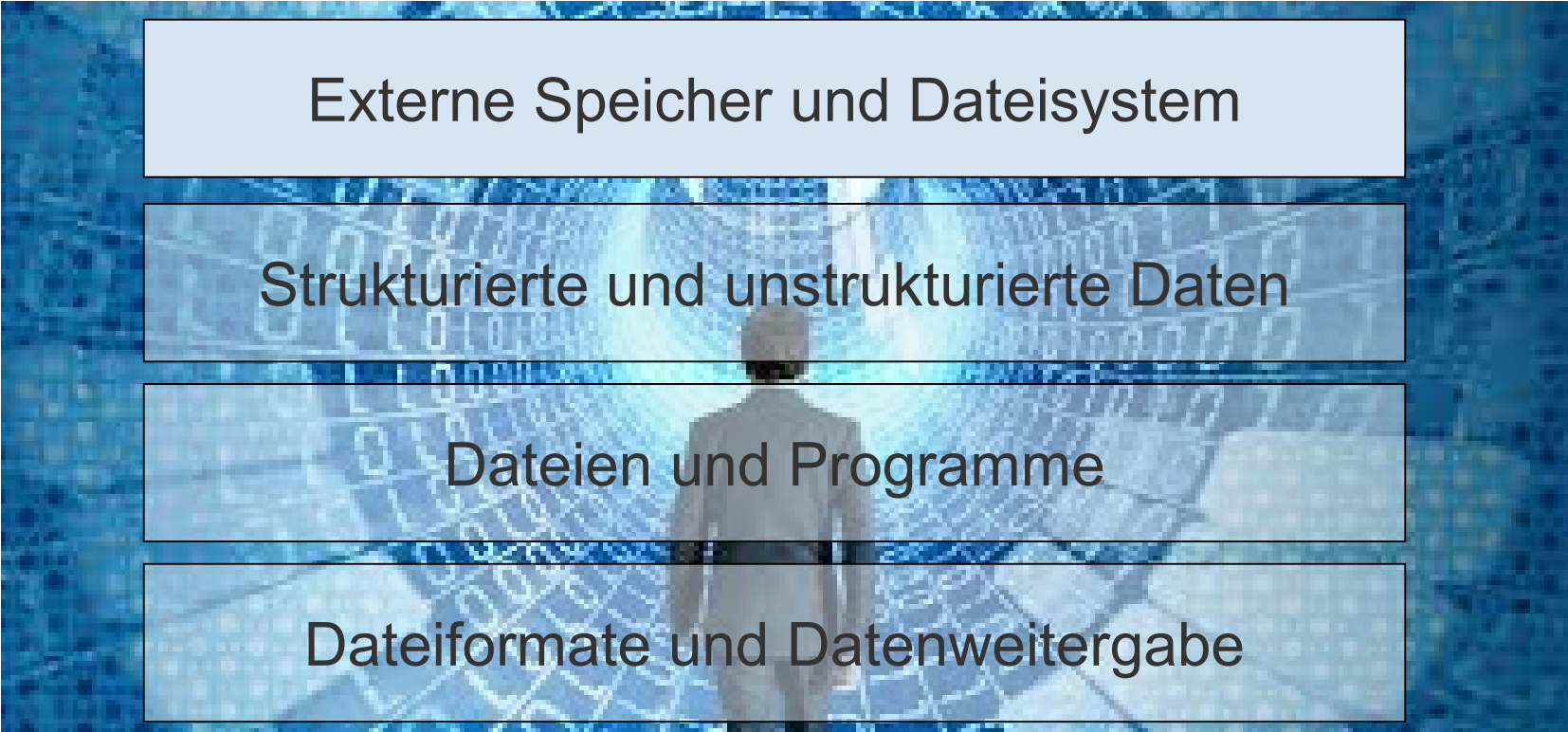


- Ablage für Aufzeichnungen in Papierform.
- Elemente:
  - Aktenschrank
  - Akte/Dossier
  - Dokument
  - Dokumentelement



- Ablage für Aufzeichnungen in digitaler Form.
- Elemente
  - Dateiverzeichnis
  - Datei
  - Datensatz
  - Datenelement

- Sie haben einen Eindruck von der Analogie herkömmlicher Dokumentablagen und Dateiverwaltungssystemen.
- Sie wissen, dass Dateien und Dateiverzeichnisse Abstraktionen von Betriebssystemen über Hardware-Komponenten sind.
- Sie kennen den Unterschied zwischen strukturierten und unstrukturierten Daten.
- Sie wissen, wie der Zusammenhang zwischen Programmen und Dateien ist.
- Sie können Beispiele für die Nutzung von Dateiformaten durch verschiedene Programme nennen.
- Sie kennen das Konzept von Open Data
- Sie können das Five-Star-Modell für die Eignung unterschiedlicher Dateiformate für Open Data beschreiben.



Externe Speicher und Dateisystem

Strukturierte und unstrukturierte Daten

Dateien und Programme

Dateiformate und Datenweitergabe

# Aufbewahrung von Aufzeichnungen



- Ablage für Aufzeichnungen in Papierform.
- Elemente:
  - Aktenschrank
  - Akte/Dossier
  - Dokument
  - Dokumentelement



- Ablage für Aufzeichnungen in digitaler Form.
- Elemente
  - Dateiverzeichnis
  - Datei
  - Datensatz
  - Datenelement

# Externer Speicher

- Im Rahmen eines Rechnersystems stehen typischerweise externe Speicher zur Verfügung.
- Dies hat verschiedene Vorteile:
  - Die beschränkte Kapazität des Hauptspeichers wird ausgeweitet.
  - Die Daten auf externen Speichern sind persistent.
- Extern heisst, dass der Prozessor nicht direkt auf den Speicher zugreifen kann.

