Masterarbeit

Erhöhen Open Data Initiativen die Datenvalidität?

Case Study am Beispiel der International Aid Transparency Initiative (IATI)

eingereicht an der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Universität Bern

Institut für Wirtschaftsinformatik
Abteilung Information Management

Prof. Dr. Thomas Myrach

Betreuender Assistent
Gabriel Abu-Tayeh

eingereicht von
Stephanie Joss
von Hasle bei Burgdorf, BE
im 10. Semester
Matrikelnummer: 09-118-811

Studienadresse
Lährenbühlstrasse 20c
8112 Otelfingen
076 412 84 69
stephanie.joss@students.unibe.ch

Bern, 29.09.2016

Zusammenfassung

Open Data ist ein wichtiges Thema in der Entwicklungszusammenarbeit. Damit Spendengelder gezielt eingesetzt werden, braucht es transparente Daten zu Projekten und Budgets, um die Finanzflüsse zu kontrollieren und Synergien zwischen den Parteien zu knüpfen. Für die Interpretation und Weiterverwendung, müssen die Daten jedoch eine angemessene Qualität aufweisen. Open Data Initiativen sind ein gutes Instrument, um solche Qualitätskriterien festzulegen. Die International Aid Transparency Initiative (IATI) als relevantes Beispiel definiert einen Standard, wie Entwicklungshilfedaten zu teilen sind und stellt zusätzlich eine Plattform für den Informationsaustausch zur Verfügung. Die folgende Arbeit soll theoretisch und empirisch zeigen, dass Open Data Initiativen über die Zeit, aufgrund der Verwendung von Koordinationsmechanismen wie Standardisierungen, zu einer besseren Datenvalidität führen. Dabei wird die Validität als Teilkriterium der Datenqualität definiert und als die Übereinstimmung eines XML-Dokuments mit einem XML-Schema interpretiert. Die empirische Analyse zeigt auf dem 1%-Signifikanzniveau einen positiven Zusammenhang zwischen dem Alter von IATI, der Anzahl Files und der Anzahl teilnehmender Organisationen mit der abhängigen Variable Anteil valider Dokumente. Die Länge der Teilnahme der Organisationen an IATI als weiteren Einflussfaktor wird hingegen nicht bestätigt. Dennoch wird verdeutlicht, dass es sich für Organisationen und Institutionen lohnt, an einer Open Data Initiative teilzunehmen, da die Datenvalidität verbessert werden kann, was wiederum Fehlinterpretationen minimiert und die Vorteile von Open Data auftreten lässt.

Abstract

Open Data is an important topic in the field of aid transparency. In order that donation funds can be used purposefully, transparent data about projects and budgets are required which allows to monitor capital flows and generates synergies. For the interpretation and use, the data needs to be of an appropriate quality. Open Data Initiatives are a useful way to set such quality standards. The International Aid Transparency Initiative (IATI) as an important example defines a standard, how data about aid need to be shared and in addition to that makes a platform available for the information exchange. The following working paper aims to show theoretically and empirically that Open Data Initiatives over time effect a better data validity, given their usage of coordination mechanisms such as standardizations. In doing so, validity shall be defined as a subcriterion of data quality and can be interpreted as the consistency of an XML document with an XML scheme. Empirical analysis proofs a positive correlation at the 1% significance level between the age of IATI, the number of files and the number of participating organisations with the share of valid documents, being the dependent variable. The duration of participation of an organisation in IATI, however, cannot be proofed to have any impact. Nevertheless, it is illustrated that joining an Open Data Initiative is worthwhile for organisations and institutions since data validity can be improved which minimise misinterpretations and brings further advantages from Open Data.

Inhaltsverzeichnis

1 EINLEITUNG 1

1.1 AUSGANGSLAGE 1

1.2 PROBLEMSTELLUNG 4

1.3 ZIELSETZUNG 6

1.4 AUFBAU DER ARBEIT, METHODISCHES VORGEHEN 7

2 HAUPTTEIL 9

2.1 LITERATURÜBERBLICK 9

2.1.1 Open Data 9

2.1.1.1 Definition Open Data 9

2.1.1.2 Open Government Data 10

2.1.1.3 Begriff „Open“ 10

2.1.1.4 Open Data Prozess 12

2.1.1.5 Linked Open Data 13

2.1.1.6 Open Data Initiativen 14

2.1.2 Aspekte der Datenqualität 18

2.1.2.1 Definition Datenqualität 19

2.1.2.2 Regeln der Datenqualität 19

2.1.2.3 Datenqualitätskonstrukt 20

2.1.2.4 Validität als Teilaspekt der Datenqualität 20

2.1.2.5 Definition Datenvalidität 22

2.1.3 Open Data und Datenqualität 23

2.1.3.1 Einfluss der Datenqualität auf den Informationsaustausch 23

2.1.3.2 „Open“ und Datenqualität 24

2.1.3.3 Linked Open Data und Datenqualität 24

2.1.3.4 Stewardship and Usefulness 26

2.2 CASE STUDY 27

2.2.1 International Aid Transparency Initiative (IATI) 27

2.2.1.1 IATI als Open Data Initiative 27

2.2.1.2 Ursprung von IATI 28

2.2.1.3 Interessensgruppen 29

2.2.1.4 Governance Struktur 30
2.2.1.5 Finanzierung 30
2.2.2 Datenqualität von IATI 31
2.2.2.1 Infrastruktur und Ökosystem 31
2.2.2.2 IATI-Standard 33
2.2.2.3 Datenvalidität 36
2.2.2.4 Umgang der Organisationen mit der Datenqualität 37
2.2.2.5 Fazit von IATI 38
2.3 EMPIRISCHE DATENANALYSE 40
2.3.1 Methodisches Vorgehen 40
2.3.2 Organisationsübergreifende Analyse: Modelle 1 bis 9 42
2.3.2.1 Beschreibung und Bereinigung der Datensätze 42
2.3.2.2 Modell 1 & 4: Validität und Alter IATI 44
2.3.2.3 Modell 2 & 5: Validität und Anzahl Files 46
2.3.2.4 Modell 3 & 6: Validität und Anzahl Organisationen 48
2.3.2.5 Modell 7 & 8: Suche nach dem besten Modell 50
2.3.2.6 Modell 9: Anzahl Organisationen über die Zeit 51
2.3.2.7 Schlussdiskussion Modelle 1 bis 9 53
2.3.3 Analyse auf Stufe Organisationen: Modelle 10 & 11 54
2.3.3.1 Beschreibung und Bereinigung des Datensatzes 54
2.3.3.2 Modell 10: Verteilung Files pro Organisation 55
2.3.3.3 Bereinigung und Beschreibung des Datensatzes 58
2.3.3.4 Modell 11: Validität und Datum erstes Upload 58
2.3.4 Analyse Kontrollvariablen: Modelle 12 & 13 60
2.3.4.1 Modell 12: Validität und Alter IATI, Vergleich Organisationen 60
2.3.4.2 Modell 13: Validität und Anzahl Files Vergleich Organisationen 63
2.3.4.3 Schlussdiskussion Modelle 12 und 13 65
2.3.5 Übersicht Resultate 65

3 ZUSAMMENFASSUNG UND AUSBLICK 67

3.1 ZUSAMMENFASSUNG 67
3.2 AUSBLICK 69
3.3 ANHANG A 70
3.3.1 R-Code Organisationsübergreifende Analyse 70

3.3.2 R-Code: Analyse auf Stufe Organisation 77

3.3.3 R-Code: Analyse Kontrollvariablen 81

ABBILDUNGSVERZEICHNIS 89

TABELLENVERZEICHNIS 90

ABKÜRZUNGSVERZEICHNIS 91

SELBSTÄNDIGKEITSERKLÄRUNG 101

VERÖFFENTLICHUNG DER ARBEIT 102

1 Einleitung

Die folgende Arbeit beschäftigt sich mit Open Data Initiativen und deren Auswirkungen auf die Datenvalidität. Für die Untersuchung dieser Thematik wird der Fokus auf Daten der Entwicklungszusammenarbeit gelegt. Als konkretes Beispiel dient die International Aid Transparency Initiative (IATI). Die Begriffe Informationen und Daten werden in diesem Dokument synonym verwendet.

1.1 Ausgangslage

Open Data ist ein relativ junges Themengebiet, welches stark an Bedeutung gewonnen hat. Der rasche technologische Fortschritt der letzten Jahre hat dazu beigetragen, eine grössere Anzahl an Daten schneller und einfacher zu produzieren und zu verwenden. Verschiedene Softwares dienen der Interpretation und Transformation von Rohdaten in Wissen. Vor allem durch das Internet können mehr Daten gesammelt und diese einer grösseren Gemeinschaft zugänglich gemacht werden (Uhlir & Schröder, 2007, S. 38). Trotz dieser Vereinfachung können viele Daten dennoch nicht genutzt werden, da sie der Öffentlichkeit nicht zur Verfügung stehen. Die Open Data Bewegung möchte dagegen vorgehen und den Zugriff auf Daten vereinfachen, um Transparenz und Vertrauen zu schaffen. Open Data steht für die Idee, allen einen freien Zugang zu Daten zu gewähren, damit diese genutzt, modifiziert und weiterverbreitet werden (Open Knowledge, 2016). Daten sollen möglichst einfach von offengelegten Quellen bezogen werden können. Durch die gemeinsame Verwendung der Daten und Partizipation verschiedener Parteien können Innovationen entstehen, die zu einer besseren Effizienz beitragen (Huijboom & Van den Broek, 2011, S. 1). Eine Studie von McKinsey und Company (2014) schätzt, dass dank Open Data zwischen drei bis fünf Billionen US-Dollar (USD) jährlich generiert werden könnten.

