Forschungsdatenmanagement

Notwendige, aber nicht hinreichende Voraussetzung für den wissenschaftlichen Erkenntnisgewinn

02. Warum Forschungsdatenmanagement?
(II) Zunahme von Datenmenge und Digitalität

Dr. habil. Till Biskup
Physikalische Chemie
Universität Rostock
19.04.2024

Zentrale Aspekte

- Automatisierung und Digitalisierung ermöglichen es, nahezu beliebige Datenmengen (relativ) einfach zu erzeugen.
- Die Zunahme von Datenmenge und Digitalität erzwingt, Strategien zu ihrem Umgang zu entwickeln und zu etablieren.
- „Big Data“ und das Arbeiten auf digitalen Daten anderer ist für die wenigsten wissenschaftlichen Disziplinen relevant.
- Die FAIR-Prinzipien zum Umgang mit Forschungsdaten werden überbewertet, missverstanden und setzen den falschen Fokus.
- Die Qualität von Forschungsdaten, obwohl entscheidend, ist nur schwer durch Kriterien bestimmbar.

These

Forschungsdatenmanagement hat nichts mit Digitalität zu tun,
Digitalität macht Forschungsdatenmanagement nur drängender.

- Die eigentliche Neuerung der zunehmenden Digitalität: Das Erzeugen nahezu beliebiger Datenmengen wird stark vereinfacht.
- Quantität ohne Qualitätssicherung konterkariert die Wissenschaft.
- Digitalität und datengetriebene Wissenschaft als Treiber des Forschungsdatenmanagements zu verstehen, greift (viel) zu kurz.
- Forschungsdaten liegen nicht notwendigerweise digital vor.
  Die unnötige Einschränkung auf digital verfügbare Daten führt zu einer Verarmung und Einschränkung wissenschaftlicher Erkenntnis.

Übersicht

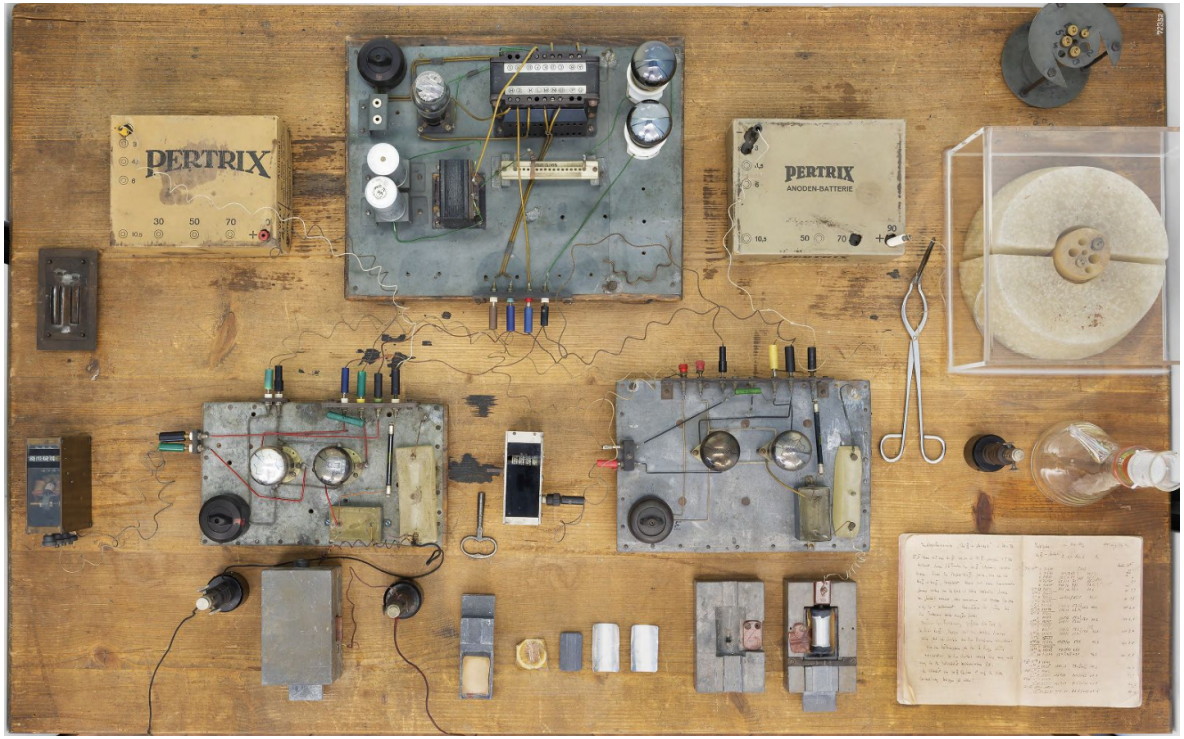
Little Science, Big Science

„Big Data“, „Datenflut“, „e-Science“ und das „vierte Paradigma“

Die FAIR-Prinzipien zum Umgang mit Forschungsdaten

Individuelle Forschende mit klein(er)en Datenmengen

Little Science. . .



. . . Big Science

Titel/Untertitel: Derek J. de Solla Price. Suhrkamp, Frankfurt 1974
Deutsches Museum / Christian Illing, CC BY-SA 4.0
Jack W. Aeby; Time Inc.