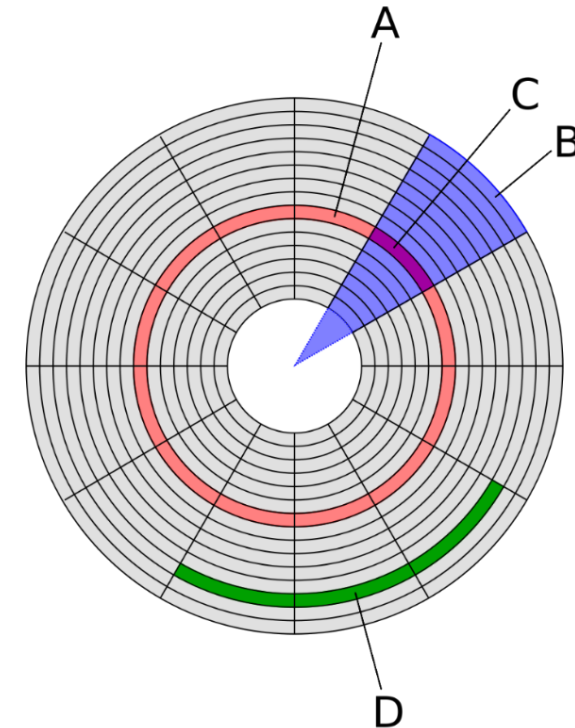




# Externer Speicher

## Beispiel: Festplatten

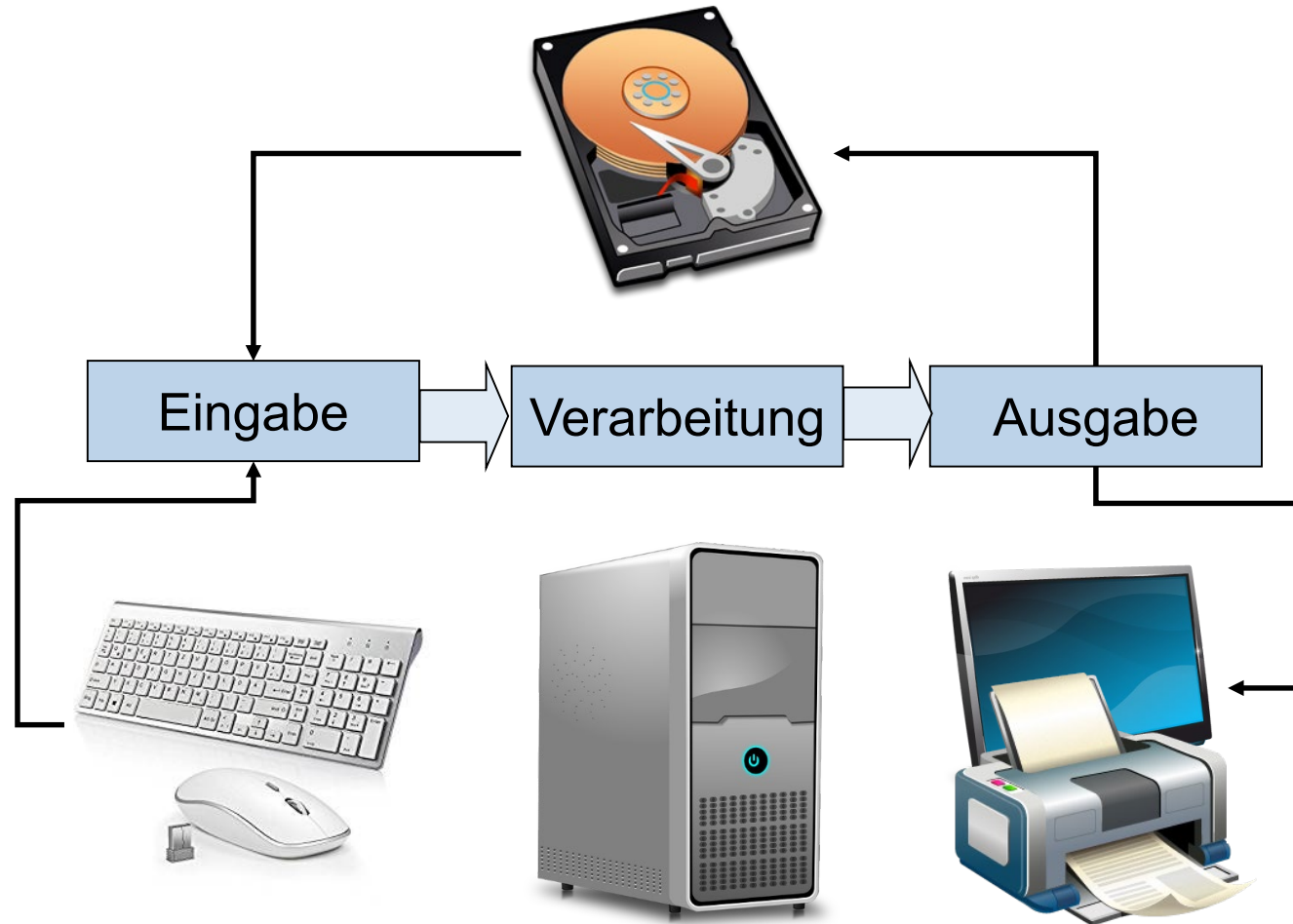
- Festplatten organisieren ihre Daten in Datenblöcken.
- Ein Block kann etwa 512, 2048 oder 4096 Byte umfassen.
- Rechner können immer nur ganze Datenblöcke oder Sektoren lesen und schreiben.
- Eine Datei kann sich über mehrere Blöcke erstrecken.
- Auf der Festplatte ist vermerkt, in welchen Block eine Datei beginnt.



- (A) Spur (auch Zylinder),
- (B) Sektor,
- (C) Block,
- (D) Cluster.

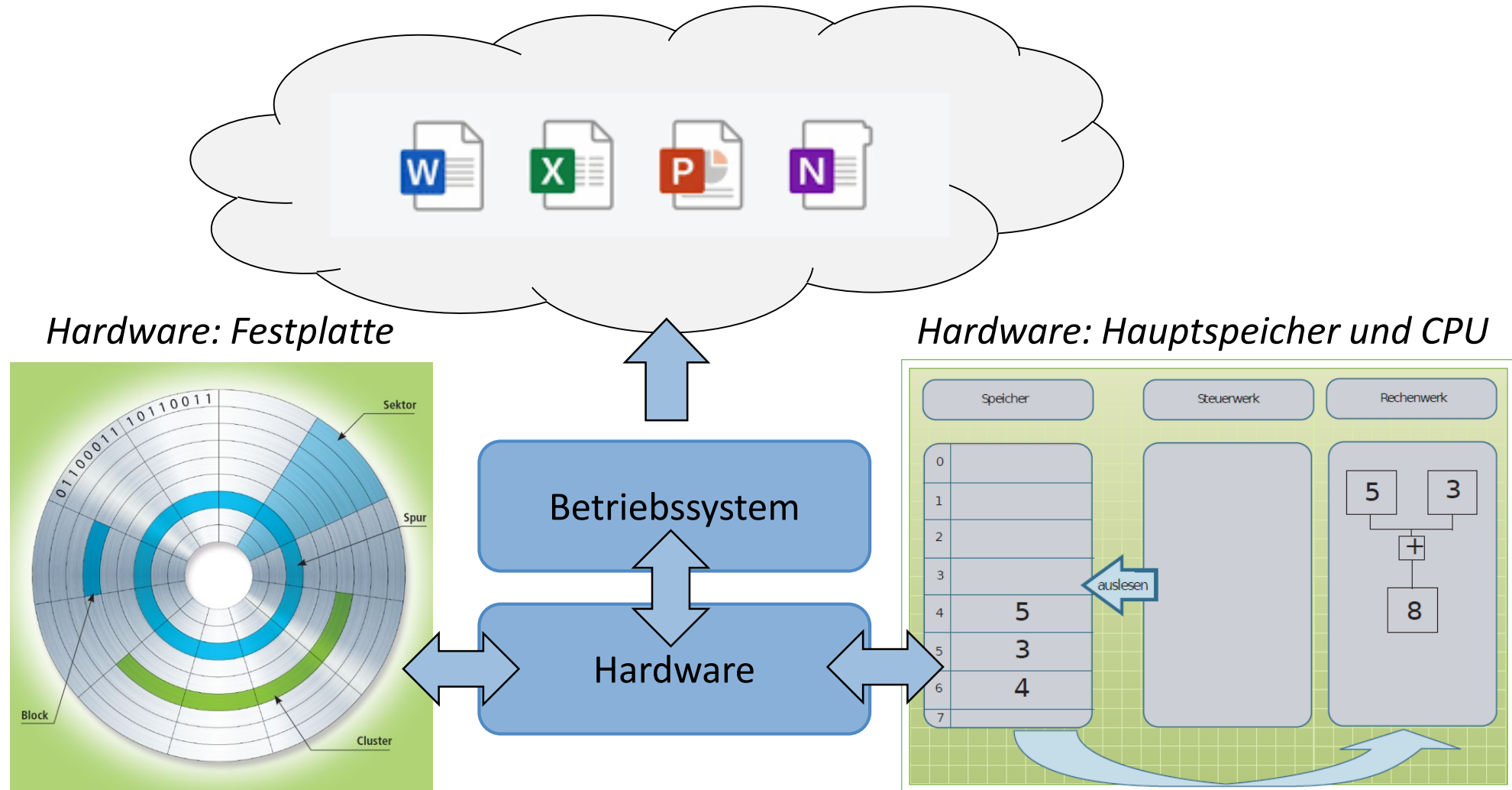
- Im Rahmen von Programmen muss es möglich sein, mit unterschiedlichen Daten zu arbeiten.
- Die in einem Programm verwendeten Daten müssen eingelesen und ausgegeben werden können.
- Dazu stehen in Programmiersprachen entsprechende Befehle zur Verfügung.
- Als Standard-Eingabemechanismus wird oftmals die Tastatur angenommen.
- Als Standard-Ausgabemechanismus gilt der Monitor.
- Die Eingabe oder Ausgabe von Daten kann auch ein Speichermedium betreffen.
- Auf diesen werden Daten dauerhaft (persistent) abgelegt.

# Input-Process-Output



- Menge zusammengehöriger gleichartiger Daten auf einem externen Speichermedium.
- Ablage und Zugriff auf Dateien erfolgen über ein Betriebssystem.
- Dateien werden durch das Betriebssystem über einen Namen adressiert.
- Die Codierung von Daten in einer Datei wird durch ein Dateiformat festgelegt.
- Das Dateiformat wird vom Betriebssystem angezeigt.
- Dies geschieht üblicherweise durch eine Namenserverweiterung (Extension).