Die Open Data Thematik ist auch in der Entwicklungszusammenarbeit relevant. Oftmals wird an der Entwicklungshilfe kritisiert, dass sie das Potential an Effektivität nicht vollständig ausschöpft. Spendengelder werden nicht dort eingesetzt, wo sie den grössten Nutzen erzielen. Die Gründe dafür sind vielseitig. Ein Problem stellt die Komplexität des Entwicklungshilfegefüges dar. Die involvierten Parteien weisen unterschiedliche Organisationsstrukturen auf und verfolgen verschiedene Ziele. Die Organisationstypen reichen von multilateralen und bilaterale Dienststellen, privaten

Stiftungen, regionalen Initiativen und globalen Funds, bis hin zu einem wachsenden Netzwerk an Nichtregierungsorganisationen (NGOs). Diese grosse Menge an Schnittstellen führt zu einer Vielzahl von Transaktionen, was wiederum zu enormen Kosten und einem grossen Koordinationsaufwand beiträgt. Ein anders Problem ist die strategische Ausrichtung von Entwicklungshilfe. Der Fokus wird oftmals zu stark auf den Projektabschluss gelegt, wobei eine längerfristige Ausrichtung auf nationaler Ebene vernachlässigt wird. Ein Ziel von Entwicklungshilfe sollte es jedoch sein, ein Land zu unterstützen, sich nachhaltig neu zu positionieren (Linders, 2013). In einigen Entwicklungsländern ist Korruption, die unter anderem auf intransparente Informationen zurückzuführen ist, nach wie vor ein aktuelles Thema (Schwegmann, 2012). Diese Schwierigkeiten begründen eine Zunahme der Wichtigkeit eines offenen Informationsaustauschs, um die Kommunikation und die Effektivität von Entwicklungshilfe zu verbessern. Hinzu kommen Motive wie die Möglichkeit der Partizipation der Stakeholder oder ein erhöhtes Verantwortungsbewusstsein der Regierungen (Linders, 2013). Die Verfügbarkeit von Internet auf Mobiltelefonen hat zudem dazu beigetragen, die Open Data Thematik in Entwicklungsländern bekannter zu machen. Die Gelegenheit weitere Daten zu sammeln und zu teilen wurde dadurch nochmals ausgeweitet (Hartung et al., 2010). Rein ökonomische Anreize wie Innovationen spielen in Entwicklungsländern eher eine untergeordnete Rolle (Schwegmann, 2012).

Die Diskussion rund um die Transparenz und Effektivität von Entwicklungshilfe hat bereits seit dem Jahr 2000 einen internationalen Kontext angenommen, worauf mit verschiedenen Vereinbarungen reagiert wurde. 2005 wurde die Paris Declaration, 2008 die Accra Agenda for Action und 2011 das Busan Partnership Agreement ins Leben gerufen (Publish What You Fund, 2016b, S. 4). Es wurde die Nachfrage nach einem offenen Standard laut, der Informationen zu Hilfsprojekten und Budgets zeitnahe, vorausschauend und reichhaltig liefern kann. Das Busan Partnership Agreement hat sich daher zum Ziel gesetzt, die Effektivität von Entwicklungszusammenarbeit mittels eines Standards zu erhöhen und einen regelmässigen Dialog zwischen den Parteien zu fördern. Die International Aid Transparency Initiative (IATI) ist eine Open Data Initiative und wurde entwickelt, um bei der Erreichung dieser Ziele zu helfen (Busan HL-4, 2011; Organisation for Economic Cooperation and Development [OECD], 2016b). Sie legt Standardisierungen fest, wie Informationen

über Entwicklungshilfe geteilt werden müssen und stellt eine Plattform zur Verfügung, von der die Daten bezogen werden können (International Aid Transparency Initiative [IATI], 2012).

Seit 2011, als erstmals Daten über die IATI Plattform veröffentlicht wurden, sind über 450 Organisationen beigetreten. Die Teilnahme an IATI ist grundsätzlich freiwillig. Dies wird teilweise kritisiert, da dadurch das Potential von IATI nicht vollständig ausgeschöpft wird. Dennoch gibt es einige Anreize, die einen Beitritt zu IATI fördern. Die Mehrheit der Datenherausgeber sind NGOs. Dies liegt daran, dass einige Spendengeber als Grundbedingung für die finanzielle Unterstützung das Reporting an IATI verlangen (IATI, 2016a). 2011 wurde zudem erstmals der Aid Transparency Index veröffentlicht. Er dient zur unabhängigen Messung der Transparenz von Entwicklungshilfe der Hauptspenderorganisationen (Publish What You Fund, 2016c). Der Index besteht aktuell aus 39 Indikatoren, die unterschiedlich gewichtet werden. Die Validität als Kriterium ist jedoch nicht vertreten. Jeder Indikator dient entweder zur Messung des Commitments oder der Publikationstätigkeit auf Stufe der Organisation oder der Aktivitäten. IATI spielt für die Bewertung eine entscheidende Rolle, da viele Daten von ihrer Webseite bezogen werden. Den Hauptspenderorganisationen wird empfohlen den IATI-Standard zu verwenden. Als Gegenzug werden dafür Zusatzpunkte verteilt (Publish What You Fund, 2016b). Dies kann wiederum einen positiven Einfluss auf das Rating haben. Je nach Anzahl Punkte werden die Organisation in unterschiedliche Kategorien eingeteilt, die das Ausmass an Transparenz definieren. Der Index bewertet zum jetzigen Zeitpunkt 46 Parteien. Dabei handelt es sich um 29 bilaterale Agenten (wie die United States Agency for International Development (USAID)), 16 multilaterale Institutionen (wie die Weltbank oder Frankreich) und um eine philanthropische Organisation. Die Teilnehmer müssen mindestens zwei der drei Bedingungen erfüllen, wie ein jährliches Spendenvolumen von einer Milliarde USD, die Unterzeichnung des Busan Agreements oder die Zugehörigkeit zu einer Vereinigungen wie die G7, die ein hohes Commitment zur Verbesserung der Transparenz ausgesprochen haben. Seit 2011, als der Index zum ersten Mal publiziert wurde, hat sich die Transparenz verbessert. Mittlerweile befindet sich ein Drittel der Organisationen im Bereich gut oder sehr gut, wobei zu Beginn von IATI keine Teilnehmer in diesen Kategorien vertreten waren (Publish What You Fund, 2016a, c). Die Erklärungen zum Aid

Transparency Index verdeutlichen das ansteigende Bewusstsein für die Wichtigkeit von transparenten Daten in der Entwicklungszusammenarbeit. Der Beitritt zu Open Data Initiativen, insbesondere zu IATI, gewinnt dadurch zunehmend an Bedeutung.

1.2 Problemstellung

Daten nur öffentlich zugänglich zu machen, reicht für die Lösung der Effektivitätsproblematik von Entwicklungshilfe jedoch nicht aus. Ein entscheidender Aspekt beim Austausch von Informationen ist der Zustand der verfügbaren Daten. Für eine zweckmässige Verwendung müssen Daten von guter Qualität sein. Risiken von Fehlinterpretationen und Datendiskrepanzen sollen reduziert werden (Dawes, 2010, S. 380). Schlechte Daten können einen negativen sozialen und ökonomischen Einfluss haben (Wang & Strong, 1996, S. 5). Das Information System (IS) Success Model von DeLone und McLean (2003) sagt aus, dass die Informationsqualität neben der System- und Servicequalität einer der zentralen Faktoren für ein erfolgreiches Informationssystem ist. Die wahrgenommene Qualität hat einen entscheidenden Einfluss auf die Zufriedenheit der User mit einem Informationssystem. Dies wiederum beeinflusst die Absicht es erneut zu nutzen. Erst eine gute Datenqualität lässt die Vorteile von Open Data auftreten.

Die Koordinationstheorie ist für die Erarbeitung der Fragestellung sehr entscheidend. Aus diesem Grund soll die Theorie an dieser Stelle kurz vorgestellt werden. Sie zeigt was hinter einem Standard steckt und wie dieser die Datenqualität beeinflusst.
Daten sind oftmals unterschiedlich aktuell, umfassend, exakt und in verschiedenen Formaten vorhanden. Damit sie durch unterschiedliche Systeme verarbeitet und weiterverwendet werden können, müssen sie in einem einheitlichen, maschinenlesbaren und umwandlungsfähigen Format verfügbar sein. Die Herausgeber kennen die Bedürfnisse der User jedoch nicht, da sie nicht miteinander kommunizieren. Deshalb braucht es zentrale Koordinationsmechanismen, die bei der Lösung solcher Probleme helfen. Die Koordinationstheorie sagt demnach aus, dass gewisse regulatorische Bedingungen gegeben sein müssen, wie Daten zu teilen sind. Erst eine bewusste Steuerung des Open Data Prozesses kann zu Vorteilen wie eine verbesserte Transparenz, Wirtschaftswachstum und Innovationen führen. Koordination bezeichnet im engeren Sinn das zielorientierte Managen von Wechselwirkungen zwischen verschiedenen Aktivitäten (Zuiderwijk & Janssen, 2013). Standards für

Daten und Metadaten, die bestimmen wie Datenelemente beschrieben, definiert und in einem System präsentiert werden müssen, ermöglichen eine bessere Datenqualität (Dawes, 2010, S. 380). Hinzu kommen Mechanismen wie Pläne oder gegenseitige Anpassungen, welche die Koordination und den Informationsaustausch vereinfachen. Die Bereitstellung von Feedbacksystemen und die Lancierung von Diskussionsrunden helfen zusätzlich die Qualität kontinuierlich zu verbessern (Zuiderwijk & Janssen, 2013). Open Data Initiativen sind ein wichtiges Instrument, um solche Rahmenbedingungen zu setzen. Aus diesem Grund wird in dieser Arbeit genauer auf Open Data Initiativen eingegangen und anhand von IATI den Einfluss auf die Datenqualität und insbesondere auf die Datenvalidität gezeigt. Damit wird ein zusätzlicher Grund geliefert, der für den Beitritt zu einer Open Data Initiative spricht.

Mit zunehmender Reichweite von IATI steigt jedoch die Herausforderung der Handhabung des Datenvolumens bezüglich der Qualität. IATI strebt danach die Datenqualität laufend zu verbessern. Um diese Entwicklung zu evaluieren, wurde im jährlichen Report 2015 von IATI (IATI, 2015a) die drei Qualitätsdimensionen Aktualität, Umfang und Vorhersage der Daten überprüft. Die Auswertung zeigt, dass im Jahr 2015 Entwicklungsgelder von über 78 Milliarden USD über IATI gemeldet wurden. Die Aktualität der Daten wird von 80% der Beteiligten mindestens vierteljährlich überprüft, in 41% der Fälle sogar mindestens monatlich. Das Qualitätsmerkmal des Umfangs betrifft die Verwendung von Kernelementen wie z. Bsp. Titel, Reporting Organisation oder die Beschreibung von Projekten, die wichtige Bestandteile für die Interpretation von Daten sind. Von 2014 auf 2015 hat sich die Anzahl dieser Angaben erhöht und zeigt damit eine wünschenswerte Entwicklung. Noch etwas mangelhaft sind die Vorhersagedaten zu Budgets. Für 2016 wurden lediglich 20% der erwarteten Spenden über IATI budgetiert. Für das Jahr 2017 sogar nur 8% (IATI, 2015a).