Großforschung

Big Science, außeruniversitär, meist in großen Forschungseinrichtungen teils quasi-industriell betriebene Form der Wissenschaft, die seit Mitte des 20. Jh. verstärkt auftritt. Meist wird der Beginn mit dem Manhattan-Projekt (Los Alamos National Laboratory) gleichgesetzt.

Charakteristika

- erhebliche staatliche Investitionen in einzelne Forschungszentren
- Forscherkollektiv statt individuelle Forschende
- erheblicher administrativer Aufwand, hinter dem die Kreativität der Forschenden und die Qualität der Forschung zurücktritt.

A. Weinberg, *Science* 134:161–164, 1961
Derek J. de Solla Price: *Little Science, Big Science*. Suhrkamp, Frankfurt 1974 (original 1963)

Little Science, Big Science

Die Gefahr von Großforschung für die Wissenschaft

> *When history looks at the 20th century, she will see science and technology as its theme; she will find in the monuments of Big Science—the huge rockets, the high-energy accelerators, the high-flux research reactors—symbols of our time just as surely as she finds in Notre Dame a symbol of the Middle Ages. [. . .]*
>
> *"Is Big Science ruining science?"*
>
> *[. . .] one sees evidence of scientists' spending money instead of thought. This is one of the most insidious effects of large-scale support of science. In the past the two commodities, thought and money, have both been hard to come by. Now that money is relatively plentiful but thought is still scarce, there is a natural rush to spend dollars rather than thought.*
>
> – Alvin Weinberg

A. Weinberg, *Science* 134:161–164, 1961

Little Science vs. Big Science

Der Unterschied ist *qualitativ*, nicht quantitativ.

de Solla Price, 1963 (1974)

- Wissenschaft ist seit ca. 1650 exponentiell gewachsen.
- Abweichung vom exponentiellen Wachstum ist nicht feststellbar, aber bis zum Ende des 20. Jahrhunderts zwangsläufig.
- Exponentielles Wachstum ist begrenzt, meist liegt ein logistisches Wachstum darunter (Sättigungskurve)
- Das „Ertrinken“ in Informationen ist kein neues Phänomen, die wissenschaftlichen Zeitschriften (ab 1660) eine Reaktion darauf.

Qualitative Unterschiede

- Großforschung konzentriert Geld auf wenige Fragestellungen.
- Wissenschaftspolitik und staatlich alimentierte Forschung dominieren den Wissenschaftsbetrieb und setzen Schwerpunkte.

Derek J. de Solla Price: *Little Science, Big Science*. Suhrkamp, Frankfurt 1974 (original 1963)

Kritische Rückschau aus heutiger Sicht

- Forschung und Wissenschaft haben sich seither stark gewandelt.
- Heute dominieren oft verteilte, daten- und rechenzentrierte Ansätze, die damals noch nicht absehbar waren.
- de Solla Price blickt sehr einseitig auf die Physik und ignoriert weite Teile der Natur- und komplett die Geisteswissenschaften.
- Der *qualitative* Unterschied gilt auch für „little data“ *vs.* „big data“.

Hat Großforschung die Wissenschaft ruiniert? (A. Weinberg)

- Ökonomische Interessen haben mit Sicherheit einen Einfluss – auch für individuelle Forschende („*publish or perish*“).
- Großforschung hat mit Sicherheit die Technologie vorangebracht – aber nicht unbedingt die Wissenschaft.

Übersicht

> *In 1963, Derek de Solla Price famously contrasted “little science” and “big science.” [. . .]*
> *“Big data” has acquired the hyperbole that “big science” did fifty years ago. Big data is on the covers of Science, Nature, the Economist, and Wired magazine and the front pages of the Wall Street Journal, New York Times, and many other publications, both mainstream and minor. Just as big science was to reveal the secrets of the universe, big data is expected to reveal the buried treasures in the bit stream of life. Big data is the oil of modern business [. . .], the glue of collaborations [. . .], and a source of friction between scholars [. . .].*
>
> – Christine L. Borgman

C. L. Borgman: Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press, Cambridge, MA, 2015. S. 3

“Big Data”

Volume

Velocity

Variety

Value

Validity

Big Data

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

(Gartner)

- drei Charakteristika: Umfang, Geschwindigkeit, Vielfalt
- machen neue Formen der Datenverarbeitung notwendig
- Versprechen: Einblicke, Entscheidungen, Automatisierung

☛ „Big Data“ ist zunächst einmal ein Konzept aus der Wirtschaft.

https://www.gartner.com/en/information-technology/glossary/big-data (Zugriff am 17.04.2024)

Ein paar Beispiele für große Datenmengen

- Protein(kristall)strukturen: PDB
  - Seit spätestens 1972 gibt es das Datenformat.
  - Eine der ältesten „digitalen“ Datensammlungen
- Gen- und andere Sequenzen: NCBI
  - Gen- und Proteinsequenzen inkl. (Online-)Werkzeugen
  - einheitliches Datenformat für Sequenzen seit Jahrzehnten
- Astronomie: SDSS *et al.*
  - komplette Himmelsdurchmusterungen
  - einheitliche Datenformate seit Jahrzehnten
- Teilchenphysik: LHC *et al.*
  - Daten werden am Detektor vorsortiert und nur teils gespeichert
  - strukturierte, *vorher* festgelegte Formate

☛ mitunter große Mengen *strukturierter* Daten

Die „Datenflut“

Daten zu erzeugen ist (zu) einfach geworden.