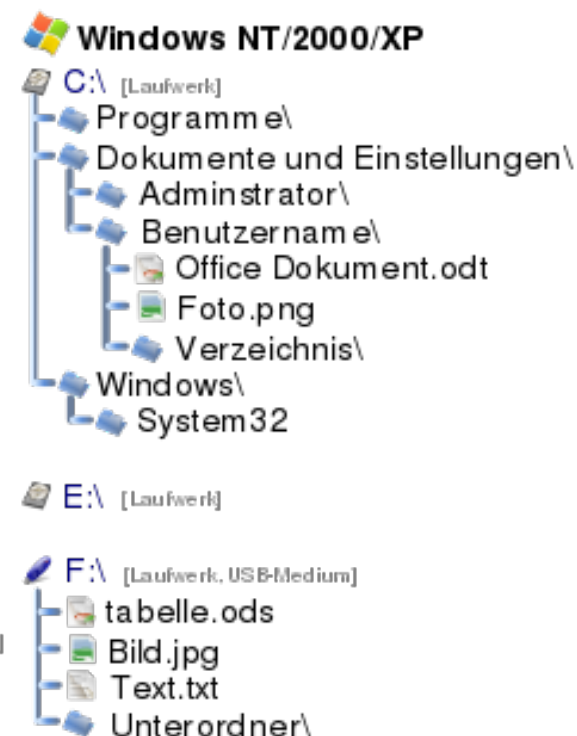
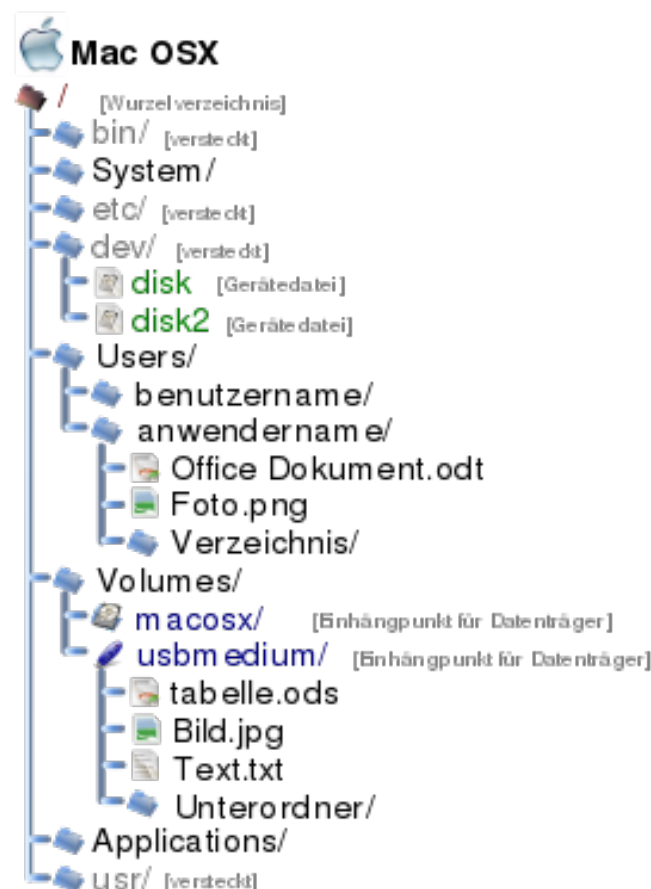
# Dateien als Abstraktion



## Dateiverzeichnisse

- Typischerweise können in Betriebssystemen hierarchische Dateiverzeichnisse aufgebaut werden.
- Jede Datei ist eindeutig in einem Dateiverzeichnis verordnet.
- Über Dateiverzeichnisse können Ordnungssysteme aufgebaut werden.
- Ordnungssysteme erleichtern das Auffinden von Dateien.
- Dateiverzeichnisse können sich über verschiedene physische und logische Speichermedien erstrecken.

# Dateiverzeichnisse in Betriebssystemen



Quelle: <https://de.wikipedia.org/wiki/Dateisystem>

- Externe Speicher sind ein zentrales Element von Rechnerarchitekturen.
- Sie erlauben eine persistente Speicherung von Daten.
- Externe Speicher müssen durch den Zentralprozessor eines Systems angesprochen werden können.
- Dies geschieht über Betriebssysteme.
- Betriebssysteme stellen Dateien als Abstraktion von zusammengehörenden Datenblöcken auf einem externen Speichermedium zur Verfügung.
- Darüber hinaus bieten Betriebssysteme typischerweise hierarchische Dateiverzeichnisse.
- Dateiverzeichnisse erlauben die Gestaltung eines Ordnungssystems für die einzelnen Dateien.





Externe Speicher und Dateisystem

Strukturierte und unstrukturierte Daten

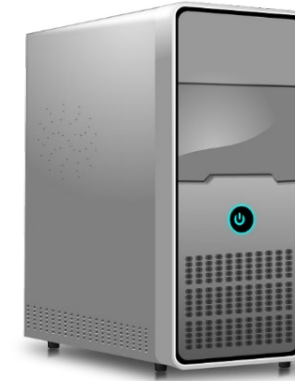
Dateien und Programme

Dateiformate und Datenweitergabe

# Aufbewahrung von Aufzeichnungen



- Ablage für Aufzeichnungen in Papierform.
- Elemente:
  - Aktenschrank
  - Akte/Dossier
  - Dokument
  - Dokumentelement



- Ablage für Aufzeichnungen in digitaler Form.
- Elemente
  - Dateiverzeichnis
  - Datei
  - Datensatz
  - Datenelement

# Dateien und Datenstrukturen

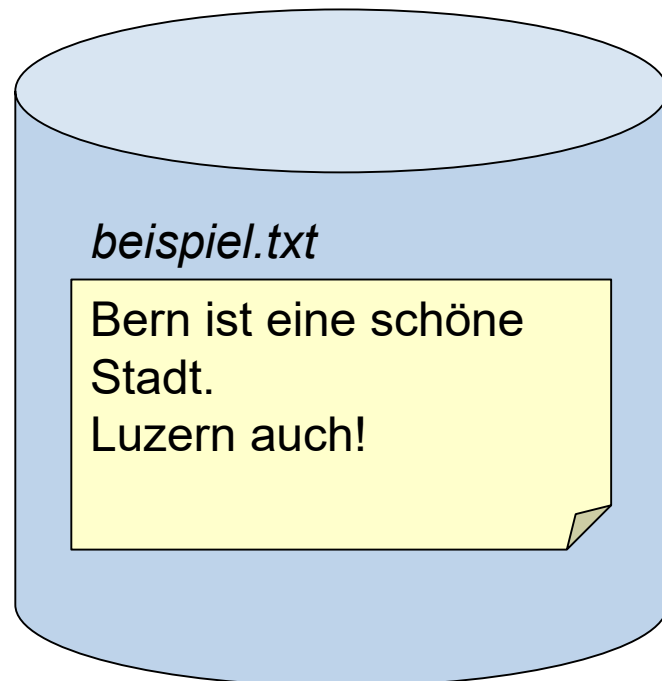
## Strukturierte und unstrukturierte Daten

- Unstrukturierte Daten
  - sind logisch nicht weiter unterteilt und sie haben keine formalisierte Struktur.
  - auf sie kann von Computerprogrammen nicht gezielt auf bestimmte Teile zugegriffen werden.
  - Die automatische Verarbeitung unstrukturierter Daten ist dadurch eingeschränkt.
- Strukturierte Daten
  - Sind in einer bestimmten Art und Weise angeordnet und verknüpft.
  - Erlauben den gezielten Zugriff auf bestimmte Teile, wie z.B. Datensätze oder Datenelemente.
  - Begünstigen den effizienten Zugriff und die Verwaltung.
  - Die automatische Verarbeitung strukturierter Daten ist möglich.

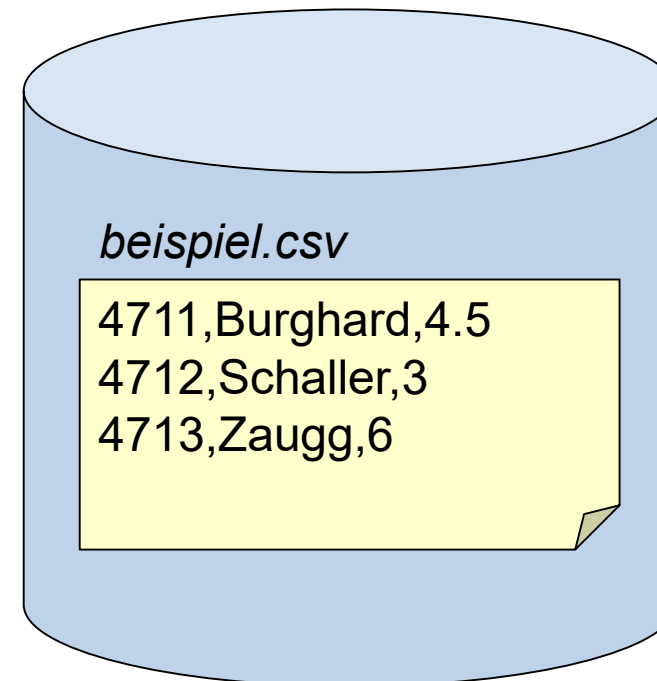
# Gegenüberstellung unstrukturierter und strukturierter Daten

## Beispiel: Textdaten versus CSV

Unstrukturiert (Text)