Das Qualitätskonstrukt besteht noch aus weiteren Eigenschaften, als diejenigen, die von IATI diskutiert werden. Da es den Rahmen einer Masterarbeit sprengen würde, alle Qualitätskriterien zu evaluieren, konzentriert sich diese Arbeit auf die Validität als Teilkriterium. In vielen Publikationen wird die Validität als wichtiges Qualitätsmerkmal erwähnt. Dabei wird Validität allgemein als die Gültigkeit der Daten für eine bestimmte Anwendung beschrieben (Bracht, Geckler, & Wenzel, 2011, S. 167).

Zu diesem Zweck braucht es definierte kontextabhängige Rahmenbedingungen (Bernard, Killworth, Kronenfeld, & Sailer, 1984, S. 505). Mittels der Validität werden Daten untereinander vergleichbar gemacht. Die Verwendung von invaliden Daten kann demzufolge zu Fehlinterpretationen führen.
Um die Validität zu messen, gibt es verschiedene Möglichkeiten. Diese Arbeit fokussiert sich auf die Validität eines Dokuments. Diese lässt sich damit bestimmen, ob ein in Extensible Markup Language (XML) verfasstes Dokument mit einer Original Dokumenten Typ Definition (DTD) oder einem XML-Schema übereinstimmt. Das Schema definiert, wie auch das DTD, die Struktur eines Dokuments und ermöglicht den Inhalt der Daten zu verifizieren. Dadurch können fehlerhafte Angaben minimiert werden (World Wide Web Consortium (W3C), 2012).

IATI erfasst zum einen wie viele hochgeladene XML-Dokumente mit dem definierten XML-Schema übereinstimmen und damit valide sind. Zum anderen misst sie die Anzahl Organisationen, die mindestens ein nicht valides Dokument hochgeladen haben (IATI, 2016h). Damit wird deutlich, dass die Validität eines Dokuments für IATI ebenfalls ein relevantes Kriterium ist, welches zu kontrollieren gilt. Wie diese Messungen jedoch interpretiert werden, wurde von IATI bisher nicht evaluiert. Aus diesem Grund soll in dieser Arbeit untersucht werden, wie sich die Validität der Dokumente seit Einführung von IATI verändert hat.

1.3 Zielsetzung

Die bisher beschriebenen Entwicklungen können zu der Annahme führen, dass sich die Validität der IATI-Dokumente seit Beginn erhöht hat. Dies könnte unter anderem mit der Zunahme des Alters der Open Data Initiative und des Lerneffekts begründet werden. Durch das Öffnen eines Systems findet ein Feedback Loop statt, der es ermöglicht Daten und das System laufend anzupassen (Janssen, Charalabidis, & Zuiderwijk, 2012, S. 259). Solche Qualitätsprozesse entwickeln sich oftmals über Jahre hinweg, da ein kontinuierliches Lernen stattfindet (Juran, 1998). Die Theorie des Lerneffekts könnte auch bei einzelnen Organisationen greifen und zur Überlegung führen, dass die Länge der Teilnahme der Organisationen an der Open Data Initiative ein zusätzlicher Grund für eine bessere Validität der Dokumente ist. Des Weiteren hat sich die Anzahl teilnehmender Organisationen im Laufe der Zeit deutlich erhöht. Dabei könnte die Weisheit der Masse zum Tragen kommen. Das Prinzip der

kollektiven Intelligenz sagt aus, dass eine Gruppe von Personen oftmals bessere Ergebnisse liefern als einzelne Individuen. Dies liegt an der vielfältigen Sichtweise auf ein Problem und der Vereinigung von Wissen und unterschiedlichen Fähigkeiten (Leimeister, 2010, S. 239). Mit zunehmender Anzahl Organisationen hat sich gleichwohl die Anzahl hochgeladener Dokumente erhöht. Um einen sauber Umgang mit den Daten zu pflegen, braucht es gewisse Erfahrungen mit dem Publikations-prozess von IATI. Durch eine grössere Anzahl an Dokumenten können mehr Erfahrungspunkte gesammelt werden, was wiederum einen positiven Einfluss auf die Validität der Dokumente haben könnte.

Mittels dieser Überlegungen werden die folgenden Hypothesen aufgestellt, die in dieser Arbeit überprüft werden sollen.

Theoretische Ebene:

- *Open Data Initiativen erhöhen die Datenvalidität über die Zeit*

Messebene:

- *Hypothese 1: Die Zunahme des Alters von IATI führt zu einer höheren Validität von Dokumenten*
- *Hypothese 2: Die Zunahme der Anzahl hochgeladener Dokumente führt zu einer höheren Validität von Dokumenten*
- *Hypothese 3: Die Zunahme der Anzahl teilnehmender Organisationen führt zu einer höheren Validität von Dokumenten*
- *Hypothese 4: Eine längere Teilnahme an IATI führt zu einer höheren Validität von Dokumenten*

Anschliessend soll das Modell gefunden werden, welches die Varianz der Validität der Dokumente am besten erklärt.

1.4 Aufbau der Arbeit, Methodisches Vorgehen

Der Hauptteil dieser Masterarbeit setzt sich aus den drei Teilbereichen Theorie, Praxis und Analyse zusammen.

Der erste Teil bildet die theoretische Grundlage und soll die Beziehung zwischen Open Data Initiativen und der Datenvalidität mittels verschiedener Studien belegen. Dazu werden die beiden Thematiken Open Data und Datenqualität genauer betrachtet. Zuerst sollen relevante Aspekte von Open Data gezeigt werden, um ein Verständnis für den Begriff zu schaffen und die Voraussetzungen für eine bessere Datenqualität zu definieren. Im nächsten Schritt soll das Konstrukt der Datenqualität erklärt werden. Dabei wird der Fokus auf die Validität als wichtiges Teilkriterium gelegt. Aufgrund der Beziehung zwischen der Qualität und der Validität kann begründet werden, dass eine Erhöhung der Datenqualität gleichzeitig zu einer besseren Datenvalidität führt. Im Anschluss sollen Studien beschrieben werden, welche die Überlegungen von Open Data und Datenvalidität vereinen. Der Theorieteil soll zusätzlich dazu dienen, die technische Messung der Validität zu stützen.

Der zweite Teil soll die theoretisch diskutierten Aspekte in die Praxis übernehmen und als Case Study bearbeiten. Zu diesem Zweck wird IATI hinzugezogen. Zu Beginn sollen einige Fakten zu IATI beschrieben werden, um relevante Hintergrundinformationen über die Open Data Initiative zu liefern. In einem nächsten Abschnitt wird die vorgegebene Qualität der IATI-Daten genauer betrachtet. Dabei soll die Validität besonders hervorgehoben werden, um die Grundlagen für die empirische Analyse zu legen. Zuletzt wird überprüft, ob IATI die Vorgaben für eine erfolgreiche Open Data Initiative erfüllt.

Im abschliessenden Teil wird eine empirische Datenanalyse durchgeführt, um die Hypothesen 1 bis 4 zu beantworten und das beste Modell zu finden. Die Validität der Dokumente soll dabei operationalisiert werden als der Anteil valider Dokumente. Wären alle hochgeladenen Dokumente valide, würde ein Anteil valider Dokumente von 1 und damit von 100% erreicht werden. Die Daten dazu können direkt von IATI bezogen werden. Der Anteil valider Dokumente soll zuerst auf organisationsübergreifender Ebene untersucht werden, um im nächsten Schritt einzelne Organisationen als Kontrollvariablen zu evaluiert. Zuletzt werden die Hypothesen anhand der ausgewählten Organisationen überprüft.

2 Hauptteil

2.1 Literaturüberblick

Der Literaturteil soll relevante Studien zu Open Data und zur Datenqualität diskutieren, um die theoretische Hypothese, dass Open Data Initiativen über die Zeit zu einer besseren Datenvalidität führen, zu bestätigen.

2.1.1 Open Data

Das folgende Kapitel behandelt wichtige Aspekte von Open Data, die für die Beantwortung der Fragestellung entscheidend sind.

2.1.1.1 Definition Open Data

Um ein gemeinsames Verständnis für Open Data zu schaffen, soll zuerst erklärt werden, was unter dem Begriff zu verstehen ist. Open Data ist vergleichbar mit der Definition von Open Source, welche oftmals in Verbindung mit Softwares auftaucht. In diesem Kontext wird Open Data als die Möglichkeit beschrieben, Daten zweckunabhängig zu nutzen, zu studieren, zu kopieren und ohne Einschränkungen weiterzuverarbeiten und die veränderte Version wieder zu teilen. Rechte müssen so weitergegeben werden, dass die modifizierten Daten wieder zu denselben Bedingungen genutzt werden können. „Open“ steht dabei für frei und soll dafür sorgen, dass Erlaubnisbarrieren abgebaut werden. Open Access weist ebenfalls gewisse Ähnlichkeiten mit Open Data auf. Bei dieser Definition wird der Begriff „Open“ jedoch öfters als kostenlos interpretiert. In der Literatur bedeutet Open Access freien Zugriff auf den vollständigen Inhalt von Dokumenten im Internet. Diese sollen ohne technische, finanzielle oder rechtliche Barrieren zugänglich sein, um sie zu lesen und auf unterschiedliche Weise zu verbreiten. Im Vergleich zu Open Access liegt der Schwerpunkt von Open Data vor allem auf der Möglichkeit der Wiederverwendung. Um die freie Nutzung der Daten zu bestätigen, werden sie oftmals mit einer Lizenz versehen. Die für diesen Zweck am weitesten verbreitete Lizenz ist die Creative Common Attribution License (CC-BY) (Murray-Rust, 2008, S. 52ff.). Sie erlaubt eine hohe Datenverwendung und anerkennt gleichzeitig den Urheber des Originals (Creative Commons, 2016).

