SWITCH, Werdstrasse 2, P.O. Box, CH-8021 Zürich
www.switch.ch

# SWITCH Innovation Lab «Nachvollziehbare Datenqualität»

**Innovation Lab Partner**
Schweizerische Akademie der Technischen Wissenschaften - SATW im Auftrag von SWITCH
Esther Koller-Meier
Manuel Kugler

**Executive Summary**
Aktuelle, frei zugängliche und nutzbare Forschungsdaten sind die wichtigste Ressource für die Sicherung der Innovationskraft und Forschungsstärke der Schweiz. Zu diesem Zweck ist eine nachvollziehbare Datenqualität der Forschungsdaten erforderlich, die sowohl einen breiteren Zugang und Austausch, als auch die (Nach-)Nutzung von Forschungsdaten ermöglicht.

Das Forschungsdatenmanagement im Bereich Open Science konfrontiert die Schweizer Hochschulen und Forschungsinstitutionen mit unterschiedlichen Herausforderungen, eröffnet ihnen aber auch zahlreiche Chancen. Zu diesen Themen laufen bereits diverse Aktivitäten parallel und je nach Institution variieren der Wissensstand sowie die entsprechenden Erwartungen stark.

Die vorliegende Kurzstudie hat zum Ziel, den aktuellen Stand des Wissens und umgesetzte Massnahmen für eine nachvollziehbare Datenqualität in unterschiedlichen Forschungsbereichen mit Hilfe von Expertenbefragungen der SATW zu erfassen. Zudem sollen in diesem Zusammenhang nationale Bedürfnisse und Probleme aufgezeigt werden.

Einige Schlüsselergebnisse dieser Studie weisen auf die Notwendigkeit der Zugänglichkeit von Forschungsdaten, automatisierten Prozessen und dokumentierten Metadaten hin. Authentizität, Integrität und Unbestreitbarkeit sind ebenfalls grundlegende Aspekte für eine nachvollziehbare Datenqualität. Dafür sollten Richtlinien und Standards festgelegt werden. All diese Aspekte sind für die Verwirklichung der Vision eines Forschungsdaten-Konnektoms für die Schweiz von zentraler Bedeutung.

Kurzbericht

# Nachvollziehbare Datenqualität

Esther Koller-Meier, Manuel Kugler

28.01.2020

Schweizerische Akademie der Technischen Wissenschaften SATW
St. Annagasse 18 | 8001 Zürich | +41 44 226 50 11 | info@satw.ch | www.satw.ch

# 1 Über diese Kurzstudie

## 1.1 Ziele des SWITCH Innovation Labs «Nachvollziehbare Datenqualität»

Die Themen Forschungsdaten und Datenverwaltung im Bereich Open Science konfrontieren die Schweizer Hochschulen und Forschungsinstitutionen mit unterschiedlichen Herausforderungen, eröffnen ihnen aber auch zahlreiche Chancen. Zu diesen Themen laufen bereits diverse parallel Aktivitäten und je nach Institution variieren der Wissensstand sowie die entsprechenden Erwartungen stark.

SWITCH - als integraler Bestandteil der schweizerischen akademischen Gemeinschaft - hat die «Nachvollziehbare Datenqualität» als ein zentrales Thema bei der gemeinsamen Nutzung, Wiederverwendung und Zusammenführung von Forschungsdaten identifiziert.

Die vorliegende Kurzstudie hat zum Ziel, den aktuellen Stand des Wissens und umgesetzte Massnahmen für eine breitgestützte nachvollziehbare Datenqualität in unterschiedlichen Forschungsbereichen zu erfassen. Zudem sollen in diesem Zusammenhang nationale Bedürfnisse und Probleme aufgezeigt werden.

## 1.2 Vorgehen

Um die beschriebenen Ziele zu erreichen, wurde die SATW damit beauftragt, die dafür notwendigen Informationen mittels einer Expertenbefragung zusammenzutragen. Die SATW hat dafür ihre internen und externen Expertennetzwerke inklusive der Kontakte zum Dachverband Akademien der Wissenschaften Schweiz a+ aktiviert.

In einer ersten Phase wurden Fachleute aus national relevanten Forschungsdomänen, der Industrie oder dem Dienstleistungssektor zum Themenfeld «nachvollziehbare Datenqualität identifiziert. Anschliessend wurde das Expertenwissen mit Hilfe eines Fragekatalogs eingeholt und die eingegangenen Antworten zum vorliegenden Kurzbericht zusammengefasst.

# 2 Resultate aus der schriftlichen Befragung

## 2.1 Charakteristik und Stellenwert von Forschungsdaten und Datenqualität

Die **Charakteristik** von Forschungsdaten ist laut der befragten Personen sehr **heterogen** und **projektabhängig**. Die Datenqualität (DQ) wird durch die jeweilige **Nutzung** der Daten geprägt: Abhängig von der Anwendung, werden unterschiedliche Ansprüche an die Qualität gestellt. Die **Verantwortung** für die DQ liegt bei den Forschenden; sie müssen wissenschaftliche Standards und Ethikrichtlinien einhalten.