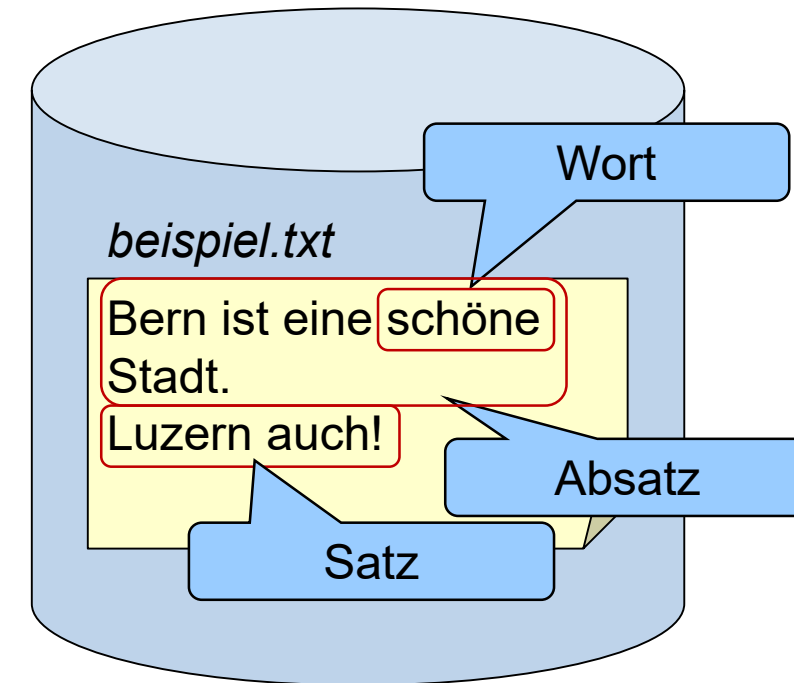
Strukturiert (CSV)



# Unstrukturierte Daten

## Beispiel: Textdateien

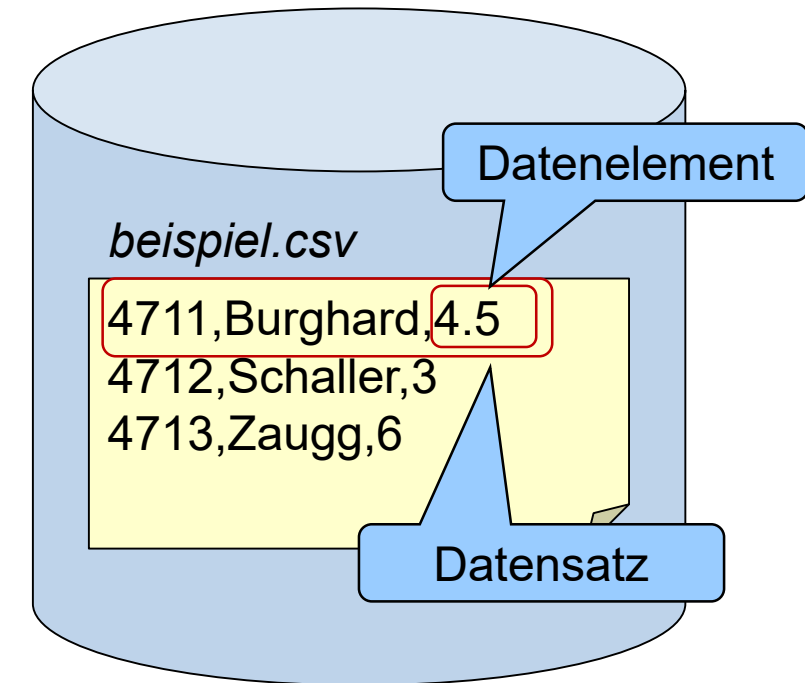
- Diese sind etwa in Windows mit der Erweiterung *\*.txt* gekennzeichnet.
- Textdateien sind eine sequentiell angeordnete Menge von Zeichen.
- Die einzelnen Zeichen sind z.B. nach ISO 7-bit codiert.
- Computerintern werden sie durch unstrukturierte Bitfolgen repräsentiert.
- Ein Computerprogramm erkennt nicht ohne weiteres, welche Zeichen ein Wort bilden oder einen Satz.
- Dies bedeutet: Der Mensch erkennt eine Struktur, der Computer nicht!



# Strukturierte Daten

## Beispiel: CSV-Dateien

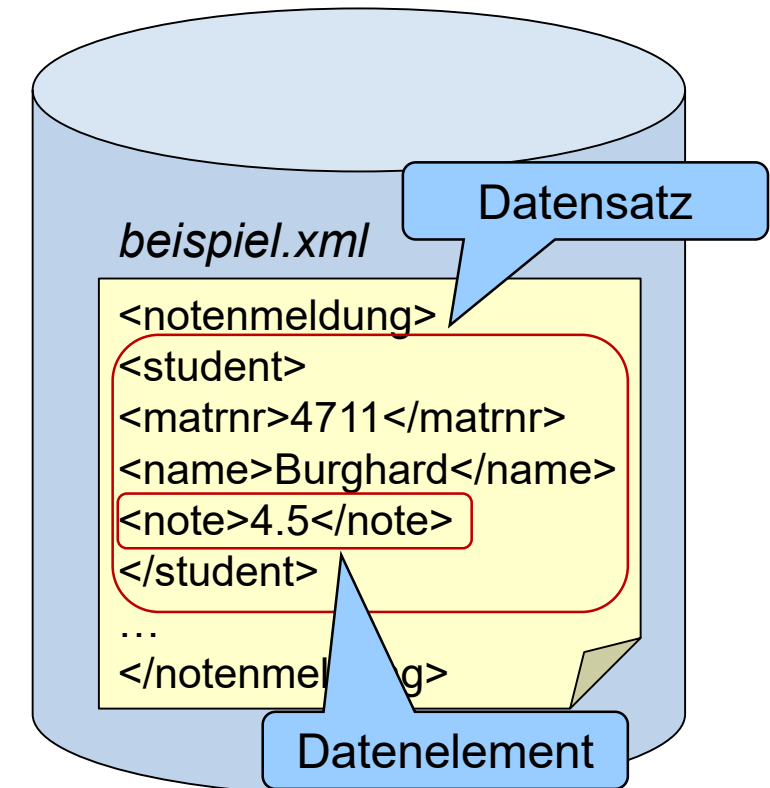
- CSV steht für Comma Separated Value.
- Die Trennung der einzelnen Datenwerte erfolgt durch ein Komma oder andere spezielle Zeichen.
- Aufgrund der Trennzeichen kann beim Einlesen von Daten erkannt werden, wann ein Datenelement beginnt bzw. endet.
- Die Semantik der Daten ist bei CSV nicht ohne weiteres ersichtlich.
- Daten werden "erkannt" auf Grund einer spezifischen Reihenfolge und der Trennzeichen.



# Strukturierte Daten mit Semantikinformation

## Beispiel: XML-Dateien

- XML steht für Extensible Markup Language.
- XML ist eine speziell für das Internet geschaffene Auszeichnungssprache.
- Datenelemente werden durch sog. Tags strukturiert.
- Die Tags können Aufschluss über die Semantik der Daten geben.
- Der Auszeichnungsmechanismus von XML ähnelt dem von HTML.
- Im Unterschied zu HTML erlaubt XML die Einführung von beliebigen Tags.



# Gegenüberstellung unstrukturierter und strukturierter Daten

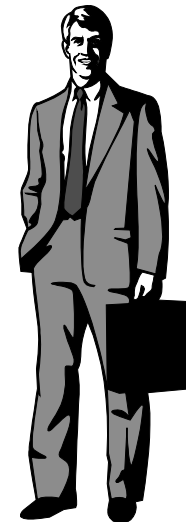
## Beispiel: Notenmeldung

### Unstrukturiert (Text)

#### NOTENMELDUNG

Vorlesnr: 4711  
 Vorlesung: *Digitale Welten*  
 Dozent: *Myrach*  
 Semester: *FS 2019*  
 ECTS: 4  
 Datum: 2019-03-20

Matrnr	Name	Vname	Note
8912307	<i>Müller</i>	<i>Jürg</i>	6.0
9056701	<i>Meier</i>	<i>Urs</i>	1.0



### Strukturiert (JSON)

```

{
  "vorlesnr": "4711",
  "vorlesung": "Digitale Welten",
  "dozent": "Myrach",
  "semester": "FS 2019",
  "ects": "4",
  "datum": "2019-03-20",
  "students": [
    { "matrnr": "8912307", "name": "Müller",
      "vname": "Jürg", "note": "6.0" },
    { "matrnr": "9056701", "name": "Meier",
      "vname": "Urs", "note": "1.0" }
  ]
}
  
```



# Eigenschaften unstrukturierter Daten

## Beispiel: Notenmeldung

Unstrukturiert (Text)

### NOTENMELDUNG

Vorlesnr: 4711  
Vorlesung: *Digitale Welten*  
Dozent: *Myrach*  
Semester: *FS 2019*  
ECTS: 4  
Datum: 2019-03-20

Matrnr	Name	Vname	Note
8912307	Müller	Jürg	6.0
9056701	Meier	Urs	1.0

- Formular als Textdokument.
- Bei Textdaten kann nur nach Zeichenketten gesucht werden.
- Beispiel:
  - Suche die Zeichenkette "Müller" oder Suche die Zeichenkette "6.0".
- Unspezifiziert:
  - "Müller" ist der Name eines Studierenden.
  - "6.0" ist die Note eines Studierenden in einer bestimmten Prüfung.
  - Ein Studierender mit dem Namen "Müller" hat die Note "6.0".