2.1.1.2 Open Government Data

Im Kontext der Entwicklungshilfe wird Open Government Data aktuell. Die Thematik beinhaltet den Austausch von Daten aus dem öffentlichen Sektor, die nicht unter das Datenschutzgesetzt fallen. Im Fokus steht die Verwaltung und Politik, die relevante Informationen benötigen, um über Verfahren und Strategien, welche die Öffentlichkeit betreffen, zu entscheiden. Daher ist es wichtig, dass sie auf eine grosse Anzahl von Daten zurückgreifen können (Barnickel & Klessmann, 2012, S. 128f.). Mittlerweile haben die meisten Regierungen Open Data als festen Bestandteil in ihre eigene Strategie integriert. Die Motive für die Veröffentlichung von Regierungsdaten sind dabei unterschiedlich. Ein zentraler Punkt ist die Möglichkeit der politischen Partizipation der Bevölkerung. Transparente Daten führen zu einem besseren Verständnis über die Tätigkeiten und Leistungen von Regierungen. Die Gesellschaft wird dadurch ermutigt, die Ausübung ihrer demokratischen Rechte wahrzunehmen. Gleichzeitig können Regierungen durch die Offenlegung zur Verantwortung gezogen werden. Plattformen, welche Daten zentral zur Verfügung stellen, fördern zusätzlich die Kollaboration der Bürger. Diese kann kreative Prozesse hervorrufen, was wiederum vorteilhaft für Innovationen ist (Huijboom & Van den Broek, 2011). Das Prinzip der kollektiven Intelligenz kann hier wieder aufgegriffen werden. Durch die Verschiedenartigkeit und Unabhängigkeit von Personen innerhalb einer Gruppe, können Probleme vielseitiger angegangen werden und zu einer besseren Lösung beitragen (Leimeister, 2010, S. 239f.). Auf die kollektive Intelligenz wird im Verlauf dieser Arbeit nicht detaillierter eingegangen, sie soll aber als Argumentationsgrundlage dienen. Ein anderer Anreiz für Open Data ist, dass durch die Veröffentlichung der Daten eine finanzielle Entlastung der Regierungen entsteht, indem die Bevölkerung dazu motiviert wird, bei der Sammlung, Analyse und Verwendung der Daten zu unterstützen (Ren & Glissmann, 2012, S. 95).

2.1.1.3 Begriff „Open“

Damit Daten als „Open“ gelten müssen sie gewisse Eigenschaften erfüllen. Gestützt wird sich auf eine offizielle Definition aus der Open Government Data Thematik. Dahinter steckt die Überlegung, dass die reine Verfügbarmachung der Daten nicht ausreicht, um eine effektivere Nutzung und damit einen Mehrwert für die Gesellschaft zu erzielen. Ein Konsortium aus 30 Anwälten, die sogenannte Open Government Working Group hat sich zusammengesetzt, um zehn Attribute zu

definieren, die offene Daten aufweisen müssen (Dawes, 2010, S. 379). Wenn öffentliche Daten als „Open“ deklariert werden, dann sind sie (Sunlight Foundation, 2010):

- *Komplett:* Im Rahmen der Bestimmungen des Datenschutzes sollen alle Daten in der Rohfassung mit allen Metadaten vollständig verfügbar sein.
- *Primär:* Die Daten sollen der ursprünglichen Quelle entnommen werden.
- *Zeitgerecht:* Die Veröffentlichung der Daten soll innerhalb eines gewissen Zeitraums geschehen. Dieser ist von der Entstehung und der Aktualisierung der Daten abhängig und im Einzelfall zu betrachten.
- *Zugänglich:* Ohne grossen Aufwand sollen Daten zugänglich sein. Jegliche Art von Barrieren müssen abgebaut werden.
- *Maschinenlesbar:* Daten sollen in maschinenlesbarer Form vorhanden sein, damit sie möglichst einfach in Softwareapplikationen überführt werden können. Die Automatisierung kann dadurch gefördert werden.
- *Nicht diskriminierend:* Personen sollen ohne Offenlegung ihrer Identität (Registrierung, Mitgliedschaft, etc.) oder des Verwendungszwecks, Zugriff auf die Daten bekommen. Es dürfen keine Personen von der Nutzung der Daten ausgeschlossen werden.
- *Nicht proprietär:* Es soll ein offener Standard für die Datennutzung verwendet werden. Die Benutzung von zusätzlichen kostenpflichtigen Programmen soll vermieden werden.
- *Lizenzfrei:* Daten müssen gemeinfrei nutzbar sein und dürfen nicht durch Lizenzen beschränkt werden. Lizenzen wie die CC-BY sind zulässig, da sie die Verwendung und Weiterverbreitung fördern.
- *Dauerhaft:* Publizierte Daten sollen offen bleiben, d.h. alle Veränderungen müssen ebenfalls wieder öffentlich gemacht werden. Sie sollen permanent verfügbar sein und regelmässig archiviert werden.
- *Kostenlos:* Der Zugriff auf die Daten soll frei von Gebühren sein, damit die Nutzung nicht durch finanzielle Barrieren gehindert wird.

Wenn diese Bedingungen erfüllt sind, dann sind die Daten offen und können von einer grösseren Anzahl Personen einfacher verwendet werden.

2.1.1.4 Open Data Prozess

Bei Open Data geht es nicht nur um die Offenlegung von Daten. Die Thematik beinhaltet alle Aktivitäten von der Produktion bis hin zur Veröffentlichung und eigentlichen Nutzung der Daten (Zuiderwijk & Janssen, 2013, S. 38). Open Data ist ein fortlaufender Prozess, der durch sein zyklisches Verhalten die Verwendung der Daten kontinuierlich weiterentwickelt. Er stellt ein wichtiges Hilfsmittel bei der Identifizierung von Barrieren dar und zeigt, wo Feedbackmechanismen greifen. Zwei unterschiedliche Sichtweisen sind dabei zu beachten. Auf der einen Seite stehen die Datennutzer, wie die Bevölkerung, Unternehmen, Forscher, Behördenbedienstete und weitere User. Auf der anderen Seite befinden sich die Organisationen und Institutionen, welche die Daten zur Verfügung stellen (Zuiderwijk, Janssen, Choenni, Meijer, & Alibaks, 2012, S. 157 f.). Wie die Abbildung 1 zeigt, gliedert sich der Open Data Prozess in fünf Phasen.

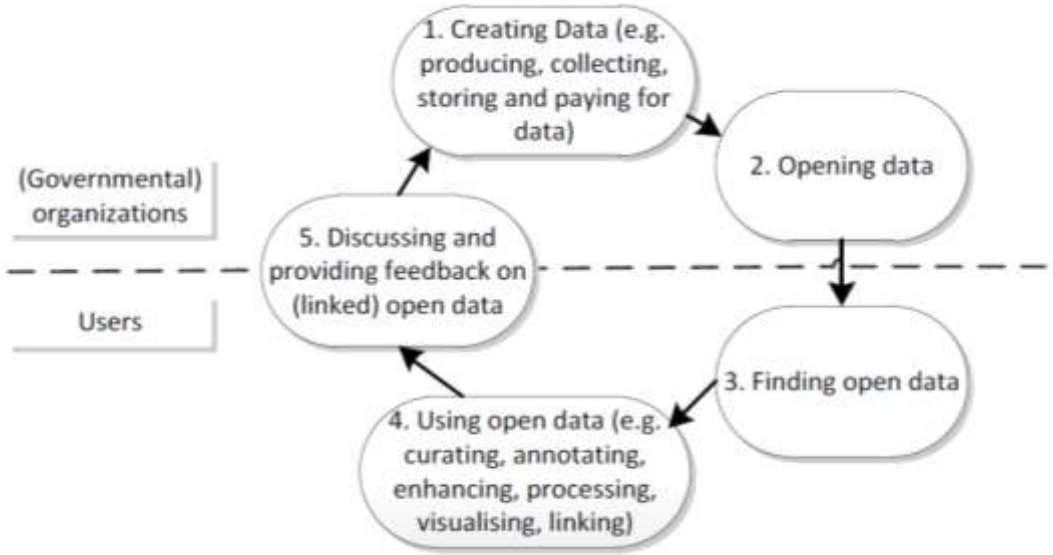

Abbildung 1: Open Data Prozess nach Zuiderwijk et al. (2012, S. 157).

Als Erstes werden Gelder den öffentlichen Organisationen gestellt, um Daten zu produzieren, zu sammeln und in die Systeme zu integrieren. Im zweiten Schritt wird über die Veröffentlichung der Daten entschieden. Alle nicht privaten Daten, die nicht unter das Datenschutzgesetz fallen, sollen im Internet zugänglich gemacht werden. Dies kann entweder auf der eigenen Webseite, auf einem nationalen Portal oder auf irgendeiner anderen Plattform sein. Wie die Daten schlussendlich genutzt werden,

hängt davon ab, in welcher Qualität sie publiziert sind. Daher sollte ein gewisses Mass an Qualität bereits zu Beginn vorhanden sein, damit die Daten zur Verwendung kommen. Koordinationsmechanismen wie Standardisierung oder Plänen helfen dabei, die Daten in adäquater Form zu veröffentlichen. In der dritten Phase geht es um die Auffindbarkeit der Informationen. Eine Herausforderung stellt die starke Fragmentierung der Daten dar. Standardisierte Kataloge oder die Verlinkung der Daten wie bei Linked Open Data können jedoch Abhilfe schaffen. Im vierten Schritt geht es darum, die Daten ohne Hindernisse von Copyrights, Kontrollmechanismen oder Patenten zu verwenden und weiterzuverarbeiten. Als letzte Phase sollte sich ein Feedback Loop einstellen, der den öffentlichen Organisationen bei der Weiterentwicklung der Daten helfen soll. Diese Phase ist besonders wichtig für die Verbesserung der Datenqualität, da die Feedbacks direkt von den Usern kommen. (Zuiderwijk et al., 2012, S. 157 f.; Zuiderwijk & Janssen, 2013). Dieses Beispiel untermauert die Theorie, dass Open Data unter Verwendung von Koordinationsmechanismen die Datenqualität über die Zeit erhöht.

2.1.1.5 Linked Open Data

Im semantischen Web ist eine Vielzahl von Daten verfügbar. Die Übersichtlichkeit geht dadurch schnell verloren und die Auffindbarkeit der Daten wird erschwert. Gemäss dem Open Data Prozess müssen die Daten, um zur Verwendung zu kommen, zuerst aufgefunden werden. Aus diesem Grund sind in der aktuellen Entwicklung Linked Open Data nicht mehr wegzudenken (Bauer & Kaltenböck, 2012). Linked Open Data bezeichnen maschinenlesbare offene Daten, die im semantischen Web miteinander verlinkt sind (Yu, 2011, S. 409). Damit die Kosten bei der Transformation möglichst tief bleiben, werden Standards für die Verlinkung verwendet. Die Nutzung von Schemas, einer gemeinsamen Syntax und Vokabulars helfen den Austausch von Informationen mittels Linked Open Data zu vereinfachen (Bauer & Kaltenböck, 2012, S. 25). Dazu kann das Resource Description Framework (RDF) hinzugezogen werden. Es unterstütz bei der Publikation von strukturierten Daten und der Zusammenführung von Daten aus unterschiedlichen Quellen (Yu, 2011, S. 410). Beim RDF handelt es sich um ein in XML geschriebenes Standardmodell. Es beschreibt die Rahmenbedingungen für die im Web verwendeten Ressourcen. Das RDF besteht aus einem Tripel und formuliert die Beziehung eines Subjekts, Prädikats und Objekts, wobei sich die Richtung immer vom Subjekt zum Objekt verhält.

