

Datenaufbereitung und -kuration im Spannungsfeld von Reproduzierbarkeit und Wiedernutzung als Leitideen der Open Sciences.

Eine Fallstudie aus der Kunstgeschichte

Niemann, Klara

klaraniemann@gmx.de
Universität zu Köln

Klammt, Anne

aklammt@hotmail.com
Deutsches Forum für Kunstgeschichte Paris

Einleitung

Zu den Anstrengungen der Europäischen Union für die Umsetzung von *Open Sciences* gehört auch eine Klärung und Förderung der Reproduzierbarkeit und der Sicherung von Integrität. Unter Reproduzierbarkeit wird dabei im *Scope Report* der Europäischen Kommission die Wiederholbarkeit des Forschungsprozesses mit denselben Daten und Methoden verstanden (Directorate-General for Research and Innovation 2020). Für die Autor*innen des Reports stellt die Reproduzierbarkeit einen spezifischen Fall der Wiedernutzbarkeit (Re-Use) von Forschungsdaten dar. Aus unserer Sicht bilden jedoch die strikte Reproduzierbarkeit und eine offene Wiedernutzung Szenarien, die in der Praxis zu widersprüchlichen Anforderungen an die Kuration von Forschungsdaten sowie ihrer Ausgabe über graphische Nutzeroberflächen und APIs führen. Im Verhältnis mit der Datenintegrität entsteht dabei ein nicht vollständig aufzulösendes Dilemma, das Versuche der Aushandlung aber in eine produktive Spannung überführen können. Dies möchten wir verdeutlichen, indem wir einerseits unseren Gebrauch der Begriffe Re-Use, Neuausrichtung und Datenkuration präzisieren, und die Problematik andererseits an einem Fallbeispiel aus der Forschungsarbeit des Deutschen Forums für Kunstgeschichte Paris (DFK Paris) veranschaulichen, für das wir nach Wegen gesucht haben, zwischen den Polen Reproduzierbarkeit und Re-Use zu vermitteln. Es handelt sich um die Aufbereitung und neue Präsentation einer 20 Jahre alten Datenbank zur wechselseitigen Rezeption des Kunstgeschehens in Tex-

ten der Kunstkritik aus Deutschland und Frankreich zwischen 1870 und 1960 (DFKV) (DFK Paris 2022b).

Fallstudie

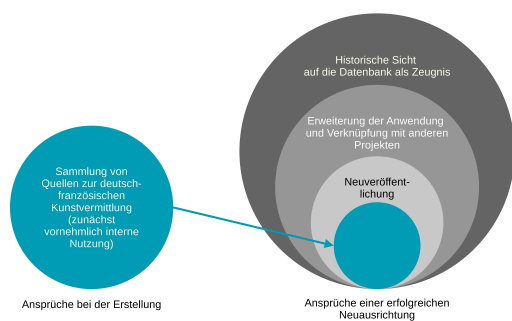
Von 1999 bis 2005 wurden in drei aufeinanderfolgenden Förderungen Quellenanthologien, Aufsätze, eine Monografie und drei bibliographische Datenbanken zur deutsch-französischen Kunstrezeption erstellt (DFK Paris 2022c), die den Blickwinkel der in den 1990er und 2000er Jahren sehr einflussreichen Kulturtransferforschung einnahmen (Espagne & Werner 1988; Espagne 1999; Gaethgens 2009). Die Datenbanken waren technologisch und im Datenmodell einheitlich angelegt, während sich die Verschlagwortung jeweils spezifisch entwickelte. Insgesamt beinhalten sie knapp 6800 mehrheitlich kommentierte und verschlagwortete Hinweise auf Artikel, Meldungen und Notizen in 314 verschiedenen Reihen deutscher, beziehungsweise französischer Kunstzeitschriften und wenigen zeitgenössischen Buchpublikationen. Ziel war es, Kunsthistoriker*innen ein Hilfsmittel zur wissenschaftlichen Arbeit anzubieten. Ab Winter 2004/2005 standen sie offen online zur Verfügung und sind bis 2016 über die Webseite des DFK Paris auffindbar gewesen. 2019 erfolgte ein erster Relaunch der zwischenzeitlich in eine MySQL Datenbank migrierten Datenbanken. Von 2021 bis 2022 wurden die Daten umfangreich kuratiert und im Juni 2022 neu veröffentlicht (DFK Paris 2022a). Ausgangspunkt war die mangelhafte Nutzbarkeit der öffentlichen Datenpräsentation, die aus der mehrfachen Migration hervorgegangen war und das Verständnis für die Zusammensetzung und Bedeutung der Daten minderte. Das Webangebot führt heute ein einstiges Werkzeug fort, dessen zugrundeliegende Forschungsfrage des Kulturtransfers inzwischen aufgrund der Weiterentwicklung hin zur Untersuchung von Mobilität und Migration nicht mehr in gleicher Weise gestellt wird.