# Eigenschaften strukturierter Daten

## Beispiel: Notenmeldung

- Daten sind gemäss einer bestimmten Syntax strukturiert.
- Daten sind als Datenelemente gegliedert und diese zu Datensätzen gruppiert.
- Es kann gezielt nach Daten in ihrem Kontext gesucht werden.
- Beispiel:
  - Suche nach dem Studierenden mit dem Namen "Müller".
  - Dem Studierenden ist die Note "6.0" zugeordnet.

### Strukturiert (JSON)

```
{
  "vorlesnr":"4711",
  "vorlesung":"Digitale Welten",
  "dozent":"Myrach",
  "semester":"FS 2019",
  "ects":"4",
  "datum":"2019-03-20",
  "students":[
    {"matrnr":"8912307", "name":"Müller",
     "vname":"Jürg", "note":"6.0"},
    {"matrnr":"9056701", "name":"Meier",
     "vname":"Urs", "note":"1.0"}]
}
```

- Bei der Unterscheidung zwischen strukturierten und unstrukturierten Daten ist zwischen der Perspektive von Menschen und Maschinen zu unterscheiden.
  - Aus der Perspektive der Menschen können Inhalte durchaus strukturiert erscheinen.
  - Aus der Perspektive von Computerprogrammen stellen diese hingegen u.U. blosse unstrukturierte Abfolgen von Zeichen dar.
- Die Strukturierung von Daten ermöglicht Computerprogrammen die gezielte Bearbeitung von Dateninhalten.
  - Datenelemente können anhand ihrer jeweiligen Metadaten gezielt angesprochen werden.
  - Datenelemente können zu Datensätze gruppiert und damit ihr Zusammenhang verdeutlicht werden.
  - Durch die Gruppierung kann von einem Datenelement auf andere geschlossen werden.

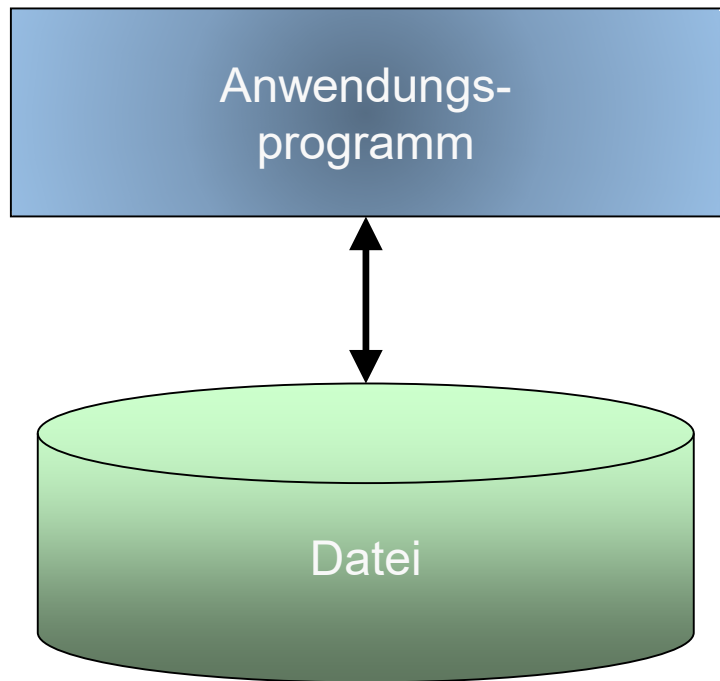


Externe Speicher und Dateisystem

Strukturierte und unstrukturierte Daten

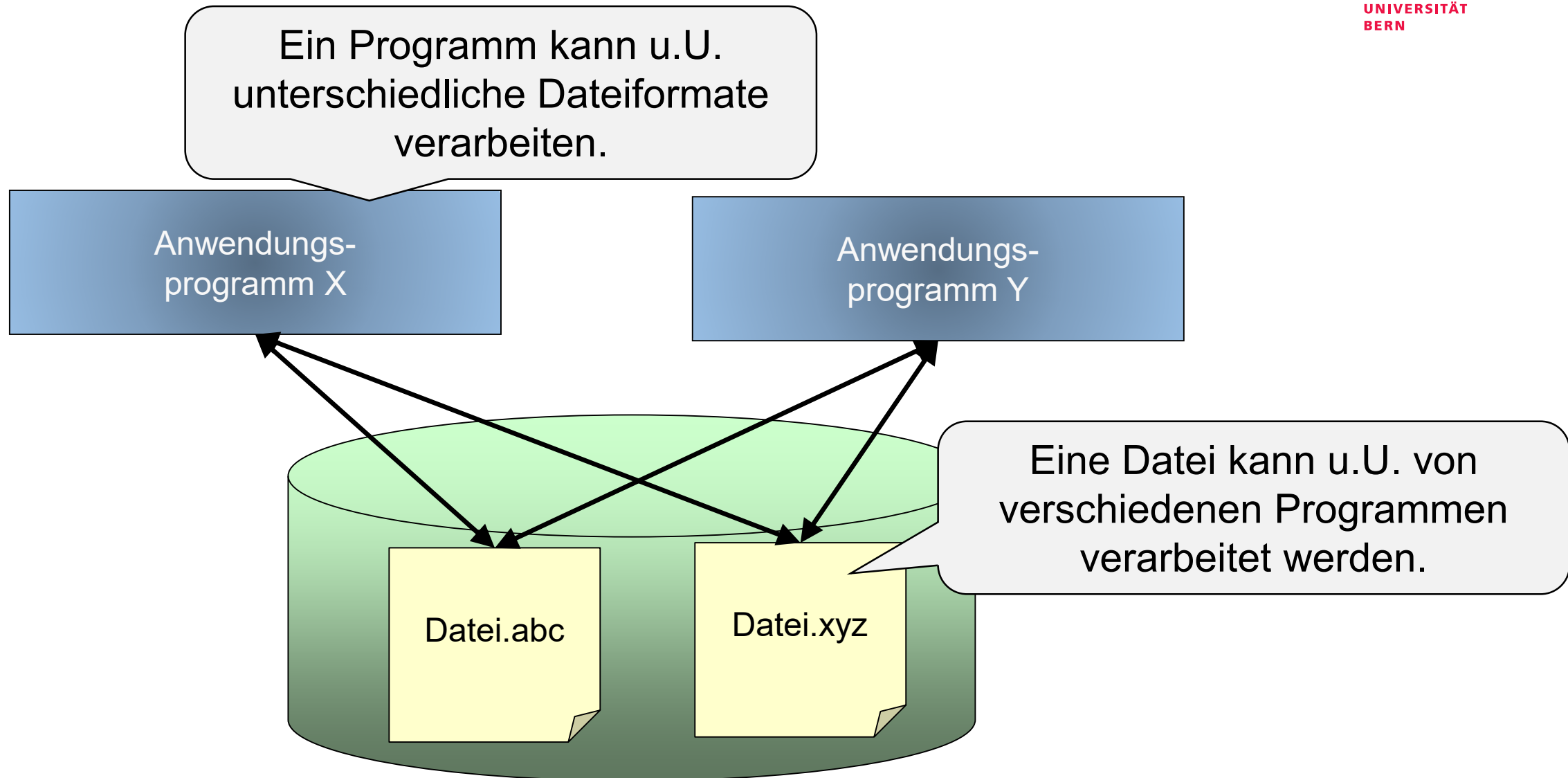
Dateien und Programme

Dateiformate und Datenweitergabe



- Die Verarbeitung von Daten erfolgt über Anwendungsprogramme.
- Anwendungsprogramme verwalten die von ihnen benötigten Daten herkömmlicherweise selber.
- Die von einem Programm verwalteten Daten entsprechen den Anforderungen des Programms.
- Das können allgemein bestimmte Dateiformate sein oder Dateien mit spezifischen Datenstrukturen.

