Programmierkonzepte in den Naturwissenschaften

32. Datenrepräsentation: Darstellungs- und Berichterstellung

PD Dr. Till Biskup
Physikalische Chemie
Universität des Saarlandes
Sommersemester 2021

Zentrale Aspekte

- Ein Bild sagt mehr als tausend Worte: Der Wert guter Repräsentationen sollte nicht unterschätzt werden.
- Charakteristika eines Datensatzes herauszuarbeiten, ist die eigentliche intellektuelle Leistung der Auswertung.
- Erkenntnisgewinn lässt sich nicht automatisieren, viele Einzelschritte auf dem Weg dahin schon.
- Berichte präsentieren übersichtlich Informationen zu einem Datensatz und lassen sich automatisch erzeugen.
- Zentraler Aspekt der Berichterstellung ist die Trennung von verarbeitenden Routinen und Darstellung.

Übersicht

Zur Bedeutung der Repräsentation von Daten

Repräsentationen sollten aus den Primärdaten
automatisch generierbar sein

Berichte: Übersicht über die Informationen zu einem Datensatz

Vorlagen: Trennung von Inhalten und Darstellung

Ein Bild sagt mehr als tausend Worte

Große Wirkung bringt große Verantwortung mit sich.

- Wirkung grafischer Darstellungen nicht unterschätzen
  - Menschen sind sehr gut darin, Zusammenhänge visuell wahrzunehmen.
  - Mitunter erkennen wir mehr, als wirklich vorhanden ist . . .
- Verantwortung des Wissenschaftlers
  - Abbildungen dürfen nicht zu viel/das Falsche implizieren.
  - Oft implizieren Bilder ungewollt und unbeabsichtigt zu viel.
  - Kriterium: nur implizieren, was die Daten real hergeben
- „Aushängeschild“ der eigenen Forschung
  - „Ein gutes Bild sagt mehr als tausend Worte.“
  - Die Bedeutung sollte sich in der aufgewandten Sorgfalt bei der Erstellung von Darstellungen widerspiegeln.