Begrifflichkeiten: Re-Use, Datenkuration und Neuausrichtung

Um die von uns verwendeten Begrifflichkeiten zu schärfen und Divergenzen zu anderen Definitionen vorzubeugen, sei eine Erläuterung des hier zugrunde gelegten Verständnisses der Begriffe *Re-Use*, *Neuausrichtung* und *Datenkuration* vorangestellt. Re-Use beschreibt die erneute, offen gedachte Nutzung von Daten in der Forschung zur Beantwortung einer neuen Fragestellung. Dies kann die Rekombination mit anderen Daten beinhalten und ist ein wesentliches Ziel der Bemühung um die Anwendung der FAIR Prinzipien (Huie et al. 2021). Als Datenkuration möchten wir hier eine Gruppe von Aktivitäten verstehen, die an der Datenhaltenden Institution angesiedelt ist, und mit dem Ziel, eine verständliche Nachnutzung durch Dritte zu ermöglichen, durchgeführt wird. Notwendige Voraussetzung sind dabei die Dokumentation und Kenntnis der Erzeugung, Prozessierung und Anzeige der Daten, wie es Flanders und Muñoz

aus Perspektive der Geisteswissenschaften zusammengestellt haben (Flanders u. Muñoz 2015). Dabei weisen die von Kim und Koh (Kim u. Koh 2021) herausgegebenen Forschungen zur Geschichte von Digital Humanities-Projekten darauf hin, dass unter „Erzeugung“ wie „Prozessierung“ ein Zusammenspiel von theoretischer Ausrichtung, methodischer Vorgehensweise und organisatorischen wie institutionellen Faktoren zu betrachten ist. Die Dokumentation nach Flanders und Muñoz ist die Voraussetzung, um Daten überhaupt auf eine neue Verwendung hin aufzubereiten. Die Herausforderung liegt nach Woodall (2017) dann darin, die Eignung der Daten für diesen neuen Anwendungsfall bestimmen zu können und sie gezielt daraufhin zu entwickeln, um Fehlinterpretationen vorzubeugen. Im Idealfall sollte die Datenkuration darauf gerichtet sein, innovative Forschungen zu ermöglichen und die Daten entsprechend für möglichst viele Ansatzpunkte öffnen, beispielsweise durch Verknüpfung mit Normdaten.

Mit dem Begriff der Neuausrichtung schlagen wir eine spezifische Auslegung der Datenkuration vor, die sich diesen Herausforderungen auch auf Ebene der Benutzeroberfläche widmet, wie es die Medienwissenschaftlerin Drucker (2021, 78) für die Aufbereitung von Forschungsdaten in den Geisteswissenschaften angeregt hat. Wir erweitern das Verständnis von der Aufbereitung damit inhaltlich gegenüber dem vorher in den *Data Sciences* auf die Daten gerichteten Fokus (Woodall u. Wainman 2015) und nähern es dem *Refashioning* nach Bolter und Grusin (1999, 45 f.) an. Ziel ist es also, einen (alten) bestehenden Datensatz visuell und textuell so zu vermitteln, dass dessen Beschaffenheit für den Re-Use verständlich wird (bspw. durch eine entsprechende Suchmaske). Indem die Neuausrichtung durch ihre spezifische kuratorische Ordnung und Anreicherung der Daten jedoch bereits Beispiele der Weiternutzung, Interpretation und Verknüpfung anbietet, nimmt sie ein ambivalentes Verhältnis sowohl zur Maxime der weitmöglichsten Öffnung der Daten zum Re-Use als auch zur historischen Gewachsenheit der Daten ein. Diese gilt es wiederum zu kommunizieren.