# Zusammenhang Dateien und Programme



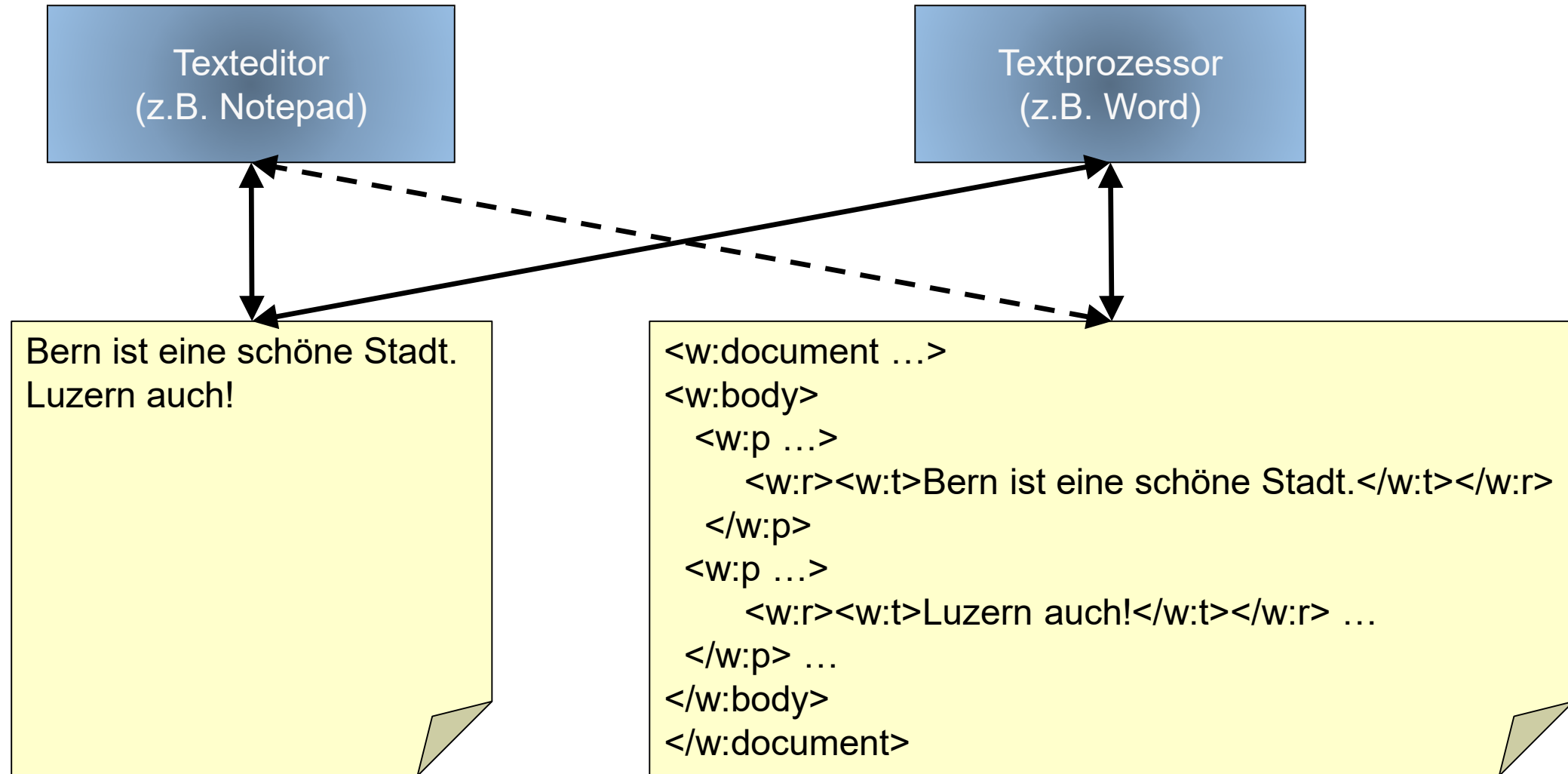
# Verarbeitung von Textdateien

## Reine und formatierte Textdaten

- Reine Textdateien enthalten lediglich die Texte in Form von codierten Zeichen.
- Sie lassen sich mit einfachen Programmen wie unter Windows etwa Notepad erstellen und bearbeiten.
- Textverarbeitungsprogramme wie Word benutzen komplexere Dateiformate, weil sie unter anderem auch Layout-Informationen abbilden.
- Das Textverarbeitungsprogramm Word arbeitet mit einem eigenen Dateiformat (DOC bzw. DOCX).
- DOCX-Dateien sind eigentlich ZIP-Dateien, in denen mehrere Dateien enthalten sind.
- Die zentrale Inhaltsdatei ist die "document.xml" Datei.
- Word unterstützt auch reine TXT-Daten.

# Verarbeitung von Textdateien

## Beispiel: TXT- und XML-Dateien





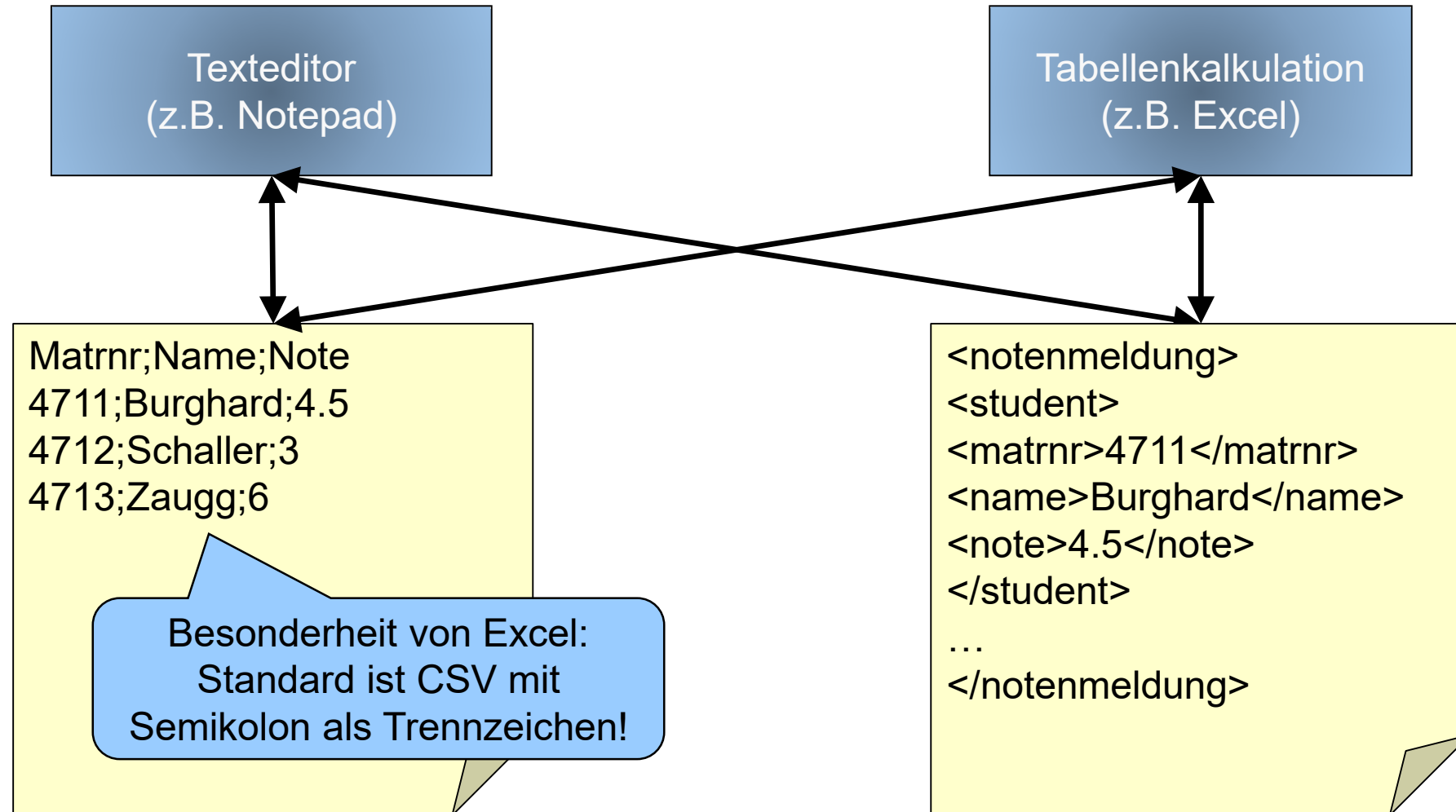
# Verarbeitung von (strukturierten) Daten

## Daten in Tabellenkalkulationsprogrammen

- Strukturierte Daten in Form von CSV- oder XML-Dateien können als reine Textdateien erstellt werden.
- Das Tabellenkalkulationsprogramm Excel arbeitet mit einem eigenen Dateiformat (XLS bzw. XLSX).
- Dieses Dateiformat wird vielfach für den Austausch von Daten verwendet.
- Excel unterstützt auch das CSV-Dateien, die erstellt und gelesen werden können.
- Dies beinhalten aber keine Formeln und Formatierungen.
- Darüber hinaus kann Excel mit gewissen Einschränkungen auch mit Formaten wie XML umgehen.

# Verarbeitung von (strukturierten) Daten

## Beispiel: CSV- und XML-Dateien



- Dateien werden durch Programme erstellt, verändert und gelesen.
- Damit Programme mit den Daten aus Dateien umgehen können, unterstellen sie eine bestimmte Codierung und ein bestimmtes Format.
- Programme können mit bestimmten Dateiformaten arbeiten, die unter Umständen auch eine spezifische Struktur aufweisen müssen.
- Viele Programme haben ein eigenes spezifisches Dateiformat.
- Sie sind unter Umständen aber auch in der Lage, Dateien mit anderen Dateiformaten zu lesen.
- Dies erhöht die Flexibilität beim Import und Export von Daten.
- Die Verwendung von "fremden" Formaten ist oftmals mit gewissen Einschränkungen verbunden.