Die DQ hat aus Sicht der befragten Personen einen sehr hohen Stellenwert. Höchste Qualität zu erreichen, ist eine grosse Herausforderung und lässt sich meist nur mit beträchtlichem Aufwand sicherstellen. Darum werden oft **Qualitätsstufen** definiert[1]. Die Sicherstellung einer angestrebten DQ beansprucht dennoch viel Zeit, bspw. aufgrund einer allenfalls benötigen Kuration.

Ebenfalls erwähnt wurde, dass der Wert der DQ nicht immer von allen Akteurinnen und Akteuren erkannt wird, z.B. seitens des Managements. Sind die Rahmenbedingungen im Zusammenhang mit der DQ unklar, kann dies einen Mehraufwand bei der Datenerfassung verursachen[2]. Abhängig von der Quelle, kann die Bedeutung von DQ auch untergeordnet sein[3]. Der allgemeine Trend in der Forschung geht allerdings in Richtung mehr und frühzeitiger Berücksichtigung von DQ[4].

## 2.2 Nachvollziehbare Datenqualität: Wichtige Aspekte und Herausforderungen

Für eine nachvollziehbare Datenqualität (DQ) sind diverse **Aspekte** zu berücksichtigen (vgl. Abbildung 1). Die Relevanz der einzelnen Aspekte hängt von der betrachteten Problemstellung ab. Für eine bestimmte Problemstellung wird üblicherweise nur eine Auswahl der Aspekte als Kriterien festgelegt. Typischerweise limitiert das am schwächsten erfüllte Kriterium die Qualität der Daten.

**Zugänglichkeit** zu Daten, wurde am häufigsten als wichtige Voraussetzungen für eine nachvollziehbare DQ genannt. Ohne Zugänglichkeit können alle restlichen Merkmale nur theoretisch abgehandelt werden. In der Forschung ist aktuell allerdings nur ein Bruchteil der produzierten Daten zugänglich. Angegebene Gründe dafür sind:

- **Forschungsfragen** sind oft so **spezifisch**, dass Forschende geeignete Daten nicht finden oder nicht veröffentlichen. Oder es mangelt an **Transparenz**, welche Informationen vorhanden und zugänglich sind.
- Bei Publikationen bestehen oft **Vorgaben** bspw. von Auftraggebenden oder Herausgeberinnen und -gebern.

---

[1] Bspw. unterscheiden die EIDAS wie auch die eCH-Standards (http://www.ech.ch/standards/48092 oder https://www.ech.ch/standards/39992) unterschiedliche Qualitätsstufen für Authentisierungen und Attribut-Bestätigungen. Die Benutzerfreundlichkeit für die sich authentifizierende Person ist typischerweise bei höherer Qualität wesentlich schlechter als bei tieferer.

[2] Bspw. durch eine aufwändigere Datenkuration oder eine erneute Datenerfassung aufgrund veränderter Anforderungen an die DQ.

[3] Bspw. steht bei Befragungen die wahrheitsgetreue Beantwortung der Fragen im Vordergrund.

[4] Zum Beispiel durch Forderungen seitens Forschungsförderorganisationen wie dem Schweizerischen Nationalfonds SNF.

- Zugang zu Daten wird oft durch **Registration** kontrolliert: Wer Daten nutzen will, muss aus rechtlichen Gründen Datennutzungsverträge abschliessen.
- Um Daten zu publizieren, muss viel Zeit für die **Dokumentation** aufgewendet werden.

**Authentizität, Integrität**[5] und **Unbestreitbarkeit** sind ebenfalls grundlegend für die DQ und die Wiederverwertbarkeit von Daten. Diese Aspekte werden aber oftmals nur spärlich überprüft. **Prozesse** zur Sicherstellung der DQ wie der Aufbau von **Qualitätsregeln, Qualitätsmassstäben** und einer **Data Governance**, ein kontinuierliches **Datenmonitoring** oder die **Fehlerbetrachtung** stellen grosse Herausforderungen in der Verwaltung von Daten dar. Bestehende Prozesse sollten optimiert und nach Möglichkeit **automatisiert** werden.

Spezifische **Herausforderungen** betreffend DQ sind in Abbildung 2 abgebildet. Am meisten genannt wurden **fehlerhafte Daten oder Dubletten, Aktualität, Konsistenz** und **Relevanz. Vollständigkeit** und **Genauigkeit** wurden ähnlich oft genannt. Daten beschreiben die Realität nur approximativ, woraus eine Unschärfe in der weiteren Auswertung resultiert. Dieser Aspekt ist bereits bei der Datenerhebung zu berücksichtigen. Weiter ist die Sicherung der Aktualität der Daten (Veränderung über die Zeit) schwierig und kostenintensiv.