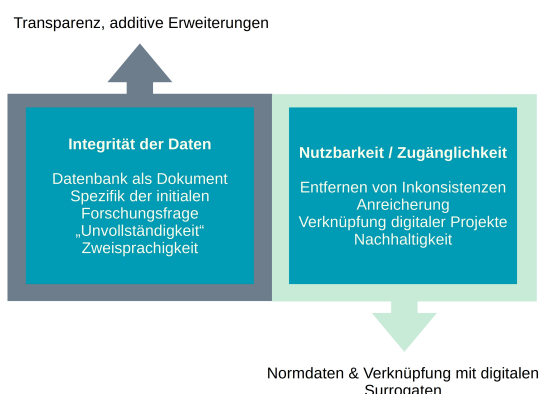
Die verschiedenen Nutzungsansprüche einer republizierten Datenbank im Vergleich zur ursprünglichen Zielsetzung am Beispiel der Datenbanken DFKV (Grafik: Klara Niemann, CC BY 4.0).

Vom Dilemma zwischen Neuausrichtung und historischer Integrität

Als Maßnahmen der Datenkuration möchten wir aus unserem Fallbeispiel die semantische Anreicherung von Daten durch die Referenzierung mit Normdaten und die Einbeziehung von digitalen Surrogaten des Quellmaterials für Datenbanken der DFKV heranziehen. Auf diese Weise wurden von 9076 in den Datenbanken DFKV genannten Personen (Autor*innen, Künstler*innen, Kurator*innen und weitere) 4686 (52 %) mit Normdaten (Getty Union List of Artists Names (ULAN), Gemeinsame Normdatei (GND), *notices d'autorité* der Bibliothèque nationale de France (BnF)) oder Wikidata referenziert. Dies hat zur Entdeckung von insgesamt 603 Personennamen geführt, die entweder in mehr als einer der Datenbanken auftreten oder auch unbemerkt jeweils mit verschiedenen Namensvarianten eingetragen wurden. Diese Varianten sind meist in den Quellen angelegt, wenn etwa Vornamen und Adelstitel in Französisch oder Deutsch übersetzt wurden. Die Gewohnheit, in den Zeitschriften Namen zu übersetzen oder auch Umlaute und Akzente anzupassen, ist ein Hinweis zur ursprünglichen Rezeption. Dass diese verschiedenen Schreibweisen bei der Datenerfassung in den 2000er Jahren übernommen, aber nicht mit übereinstimmenden Personen assoziiert wurden, ist wiederum eine wichtige Information zur Einschätzung der Qualität der Daten. Darüber hinaus konnten einige Aliasnamen aufgedeckt werden, die den Erstbearbeiter*innen nicht bekannt waren.¹ Zusammengefasst beträgt der Anteil an Dubletten damit rund 9 %. Für 2548 der 5735 in der Datenbank beschriebenen Zeitschriftenbeiträge (Zeitraum vor 1940) konnte eine Verlinkung auf ein Digitalisat in den Angeboten der Universitätsbibliothek Heidelberg oder der BnF erstellt werden. Wo ein Dateneintrag zusammenfassend auf mehrere verschiedene Zeitschriftenbeiträge verweist, musste die Datengrundlage erweitert werden, sodass die Gesamtanzahl bibliografischer Attribute in der Datenbank durch die Kuration von 5948 auf 6194 angestiegen ist.

Im Sinne der Neuausrichtung sollten die Mehrfachnennungen der Personen zusammengefasst werden und die Qualität der Daten einschätzbar sein, aber zugleich die historisch bedingte Ambiguität erhalten bleiben. Dieses Dilemma zwischen der Neuausrichtung und dem Erhalt der ursprünglichen Datenbanken als Artefakt hat uns zur Frage geführt, was die historische Integrität dieser Daten ausmacht. Medienarchäologische Studien und Datenzentren haben von unterschiedlichen Ausgangspunkten ausgehend wahlweise die vollständige Emulation oder die Erhaltung der Präsentationsschichten bei einem Wechsel der zugrundeliegenden Technik vorgeschlagen (Waelder 2017; Steiner et al. 2022). Im Falle der Datenbanken der DFKV sind jedoch zum einen bereits unterschiedliche Softwares und Ansichten zur Dateneingabe und für die Ausgabe im Internet verwendet worden, sodass mehrere Versionen emuliert werden müssten. Zum anderen hat sich aus den Befragungen von ehemaligen Mitarbeiter*innen zum Gebrauch der Datenbanken in den 2000er Jahren ergeben, dass die Software als