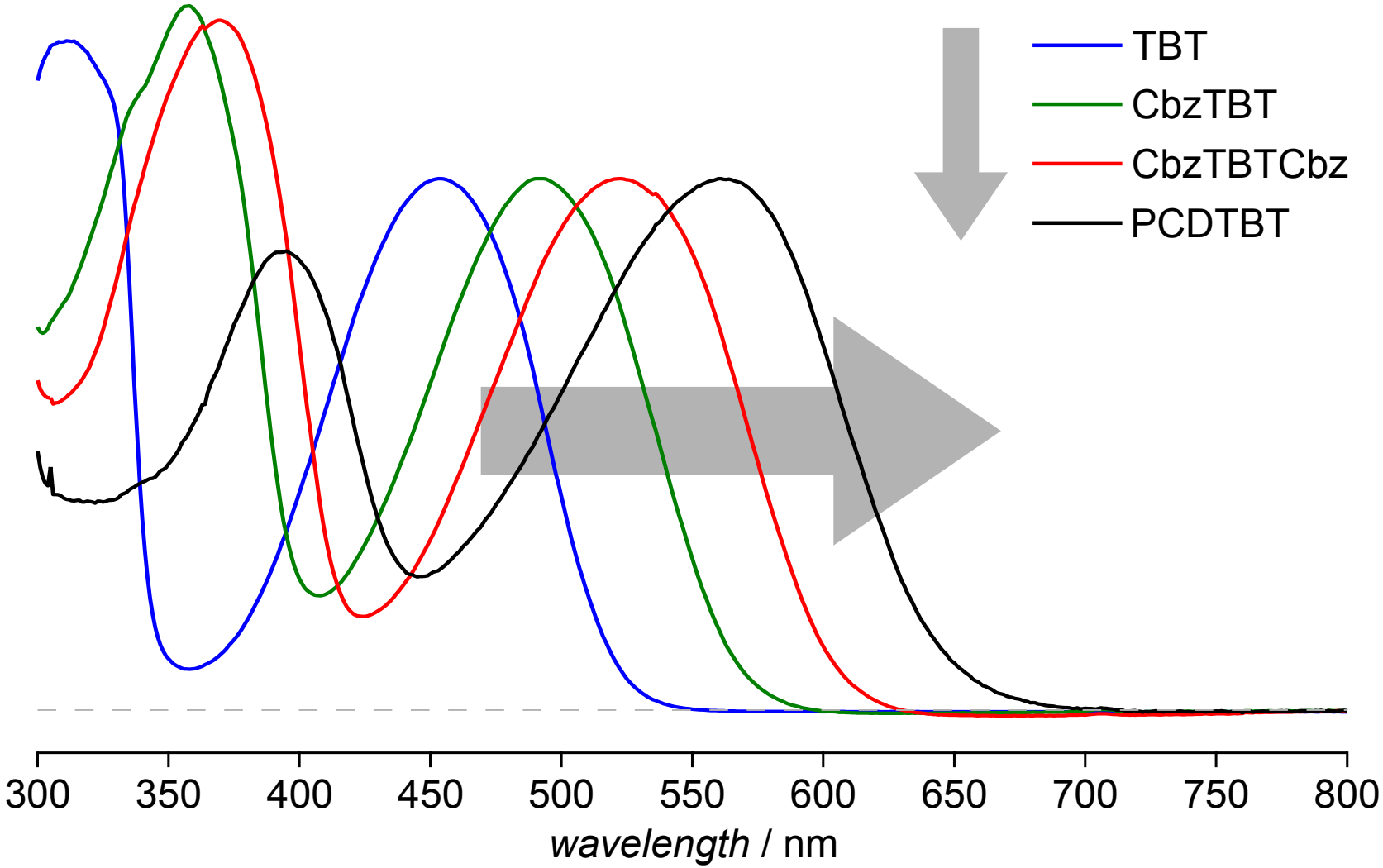
Repräsentationen sind nicht nur Bilder

Tabellarische Darstellungen können für Überblick sorgen.

- Repräsentationen können sowohl Abbildungen als auch Tabellen etc. sein.
  - entscheidend: Darstellung charakteristischer Parameter
- Repräsentationen können mehrere Datensätze umfassen.
  - Oft werden Zusammenhänge erst im Vergleich mehrerer Datensätze und ihrer Charakteristika deutlich.
- Beispiele für Aussagen aufgrund von Vergleichen
  - Verschiebung von Absorptionsbanden mit zunehmender Delokalisierung des aromatischen Systems
  - Linienbreiten in der EPR-Spektroskopie in Abhängigkeit von der der Temperatur
  - Kinetiken: Absorbanz bei einer Wellenlänge
  - Kinetiken: Vergleich der Zeitkonstanten (Tabelle)

Beispiel: Vergleichende Abbildung

Aussage: Rotverschiebung einer Absorptionsbande



Matt *et al.*, *Macromolecules* **51**:4341–4349, 2018

Beispiel: Vergleichende Tabelle

Aussage: Abnehmende Werte für $|D|$ und Linienbreite

λ / nm	$\lvert D\rvert$ / MHz	Γ / mT
492	1361.6 ± 3.0	3.42
630	1344.7 ± 1.5	2.08
650	1317.2 ± 1.4	1.83
680	1288.5 ± 1.3	1.54

- ☛ Wichtig: Reduktion auf das Wesentliche
  (es gäbe noch deutlich mehr Parameter)

Meyer *et al.*, *Journal of Physical Chemistry Letters* 8:1677–1682, 2017

Auswertung ist (meist) nicht automatisierbar.

Viele Schritte auf dem Weg dorthin allerdings schon...

These

Eine aussagekräftige Darstellung eines Datensatzes
ist das Ergebnis der intensiven Beschäftigung mit den Daten
und oft die eigentliche intellektuelle Leistung.

- Daten müssen meist vorverarbeitet werden.
  - Vorverarbeitung lässt sich automatisieren.
  - Historie sorgt für Nachvollziehbarkeit und Transparenz
- Sichten auf Daten lassen sich formalisiert ablegen.
  - intellektuelle Leistung: Welche Sicht ist relevant?
  - Die eigentliche Darstellung ist vollständig automatisierbar.

Der Imperativ: „Kenne deine Daten“

Eine einheitliche Darstellung hilft bei der Übersicht.

- Verantwortung des Wissenschaftlers
  - solider und umfassender Überblick über die eigenen Daten
  - Dazu gehören auch repräsentative Darstellungen.
- Ausgangspunkt: die „Briefmarkensammlung“
  - gleichartig formatierte Darstellungen
  - hilft, auf die Unterschiede in den Daten zu fokussieren
  - Grundlage für Vergleiche zwischen Datensätzen
- Ziel: Verständnis von Zusammenhängen
  - setzt intime Kenntnis der Daten voraus
  - Hypothesen einfach anhand der Ergebnisse überprüfbar
  - Ideal: Daten verinnerlicht und vor dem inneren Auge

☛ wird im Kontext von Berichten noch bedeutsam

Ein Wort zur Formatierung

Es gibt feste Regeln – auch wenn Wenige sie kennen.