Externe Speicher und Dateisystem

Strukturierte und unstrukturierte Daten

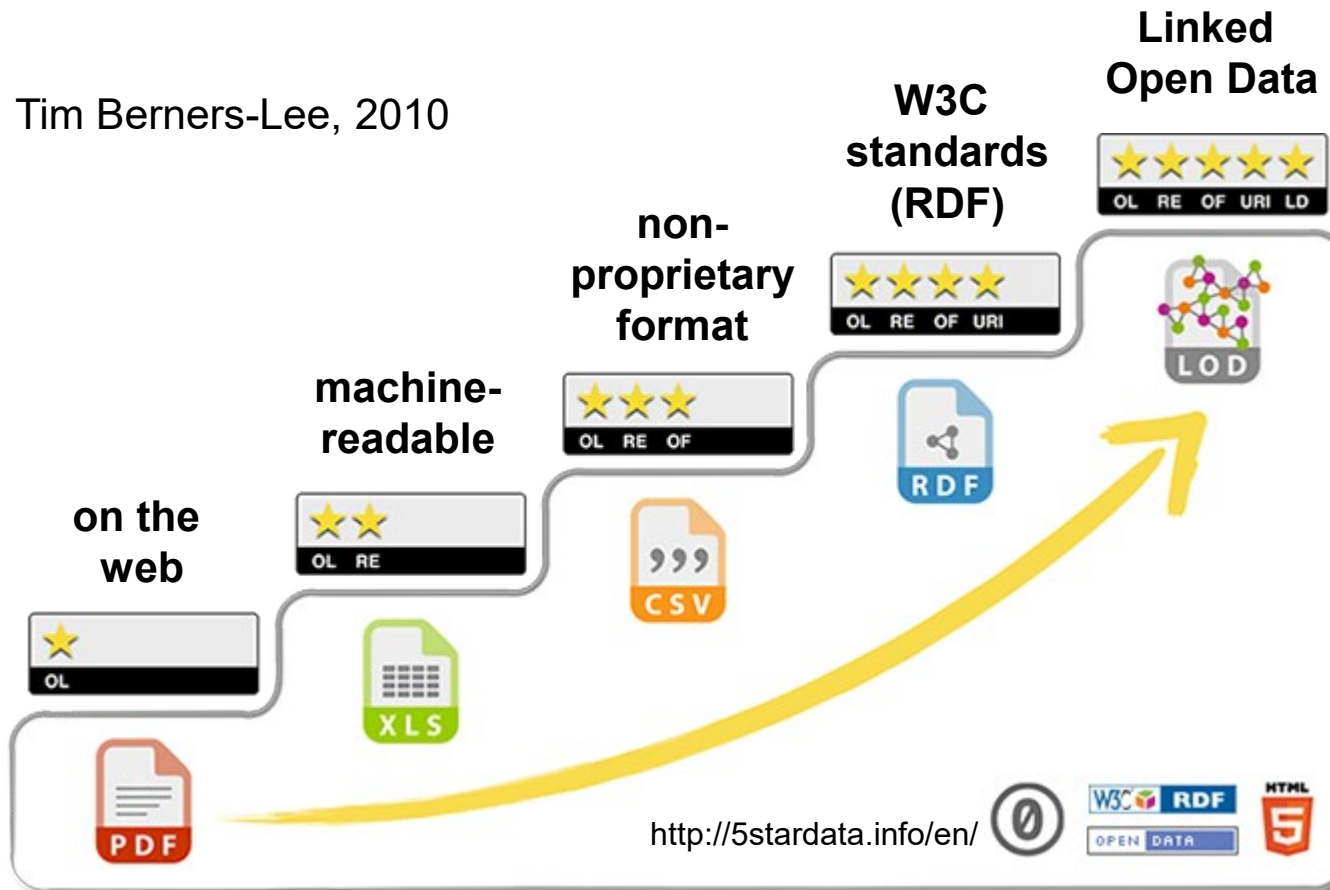
Dateien und Programme

Dateiformate und Datenweitergabe

- Dateien werden nicht nur von einem Programm angelegt und von diesem Programm wieder gelesen.
- Häufig dienen Dateien der Weitergabe von Daten zwischen verschiedenen Programmen.
- Dies geschieht etwa beim Datenaustausch im Zuge des E-Business.
- Das Konzept des Open Data postuliert, dass Daten der Öffentlichkeit frei verfügbar gemacht werden.
- Ziel ist dabei, dass die offengelegten Daten möglichst leicht für verschiedene Zwecke verwendet werden können.
- Das Dateiformat beeinflusst, in welchen Ausmass die flexible Verwendung von weitergegebenen Daten möglich ist.

# Das Five-Stars-Modell für Open Data

## Entwicklungsstufen für Open Data



# Das Five-Stars-Modell für Open Data

## Bedeutung der fünf Sterne

- ★ stelle deine Daten im *Web* unter einer offenen Lizenz bereit. Das Format ist dabei egal<sup>1</sup>
- ★★ stelle Daten in einem strukturierten Format bereit (z. B. Excel anstelle eines eingescannten Bildes einer Tabelle)<sup>2</sup>
- ★★★ verwende offene, nicht proprietäre Formate (z. B. CSV statt Excel)<sup>3</sup>
- ★★★★ verwende URIs um Dinge zu bezeichnen, damit deine Daten verlinkt werden können<sup>4</sup>
- ★★★★★ verlinke deine Daten mit anderen Daten um Kontexte herzustellen<sup>5</sup>

# Von einem (★) zu zwei (★★) Sternen

## Daten als PDF- und als Excel-Datei

### Ständige Wohnbevölkerung der Kantone, 2015 T 2

	Total	Männer	Frauen	Schweiz	A	B	C	D	E	F	G	H	I	J	
Schweiz	8 327 126	4 121 471	4 205 655	6 278 1	cc-d-1.1.4	Ständige und nichtständige Wohnbevölkerung nach Staatsangehörigkeitskategorie, Geschlecht und Quartals 2016									
Zürich	1 466 424	728 517	737 907	1 083 3	2										
Bern	1 017 483	498 258	519 225	860 4	3	Grossregion	Ständige Wohnbevölkerung								
					4		Total			Schweizer Staatsangehörigkeit			Ausländische Staatsangehörigkeit 1)		
					5	Kanton	Total	Mann	Frau	Total	Mann	Frau	Total	Mann	Frau
Luzern	398 762	198 192	200 570	327 7	7	Total	8 391 973	4 157 939	4 234 034	6 309 021	3 051 929	3 257 092	2 082 952	1 106 010	976 942
Uri	35 973	18 348	17 625	31 9	9	Genferseeregion	1 606 172	786 706	819 466	1 069 767	507 176	562 591	536 405	279 530	256 875
					10	Waadt	779 609	382 529	397 080	519 289	245 916	273 373	260 320	136 613	123 707
Schwyz	154 093	78 825	75 268	122 11	11	Wallis	337 590	167 358	170 232	259 197	125 701	133 496	78 393	41 657	36 736
					12	Genf	488 973	236 819	252 154	291 281	135 559	155 722	197 692	101 260	96 432
Obwalden	37 076	18 801	18 275	31 13	13	Espace Mittelland	1 854 992	915 009	939 983	1 509 103	730 497	778 606	345 889	184 512	161 377
Nidwalden	42 420	21 705	20 715	36 14	14	Bern	1 024 192	502 264	521 928	861 927	416 179	445 748	162 265	86 085	76 180
					15	Freiburg	310 466	155 494	154 972	241 568	118 530	123 038	68 898	36 964	31 934
Glarus	40 028	20 309	19 719	30 16	16	Solothurn	268 639	133 773	134 866	210 214	102 582	107 632	58 425	31 191	27 234
					17	Neuenburg	178 660	87 367	91 293	132 907	62 780	70 127	45 753	24 587	21 166
Zug	122 134	61 708	60 426	89 18	18	Jura	73 035	36 111	36 924	62 487	30 426	32 061	10 548	5 685	4 863
Freiburg	307 461	153 729	153 732	240 19	19	Nordwestschweiz	1 138 566	564 665	573 901	844 950	408 567	436 383	293 616	156 098	137 518
					20	Basel-Stadt	193 212	93 295	99 917	124 479	57 705	66 774	68 733	35 590	33 143
Solothurn	266 418	132 439	133 979	209 21	21	Basel-Landschaft	284 717	139 623	145 094	221 738	106 529	115 209	62 979	33 094	29 885
					22	Aargau	660 637	331 747	328 890	498 733	244 333	254 400	161 904	87 414	74 490
					23	Zürich	1 482 650	737 009	745 641	1 090 469	528 584	561 885	392 181	208 425	183 756