solche kaum wahrgenommen wurde und weder Funktionen zur Filterung noch des Exports später beschrieben werden konnten.² In dieser Situation haben wir uns entschlossen, nicht die historische Integrität der Datenbanken als Ganzes zu betrachten, sondern auf die Ebene der einzelnen Datensätze zu gehen und sie abstrakter als einen definierten Zustand der Daten anzusehen.³ Dadurch sind wir dazu gelangt, die Datenbanken orientiert an CIDOC CRM als Konvolute von Dokumenten (E31; Bekiari et al. 2022, 83 f.) zu verstehen, deren Erzeugung wie auch Anreicherung diskrete Ereignisse (E5; Bekiari et al. 2022, 63 f.) ihrer Objektgeschichte bilden. Mit den Ereignissen sind ein Zeitraum (1999–2004 und 2021–2022), die ausführenden Personen, die Art der Aktivitäten und damit die Zusätze und Streichungen beschreibbar.



Das Dilemma zwischen Integrität und Nutzbarkeit der Daten am Beispiel der Datenbanken DFKV (Grafik: Klara Niemann, CC BY 4.0).

Gestaltung des GUI

Ausgehend von der Idee, die Zustände und Anreicherungen der Daten selbst erfahrbar zu machen, haben wir das GUI gestaltet (Niemann 2021). Es ist in drei Funktionsbereiche aufgeteilt: die Übersichtsseite zur Suche und Auswahl der in verkürzter Ansicht gezeigten Dateneinträge, die vollständige Ansicht der einzelnen Dateneinträge und eine Merkliste mit individuell von den Nutzer*innen ausgewählten Einträgen.

In der vollständigen Ansicht der einzelnen Dateneinträge wurde mit Farbhintergründen und Schichten gearbeitet, die die Inhalte den verschiedenen Phasen der Datenbanken und ihrer Bearbeitung zuordnen.⁴ Auf neutralem Grund sind bibliografische Angaben und Textauszüge angelegt, die faktische Informationen zum aufgenommenen Quelltext liefern. Farblich hinterlegt sind die Schlagworte, Kommentierungen und weitere Anreicherungen der Datenerfassung und somit des ersten objektgeschichtlichen Ereignisses. Als Widgets können für die Autor*innen und die als genannte Personen angegebenen Namen die Referenzierungen auf Normdaten oder Wikidata und Hinweise auf weitere in der Datenbank vorhandene Schreibweisen aufgerufen werden. Sie bilden somit das Ereignis der Kuration ab. In gleicher Weise sind Informationen zur Zeitschrift bei der BnF oder der

GND aufrufbar. Als Fly-in schiebt sich von rechts über die Grundebene ein Widget mit Erläuterungen zur Nutzung, die anhand von Symbolerklärungen die Hintergründe der objektgeschichtlichen Stationen und die Entscheidungen der Kuration transparent vermitteln. Als äußerste Schicht kann jeder Dateneintrag als JSON in einem weiteren Widget aufgerufen werden, um die Anreicherungen und Verknüpfungen der Informationen per IDs (als Spuren der ursprünglichen Erfassung in einer relationalen Datenbank) nachzuvollziehen. Insgesamt stellen diese Gestaltungsentscheidungen eine Reaktion auf den Anspruch der Nachvollziehbarkeit und damit der Reproduzierbarkeit dar.

Die Suchfunktionen auf der Übersichtsseite sind hingegen auf den Re-Use ausgerichtet. Die verschiedenen Optionen, über die Datenbankzugehörigkeit, mit dynamischen Filtern in vorgegebenen Kategorien, dem interaktiven Zeitstrahl oder der Freitextsuche zu suchen, regen eine Entdeckungstour durch die Daten an, bei der die Nutzer*innen weniger über eine spezifische Suche als ein Schweifen in das Material einsteigen. Um dies zu fördern, kann man sich auf der bereits beschriebenen Vollanzeige der Dateneinträge außerdem horizontal per Klick von einem zum nächsten bewegen. Die digitalen Surrogate werden schließlich über ein Icon aufgerufen und öffnen sich in einem IIIF-Viewer, der sich in einem neuen Browsertab öffnet. Indem die Manifeste der Zeitschriftenbände verknüpft sind, können die referenzierten Artikel und Beiträge vollständig gelesen werden, in dem Band geblättert werden und weitere von den Bibliotheken erstellte Metadaten aufgerufen werden.