These

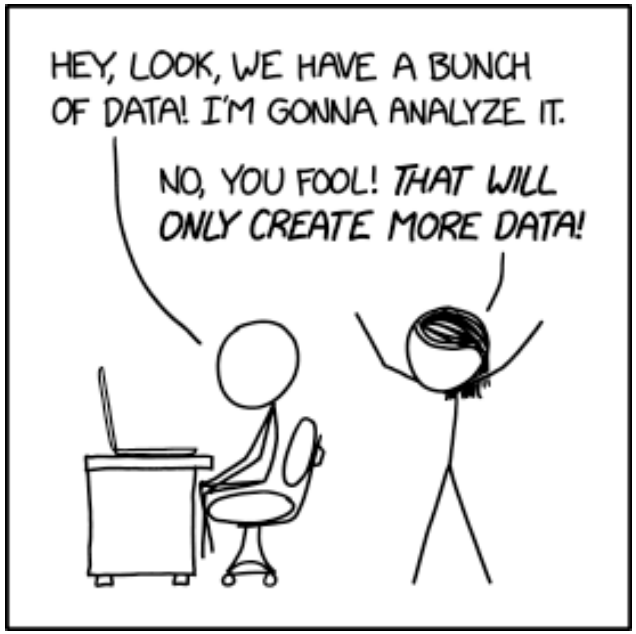
Das Erzeugen (fast beliebiger Mengen) von Daten ist heutzutage (in den Naturwissenschaften) meist kein Problem mehr. Der limitierende Schritt des wissenschaftlichen Erkenntnisgewinns ist die Datenauswertung.

Gründe

- Datenerhebung ist heute meist digital.
- Datenspeicher sind mittlerweile meist ausreichend groß.

Konsequenzen

- riesige Mengen an Daten, meist ohne strukturierte Ablage
- Daten werden oft erst erhoben und dann nachgedacht.



“It’s important to make sure your analysis destroys as much information as it produces.”

Die „Datenflut“

Eine Parallele aus der Frühzeit der Softwareentwicklung

> *. . . so long as there were no machines, programming was no problem at all; when we had a few weak computers, programming became a mild problem, and now we have gigantic computers, programming has become an equally gigantic problem. [. . .]*
>
> *The increased power of the hardware, together with the perhaps even more dramatic increase in its reliability, made solutions feasible that the programmer had not dared to dream about a few years before. And now, a few years later, he had to dream about them and, even worse, he had to transform such dreams into reality!*
>
> – Edsger Dijkstra

☛ Bei (mindestens) exponentiell wachsenden Datenmengen brauchen wir eindeutig Werkzeuge und Strategien zum Umgang mit ihnen.

E. W. Dijkstra, *Commun. ACM* 15:859, 1972

Erkenntnis aus der Informatik (Jim Gray)

- Datenbanken wurden dafür entwickelt,
  große strukturierte Datenmengen einfach durchsuchbar zu machen.
- Datenbanken erlauben komplexe verknüpfte Abfragen/Suchen.
- Datenbanken haben das Potential, Fragestellungen zu beantworten,
  für die bislang schlicht die Werkzeuge fehlten.

e-Science: Informatik trifft auf Wissenschaft

- frühes Beispiel: Bioinformatik
- eigene Teildisziplin einer jeweiligen wissenschaftlichen Disziplin,
  die digitale Werkzeuge für die Datenanalyse entwickelt
- Entwicklung der notwendigen digitalen Infrastrukturen

J. Gray in: Hey, Tansley, Tolle (Hg.): The Fourth Paradigm. Microsoft Research, Redmont 2009

Die vier Paradigmen nach Bell/Hey/Szalay/Gray

Eine erste Vorstellung

1. Experiment
   - Beobachtung steht im Zentrum
   - Daten als Grundlage der empirischen Wissenschaften
2. Theorie
   - Erklärungen für beobachtete Phänomene, Generalisierungen
   - Theoretische Vorhersagen als Ausgangspunkt neuer Experimente
3. Simulation
   - Zunehmende Komplexität der Theorien erzwingt Berechnungen
   - Im Wesentlichen erst seit Einführung des Computers möglich
4. datengetriebene Wissenschaft
   - Verfügbarkeit großer Mengen digitaler Daten
   - Datenverfügbarkeit definiert die Fragestellung

G. Bell, T. Hey, A. Szalay, *Science* 323:1297, 2009
J. Gray in: Hey, Tansley, Tolle (Hg.): The Fourth Paradigm. Microsoft Research, Redmont 2009

Die vier Paradigmen nach Bell/Hey/Szalay/Gray

Eine kritischere Betrachtung

- Reihenfolge von Theorie und Experiment diskussionswürdig
  - Experimente in der Antike verpönt, im Mittelalter teils verboten
  - Andererseits existieren astronomische Aufzeichnungen der Sumerer und Babylonier über viele Jahrhunderte.
- Der Begriff „Paradigma“ ist vermutlich schlecht gewählt.
  - Theorie, Experiment und Simulation sind keine Paradigmen i.e.S.