Im Zusammenhang mit bestimmten Methoden wie maschinellem Lernen stehen die Qualität und Verwendung von hochwertigen **Labels** bzw. **Metadaten** im Vordergrund. Nicht zu unterschätzen in dem Zusammenhang ist auch die **Skalierbarkeit**: Heute werden extrem grosse Datensätze benötigt, um Modelle zu trainieren; die Verwendung hochwertiger Labels für solche Datensätze ist eine grosse Herausforderung.


Abbildung 1: In der Befragung gesammelte Kriterien für eine nachvollziehbare Datenqualität.

[5] Einigen Experten war der Unterschied zwischen Authentizität und Integrität nicht klar, sie sehen die Authentizität als Teil der Datenintegrität.


Abbildung 2: Von den befragten Expertinnen und Experten genannte Herausforderungen im Zusammenhang mit einer nachvollziehbaren Datenqualität.

## 2.3 Auswirkungen von Richtlinien und Standards auf die Datenqualität

Allgemein wirken sich Richtlinien, Standards und Best Practises positiv auf die DQ aus. Die Verantwortung für die Einhaltung wissenschaftlicher Standards und (Ethik)Richtlinien liegt weitgehend bei den Forschenden. In Einzelfällen monitoren Datenarchivfachleute in den Organisationen verschiedene Standards und implementieren solche, die sinnvoll erscheinen. Einige Befragte gaben an, keine Metadaten-Standards zu kennen und anzuwenden.

**FAIR-Prinzipien:** Durch Einhaltung der FAIR-Prinzipien (Findable, Accessible, Interoperable und Reusable) werden Daten wiederverwendbar. Die meisten Befragten gaben an, die FAIR-Prinzipien zu kennen. Gemäss Aussagen versuchen auch einige Forschungsgruppen, diese so gut wie möglich umzusetzen. Gewisse Institutionen bieten dafür gar Kurse, Workshops und Präsentationen an. Oft werden die Prinzipien aber aus Zeitgründen nicht angewendet. Ein wichtiger Aspekt im Zusammenhang mit FAIR ist auch die langfristige Aufbewahrung von Forschungsdaten.

**Richtlinien:** Teilweise verwenden Datenverwalterinnen und -verwalter **domänenspezifische Vorgaben** oder **Instrumente und methodische Vorgaben** aus internationalen Projekten.

**Standards:** Abhängig davon, ob Daten im Rahmen von Forschungsprojekten, in der Verwaltung oder in der Industrie generiert werden, kommen unterschiedliche **Standards** zur Anwendung, beispielsweise für den Austausch von Metadaten. Speziell für Umfragen gibt es umfangreiche Standards von Branchenverbänden[6]. Allgemein anerkannte Standards für das Datenmanagement haben sich bis heute jedoch nicht durchgesetzt. Werden Daten nach nicht-standardisierten Verfahren erhoben, sind sie für Dritte kaum nutzbar[7].

[6] Bspw. die Standards von ESOMAR zu Datenerhebungen.
[7] Trotzdem sollten sie publiziert werden, um den Standardisierungsprozess voranzubringen.

Amtliche Statistiken folgen dem **Europäischen Statistischen Verhaltenskodex**[8i]. Die entsprechende Organisation und Pflege der Daten sowie der Verhaltenskodex selbst könnten als Vorbild für ähnliche Bestrebungen in der Forschung dienen.

Der Verein eCH fördert, entwickelt und verabschiedet Standards im Bereich E-Government, für eine effiziente elektronische Zusammenarbeit zwischen Behörden, Unternehmen und Privaten[9]. Für Stammdaten existieren für Data Governance zudem Best Practices[10]. Seitens Behörden scheinen die Metadaten des Bundesamts für Statistik und deren Nomenklaturen am breitesten abgestützt. Eine internationale Abstimmung sichert aber nicht zwingend eine hohe Nutzbarkeit.

## 2.4 Veröffentlichung von Forschungsdaten

**Rohdaten** werden meist nur in **anonymisierter Form** veröffentlicht. Die Einhaltung des Datenschutzes und die korrekte Angabe der Quellen spielen eine wichtige Rolle. Einige Forschende veröffentlichen generell keine Rohdaten. In den meisten Fällen werden nur **aggregierte Daten** veröffentlicht, in denen die ursprüngliche Information nicht mehr in ihrer Detailliertheit vorhanden ist. Die Wiederverwendung aggregierter Daten kann eine Herausforderung darstellen, da nicht immer klar ist, wie die Aggregation durchgeführt wurde.

Journals verlangen zunehmend bereinigte Rohdaten, auf denen aggregierte Resultate basieren. Eine **Verpflichtung**, Rohdaten aus Forschungsprojekten publizieren zu müssen, wirkt sich grundsätzlich positiv auf die DQ aus. Forschenden ist so von Anfang an bewusst, dass andere ihre Daten nutzen und allenfalls verifizieren können. Dies kann zu besserer **Dokumentation** und Sorgfalt im Umgang mit Daten und Metadaten führen. Schliesslich will in der Forschung niemand kritisiert werden, mit Daten gepfuscht zu haben. Analoge Erfahrungen liegen auch in der Verwaltung vor[ii].