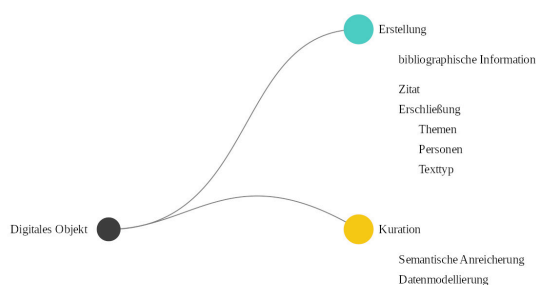
Umsetzung in der Datenmodellierung

Nicht implementiert haben wir eine vorerst prototypische Umsetzung dieser Schichtung in einem Datenmodell, das wir mit dem Linked Art Data Model erstellt haben (Klammt 2022). Das Linked Art Data Model (LADM) ist eine Anwendung des CIDOC CRM (Linked Art), das den Re-Use von Kulturdaten unterstützen möchte. Anders als LIDO XML geht es dabei nicht um ein Transferformat, mit dem Metadaten verschiedener Quellen zusammengeführt werden können, sondern darum, Kulturdaten zu einfach nutzbaren Linked Open Data zu machen. Dieser Fokus spiegelt sich auch in der Wahl von JSON als Dokumentformat. In einem ersten Prototyp haben wir die einzelnen Dateneinträge als Informationsobjekt modelliert, das auf meist einen Zeitschriftenbeitrag referenziert, dessen bibliografische Angaben mit dem LADM ausgedrückt werden können. Jeder Dateneintrag hat als Ereignisse seine Erstellung und die Kuration eingetragen, mit den Zeiträumen und den jeweils verantwortlichen Projektleitern. Über verschiedene definierte Eigenschaften sind die Verlinkung zum IIIF-Manifest, die Kommentierungen und Verschlagwortung genauso wie die Referenzierung auf Normdaten im Zuge der Neuausrichtung als LOD erklärt. Die Verwendung des Modells für die Beschreibung von Forschungsdaten liegt außerhalb seiner ursprünglichen Intention. Sie erlaubt aber auf Datenebene transparent zu dokumentieren, welche Maßnahmen zur Ausrichtung der Daten auf den Re-Use

ausgeführt wurden. Gleichzeitig können diese Veränderungen reversibel eingeschrieben werden.

Resümee

Im Prozess der Neuausrichtung der DFKV-Datenbanken durch verschiedene Anreicherungsprozesse und das Einbetten in ein neues GUI sahen wir uns in der Praxis mit der Aufgabe einer offen gedachten Reproduzierbarkeit gegenübergestellt. Reproduzierbarkeit ist dabei sowohl der Weiterverwendung als auch der Integrität der Daten verpflichtet. Wenn es darum geht, geisteswissenschaftliche Datenbanken langfristig zu erhalten, heißt das, diese auch jenseits der Langzeitarchivierung im Sinne der dynamischen Entwicklung der Forschungsfragen anschlussfähig an die wissenschaftliche Community zu halten. Entscheidungen der Datenkuration müssen in diesem Prozess individuell entsprechend des Einzelfalls, aber immer respektive der Geschichte und konkreten Beschaffenheit der Daten getroffen werden. Versteht man alle Eingriffe und Veränderungen als spezifische Ereignisse der Datenhistorie, gilt es, diese für den Re-Use transparent zu kommunizieren. Im Angesicht der ersten, nicht mehr abrufbaren Datenbanken aus den späten 1990ern und frühen 2000ern wird der Handlungsbedarf in diesem Bereich deutlich.



Datensätze als digitales Objekt, das durch Erstellung und Kuration geformt wird (Grafik: Anne Klammt; Lizenz: CC BY 4.0).

Fußnoten

1. Ein Beispiel ist die Kunsthistorikerin Lina Boelsche, die unter dem männlichen Synonym Hermann Billung Zeitschriftenartikel veröffentlichte und dementsprechend unter diesem Namen in den Datenbanken geführt wurde. Die Verknüpfung mit Wikidata (<https://www.wikidata.org/wiki/Q95196696>) ermöglichte die nachträgliche Identifikation.
2. Die Befragung per Interviews und Fragebögen wurde von Deborah Schlauch, DFK Paris, von Januar bis Mai 2022 durchgeführt.
3. Unbemerkt haben wir uns damit der Frage nach den „signifikanten Eigenschaften“ nach Giaretti et al. (2009) und Recker (2021) angenähert. Für den freundlichen Hinweis danken wir T. Staecker, Darmstadt.
4. Als Beispiel eines ausführlichen Datenbankeintrags siehe: https://dfk-paris.org/de/page/deutsch-franzoesische-kunstvermittlung-1870_1940-und-1945_1960-datenbank-2391.html#/records/10720.

sische-kunstvermittlung-1870_1940-und-1945_1960-datenbank-2391.html#/records/10720.