Von zwei (★★) zu drei (★★★) Sternen

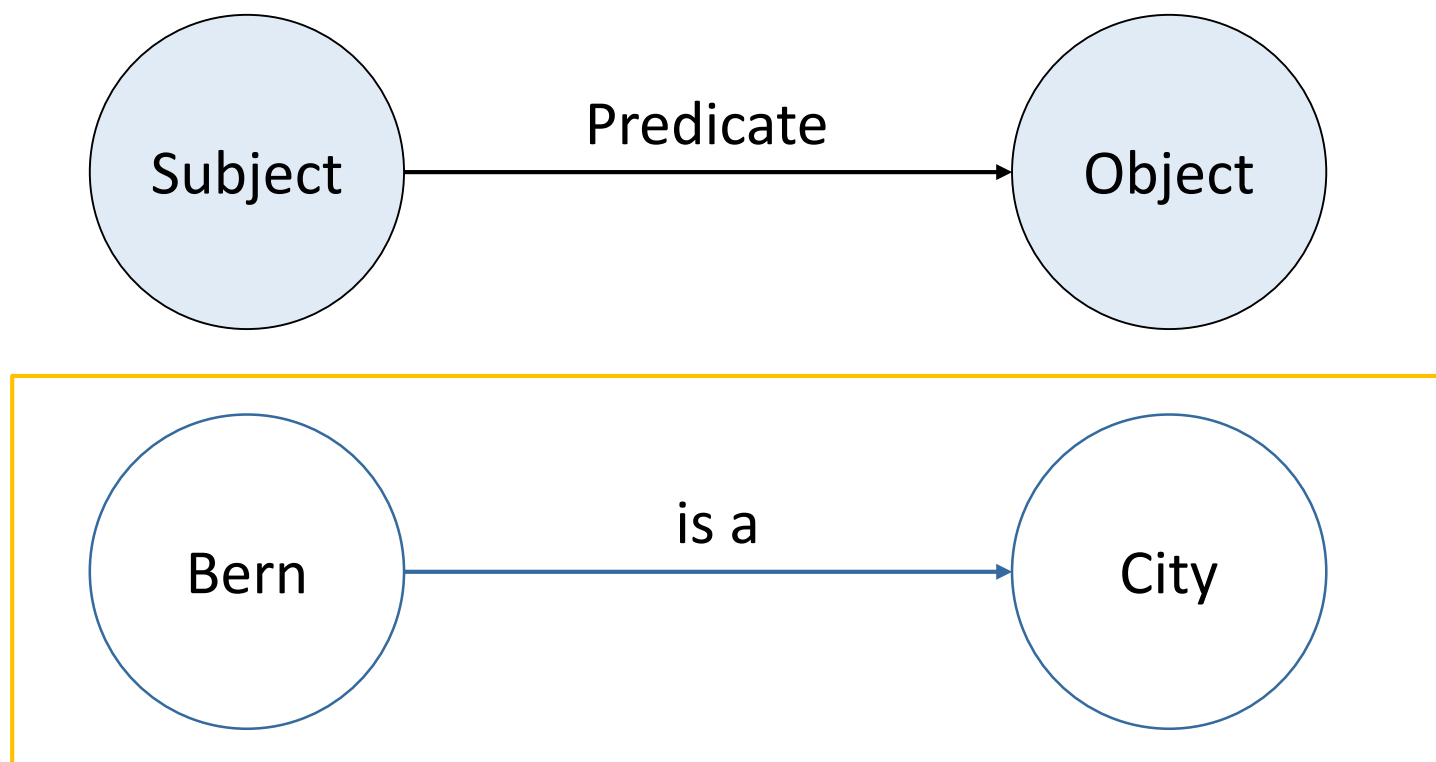
Daten als Excel- und CSV-Daten

	A	B	C	D	E	F
1	Vorname	Nachname	Geschlecht	Jahrgang	Land	
2	Julia	Tinner	weiblich	1970	Italien	
3	Tim	Foster	männlich	1987	Irland	
4	Michael	Dobler	männlich	1959	Deutschland	
5	Caroline	Messmer	weiblich			
6						
7						

```
student.csv - Notepad
File Edit Format View Help
Vorname;Nachname;Geschlecht;Jahrgang;Land
Julia;Tinner;weiblich;1970;Italien
Tim;Gutweniger;männlich;1987;Irland
Michael;Dobler;männlich;1959;Deutschland
Caroline;Messmer;weiblich;1964;Schweiz
```

# Linked (Open) Data (★★★★★)

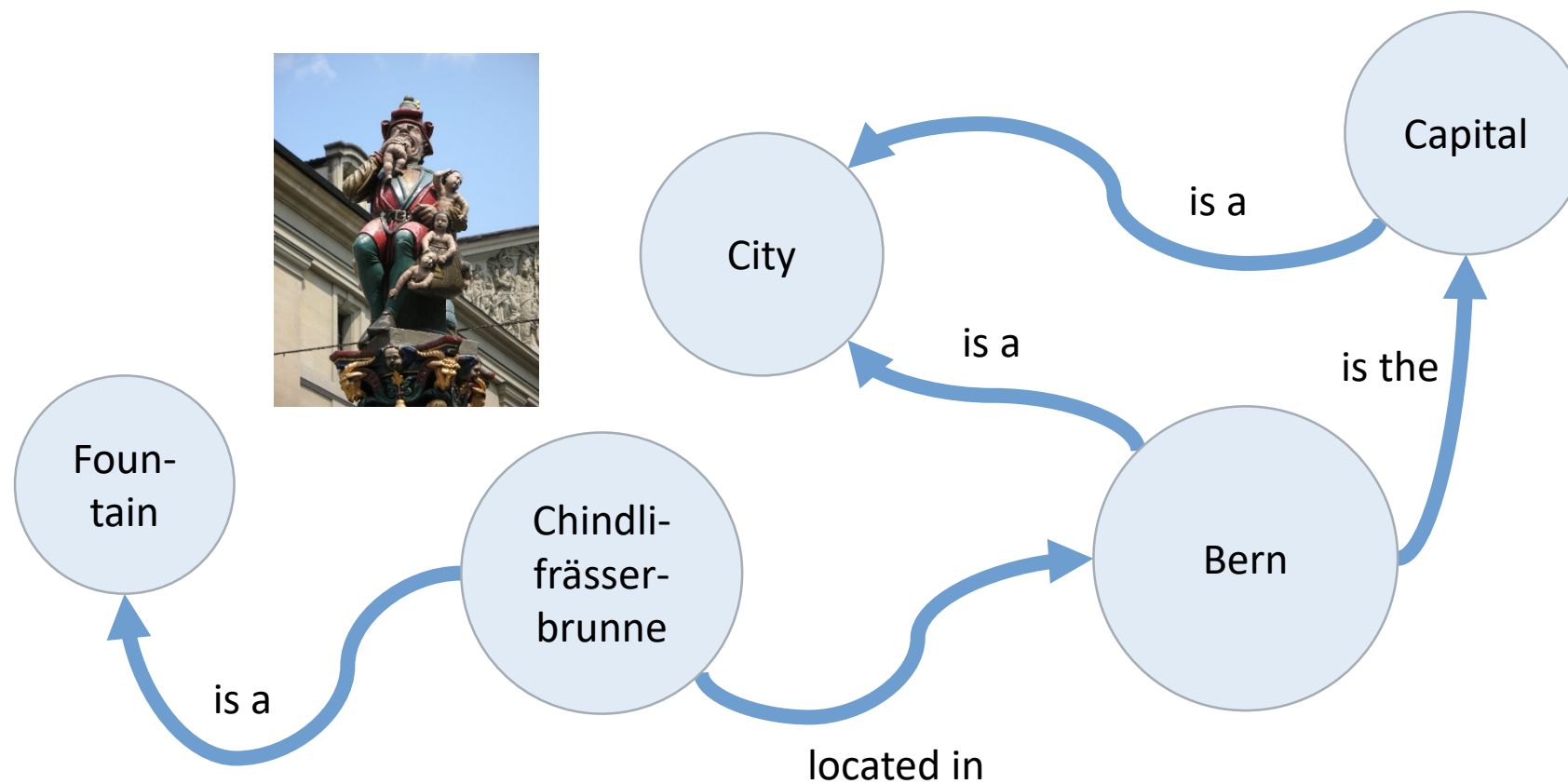
## Linked-Data-Triple



Technische Spezifikationen im **Resource Description Framework (RDF)**

# Linked (Open) Data (★★★★)

## Begriffsnetzwerke



- Dateien werden auch zur Weitergabe von Informationen verwendet.
- Ein besonderer Anspruch an die Datenweitergabe ist mit dem Konzept von Open Data gegeben.
- Bei der Datenweitergabe werden Daten zwischen Programmen ausgetauscht.
- Die jeweiligen Programme müssen in der Lage sein, die Daten zu verarbeiten.
- Die Verarbeitbarkeit der Daten wird beeinflusst von den verwendeten Dateiformaten.
- Mit Bezug auf Open Data wird dies im Five-Star-Modell formalisiert.
- Dies legt eine Rangreihung von Dateiformaten mit Bezug auf das Ziel der Offenheit von Daten fest.