Uniform Resource Identifiers (URIs) werden für die Identifizierung solcher Ressourcen benötigt (W3C, 2004).

Das in der Abbildung 2 dargestellte fünf Sterne Modell von Sir Tim Berners-Lee beschreibt die Entwicklung von Open Data im Sinne der Veröffentlichung von Daten in irgendeinem Format, bis hin zu Linked Open Data. Es wird zwischen verschiedenen Stufen von Formaten unterschieden, die mit Zunahme der Anzahl Sterne eine vereinfachtere Datenverwendung bedeuten, da die Formate von mehreren Systemen interpretiert werden können. Vom ursprünglichen Portable Document Format (PDF) hat sich Open Data über das Excel-Format, bis hin zum Comma-Separated Values (CSV) und zum RDF, schlussendlich zu Linked Open Data entwickelt. Linked Open Data ist damit aktuell die beste Form, um Informationen im Web verfügbar zu machen (Bauer & Kaltenböck, 2012, S. 17).

★	Information is available on the Web (any format) under an open license
★★	Information is available as structured data (e.g. Excel instead of an image scan of a table)
★★★	Non-proprietary formats are used (e.g. CSV instead of Excel)
★★★★	URI identification is used so that people can point at individual data
★★★★★	Data is linked to other data to provide context

Abbildung 2: Fünf Sterne Modell nach Bauer & Kaltenböck (2012).

2.1.1.6 Open Data Initiativen

Da diese Masterarbeit die Hypothese aufstellt, dass Open Data Initiativen über die Zeit zu einer besseren Validität führen, soll im folgenden Abschnitt aufgezeigt werden, welche Komponenten den Erfolg einer Initiative begünstigen. Die Koordinationstheorie sagt aus, dass die Vorteile von Open Data wie Transparenz und Innovation dann auftreten, wenn Koordinationsmechanismen den Open Data Prozess steuern (Zuiderwijk & Janssen, 2013). Solche Mechanismen sind Teil von Open Data Initiativen. Kernelemente sind vor allem offene Formate, Standards für den Publikationsprozess und die Schaffung eines Lizenzrahmens für die einfachere Daten-

verwendung (Davies, 2011, S. 1). Anbei sollen noch weitere Aspekte von Open Data Initiativen gezeigt werden.

2.1.1.6.1 Aufbau einer E-Infrastruktur

Wie bereits angesprochen, ist das Internet ein wichtiges Hilfsmittel für den Informationsaustausch. Webseiten, die nach festgelegten Richtlinien designt werden, können einen optimalen Rahmen für die vorteilhafte Nutzung von Open Data liefern. Im Zentrum von Open Data Initiativen steht daher der Aufbau einer E-Infrastruktur. Es müssen Strukturen geschaffen werden, die es den unterschiedlichen Interessensgruppen ermöglichen, sich untereinander auszutauschen (De Cindio, 2012). In Webseiten integrierte Diskussionsforen oder Blogs helfen dabei Feedbacks zu fördern (De Cindio, 2012; Zuiderwijk & Janssen, 2013). Mittels Anwendungsprogrammierschnittstellen (API) werden die Daten und Prozesse in die eigenen Systeme der Beteiligten integriert, um damit den Informationsaustausch zu automatisieren. Dies ermöglicht das Userverhalten zu evaluieren und zu prüfen, in welcher Phase des Open Data Prozesses sich die Stakeholder befinden. Des Weiteren helfen die Verknüpfungen, die Auffindbarkeit der Daten zu erleichtern (Zuiderwijk & Janssen, 2013).

Eine gute E-Infrastruktur macht aber nicht nur technische Anforderungen aus. Für die User müssen Bedingungen geschaffen werden, die sie anreizen in einem bestimmten Umfeld partizipieren zu wollen. Features können dafür ein gutes Hilfsmittel sein. Das Ziel einer Open Data Initiative sollte es insgesamt sein eine gemeinsame Identität aufzubauen, damit die Beteiligten als Gemeinschaft agieren. Die Verwendung eines Slogans oder die Gestaltung eines speziellen Logos können eine Verbundenheit mit der Initiative erzeugen. Eine E-Infrastruktur sollte zudem flexibel genug sein, um auf spezielle Ereignisse reagieren zu können, da das Partizipationsverhalten stark von aktuellen Geschehnissen abhängt (De Cindio, 2012).

Neben dem digitalen Lebensraum, der für den Informationsaustausch kreiert wird, sollten auch formale Bedingungen festgelegt werden. Abkommen zwischen den Parteien können das gegenseitige Commitment aussprechen. Eine Art „Code of Conduct“ soll zudem ein geregeltes Verhalten zwischen den Partizipierenden spezifizieren. Ein Moderator soll als vertrauenswürdige Partei dafür sorgen, dass die Vorgaben eingehalten werden (De Cindio, 2012). Eine Open Government Data Initiative kann zentral von der Regierung, einem Departement, oder von aussenstehenden Mit-

streitern geführt werden (Davies, 2011, S. 1). Für eine saubere Definition der Regeln und Standards können Intermediäre die Interessen der Regierung und den Partizipierenden ausbalancieren. Diese Partei sollte einen nicht profitablen Charakter aufweisen (De Cindio, 2012).

Damit die User einfacher auf Daten zugreifen können, sollten Eintrittsbarrieren durch Registrierungen möglichst vermieden werden. Nicknamen oder E-Mail-Adressen können in einigen Fällen abgefragt werden. Die Anonymität der User sollte jedoch geschützt bleiben. Bei öffentlichen Organisationen, die Daten zur Verfügung stellen, ist es hingegen sinnvoll ihre Identität preiszugeben, damit sie Verantwortung für die publizierten Daten übernehmen können (De Cindio, 2012).

Der Erfolg einer Open Data Initiative hängt von vielen Faktoren ab. Eine abschliessende Aufzählung aller Voraussetzungen kann in dieser Arbeit nicht gemacht werden. Ebenso wenig sollen die Bedingungen nach ihrer Wichtigkeit sortiert werden. Das Unterkapitel soll jedoch zeigen, dass gewisse Voraussetzungen gegeben sein müssen, damit Open Data Initiativen erfolgreich sind und die Vorteile von Open Data auftreten.

2.1.1.6.2 Relevante Informationen in einer Open Data Initiative

Zusätzlich zum Aufbau einer E-Infrastruktur müssen im Vorfeld bestimmte Informationen identifiziert werden, die für eine erfolgreiche Open Data Initiative entscheidend sind. Der Abbildung 3 ist zu entnehmen, dass die Identifizierung in fünf Phasen abläuft (Ren & Glissmann, 2012).

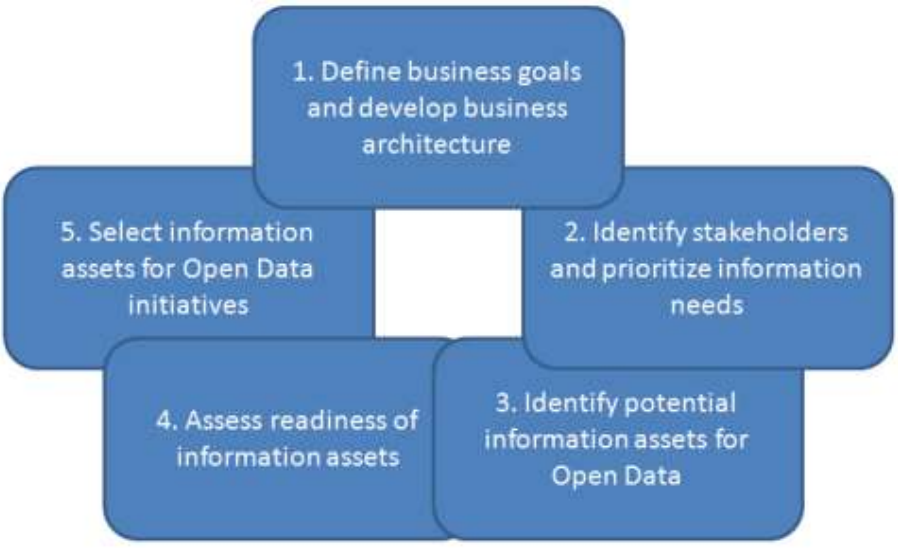

Abbildung 3: Informationsidentifizierung nach Ren & Glissmann (2012, S. 96).

In der ersten Phase geht es darum Geschäftsziele zu definieren. Bekannte Beispiele dafür sind Kostenersparnisse oder höhere Einnahmen. Um jedoch die effektiven Ziele und Massnahmen für eine Open Data Initiative abzuleiten, müssen zuerst die einzelnen Organisationskomponenten identifiziert werden. Der Aufbau einer Geschäftsarchitektur kann dabei helfen, die relevanten Elemente zu evaluieren und ihre Wichtigkeit zu priorisieren. Neben der Erfüllung von Geschäftszielen müssen Open Data Initiativen zusätzlich die Bedürfnisse der Stakeholder abdecken. Diese können in der zweiten Phase ermittelt werden. Die spezifische Nachfrage nach Informationen kann in den einzelnen Geschäftskomponenten lokalisiert werden. Für eine möglichst effektive Priorisierung sollten die Bedürfnisse der Stakeholder und die der Geschäftsziele im gleichen Ausmass berücksichtig werden. In der dritten Phase geht es um das Erkennen von Potenzialen von hoch priorisierten Informationen. Oftmals können Informationen für weitere Zwecke als jene, die auf den ersten Blick ersichtlich sind, genutzt werden. Im vierten Schritt soll die Datenqualität evaluiert werden. Dazu werden theoretisch definierte Datenqualitätsmerkmale herangezogen. Mittels Interviews und Studien können die Stakeholder die Informationsqualität anhand der Merkmale bewerten. Dabei wird unterteilt in die Datenqualität, die sein sollte und diejenige, die zurzeit gegeben ist. Dieser Gap wird in der letzten Phase genauer betrachtet. Die Informationen werden abschliessend nach ihrem Kosten-Nutzen-Verhältnis ausgewählt (Ren & Glissmann, 2012). Die Theorie zeigt, dass Open Data Initiativen frühzeitig bemüht sind eine angemessene Informationsqualität zu erreichen.