Aggregierte und prozessierte[11] Daten werden oftmals über Webseiten und öffentliche Dienste[12], auf Archivplattformen[13] oder in relationalen Datenbanken z.B. mit einer **CC-BY-Lizenz**[14] veröffentlicht, angeboten und zugänglich gemacht. **Sensitive Daten** können via Zugriffsbeschränkungen verwaltet werden.

---

[8] Der Verhaltenskodex für europäische Statistiken setzt den Standard für die Entwicklung, Erstellung und Verbreitung europäischer Statistiken. Er baut auf einer einheitlichen ESS-Definition der Qualität in der Statistik auf und richtet sich an alle relevanten Bereiche im institutionellen Umfeld über die statistischen Produktionsprozesse bis hin zu unserem Output: Europäische amtliche Statistiken.

[9] Die Verwaltung verwendet beispielsweise den Standard eCH-0170, der auf entsprechenden europäischen und amerikanischen Standards basiert.

[10] Datenmanagementprozesse gemäss Otto, Boris (HSG) «Stammdatenverwaltung». Eine interne Bewertung zweier Referenzsysteme hat gezeigt, dass diese unterschiedlich gut, in der Summe aber recht gut angewandt werden. Verbesserungen werden nun angestossen.

[11] Bspw. kuratierte Daten.

[12] z.B. UniProtKB, https://www.uniprot.org/help/uniprotkb

[13] Bspw. veröffentlicht FORS für alle Umfragen die anonymisierten Rohdaten auf ihrer Archivplattform https://forsbase.unil.ch/ mit allen dazugehörigen Dokumentationen.

[14] Diese Lizenzform ist die freiest mögliche und erlaubt die Daten auch kommerziell zu nutzen und zu verarbeiten, diese zu verbreiten und darauf aufzubauen, solange der Urheber des Originals genannt wird.

## 2.5 Verwendung von Forschungsdaten anderer Domänen für die eigene Forschung

Daten, die von Forschenden selbst erhoben werden, sind i.A. gut dokumentiert. Bei externen Daten ist die Qualität der Dokumentation hingegen sehr unterschiedlich. Der **Kontext der primären Datenerhebung** ist wichtig: Wofür wurden die Daten erhoben? Dieser gibt sekundären Datennutzerinnen und -nutzern Anhaltspunkte, inwiefern und wofür die Daten nutzbar sind. Die Ausgangslage und Überlegungen zur Datenerfassung sollten in Metadaten dokumentiert werden.

Wichtige Beurteilungskriterien für die Korrektheit externer Daten sind deren **Herkunft**: Datennutzung hat immer auch mit **Vertrauen** in deren Qualität und somit auch in den entsprechenden Datenlieferanten[15] zu tun. Gewissen Produzentinnen und Produzenten von Daten traut man mehr – insbesondere, wenn sie sich schon länger mit Datenerhebungen in einem entsprechenden Kontext bewährt haben (Authentizität). Die **Sensibilität** der Forschenden bzw. der Produzentinnen oder Produzenten der Daten hinsichtlich DQ ist zentral. Zusätzlich braucht es entsprechende **Kompetenzen**, um Quellen kritisch zu betrachten, Daten zu nutzen, zu analysieren und Ergebnisse zu Interpretieren.

In der Forschung werden i.A. noch relativ wenig wissenschaftliche Daten ausgetauscht[16]: Viele Forschende arbeiten bislang nur mit eigenen Daten oder mit Daten von Unternehmen. Forschende stellen ihre Daten aber bei Veröffentlichung oft zur Verfügung und bereiten sie für Interessenten teilweise auch auf individueller Basis auf. Die Geisteswissenschaften scheinen in der Vernetzung und Nachnutzung von Forschungsdaten weiter fortgeschritten als andere Disziplinen[17]. Eine solche Nachnutzung kann die Visibilität der eigenen Forschung steigern. Zudem können Forschende dadurch von den Schwerpunkten, Kompetenzen und Aufträgen ihrer Forscherkolleginnen und -kollegen profitieren und die erhobenen Daten werden produktiver genutzt.

Einige Forschende verwenden auch Daten der öffentlichen Statistik (Bundesamt für Statistik (BFS), Kantone, etc.)[18]. Der Zugang zu administrativen Daten ist oft möglich, teils jedoch kompliziert und langwierig. Die zugehörigen Metadaten sind sehr heterogen: Einige Datensätze sind einfach zu finden, von den Anbietern gut beschrieben, validiert und versioniert; andere weniger.