Paradigmata

allgemein anerkannte wissenschaftliche Leistungen,
die für eine gewisse Zeit einer Gemeinschaft von Fachleuten
maßgebende Probleme und Lösungen liefern

(Thomas S. Kuhn)

Thomas S. Kuhn: Die Struktur wissenschaftlicher Revolutionen, Suhrkamp, Frankfurt 1976, S. 10

Ein paar wichtige Punkte

- e-Science ist *nicht* das „vierte Paradigma“.
  - e-Science: digitale Werkzeuge und Infrastruktur
  - „viertes Paradigma“: datengetriebene Wissenschaft
- Datengetriebene Wissenschaft stellt das bewährte wissenschaftliche Vorgehen auf den Kopf.
  - Daten sollen ursprünglich der Überprüfung von Theorien dienen.
  - Datengetriebene Wissenschaft entwickelt Fragen aus den Daten.
- Datengetriebene Wissenschaft erzwingt Digitalität von Daten.
  - Nur digitale Daten sind für diese Form der Wissenschaft existent.
  - Aber: Nicht alle Daten liegen digital vor/sind zugänglich.

☛ Datengetriebene Wissenschaft ist ein grundlegend anderes/neues Vorgehen mit potentiell großen Auswirkungen auf die Wissenschaft.

Übersicht

Little Science, Big Science

„Big Data“, „Datenflut“, „e-Science“ und das „vierte Paradigma“

Die FAIR-Prinzipien zum Umgang mit Forschungsdaten

Individuelle Forschende mit klein(er)en Datenmengen

These

Die FAIR-Prinzipien zum Umgang mit Forschungsdaten werden überbewertet, missverstanden und setzen den falschen Fokus.
Sie taugen nicht als Begründung für Forschungsdatenmanagement und tragen vermutlich zum Niedergang der Wissenschaft bei.

Warum sich mit den FAIR-Prinzipien beschäftigen?

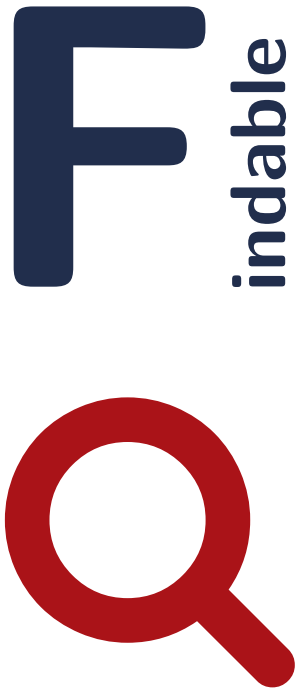
- Weil sie politisch relevant sind und man mitreden können sollte.
  - Forschungsdatenmanagement wird fast immer damit begründet.
- Weil sich fast niemand kritisch mit ihnen auseinandersetzt.
- Weil sie eine Gefahr für die Wissenschaft(lichkeit) sind.

Die FAIR-Prinzipien

Meist irrelevant – und gefährlich

„The FAIR Guiding Principles for scientific data management and stewardship“
Wilkinson *et al.*, *Scientific Data* 3:160018, 2016

Findable – Auffindbar



- F1 (Meta)daten sind mit einer weltweit eindeutigen und dauerhaften Kennung (*persistent identifier*, PID) versehen.
- F2 Daten sind durch aussagekräftige Metadaten beschrieben.
- F3 Metadaten enthalten klar und explizit die Kennung (PID) der Daten, die sie beschreiben.
- F4 (Meta)daten sind in einem durchsuchbaren Katalog registriert und indiziert.

☛ Konzepte: Metadaten, PID, Katalog

Wilkinson *et al.*, *Scientific Data* 3:160018, 2016

Accessible – Zugreifbar

A1 Der Zugriff auf (Meta)daten ist über ihre Kennung (PID) unter Nutzung standardisierter Kommunikationsprotokolle möglich.

- A1.1 Das Protokoll ist offen, frei und universell implementierbar.
- A1.2 Das Protokoll sieht ein Authentifizierungs- und Autorisierungsverfahren vor, falls erforderlich.

A2 Die Metadaten sind zugreifbar, selbst wenn die Daten nicht mehr verfügbar sind.

☛ Konzepte: digitales Protokoll, Autorisierung, Authentifizierung

Wilkinson *et al.*, *Scientific Data* 3:160018, 2016

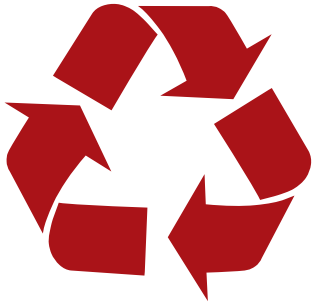
Interoperable – Interoperabel

- I1 (Meta)daten nutzen eine formale, zugängliche, geteilte und breit anwendbare Sprache für die Wissensdarstellung.
- I2 (Meta)daten nutzen ein Vokabular, das den FAIR-Prinzipien folgt.
- I3 (Meta)daten enthalten qualifizierte Referenzen zu anderen (Meta)daten.