- Beschriftung von Abbildungen und Tabellen
  - Abbildungen haben *Unter*schriften, Tabellen *Über*schriften.
  - Abbildungen und Tabellen werden fortlaufend nummeriert.
  - Auf jede Abbildung/Tabelle wird aus dem Text verwiesen.
- Größen und Einheiten
  - Größe kursiv, Einheit aufrecht, Schrägstrich dazwischen
  - Einheiten *niemals* in eckigen Klammern
- Abbildungen
  - Achsenbeschriftungen in lesbarer Größe und konsistent
  - nie auf Farbe verlassen (Graustufen sollten funktionieren)
- Tabellen
  - nur horizontale Linien, sparsam eingesetzt
  - Linien zur Gruppierung von Inhalten verwenden

Übersicht

Zur Bedeutung der Repräsentation von Daten

Repräsentationen sollten aus den Primärdaten automatisch generierbar sein

Berichte: Übersicht über die Informationen zu einem Datensatz

Vorlagen: Trennung von Inhalten und Darstellung

Bedeutung automatischer Generierung

Nicht nur eine Frage der Bequemlichkeit

- Zeitersparnis
  - Die eigentliche Darstellung ist meist reine Routine.
  - Nicht nachdenken zu müssen, erspart Zeit.
- Konsistenz
  - einheitliche Formatierung
  - bessere Vergleichbarkeit untereinander
- Fokussierung auf das eigentlich Wesentliche
  - Analyse und Verständnis stehen im Mittelpunkt.
  - Charakteristika lassen sich abstrakt ablegen.
- Voraussetzung für Berichterstellung
  - Zusammenfassung der Charakteristika und Informationen
  - involviert häufig (grafische) Darstellungen

Metadaten für Repräsentationen

Repräsentationen lassen sich abstrakt beschreiben.

- Idee
  - Formalisierung von Repräsentationen
  - Fokussierung auf das „Was“ statt auf das „Wie“
  - Ablage charakteristischer „Sichten“ auf Daten
  - Minimierung der Notwendigkeit zur Nachbearbeitung
- Vorteile
  - identische Darstellung unterschiedlicher Datensätze
  - sorgt ggf. für Konsistenz und bessere Vergleichbarkeit
  - Entkopplung von Datenquelle und Darstellung
- Voraussetzungen
  - Routinen zur Erzeugung von Darstellungen
  - weitgehende Konfigurierbarkeit
  - Kontrolle über die einzelnen Darstellungsoptionen

Metadaten auf unterschiedlichen Ebenen

Drei Ebenen: Daten, (Vor-)Verarbeitung, Darstellung

- Daten
  - Liste von Datensätzen
  - wo möglich auf Rohdaten verweisen
- (Vor-)Verarbeitung
  - notwendige Vorverarbeitungsschritte
  - ggf. Extraktion der Charakteristika
  - wird auf jeden Datensatz aus der Liste angewendet
- Darstellung
  - genereller Typ (Abbildung: 1D, 2D, ...)
  - allgemeine Formatierungen und Beschriftungen
  - Details zur Formatierung

☛ auf Abbildungen und Tabellen gleichermaßen anwendbar

Übersicht

Zur Bedeutung der Repräsentation von Daten

Repräsentationen sollten aus den Primärdaten automatisch generierbar sein

Berichte: Übersicht über die Informationen zu einem Datensatz

Vorlagen: Trennung von Inhalten und Darstellung

Idee: Informationen zugänglich machen

Die Informationen sind vorhanden und wollen genutzt werden.

- Ein Gesamtsystem zur Datenverarbeitung erzeugt große Mengen an Informationen.
  - Informationen zur Datenaufnahme (z.B. Infodatei)
  - Informationen zur Datenverarbeitung (was, wie)
  - Informationen zu Charakteristika der Daten
- Ziel: Informationen zugänglich machen
  - Zusammenfassung der Charakteristika und Informationen
  - übersichtlich und lesbar aufbereitet
  - hilfreich für den Vergleich von Datensätzen untereinander
- Berichte sind oft hochspezifisch.
  - Übersichtliche Darstellung ist eine intellektuelle Leistung.
  - Berichte sollten strukturell unabhängig vom Datensatz sein.