Für bestimmte Forschungsdisziplinen ist die Verknüpfung[19] eigener Daten mit administrativen Daten oder solchen von privaten und staatlichen Anbietern besonders interessant[20]. Das BFS kann mittels Identifikationsnummern eine Vielzahl statistischer Informationen verknüpfen. Eine solche **Datenverknüpfung** erlaubt es, Informationen zu erweitern, neue statistische Analysen anzuwenden

---

[15] In der Forschung versteht man unter Vertrauen die wissenschaftliche Redlichkeit und Sorgfalt der Kollegen sowie die Herkunft der Datenquellen.
[16] Als Vorreiter für den Datenaustausch gilt beispielsweise das Swiss Personalized Health Network SPHN, das biomedizinische Daten austauscht.
[17] Dodis beteiligt sich bspw. aktiv an Metagrid und histHub.
[18] Bspw. Geodaten des Bundes (z.B. von swisstopo, BFE, BFS) und von Kantonen, Naturrisiko- oder Bevölkerungszahlen der Schweiz
[19] Ein grosses Problem des BFS ist das Fehlen einer Interoperabilitätsplattform, auf der alle Bundes-, Kantons- und Gemeindebehörden ihre Metadaten hochladen können, welche die vorhandenen Daten beschreiben. Der Bundesrat hat Ende September 2019 den Auftrag erteilt, eine Interoperabilitätsplattform aufzubauen und zu verwalten, die als öffentliches Metadatensystem für alle Bundesämter und später für alle öffentlichen Verwaltungen in der Schweiz dienen soll.
[20] Siehe Projekt linhub.ch

und dadurch neue Erkenntnisse aus vorhandenen Daten zu gewinnen. Dubletten werden vermieden, Kosten auf ein Minimum reduziert und Synergien können genutzt werden. Solche Verknüpfungen mindern den Aufwand für die Datenerfassung, da weniger Personen direkt konsultiert werden müssen. Die Datenverknüpfung unterliegt strengen Regeln in Bezug auf Datenschutz und -sicherheit: Deren Gewährleistung hat in diesem Zusammenhang höchste Priorität. Sofern bestimmte Voraussetzungen erfüllt sind, kann das BFS für Forschungs-, Planungs- und statistische Zwecke Daten ohne Personenbezug im Rahmen eines Verknüpfungs- und Datenschutzvertrags zusammenführen. Dazu befugt sind Organisationen von Bund, Kantonen und Gemeinden sowie anerkannte Forschungseinrichtungen wie Universitäten und Fachhochschulen.

Weitere genutzte Datenlieferanten sind andere vertrauenswürdige Institutionen[21] oder auch Onlinequellen[22]. Für Forschende besonders interessant sind online publizierte Daten[23], auf die sie sonst nicht oder nur sehr umständlich zugreifen[24] können. In bestimmten Bereichen ist der Zugriff selbst kein Problem, abgesehen von vereinzelt kostenpflichtigen Angeboten. Allerdings stossen Forschende häufig erst mit grosser Verzögerung auf relevante Publikationen.

Werden keine Primärquellen verwendet, kann die korrekte Interpretation der Dokumentation, bspw. von Metadaten, eine Herausforderung darstellen: Eine klare Definition der Grundgesamtheit[25] ist entscheidend. **Rahmenbedingungen** sind stets projektabhängig und alle Faktoren, welche die Interpretation von Daten beeinflussen, müssen festgehalten sein[26]. Zudem ist eine umfangreiche Dokumentation[27] wichtig.

Bevorzugt werden **validierte Daten**[28] verwendet. Vereinzelt stehen bei veröffentlichten Daten **systematische Berichte**[29] zur Verfügung. Auf **Standards** basierende Vorgaben an Informationslieferanten sind eine weitere Möglichkeit der Qualitätssicherung. **Transparenz** hinsichtlich der Zugänglichkeit zu Daten erhöht das Vertrauen, da eine **häufige Nutzung** durch möglichst viele unterschiedliche Nutzerinnen und Nutzer einen guten Schutz bietet: Je mehr und je häufiger die Daten verwendet werden, desto höher wird die DQ implizit angenommen.

---

[21] z.B. MeteoSuisse
[22] z.B. Webseiten, Tweets, Zeitungsartikel etc.
[23] Dies umfasst unter anderem Umweltdaten, Geodaten, medizinische Daten wie auch Daten im Bereich «Open Government».
[24] z.B. via Archivbesuche.
[25] Bspw. sampling frames, sampling von Prozessen, Ausschöpfung oder weitere Parameter der Datengenerierung. Bei Umfragen z.B. Art der Interviews, Zeitpunkt der Umfrage, Felddauer, Grundpopulation, Rekrutierung, etc.
[26] Bspw. hängt die Verwendbarkeit von anonymisierten und pseudonymisierten Daten von deren Art und Verwendung ab. Eingriffe oder Auslassungen sind in den Metadaten zu dokumentieren.
[27] Bei technischen Messungen z.B. Typ der Messgeräte, die Institution, welche die Messung durchführt, die Person, welche die Messungen betreut oder die Beschreibung der angewandten Kalibration. Bei Umfragen z.B. Codebücher, Fragebogen, Berechnungen von Gewichtungen.
[28] Eine mögliche Validierung kann bspw. mit statistischen Verfahren ermittelt werden, indem die Verteilung der Daten oder Anzahl Beobachtungen gemäss weiteren Kriterien wie Standort untersucht werden.
[29] Diese beschreiben die Methoden und Verfahren, auf denen die Erhebungen, Ergebnisse und Analysen der öffentlichen Daten basieren. Siehe z.B. https://www.bfs.admin.ch/bfs/en/home/services/recherche/methodological-reports.html