☛ Konzepte: formale Sprache, Ontologie, Querverweis

Wilkinson *et al.*, *Scientific Data* 3:160018, 2016

Reusable – Wiederverwendbar



- R1 (Meta)daten werden durch eine Vielzahl genauer und relevanter Attribute reichhaltig beschrieben.
  - R1.1 (Meta)daten sind unter einer klaren und zugänglichen Datennutzungslizenz veröffentlicht.
  - R1.2 (Meta)daten sind mit detaillierten Herkunftsangaben verbunden.
  - R1.3 (Meta)daten entsprechen einschlägigen Standards des jeweiligen Fachs.

☛ Konzepte: Lizenz, Provenienz, fachspezifische Standards

Wilkinson *et al.*, *Scientific Data* 3:160018, 2016

Die FAIR-Prinzipien

Meist irrelevant – und gefährlich

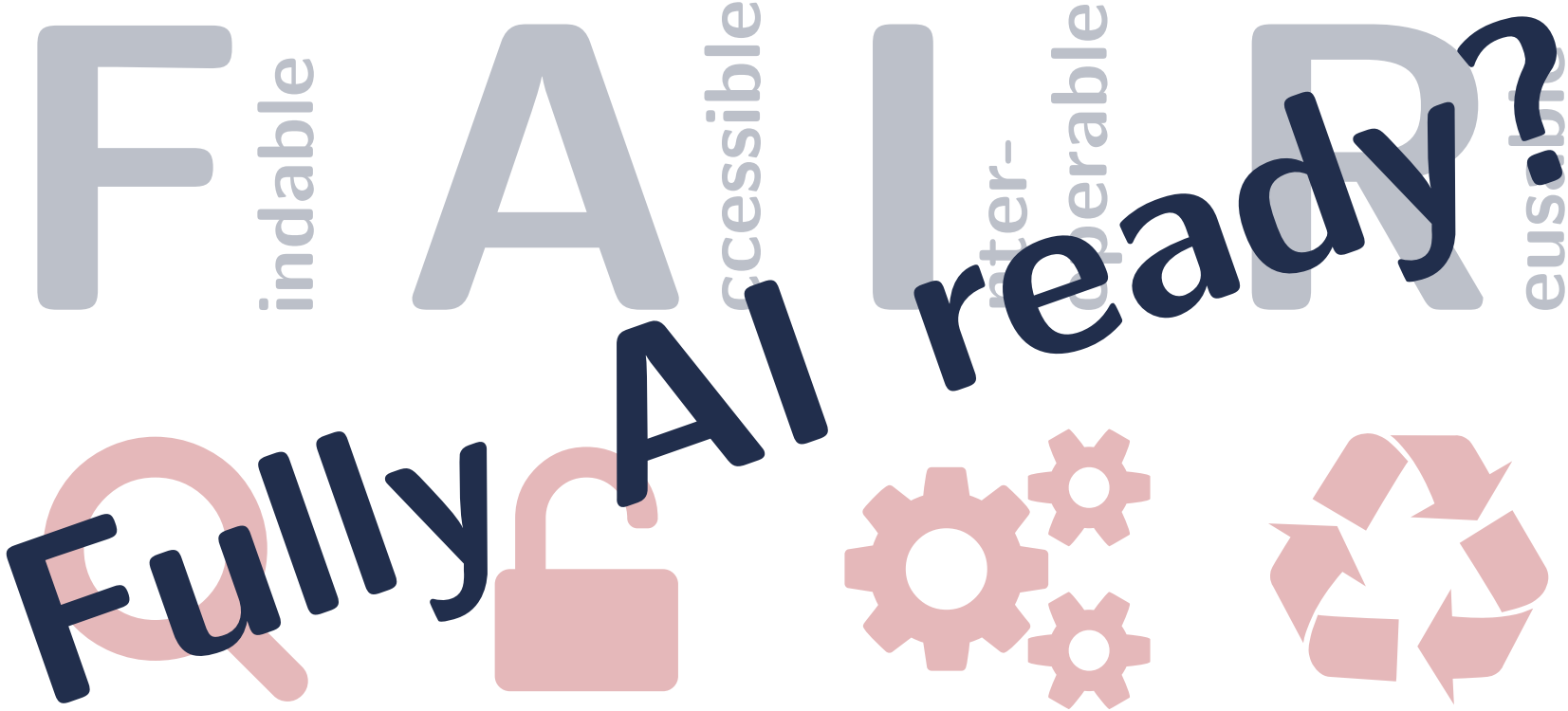
> *There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. [. . .]*
> *Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.*
>
> – Wilkinson *et al.*, Abstract

- klarer Fokus auf Maschinen statt Menschen
  - aber: Ziel der Wissenschaft ist Erkenntnisgewinn
  - Erkenntnis setzt ein erkennendes Subjekt voraus.
- Forschungssdatenmanagement thematisiert den Umgang der Forschenden mit ihren Daten.