Der Vorteil gleichartiger Berichte

Übersichtlichkeit erleichtert Vergleichbarkeit

- Verantwortung des Wissenschaftlers
  - solider und umfassender Überblick über die eigenen Daten
- Vergleich von Daten hilft beim Verständnis.
  - Muster oft nur durch Vergleich von Datensätzen erkennbar
  - Berichte möglichst gleichförmig und informativ
  - Fokus des Betrachters auf dem Vergleich,
    nicht auf dem Zusammensuchen der Informationen
  - Unterschiede treten entsprechend deutlich zutage.
- Beispiel: Ergebnisse von Kurvenanpassungen
  - meist relativ viele Parameter
  - Vergleich unterschiedlicher Anpassungen oft wichtig

☛ Berichte sollten automatisch erzeugt werden.

Übersicht

Zur Bedeutung der Repräsentation von Daten

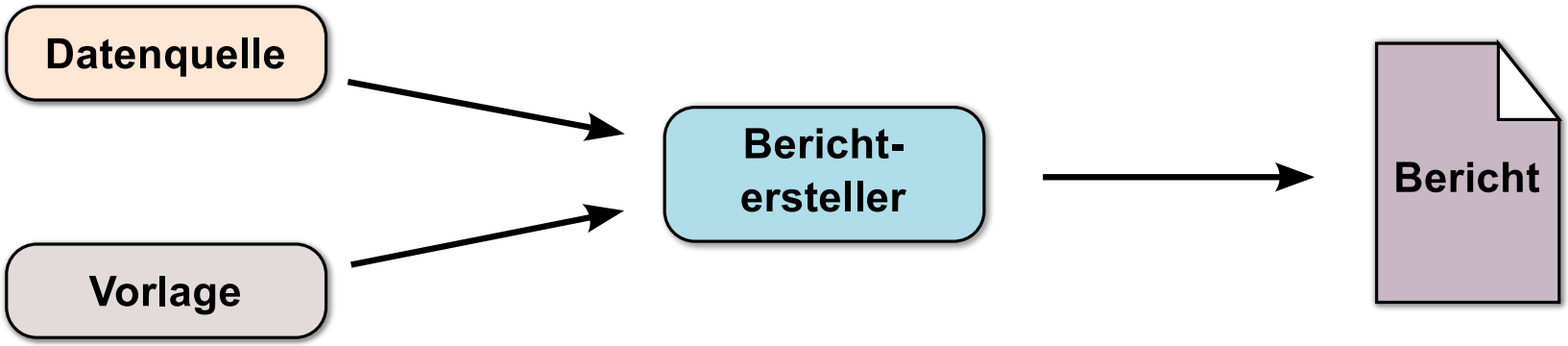
Repräsentationen sollten aus den Primärdaten automatisch generierbar sein

Berichte: Übersicht über die Informationen zu einem Datensatz

Vorlagen: Trennung von Inhalten und Darstellung

Vorlagen zur automatischen Berichterzeugung

Trennung von Informationsverarbeitung und Präsentation



- vier Grundbestandteile
  - Datenquelle
  - Vorlage (*template*)
  - Verarbeitungslogik
  - Bericht

☛ Grundidee: Trennung der Präsentation von den darzustellenden Informationen und deren Erzeugung

Grundaspekte und Vorteile von Vorlagen

Trennung der Verantwortlichkeiten sorgt für Flexibilität

- einfachstes Bild: Vorlage als „Lückentext“
  - Die „Lücken“ werden speziell markiert und gefüllt.
  - Das (Datei-)Format spielt (fast) keine Rolle.
  - intellektuelle Leistung: Ausarbeitung guter Vorlagen zur übersichtlichen Darstellung wichtiger Informationen
- Verarbeitungslogik für Vorlagen (*template engine*)
  - liest eine Vorlage und füllt die „Lücken“
  - beherrscht komplexere Ersetzungen
  - benötigt keinerlei Wissen über das Format der Vorlage
- Vorteile der Verwendung von Vorlagen
  - Trennung von Datenquelle, Verarbeitungslogik, Darstellung
  - Berichte mit gleicher Information in einer Vielzahl von Formaten bzw. Sprachen automatisiert generierbar