2.1.1.6.3 Open Data Initiativen in Entwicklungsländern

Das folgende Unterkapitel soll als kleiner Einschub weitere Open Data Initiativen aus der Entwicklungshilfe vorstellen. Damit soll der Kontext der Arbeit nochmals verdeutlicht werden. Obwohl die Anzahl Open Data Initiativen in Entwicklungsländern zum jetzigen Zeitpunkt noch nicht so hoch ist, wird mit einer Zunahme gerechnet. Der öffentliche Druck auf die Regierungen und der Reputationsgewinn, der durch die Lancierung von Open Data Initiativen erzielt werden kann, leisten einen entscheidenden Beitrag dazu. Dies zeigt die politische Motivation solcher Initiativen. Oftmals werden sie durch internationale Organisationen unterstützt (Schwegmann, 2012). Die Weltbank als Beispiel war die erste multilaterale Organisation, die ihre Daten über IATI publiziert hat (The World Bank Group, 2012, S. 1).

Sie verfolgt das Ziel, die Armut zu reduzieren und den Wohlstand von vielen Nationen zu verbessern. Neben der Bereitstellung von finanziellen Mitteln für Investitionen in Bildung, Gesundheit, Agrikultur und sonstige öffentliche Sektoren, möchte sie als technischer Assistent den Zugang zu Informationen vereinfachen. Zu diesem Zweck können Daten über Entwicklungsprojekte direkt über die Webseite der Weltbank bezogen werden. Mit Hilfe von Weltbank Live werden den Partizipierenden zudem Möglichkeiten für einen direkten Austausch geboten (The World Bank, 2016b). Um die Datenqualität laufend zu verbessern, greift die Weltbank auf international anerkannte Standards, Methoden, Definitionen, Klassifikationen und Quellen zurück (The World Bank, 2016a). Entwicklungsländern bieten solche Initiativen neue Möglichkeiten, Finanzierungsquellen zu akquirieren. Die Bedingungen, die für eine erfolgreiche Open Data Initiative gegeben sein müssen, treffen auch auf Initiativen in Entwicklungsländern zu. Ein Problem ist jedoch, dass die Internetverbindung auf Computern nicht immer ausreichend oder zu teuer ist. Ebenso sind viele Systeme mangelhaft (Schwegmann, 2012). Durch den Zugang zum Internet über das Mobil-telefon, sind hingegen neue Möglichkeiten der Datennutzung hinzugekommen (Hartung et al., 2010). Dies macht sich die Mobile and Development Intelligence (MDI) Groupe Speciale Mobile Association (GSMA) zu nutze. Verschiedene Parteien wie die Industrie oder Mobilfunkbetreiber aus aller Welt sollen mittels dieser Open Data Plattform vernetzt werden, um gemeinsam mobile Serviceleistungen für die Bevölkerung zu gestalten. Oftmals ist in Entwicklungsländern das Mobiltelefon das einzige Kommunikationsmittel. Die MDI hat sich zum Ziel gesetzt, den Bürgern über das Mobiltelefon einen schnellen Zugang zu existenziell kritischen Informationen wie Katastrophenmeldungen, medizinische Versorgung oder landwirtschaftliche Entwicklungen zu geben. Ebenso sollen ökonomische und soziale Vorteile, wie eine verbesserte Kommunikation innerhalb der Bevölkerung, und die Möglichkeit für Zahlungen direkt über das Mobiltelefon geschaffen werden (GSMA, 2016).

2.1.2 Aspekte der Datenqualität

Nachdem nun einige Aspekte von Open Data diskutiert wurden, soll das folgende Kapitel einige Theorien der Datenqualität aufzeigen. Im Fokus steht die Validität als wichtiger Teilaspekt des Qualitätskonstrukts. Dadurch soll begründet werden, dass eine bessere Datenqualität, aufgrund des Beziehungsverhältnisses gleichzeitig zu

einer besseren Datenvalidität führt. Einen Konsens darüber, wie Datenqualität am besten definiert wird, soll jedoch nicht gebildet werden (Ren & Glissmann, 2012, S. 95). Die ausgewählten Theorien sollen ein Grundverständnis für die Definition der Qualität und Validität schaffen.

2.1.2.1 Definition Datenqualität

Durch den direkten Zugriff von Usern auf Informationen und der Zunahme von universellen Datenbanken, wurde die Nachfrage nach hochqualifizierten Daten grösser (Lee et al., 2002, S. 133). Generell werden Daten als Schlüsselressourcen gesehen und als nur so wertvoll empfunden, wie ihre Qualität (Zaveri et al., 2012, S. 1). Diese zeigt sich dabei erst bei der eigentlichen Nutzung (Tayi, Ballou & Guest Editors, 1998, S. 56). Die Datenqualität ist ein mehrdimensionales Konstrukt und kann definiert werden als den Zustand der Daten, die „fit für die Verwendung“ sind. Damit ist unter anderem gemeint, dass die Qualität abhängig vom Kontext ist, in dem die Daten verwendet werden, wobei unterschiedliche Ausprägungen in verschiedenen Situationen relevant sind (Tayi, Ballou & Guest Editors, 1998, S 57). Insgesamt müssen die Daten in einer Qualität sein, die den Kundenbedürfnissen entspricht und die es ermöglicht, gute Entscheidungen zu treffen (Orr, 1998, S. 2; Stvilia et al., 2007, S. 1721).

2.1.2.2 Regeln der Datenqualität

In seinem Artikel formuliert Orr (1998) einige Regeln in Bezug auf die Datenqualität. Die Datenqualität wird definiert als der Grad an Übereinstimmung zwischen den Daten in der realen Welt und diesen, die durch ein Informationssystem präsentiert werden. Dabei sollen die Rahmenbedingungen, die für die Datenqualität aufgesetzt werden, gleichermassen für die Daten selbst wie für die Metadaten gelten. Der Kerngedanke ist jedoch die Notwendigkeit der Nutzung und nicht der eigentlichen Sammlung der Daten. Es geht darum, dass Daten ihre Gültigkeit verlieren, wenn sie über längere Zeit nicht genutzt werden. Änderungen in der realen Umwelt werden durch die Inaktivität nicht mehr aufgenommen. Deshalb entstehen grössere Abweichungen zu den Daten im System. Feedbackmechanismen, welche die Daten miteinander abgleichen und verbessern, fehlen in diesem Teil. Daten im System müssen daher eng mit der Aussenwelt verknüpft sein, um ihre Qualität zu kontrollieren. Insgesamt kann die Datenqualität jedoch nicht besser sein, als zum

Zeitpunkt der höchsten Nutzungsfrequenz. Hinzu kommt, je älter ein System ist, desto schlimmer werden die Probleme mit der Datenqualität. Orr (1998) betont damit die Relevanz der Verwendung von Daten für die Verbesserung der Qualität, was wiederum den positiven Einfluss von Open Data bestätigt.

2.1.2.3 Datenqualitätskonstrukt

Oft wird die Exaktheit der Daten als zentrales Qualitätskriterium hervorgehoben. Das Konstrukt der Datenqualität besteht jedoch aus vielen Aspekten. Wang und Strong (1996, S. 6ff.) haben dazu eine Studie durchgeführt und bei Datenkonsumenten nachgefragt, welche Eigenschaften eine gute Datenqualität ausmachen. Die Auswertung der Befragung hat ergeben, dass die Attribute in vier Kategorien eingeteilt werden können. Die Daten sollten intrinsisch gut, konzeptionell passend, übersichtlich dargestellt und zugänglich sein. Eine gute Datenqualität macht insgesamt aus, wenn die Daten „fit für die Verwendung“ sind. Die intrinsische Datenqualität beschreibt dabei den inneren Charakter von Daten, d.h. Eigenschaften, die Daten unabhängig von externen Einflüssen erfüllen müssen. Bei der kontextuellen Ansicht geht es um die Eignung von Daten für eine bestimmte Aufgabe. Die repräsentative Kategorie verlangt eine übersichtliche und klare Darstellung, um die Interpretation der Daten zu erleichtern. Zuletzt macht eine gute Qualität aus, wenn die Daten für die User leicht zugänglich und nutzbar sind. Das Kriterium der Zugänglichkeit und der Datenpräsentation zeigt die Notwendigkeit eines Systems. Die Einteilung in die vier Klassen von Wang und Strong (1996, 6ff.) hilft die verschiedenen Aspekte der Datenqualität besser zu verstehen. Sie dienen als Grundlage für weitere Studien und werden sowohl in der Industrie wie auch in der Government Thematik verwendet.

2.1.2.4 Validität als Teilaspekt der Datenqualität

Im folgenden Abschnitt soll die Validität als Teilaspekt der Datenqualität hervorgehoben werden. Fokussiert wird sich dabei auf allgemeine Definitionen und nicht spezifisch auf die Validität von Dokumenten.

Die Tabellen 1 und 2 liefern eine Übersicht verschiedener Informationsqualitätsdefinitionen, die auf den vier Kategorien von Wang und Strong (1996) basieren. Der Unterschied der einzelnen Studien besteht in der Zuordnung von verschiedenen Attributen den einzelnen Klassen. Die zwei Tabellen zeigen die akademische und die praktische Sichtweise der Informationsqualität (Lee et al., 2002, S. 134ff.).

The academics' view of information quality

	Intrinsic IQ	Contextual IQ	Representational IQ	Accessibility IQ
Wang and Strong [39]	Accuracy, believability, reputation, objectivity	Value-added, relevance, completeness, timeliness, appropriate amount	Understandability, interpretability, concise representation, consistent representation	Accessibility, ease of operations, security
Zmud [41]	Accurate, factual	Quantity, reliable/timely	Arrangement, readable, reasonable	
Jarke and Vassiliou [16]	Believability, accuracy, credibility, consistency, completeness	Relevance, usage, timeliness, source currency, data warehouse currency, non-volatility	Interpretability, syntax, version control, semantics, aliases, origin	Accessibility, system availability, transaction availability, privileges
Delone and McLean [11]	Accuracy, precision, reliability, freedom from bias	Importance, relevance, usefulness, informativeness, content, sufficiency, completeness, currency, timeliness	Understandability, readability, clarity, format, appearance, conciseness, uniqueness, comparability	Usableness, quantitativeness, convenience of access[a]
Goodhue [14]	Accuracy, reliability	Currency, level of detail	Compatibility, meaning, presentation, lack of confusion	Accessibility, assistance, ease of use (of h/w, s/w), locatability
Ballou and Pazer [4]	Accuracy, consistency	Completeness, timeliness		
Wand and Wang [37]	Correctness, unambiguous	Completeness	Meaningfulness	

[a]Classified as system quality rather than information quality by Delone and McLean.

Tabelle 1: Informationsqualität: Akademische Sicht nach Lee et al. (2002, S. 134).