Optimalerweise gibt es die Möglichkeit, **Feedback** anzubringen. **Rückmeldungen**[30] zu Dateninstanzen und die **Semantik** der publizierten Daten helfen enorm, die Qualität zu erhöhen.

Durch die zunehmende Verbreitung von Open Access und Open Data, nimmt die Verfügbarkeit von Forschungsdaten seit einigen Jahren stetig zu. Personalisierte private Daten[31] sind hingegen noch immer kaum verfügbar.

# 3 Kernaussagen und Empfehlungen für weitere Schritte

## 3.1 Kernaussagen

- Die DQ hat einen sehr hohen Stellenwert: Der allgemeine Trend in der Forschung geht in Richtung mehr und frühzeitiger Berücksichtigung von DQ. DQ hängt von der jeweiligen Datennutzung ab. Die Verantwortung liegt bei den Forschenden, welche die Daten erheben.
- Höchste Qualität ist meist nur mit beträchtlichem Aufwand zu erreichen. Verschiedene **Qualitätsstufen** zu definieren, ist meist zweckmässiger. Die Sicherstellung einer angestrebten DQ beansprucht dennoch viel Zeit. **Prozesse** zur Sicherstellung der DQ sollten soweit möglich **automatisiert** werden.
- **Zugänglichkeit** ist die wichtigste Voraussetzung für eine nachvollziehbare DQ. **Authentizität, Integrität** und **Unbestreitbarkeit** sind ebenfalls grundlegend für die DQ und die Wiederverwertbarkeit von Daten. Diese Aspekte werden aber oftmals nur spärlich überprüft. Die **Aktualität** der Daten zu gewährleisten ist schwierig und kostenintensiv.
- **Richtlinien und Standards** wirken sich positiv auf die DQ aus. Sind die Rahmenbedingungen unklar, kann dies zu einem Mehraufwand führen. Allgemein anerkannte Standards für das Datenmanagement gibt es bislang keine. Werden Daten durch nicht-standardisierte Verfahren erhoben, sind sie für Dritte kaum nutzbar.
- Die **FAIR-Prinzipien** sind bekannt, werden aber aus Zeitgründen noch selten angewendet.
- **Rohdaten** werden meist anonymisiert veröffentlicht, häufiger werden aber nur **aggregierte Daten** zugänglich gemacht. In **Metadaten** müssen alle Faktoren dokumentiert sein, welche die Interpretation von Daten beeinflussen. Sind bspw. der Kontext der primären Erhebung oder die Aggregation unzureichend dokumentiert, ist eine Wiederverwendung schwierig.
- In der Forschung werden noch relativ wenig wissenschaftliche Daten ausgetauscht. Meistens erfährt man nicht oder erst viel später von einer Online-Publikation. Eine **Verpflichtung**, Rohdaten zugänglich machen zu müssen, wirkt sich positiv auf die **Dokumentation** und damit auf die DQ aus. Es braucht entsprechende **Kompetenzen**, um Quellen kritisch zu betrachten, Daten zu nutzen, zu analysieren und Ergebnisse zu Interpretieren.
- Die **Datenverknüpfung** ist für einige Forschungsdisziplinen besonders interessant und erlaubt, neue Erkenntnisse aus vorhandenen Daten zu gewinnen. Metadaten der öffentlichen Statistik sind heterogen: Einige Datensätze sind einfach zu finden, gut beschrieben, validiert und versioniert, andere weniger.

[30] Am hilfreichsten ist dabei möglichst positives wie auch negatives bzw. ggf. korrigierendes Feedback.
[31] z.B. Facebook, Google

- Die **Herkunft** ist ein wichtiges Kriterium für die Korrektheit von Daten: Datennutzung hat mit **Vertrauen** in Lieferantinnen und Lieferanten zu tun zu tun, deren **Sensibilität** für DQ ist zentral. Je mehr und je häufiger die Daten verwendet werden, desto höher wird die DQ implizit angenommen. Die Möglichkeit, **Feedback** anzubringen, hilft, die Qualität zu erhöhen.