Voraussetzungen für ein Vorlagensystem

Geeignete Dateiformate und strikte Trennung der Zuständigkeiten

- Vorlagen parsbar und einfach veränderbar
  - Format muss Platzhalter und Steuercodes ermöglichen
  - ohne Kenntnis des Formats einlesbar und verarbeitbar
  - (beliebige) Textdateien geeignet, Binärformate i.d.R. nicht
  - Beispiele geeigneter Dateiformate: LaTeX, ODF, HTML/XML, Markup-Formate (MarkDown, reStructuredText, DokuWiki, ...)
- Vorlagenverarbeitung unabhängig vom Vorlagenformat
  - keine Erzeugung formatspezifischer Ausgaben
  - Verarbeitung ausschließlich über Steuercodes
- Trennung von Vorlagenverarbeitung und Inhalteerzeugung
  - Erzeugung von Darstellungen auslagern
  - Berichterzeugung greift auf Vorlagenverarbeitung zurück, generiert aber ggf. zusätzlich (Daten für) Darstellungen

Mindestansprüche an die Vorlagenverarbeitung

Von einfachen Ersetzungen bis zu Formatierungshinweisen

- einfache Ersetzungen
  - Grundlage jedes Vorlagensystems
- Einbinden zusätzlicher Vorlagen
  - sorgt für Modularität und Flexibilität
- Schleifen
  - besonders wichtig für das Ausfüllen von Tabellen
- (einfache) Bedingungen
  - erhöht die Flexibilität von Berichten
- Formatierungsangaben für numerische Werte
  - insbesondere für Gleitkommazahlen
- freie Wahl der Begrenzer für Ersetzungs- und Steuercodes
  - abhängig vom Format der Vorlage

Listing 1: Beispiel für einfache Ersetzungen

```
Das Programm \texttt{Tsim} hat unter Verwendung der Simulationsroutine
\texttt{[[@TSim.sim.routine]]} eine Triplett-Simulation durchgeführt.
```

Listing 2: Beispiel für Schleifen

```
[[foreach @Tsim.acknowledgement.sim]]
[[@Tsim.acknowledgement.sim]]\par
[[end]]
```

Listing 3: Beispiel für Bedingungen

```
[[if ~isempty(this.assignments.Tsim.remarks.purpose)]]
\textbf{Zielstellung:} [[@Tsim.remarks.purpose]]
[[end]]
```

Vorteil der Trennung von Vorlage und Daten

Unter besonderer Berücksichtigung der SOLID-Prinzipien

- Single Responsibility
  - Die Verarbeitungslogik weiß nichts von Datenmodell oder Bericht, kann aber (komplexe) Ersetzungen vornehmen.
  - Die Verarbeitungslogik ist modular einsetzbar.
- Liskov Substitution
  - Auch wenn die Datenquelle ein abgeleiteter Datentyp ist, funktioniert eine generische Berichterstellungsroutine.
  - Erweiterungen sind abwärtskompatibel.
- Dependency Inversion
  - Das Datenmodell weiß nichts vom Bericht, der Bericht aber vom Datenmodell.
  - Berichte sind peripher, das Datenmodell zentral.
  - Das Dateiformat des Berichtes spielt keine Rolle.

Berichterstellung

Ein komplexes Zusammenspiel

- Berichterstellungsroutine
  - nutzt Datensätze als primäre Datenquelle
  - greift auf Verarbeitungslogik für Vorlagen zurück
  - sorgt ggf. für die Erzeugung von Darstellungen etc.
  - sorgt für die Speicherung des fertigen Berichtes
  - greift ausschließlich auf separate Routinen zurück
- Modularität sorgt für Flexibilität.
  - Darstellungen als Metadaten abgelegt
  - Vorlagen für Teile von Berichten einsetzbar

☛ Berichte sorgen für Überblick und Vergleichbarkeit.

☛ Eigentliche intellektuelle Herausforderung:
aussagekräftige, übersichtliche, zugängliche Struktur

Zentrale Aspekte

- Ein Bild sagt mehr als tausend Worte: Der Wert guter Repräsentationen sollte nicht unterschätzt werden.
- Charakteristika eines Datensatzes herauszuarbeiten, ist die eigentliche intellektuelle Leistung der Auswertung.
- Erkenntnisgewinn lässt sich nicht automatisieren, viele Einzelschritte auf dem Weg dahin schon.
- Berichte präsentieren übersichtlich Informationen zu einem Datensatz und lassen sich automatisch erzeugen.
- Zentraler Aspekt der Berichterstellung ist die Trennung von verarbeitenden Routinen und Darstellung.