Wilkinson *et al.*, *Scientific Data* 3:160018, 2016

Die FAIR-Prinzipien

Meist irrelevant – und gefährlich



Mons, *Nature* **578**:491, 2020

Die FAIR-Prinzipien

Meist irrelevant – und gefährlich

Probleme mit den FAIR-Prinzipien

- fehlender Fokus auf der Qualität der Daten (Datenkuration)
- Daten sind nur die Grundlage wissenschaftlicher Erkenntnis.
- Veröffentlichung von Forschungsdaten ist kein Selbstzweck.

Probleme der Veröffentlichung der FAIR-Prinzipien

- ignoriert komplett die Vorarbeiten und ist damit unwissenschaftlich

Probleme mit KI/ML in den Wissenschaften

- KI/ML ist ein Werkzeug, produziert aber keine Erkenntnis.
- KI/ML kann nur auf zugänglichen Daten operieren.
  Die Frage der Repräsentativität der Daten wird (fast) nie gestellt.
- KI/ML liefert Korrelationen, keine Kausalitäten.

Probleme mit den FAIR-Prinzipien

Falscher Fokus und damit Gefahr für die Wissenschaft(lichkeit)

> *Big data begets big attention these days, but little data are equally essential to scholarly inquiry. [. . .] However, big data is not necessarily better data. The farther the observer is from the point of origin, the more difficult it can be to determine what those observations mean—how they were collected; how they were handled, reduced, and transformed; and with what assumptions and what purposes in mind. Scholars often prefer smaller amounts of data that they can inspect closely.*
>
> – Christine L. Borgman

- Datenqualität ist wichtiger als Datenmenge.
- Fehlende Datenqualität kann dazu führen, dass gar keine Daten verfügbar sind.

C. L. Borgman: Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press, Cambridge, MA, 2015. S. xvii

Probleme mit den FAIR-Prinzipien

Falscher Fokus und damit Gefahr für die Wissenschaft(lichkeit)

> *Bei der Entwicklung der Wissenschaft spielen Beobachtungen und Experimente nur die Rolle von kritischen Argumenten.*
>
> – Karl Popper

> *Experimental observations are only experience carefully planned in advance, and designed to form a secure basis of new knowledge ; that is, they are systematically related to the body of knowledge already acquired, and the results are deliberately observed, and put on record accurately. As the art of experimentation advances the principles should become clear by virtue of which this planning and designing achieve their purpose.*
>
> – R. A. Fisher

☛ Daten sind nur die Grundlage wissenschaftlicher Erkenntnis.

K. Popper, in: D. Miller (Hg): Karl Popper Lesebuch. Mohr Siebeck, Tübingen 1997, S. 9
R. A. Fisher: The Design of Experiments. 7. Aufl. Hafner Press, London 1971, S. 8

Probleme mit den FAIR-Prinzipien

Falscher Fokus und damit Gefahr für die Wissenschaft(lichkeit)

> *Scholarship and data have long and deeply intertwined histories. Neither are new concepts. What is new are efforts to extract data from scholarly processes and to exploit them for other purposes. Costs, benefits, risks, and rewards associated with the use of research data are being redistributed among competing stakeholders. The goal of this book is to provoke a much fuller, and more fully informed, discussion among those parties.*
> *At stake is the future of scholarship.*
>
> – Christine L. Borgman

- Das Herausreißen von Forschungsdaten aus ihrem Kontext ist ein großes Problem für den Erkenntnisgewinn.

C. L. Borgman: Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press, Cambridge, MA, 2015. S. xix; Hervorhebung nicht im Original.

Probleme mit den FAIR-Prinzipien

Ignoranz der Vorarbeiten und damit Unwissenschaftlichkeit

> *Publications that report findings set them in the context of the domain, grounding them in the expertise of the audience. Information necessary to understand the argument, methods, and conclusions are presented.*
>
> – Christine L. Borgman

Kernaspekt der Wissenschaftlichkeit: Kontextualisierung

- Voraussetzung: Kenntnis des Standes der Forschung
- Direkte Bezugnahme auf relevante Arbeiten

☛ Literaturrecherche ist eine notwendige Grundqualifikation.

❗ Die älteste Referenz in der FAIR-Publikation ist von 1997.

❗ Zentrale politische Richtlinien (z.B. OECD, 2007) fehlen ganz.

C. L. Borgman: Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press, Cambridge, MA, 2015. S. xviii

Probleme mit den FAIR-Prinzipien

Ignoranz der Vorarbeiten und damit Unwissenschaftlichkeit

- Zentrale Aspekte der FAIR-Prinzipien
  - maschinell verarbeitbare Metadaten
  - Verknüpfte Metadaten ($\rightarrow$ Wissensgraph)
- Ein bisschen (ausgesuchte) Vorgeschichte

1945	„As we may think“ (Vannevar Bush)
1950er	Ontologien, . . . in der KI-Forschung
1960er	Expertensysteme in der KI-Forschung (Feigenbaum *et al.*)
1960er	Vorführung u.a. von Hypertext-Links (Doug Engelbart)
1990	„World Wide Web“ (Tim Berners-Lee)
2001	„semantic web“ (Tim Berners-Lee)
2006	„linked data“ (Tim Berners-Lee)
2007	OECD: Forschungsdaten zugänglich machen
2016	FAIR-Prinzipien (Wilkinson *et al.*)

☛ Die anderen wussten voneinander und nahmen Bezug aufeinander.