The practitioners' view of information quality

	Intrinsic IQ	Contextual IQ	Representational IQ	Accessibility IQ
DoD [10]	Accuracy, completeness, consistency, validity	Timeliness	Uniqueness	
MITRE [25]	Same as [39]	Same as [39]	Same as [39]	Same as [39]
IRI [20]	Accuracy	Timeliness		Reliability (of delivery)
Unitech [23]	Accuracy, consistency, reliability	Completeness, timeliness		Security, privacy
Diamond Technology Partners [24]	Accuracy			Accessibility
HSBC Asset Management [13]	Correctness	Completeness, currency	Consistency	Accessibility
AT&T and Redman [29]	Accuracy, consistency	Completeness, relevance, comprehensiveness, essentialness, attribute granularity, currency/cycle time	Clarity of definition, precision of domains, naturalness, homogeneity, identifiability, minimum unnecessary redundancy, semantic consistency, structural consistency, appropriate representation, interpretability, portability, format precision, format flexibility, ability to represent null values, efficient use of storage, representation consistency	Obtainability, flexibility, robustness
Vality [8]			Metadata characteristics	

Tabelle 2: Informationsqualität: Praktische Sicht nach Lee et al. (2002, S. 136).

Wie der Tabelle 2 zu entnehmen ist, formuliert das Departement of Defense (DoD) aus der praktischen Sichtweise als Bedingung für eine gute intrinsische Informationsqualität die Exaktheit, die Vollständigkeit, die Konsistenz und die Validität

der Daten (Lee et al., 2002, S. 136.). Diese Definition stützt die Relevanz der Validität als Qualitätskriterium.

Die Betonung der Validität als Teilaspekt der Datenqualität, wird auch bei der Forschung von Taylor (1986, S. 50) ersichtlich. Die Exaktheit, die Vollständigkeit, die Aktualität, die Reliabilität und die Validität der Daten werden als die 5 zentralen Eigenschaften definiert, die Informationen wertvoll machen. Diese Attribute verkörpern die Bedürfnisse der User, wie die Datenqualität in einem System zu gestalten ist. Obwohl die Studie schon etwas älter ist, zeigt sie dennoch eine aktuelle Thematik auf.

Stvilia et al. (2007) haben in ihrer Studie 22 Informationsqualitätsmerkmale evaluiert. Diese ordnen sie den drei Kategorien der intrinsischen, der relationalen/kontextuellen und der reputationsfördernden Informationsqualität zu, welche Ähnlichkeit mit den Klassen von Wang und Strong (1996) haben. Die intrinsische Kategorie beschreibt interne Eigenschaften der Daten, die kulturellen Normen und Konventionen entsprechen. Die relationale/kontextuelle Dimension hingegen umfasst die Attribute, die Daten in einen bestimmten Kontext setzen. Die letzte Kategorie beinhaltet Merkmale, die eine gewisse Position in einer bestimmten Struktur definieren. Dieser Kategorie wird lediglich die Autorität, als Grad der Reputation einer Information in einer Situation, zugeordnet. Als eine intrinsische Ausprägung wird in diesem Ansatz die Genauigkeit zusammen mit der Validität genannt. Die Begriffe definieren in welchem Ausmass die Informationen zuverlässig zu einer stabilen Quelle, wie ein Wörterbuch referenzieren. Die Studie bestätigt damit ebenfalls die Validität als Qualitätskriterium von Daten.

2.1.2.5 Definition Datenvalidität

Die empirische Analyse, welche im dritten Teil durchgeführt wird, konzentriert sich auf die Validität von Dokumenten. Aus diesem Grund soll im folgenden Unterkapitel die Validität für diesen Kontext noch etwas genauer definiert werden.

Die Validität ist ein Gültigkeitskriterium und ein Mass für die Genauigkeit. Dabei kann zwischen der internen und externen Validität differenziert werden. Die interne Validität begründet die Gültigkeit eines Resultats, d.h. die Werteschwankungen der

abhängigen Variable, mittels der Manipulation der unabhängigen Variablen im gegebenen Experiment. Die externe Validität hingegen lässt die Verallgemeinerung von gefundenen Resultaten zu (Salkind, 2011, S. 147).

Diese Masterarbeit konzentriert sich auf die Validität von XML-Dokumenten und damit auf die interne Validität. Entspricht ein XML-Dokument einem XML-Schema, dann ist es valide und gleichzeitig wohlgeformt. Wohlgeformtheit bedeutet, dass das Dokument die festgelegte Syntax von XML erfüllt (World Wide Web Consortium [W3C], 2012). Dies kann zutreffen, ohne dass ein XML-Schema existiert. Validität ist hingegen nur gegeben, wenn die Struktur durch ein XML-Schema vorgegeben ist und der Inhalt eines XML-Dokuments dadurch qualifiziert werden kann. Das XML-Schema und das XML-Dokument können mittels des Wurzelelements miteinander verknüpft werden (Guerrini, Mesiti, & Sorrenti, 2007, S. 92).

Da sich die Anforderungen an Systeme laufend ändern, um die Entwicklungen der realen Welt abzubilden, muss das XML-Schema regelmässig angepasst werden. Mittels Aktualisierungen von Elementen, sowie einfachen und komplexen Typen, kann die Effektivität und Wirksamkeit von XML-Dokumenten gewährleistet werden. Dabei kann zwischen zwei verschiedenen Typen von Modifizierungen differenziert werden. Geht es darum eine einfache Namensänderung von Elementen und Typen zu machen, dann handelt es sich um eine Umbenennungsmodifikation. Von einer strukturellen Modifizierung wird gesprochen, wenn grundlegende Strukturen wie Subelemente, Operatoren und Kardinalitäten angepasst werden. Aufgrund dieser Anpassungen sollte nach jeder Aktualisierung die XML-Dokumente wieder neu mit dem XML-Schema abgeglichen werden (Guerrini, Mesiti, & Sorrenti, 2007).

2.1.3 Open Data und Datenqualität

Als abschliessender Theorieteil sollen konkrete Studien vorgestellt werden, die den Zusammenhang zwischen Open Data und der Datenqualität direkt begründen.

2.1.3.1 Einfluss der Datenqualität auf den Informationsaustausch

Eine Studie von Nicolaous und McKnight (2006) knüpft am IS Success Modell von DeLone und McLean (2003) an und befasst sich mit den Faktoren, welche die Absicht eines interorganisationalen Datenaustauschs beeinflussen. Im Fokus liegt die

wahrgenommene Informationsqualität. Diese wird signifikant positiv durch Transparenzkontrolle beeinflusst, da der Eindruck entsteht, dass der Datenaustausch zeitgerecht, komplett und genau ist. Die wahrgenommene Informationsqualität hat gleichzeitig eine signifikant positive Wirkung auf das Vertrauen in den Datenaustausch, was wiederum die Absicht eines Datenaustauschs erhöht. Ebenso führt eine bessere Datenqualität zu einem tieferen Risikoempfinden, was ebenfalls einen positiven Einfluss auf die Absicht eines Datenaustauschs hat. Die Studie zeigt die Wechselbeziehung, die zwischen einem offenen Datenaustausch und der Datenqualität besteht und hebt den positiven Einfluss von Transparenzkontrolle hervor. Die Datenqualität wird nicht nur durch Open Data verbessert, ebenso kann eine gute Datenqualität den Informationsaustausch anregen.

2.1.3.2 „Open“ und Datenqualität

In einigen Studien zur Informationsqualität werden Attribute wie die Zugänglichkeit, die Vollständigkeit und die Aktualität der Daten als zentrale Qualitätskriterien erwähnt (Lee et al., 2002, S. 134ff.). Sie definieren damit wichtige Voraussetzungen für eine zweckmässige Datennutzung. In der Open Data Thematik gehören diese Eigenschaften zur Grundbedingung, damit Daten überhaupt als „Open“ deklariert werden (Sunlight Foundation, 2010). Daraus kann gefolgert werden, dass Daten, welche die Voraussetzungen für Open Data erfüllen, bereits ein gewisses Mass an adäquater Datenqualität aufweisen. Diese Überlegung liefert ein weiterer Grundstein für die Hypothese, dass Open Data Initiativen zu einer besseren Datenqualität beitragen und damit auch die Validität erhöhen.

2.1.3.3 Linked Open Data und Datenqualität

Aufgrund des enormen Volumens und der Komplexität von Informationen im semantischen Web wird die Verlinkung von Daten zunehmend wichtiger. Die Publikation von Zaveri et al. (2012) zeigt, wie die von Wang und Strong (1996) definierten Qualitätskriterien in einen aktuellen Kontext wie Linked Open Data übernommen werden können. Neben den bereits vorgestellten vier Kategorien, wird das Modell noch um zwei weitere Dimensionen erweitert. Eine Herausforderung für die Datenqualität von Linked Open Data stellen besonders die unterschiedlichen Quellen dar, von denen Informationen bezogen werden. Aus diesem Grund ist die

Verlinkung als solches im Kontext von Linked Open Data ein wichtiges Kriterium für die Beurteilung der Qualität.

Die Abbildung 4 gibt einen Überblick der verschiedenen Qualitätsmerkmale von verlinkten Daten. Neben den bereits erwähnten kontextuellen, intrinsischen und repräsentativen Dimension und dem Kriterium der Zugänglichkeit, wird die Qualität noch weiter in die Trust Sphäre und die Datensatzdynamik gegliedert (Zaveri et al., 2012, S. 6).

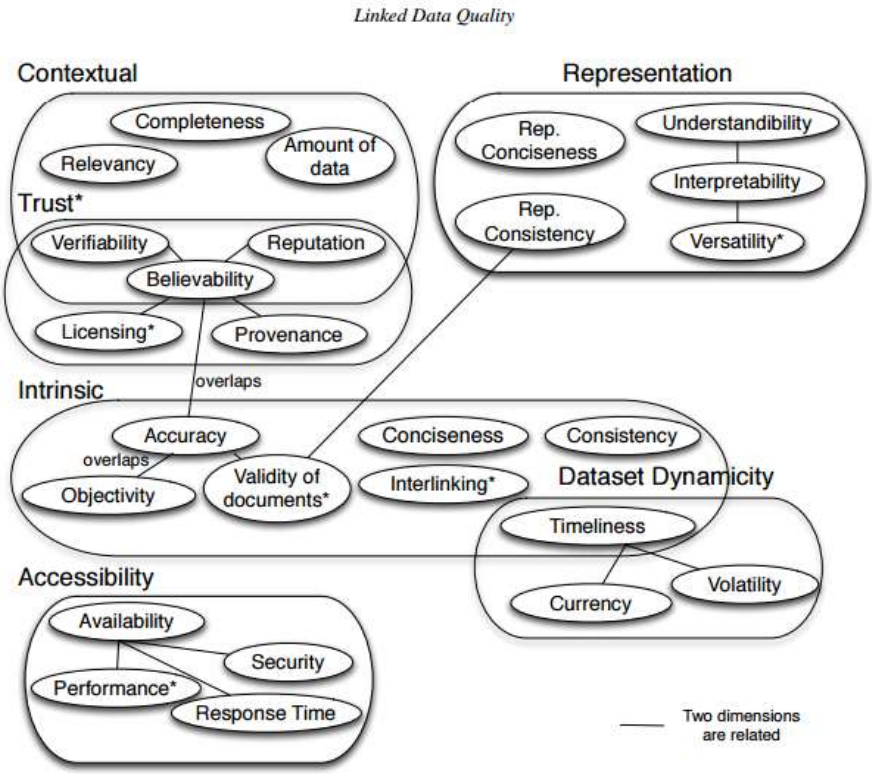

Abbildung 4: Qualität verlinkter Daten nach Zaveri et al. (2012, S. 6).