Bibliographie

Bekiari, Chryssooula et al. 2022. "Definition of the CIDOC Conceptual Reference Model Version 7.2.1. Volume A: Definition of the CIDOC Conceptual Reference Model. Version 7.2.1".

Bolter, Jay David und Richard Grusin. 1999. *Remediation. Understanding New Media*. Cambridge: MIT Press.

DFK Paris. 2022. "Datenkuration am Beispiel der Datenbank Deutsch-Französische Kunstvermittlung 1870-1940 und 1945-1960". <https://dfk-paris.org/de/research-project/datenkuration-am-beispiel-der-datenbank-deutsch-franz%C3%B6sische-kunstvermittlung-1871> (zugegriffen 2. August 2022).

DFK Paris. 2022. "Deutsch-Französische Kunstvermittlung 1870-1940 und 1945-1960". https://dfk-paris.org/de/page/deutsch-franzoesische_kunstvermittlung-1870%E2%80%931940_und-1945%E2%80%931960-2389.html (zugegriffen 2. August 2022).

DFK Paris. 2022. "Publikationen des Projekts 'Deutsch-Französische Kunstvermittlung'". https://dfk-paris.org/de/page/dfkv_publicationen-3307.html (zugegriffen 2. August 2022).

Flanders, Julia, und Trevor Muñoz. 2015. „An Introduction to Humanities Data Curation | DH Curation Guide“. <https://web.archive.org/web/20150822055930/http://guide.dhcurator.org/contents/intro/> (zugegriffen 2. August 2022).

Gaetgens, Thomas W. 2009. "Introduction: De la réception de l'art moderne français en Allemagne entre 1870 et 1945". In *Perspectives croisées. La critique d'art franco-allemande 1870-1945*, hg. von Thomas W. Gaetgens, Mathilde Arnoux und Friederike Kitschen, 3-26. Paris: Éd. de la Maison des sciences de l'homme.

Huie, J. Russell et al. 2021. "FAIR Data Reuse in Traumatic Brain Injury: Exploring Inflammation and Age as Moderators of Recovery in the TRACK-TBI Pilot." In *Frontiers in neurology* 10.3389/fneur.2021.768735.

Kim, Dorothy und Adeline Koh. 2021. *Alternative Historiographies of the Digital Humanities*. punctum Books 10.53288/0274.1.00.

Klammt, Anne. 2022. "DFKV - Data Model". https://github.com/archaeoklammt/DFKV_data_model (zugegriffen 2. August 2022).

Linked Art. <https://linked.art/> (2. August 2022).

Niemann, Klara. 2021. "Die Aufbereitung der Datenbank Deutsch-Französische Kunstvermittlung 1870-1940 und 1945-1960 und ihre zukünftigen Nutzungsmöglichkeiten". In *Jahresbericht des DFK Paris 2020/2021*, 110-111.

Steiner, Elisabeth, Gunter Vasold und Martina Scholger. 2022. "Repositorien als digitale Gedächtnisträger zwischen Evolution und Langzeitplanung". In *DHd2022: Kulturen des digitalen Gedächtnisses*, hg. von Michaela Geierhos et al. 10.5281/zenodo.6304590.

Waelder, Paul. 2017. "Summary: Media Archaeologies Evening. First December 2017, Bar-

celona“. <http://catedratelefonica.uoc.edu/wp-content/uploads/2018/01/Media-Archeologies-BCN.pdf> (zugegriffen 2. August 2022).

Woodall, Philipp. 2017. "The Data Repurposing Challenge: New Pressures from Data Analytics". In *Journal of Data and Information Quality* 8 10.1145/3022698.

Woodall, Philip, und Anthony Wainman. 2015. "Data quality in analytics: key problems arising from the repurposing of manufacturing data". In *Proceeding of the International Conference on Information Quality (ICIQ'15)*, 174–184.