Ein paar Anmerkungen zu KI/ML

- Es gibt keine anerkannte Definition von „Intelligenz“, umso weniger eine von „künstlicher Intelligenz“.
- Statt „KI“ haben wir Algorithmen des maschinellen Lernens (ML) – effektiv (komplexere) Statistik auf großen Zahlen.
- LLMs (ChatGPT etc.) sind Markov-Prozesse (erster Ordnung) – und damit weder kreativ noch zu Erkenntnis fähig.
- Die Qualität der Mustererkennung reicht für Online-Werbung – aber nicht als Werkzeug für wissenschaftlichen Erkenntnisgewinn.
- Muster sind erst einmal ein Phänomen – aber keine Erkenntnis.

☛ KI/ML mag als Werkzeug funktionieren – oder auch nicht.

☛ Es ist an uns, kritisch und überlegt damit umzugehen.

Exkurs: Korrelation, Koinzidenz, Kausalität

Eine mathematische Eigenschaft und ihre Interpretation

Korrelation

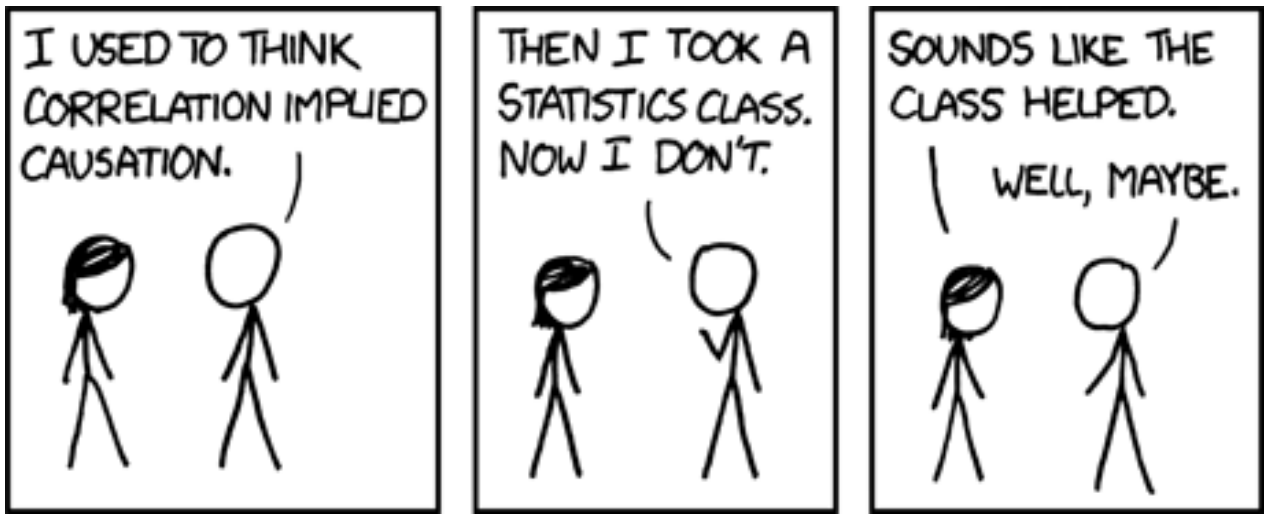
Maß für den Zusammenhang zweier Größen; mathematisch i.d. R. auf $|[0..1]|$ beschränkt, wobei negative Werte eine Antikorrelation anzeigen.

Koinzidenz

Zeitliches oder/und räumliches Zusammentreffen von Ereignissen oder Objekten; deskriptiver Begriff ohne Implikation von Zusammenhängen.

Kausalität

Ursache-Wirkungs-Beziehung zwischen Ereignissen und Zuständen: A ist die Ursache von B, wenn B von A erzeugt wird.



"Correlation doesn't imply causation,
but it does waggle its eyebrows suggestively
and gesture furtively while mouthing 'look over there'."

Das „Problem“ mit Korrelation und Signifikanz

Mathematische Algorithmen mit garantiertem Ergebnis

Grundsätzliche Eigenschaften von Korrelation und Signifikanz

- Korrelation und Signifikanz sind für beliebige Größen berechenbar.
- Die Algorithmen liefern *immer* ein Ergebnis.

„Probleme“ mit dieser Art von Algorithmen

- Ein Ergebnis ist noch keine Aussage.
- Korrelation hat keinerlei Erklärungspotential.
- Signifikanz ist ein rein statistischer Begriff.

Probleme bei der Verwendung statistischer Verfahren

- Die Wenigsten sind sich der (impliziten) Annahmen bewusst.
- Oft genug sind die impliziten Annahmen nicht gerechtfertigt – ohne dass jemand darüber nachdenkt (Bsp: Normalverteilung).

Datenqualität

- Können nur diejenigen entscheiden, die Daten und Kontext kennen
- Daten ohne Metadaten sind wertlos.
- Metadaten machen Daten noch nicht wertvoll.
- Qualität lässt sich nicht messen (quantitizieren).