### 3.2 Empfehlungen für weitere Schritte

- Überblick über bestehende Initiativen im Bereich Open Science und Open Data verschaffen und Synergien nützen. Insbesondere Entwicklungen rund um FAIR aktiv mitverfolgen und mitgestalten.
- Die wichtigsten Datenrepositorien (Forschung, Verwaltung und Industrie) identifizieren und klären, welche Daten vorhanden sind und welche Metadaten-Standards angewandt werden.
- Auffindbarkeit von Forschungsdaten durch automatisierte Tools fördern und Zugang erleichtern (bspw. im Rahmen eines Forschungsdatenkonnektoms).
- Abklären, inwiefern einheitliche (minimale) Standards und Rahmenbedingungen für Forschungsdaten entwickeln werden können (Dokumentation, Metadaten, Aggregation, …).
- Prozesse zur Sicherstellung von DQ definieren und automatisieren, um Forscher zu entlasten
- Awareness bei Forschenden hinsichtlich DQ schaffen und Open Data fördern.

## 4 Anhang

### 4.1 Befragte Fachleute

| | | | |
|---|---|---|---|
| Prof. | Andreas | Spichiger | Berner Fachhochschule (BFH) |
| Herr | Bertrand | Loison | Bundesamt für Statistik (BFS) |
| Dr. | Christiane | Sibille | Diplomatische Dokumente der Schweiz (Dodis) |
| Dr. | Ursin | Lutz | Dicziunari Rumantsch Grischun (DRG) |
| Prof. | Georg | Lutz | FORS |
| Herr | Adrian | Meyer | Mobiliar |
| Herr | Heinz | Stockinger | Swiss Institute of Bioinformatics (SIB) |
| Herr | Mario | Valle | Swiss National Supercomputing Centre (CSCS) |
| Prof. | Philippe | Cudre-Mauroux | Universität Fribourg (UNI FR) |
| Dr. | Markus | Christen | Universität Zürich (UZH) |
| Herr | Andreas | Fürholz | Zürcher Hochschule für Angewandte Wissenschaften (ZHAW) |
| Dr. | René | Locher | Zürcher Hochschule für Angewandte Wissenschaften (ZHAW) |

# 5 Glossar

| | |
|---|---|
| Aggregierte Daten | Aggregierte Daten sind im Wesentlichen zu grösseren Einheiten zusammengefasste Einzelwerte. |
| Authentizität | Authentizität bedeutet Echtheit der Daten im Sinne von «als Original befunden». |
| Best Practice | Unter dem Begriff Best Practice wird eine bereits erprobte und bewährte Methode zum Ablauf eines Arbeitsprozesses verstanden. Sie ist eine Technik oder Methodik, die sich durch Erfahrung und Forschung als zuverlässig erwiesen hat, zu einem gewünschten Ergebnis zu führen. |
| Datenformat | Das Datenformat legt fest, wie Daten strukturiert und dargestellt werden und wie sie bei ihrer Verarbeitung zu interpretieren sind. Es gibt somit die Syntax und Semantik von Daten innerhalb einer Datei an. |
| Datentyp | Der Datentyp beschreibt die Art der Daten und welche logischen Operationen damit ausgeführt werden können. |
| Datenkuration | Der Begriff Kuration wird im Sinne von «Bewahren» und «Behandeln» eingesetzt. Datenkuration beschreibt somit, welche Managementaktivitäten erforderlich sind, um Daten langfristig zu pflegen, sodass sie für die Wiederverwendung verfügbar sind. |
| Datenverknüpfung | Daten können durch die Verwendung von Identifikationsnummern in verschiedenen Datenbeständen verknüpft werden.<br>Datenverknüpfungen haben zum Ziel, Informationen aus bestehenden Daten zu gewinnen, Doppelspurigkeiten zu vermeiden, Kosten zu minimieren und Synergien zu erzielen<br>Die Sicherstellung des Datenschutzes hat dabei höchste Priorität. Daher unterliegt die Verknüpfung von Daten strengen Auflagen hinsichtlich Datenschutz und Datensicherheit |
| FAIR-Prinzipien | Der Begriff FAIR (Findable, Accessible, Interoperable und Reusable) Data wurde 2016 von der FORCE 11-Community für ein nachhaltiges Forschungsdatenmanagement geprägt. Hauptziel der FAIR-Prinzipien ist eine optimale Aufbereitung der Forschungsdaten, die demnach auffindbar, zugänglich, interoperabel und wiederverwendbar sein sollen. |
| Integrität | Integrität bezeichnet die Korrektheit bzw. Unversehrtheit von Daten, d.h. dass es nicht möglich sein darf Daten unerkannt bzw. unbemerkt zu ändern. |
| Metadaten | Metadaten sind strukturierte Daten, die Informationen über Merkmale anderer Daten enthalten. Bei den durch Metadaten beschriebenen Daten handelt es sich oft um grössere Datensammlungen wie Dokumente oder Dateien. |

| | |
|---|---|
| Open Access | Open Access ist der freie Zugang zu wissenschaftlicher Literatur und anderen Materialien (einschliesslich Primär- und Metadaten) im Internet. |
| Open Data | Als Open Data werden Daten bezeichnet, die von allen uneingeschränkt genutzt und verbreitet werden dürfen. |
| Richtlinien | Richtlinien sollen für alle Mitarbeitenden einer Institution festschreiben, welche Verfahren beim Datenmanagement eingesetzt werden sollen und wie mit Daten umzugehen ist. |
| Primärdaten | Bei Primärdaten (auch Rohdaten oder Urdaten) handelt es sich um jene Daten, die unmittelbar bei der Datenerhebung gewonnen werden. |
| Rohdaten | Bei Rohdaten (auch Primärdaten oder Urdaten) handelt es sich um jene Daten, die unmittelbar bei der Datenerhebung gewonnen werden. |
| Semantik | Die Semantik befasst sich mit der Bedeutung von Zeichen und Zeichenfolgen. |
| Standard | Ein Standard ist eine vergleichsweise einheitliche oder vereinheitlichte, weithin anerkannte und meist angewandte Art oder Weise, etwas zu beschreiben. |
| Syntax | Die Syntax bezeichnet ein Regelsystem zur Kombination elementarer Zeichen zu zusammengesetzten Zeichen in natürlichen oder künstlichen Zeichensystemen. Bezogen auf Sprachen ist die übliche Verbindung von Wörtern zu Wortgruppen und Sätzen gemeint bzw. die korrekte Verknüpfung sprachlicher Einheiten in einem Satz. |
| Urdaten | Bei Urdaten (auch Primärdaten oder Rohdaten) handelt es sich um jene Daten, die unmittelbar bei der Datenerhebung gewonnen werden. |
| Zugänglichkeit | Daten und Metadaten sollen langzeitarchiviert und verfügbar gemacht werden, sodass sie leicht von Menschen und Maschinen heruntergeladen und genutzt werden können. |

[i] Der Verhaltenskodex für die Europäische Statistik enthält 16 Grundprinzipien für die Erstellung und Verbreitung der europäischen amtlichen Statistiken und das institutionelle Umfeld, in dem die nationalen und gemeinschaftlichen Statistikbehörden tätig sind. Eine Reihe von Indikatoren für bewährte Verfahren (good practices) für jeden der 16 Grundsätze bildet eine Orientierungshilfe für die Überprüfung der Umsetzung des Kodex.
Der Verhaltenskodex für europäische Statistiken wurde am 24. Februar 2005 vom Ausschuss für das Statistische Programm angenommen und im September 2011 sowie November 2017 vom Ausschuss für das

«European Statistical System» überarbeitet. Zusammen mit der Version 2011 des Verhaltenskodex verabschiedete der Ausschuss des Europäischen Statistischen Systems den Richtlinienkatalog für die Qualitätssicherung. Dieser dient als Leitfaden für die Umsetzung des Verhaltenskodex für europäische Statistiken.
Eurostat hat das Protokoll über «impartial Access to Eurostat data» (den unparteiischen Zugang der Nutzerinnen und Nutzer zu Eurostat-Daten) angenommen, um die Umsetzung des Kodex zu unterstützen. Es richtet sich an Eurostat-Nutzer, -Personal und -Partner bei der Erstellung von Europäischen Statistiken. Jedes Jahr überwacht Eurostat die Einhaltung des Verhaltenskodex im gesamten ESS. Das European Statistical Governance Advisory Board (ESGAB) erhält zusammenfassende Informationen für seinen Jahresbericht an das Europäische Parlament und den Rat. Das ESGAB berichtet über die Umsetzung des Verhaltenskodex, soweit sie sich auf Eurostat bezieht, und enthält eine Bewertung der Umsetzung des Verhaltenskodex im gesamten ESS. Der ESGAB-Jahresbericht, welcher ab 2009 erscheint, ist über eine entsprechende Website zugänglich.
Alle fünf bis sechs Jahre werden die im oben genannten Verhaltenskodex beschriebenen Grundsätze von internationalen Expertinnen und Experten überprüft, die von Eurostat ausgewählt werden.

[ii] Für Unternehmensdaten gibt es gemäss dem Bundesgesetz über die Unternehmens-Identifikationsnummer (UID-Gesetz) eine Abstufung bezüglich der Wichtigkeit der Datenquellen. So haben bspw. die Daten im Handelsregister höchste Priorität. Weil man bei der Einführung des UID-Registers wusste, dass nicht unbedingt das Register mit der höchsten Priorität die höchste Qualität an Daten hält, wurde für das (öffentliche) Feld «Adresse» ein zweites (nicht-öffentliches) Feld «letztbekannte Adresse» eingeführt, um bei Problemen möglichst aktuelle Informationen verfügbar zu haben. Die Öffentlichkeit des UID-Registers hat dazu geführt, dass man dieses Feld nicht mehr braucht, weil der Druck zur Meldung der Korrektur relevant erhöht wurde.