Repräsentativität

- „Big Data“, datengetriebene Wissenschaft und FAIR-Prinzipien berücksichtigen alle nur digital vorliegende Daten.
- Nicht alle Daten sind digital, nicht alle digitalen Daten zugänglich.
- Es gibt Disziplinen, die sich mit Repräsentativität auskennen...

☛ Die Gefahr: falsche Antworten auf irrelevante Fragen basierend auf nichtrepräsentativen und qualitativ schlechten Daten

Übersicht

Little Science, Big Science

„Big Data“, „Datenflut“, „e-Science“ und das „vierte Paradigma“

Die FAIR-Prinzipien zum Umgang mit Forschungsdaten

Individuelle Forschende mit klein(er)en Datenmengen

Was geht mich das alles an?

Individuelle Forschende mit klein(er)en Datenmengen

- Mit welchen Datenmengen und -arten arbeiten Sie typischerweise?
- Haben „Big Data“, „e-Science“ und datengetriebene Wissenschaft für Sie eine Relevanz?

☛ Individuelle Forschende sind für den sorgsamen Umgang mit ihren Forschungsdaten verantwortlich.

Was geht mich das alles an?

Verantwortungsvoller Umgang mit den eigenen Daten

- Haben Sie eine vollständige Übersicht über Ihre Daten?
- Haben Sie ein System für Ihre Datenablage?
- Ist dieses System erweiterbar und für andere verständlich?
- Können Sie nachvollziehen, wie Ihre Daten entstanden sind?
- Haben Sie Mechanismen, um die Qualität Ihrer Daten zu bestimmen und dann auch zu dokumentieren?
- Haben Sie geeignete Werkzeuge, um über mehrere/viele Dateien/Datensätze hinweg Daten auszuwerten?

- Wir müssen strukturiert und überlegt mit unseren Daten umgehen.
- Wir müssen unsere Arbeit *hinreichend detailliert* dokumentieren.

Nochmal: die FAIR-Prinzipien

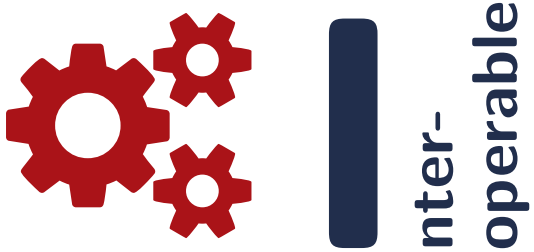
Wie lassen sie sich auf den individuellen Forschungskontext anwenden?

Weiß ich, wo genau sich dieser eine Datensatz befindet, den ich aufgezeichnet zu haben glaube und JETZT dringend für eine Veröffentlichung benötige? Und wenn nicht, hätte ich eine Chance, ihn zu finden?

Habe ich Zugang zu den Daten, oder befinden sie sich auf dem Computer, der Festplatte, dem Speicherstick eines anderen – oder gar bei meiner alten Institution (ohne dass ich dort anrufen und um Hilfe bitten könnte)?

Nochmal: die FAIR-Prinzipien

Wie lassen sie sich auf den individuellen Forschungskontext anwenden?



Habe ich die Daten in ein Format exportiert, mit dem ich arbeiten kann – ohne Zugriff auf die Software dieses alten Geräts, das vor fünf Jahren ausgemustert wurde – oder alternativ weit weg oder in meiner alten Einrichtung?

Habe ich alle notwendigen Informationen, d.h. Metadaten, um alle Fragen zu beantworten, die ich jetzt – mit viel mehr Erfahrung und Kontext und zum ersten Mal wirklich mit Blick auf die Daten – haben könnte?

Nochmal: die FAIR-Prinzipien

Wie lassen sie sich auf den individuellen Forschungskontext anwenden?

> *Note that there is no way to email yourself in the past to ask for clarifications.*
>
> – Allesina & Wilmes 2019, p. 2

☛ Sei fair zu dir selbst – und denke an dein Zukunfts-Ich.

Entscheidende Aspekte

- Die FAIR-Prinzipien lassen sich umdeuten und auf den individuellen Forschungskontext anwenden.
- Wiederverwendung bedeutet zuallererst (und meist ausschließlich): Wiederverwendung durch mein Zukunfts-Ich.
- FAIR ist nur ein Werkzeug, um die richtigen Fragen zu stellen.

Allesina & Wilmes: Computing Skills for Biologists. Princeton University Press, Princeton & Oxford 2019

Zentrale Aspekte

- Automatisierung und Digitalisierung ermöglichen es, nahezu beliebige Datenmengen (relativ) einfach zu erzeugen.
- Die Zunahme von Datenmenge und Digitalität erzwingt, Strategien zu ihrem Umgang zu entwickeln und zu etablieren.
- „Big Data“ und das Arbeiten auf digitalen Daten anderer ist für die wenigsten wissenschaftlichen Disziplinen relevant.
- Die FAIR-Prinzipien zum Umgang mit Forschungsdaten werden überbewertet, missverstanden und setzen den falschen Fokus.
- Die Qualität von Forschungsdaten, obwohl entscheidend, ist nur schwer durch Kriterien bestimmbar.