Die Trust Kategorie besteht aus Attributen, welche die Zuverlässigkeit von Daten beschreiben. Sie überschneidet sich teilweise mit der kontextuellen Dimension, da eine klare Abgrenzung nicht möglich ist. Die Sphäre der Datensatzdynamik rückt die Relevanz der regelmässigen Aktualisierung der Daten aufgrund von Volatilitäten in den Vordergrund. Umweltbedingungen sind kontinuierlichen Änderungen unterworden, welche wiederum einen Einfluss auf die Daten haben. Dies erfordert eine regelmässige Überprüfung ihrer Aktualität (Zaveri et al., 2012, S. 6ff.).

Als Teil der intrinsischen Datenqualität von Linked Data wird die Validität eines Dokuments genannt. Dabei bezieht sich die Studie auf den technischen Ansatz der

Validierung, was im Zentrum dieser Arbeit steht. In diesem Kontext geht es um die syntaktische Richtigkeit, d.h. um die Wohlgeformtheit eines Dokuments. Gleichzeitig sollte ein einheitliches Vokabular verwendet werden, um richtig zu referenzieren. Wenn diese Bedingungen nicht erfüllt sind, können die Daten nicht zweckmässig genutzt werden. Ein Validierungstool hilft dabei, fehlerhafte Notationen eines Dokuments aufzudecken (Zaveri et al., 2012, S. 13ff.). Auf diese Überlegung greift auch IATI zurück und misst die Übereinstimmung der hochgeladenen XML-Dokumente mit dem definierten IATI-Schema. Die Studie bestätigt damit auf theoretischer Ebene, dass in der Open Data Thematik die Validität eines Dokuments als Qualitätskriterium klassifiziert werden kann.

2.1.3.4 Stewardship and Usefulness

Zuletzt fasst die Theorie von Stewardship und Usefulness einige bereits angesprochene Überlegungen von Open Data und Datenqualität zusammen. Der Stewardship Ansatz und das Prinzip der Usefulness sind komplementär und liefern Argumente, wieso Open Data zu einer besseren Datenqualität, bzw. Validität führen kann (Dawes, 2010).

Das Stewardship Prinzip sagt aus, dass Daten des öffentlichen Sektors vertrauensvoll sind und geschützt werden müssen, um das Vertrauen der Bevölkerung zu bewahren. Daten sollen zwar öffentlich zugänglich sein, jedoch müssen öffentliche Organisationen die Verantwortung übernehmen einen sorgfältigen Umgang mit den Informationen zu pflegen. Daten sind Ressourcen, die einem bestimmten Verwendungszweck dienen und einen sozialen und organisationalen Wert aufweisen. Daher müssen sie vor Schaden und Verlust geschützt und „fit für den Gebrauch“ gemacht werden. Das Prinzip von Stewardship sorgt für den Erhalt von genauen, validen und sicheren Informationen. Durch die Verantwortungsübernahme von öffentlichen Organisationen gegenüber ihren Bürgern im Open Data Prozess, wird ein Anreiz geschaffen die Datenqualität zu erhöhen, um damit die Datenvalidität zu gewährleisten (Dawes, 2010, S. 380).

Der Ansatz der Usefulness als zweites Prinzip erklärt, dass durch die Offenlegung von Regierungsdaten, diese genutzt werden können und damit zu einem sozialen und ökonomischen Nutzen in Form von Innovationen führen. Dieser Grundgedanke liefert ein Argument, Daten zu teilen und Investitionen in bessere Informationssysteme zu tätigen, um wiederum bessere Daten zu erhalten (Dawes, 2010, S. 380f.).

2.2 Case Study

Die aufgezeigten Studien bestätigen auf theoretischer Ebene, dass Open Data Initiativen zu einer besseren Datenvalidität führen. Im folgenden Teil sollen einige der dargestellten Theorien in die Praxis übernommen werden. Aufgrund der Relevanz von IATI im Kontext der Entwicklungshilfe, soll die Open Data Initiative als Praxisbeispiel fungieren. Zuerst werden einige Fakten zu IATI beschrieben, um ein besseres Verständnis für die Initiative zu schaffen. Anschliessend wird die Datenqualität von IATI genauer betrachtet.

2.2.1 International Aid Transparency Initiative (IATI)

IATI steht für International Aid Transparency Initiative und ist eine freiwillige Open Data Multi-Stakeholder-Initiative, die mittels eines definierten Standards den Informationsaustausch von Entwicklungshilfe zwischen verschiedenen Interessensgruppen verbessern möchte. Das Ziel ist es, einen transparenten Umgang mit Daten zu pflegen, um die Effektivität der eingesetzten Ressourcen zu erhöhen und damit das Wohlbefinden der Bevölkerung zu verbessern. Die Verwendung von Spendengeldern und die damit verbundenen Ziele sollen offengelegt werden. Insgesamt soll eine Öffnung der Organisationen erreicht werden. Aktuell publizieren über 450 Organisationen ihre Daten über IATI (IATI, 2014, 2016a, b). IATI übernimmt dabei die technische Verantwortung, indem sie die nötige Infrastruktur für den Informationsaustausch zur Verfügung stellt (Davies, 2011, S. 2). Gleichzeitig agiert IATI als politischer Akteur, der einen offenen Standard formuliert und das Commitment der Mitglieder gegenüber von Transparenz stärkt (Davies, 2011, S. 2).

2.2.1.1 IATI als Open Data Initiative

Bevor genauer erläutert wird, was hinter IATI steckt, soll zuerst gezeigt werden, dass die Initiative die 10 Bedingungen erfüllt, die Daten als „Open“ klassifizieren (Sunlight Foundation, 2010). Damit soll bestätigt werden, dass IATI zu Recht als Open Data Initiative betitelt wird. IATI deckt die Eigenschaften folgendermassen ab:

- *Komplett:* IATI veröffentlicht die Rohdaten zusammen mit den Metadaten und erleichtert somit die Interpretation der Daten (IATI, 2016f).

- *Primär:* Das Registry verlinkt direkt zu den Rohdaten des jeweiligen Herausgebers (IATI, 2016f).
- *Zeitgerecht:* Der IATI Standard verlangt eine regelmässige Aktualisierung in monatlichen, viertel- oder mindestens halbjährlichen Abständen, um die Gültigkeit der Daten zu bestätigen (IATI, 2016a).
- *Zugänglich:* Verschiedene Verwendungsbarrieren, die vor allem von technischer Natur sind, werden durch IATI abgebaut. Der XML-Standard und eine API, helfen den Zugang zu Daten zu vereinfachen (IATI, 2016e).
- *Maschinenlesbar:* Die beteiligten Organisationen laden ihre Daten im definierten XML-Format hoch. Dieser Standard ermöglicht eine Transformation der Daten in andere Formate (IATI, 2016f).
- *Nicht diskriminierend:* Die Datensätze können ohne Registrierung vom Registry bezogen werden (IATI, 2016e).
- *Nicht proprietär:* IATI verwendet einen offenen Standard, der keine besonderen Programme verlangt (IATI, 2016a).
- *Lizenzfrei:* IATI steht unter der Creative Common Attribution License (CC-BY) (IATI, 2016m). Daten dürfen demnach für jeglichen Zweck kopiert, geändert und weiterverbreitet werden (Creative Commons, 2016).
- *Dauerhaft:* Die Daten von IATI sind permanent online verfügbar. Vom Registry können alle bisher hochgeladenen Datensets bezogen werden (IATI, 2016f).
- *Kostenlos:* IATI stellt die Daten kostenlos zur Verfügung (IATI, 2016f).

Die Erfüllung dieser Bedingungen zeigt, dass IATI als Open Data Initiative bereits ein gewisses Mass an Datenqualität mitbringt.

2.2.1.2 Ursprung von IATI

IATI wurde 2008 in Accra am dritten High Level Forum on Aid Effectiveness in Verbindung mit der Accra Agenda for Action ins Leben gerufen. Im Fokus stand das politische Commitment zu mehr Transparenz von Entwicklungshilfe und dem Wunsch nach Kooperationen zwischen den Parteien (IATI, 2016a; OECD, 2016a). Als Vorläufer gilt die Paris Declaration von 2005. Diese enthält eine Roadmap, die aus grundlegenden Erfahrungen mit der Entwicklungshilfe besteht und Umsetzungsmassnahmen für eine effektivere Entwicklungszusammenarbeit definiert. Zent-

rale Punkte sind dabei die Entwicklung von Strategien zur Verringerung von Armut und Korruption, die Verwendung von lokalen Systemen und der Austausch von Informationen, sowie die Messung von Resultaten. Die Accra Agenda for Action versucht in einem weiteren Schritt, diese Vorgaben zu vertiefen. Sie betont die notwendige Partizipation der Beteiligten, die Wichtigkeit einer guten Führung und Entwicklung von Fähigkeiten, wie die intensivere Verwendung von Systemen (OECD, 2016a). 2011 wurde in Busan am High Level Forum on Aid Effectiveness erste Vereinbarungen über einen offenen Standard für elektronische Publikationen getroffen. Es wurde die Nachfrage nach Vorgaben laut, die helfen sollen Daten zeitnahe, umfassend und vorausschauend zu sammeln (IATI, 2016a). Der Erfolg von IATI ist darauf zurückzuführen, dass sie frühzeitig ihre Chance nutzte und Abkommen mit Partnern getroffen hat, die Tools zur Datenvisualisierung liefern konnten. Dies war ein wichtiger Schachzug, um politische Unterstützung zu holen und zu zeigen, dass weitere Vorteile, wie die Nutzung der Metadaten, durch IATI möglich wären (Davies, 2011, S. 2).

2.2.1.3 Interessensgruppen

In der Entwicklungszusammenarbeit sind verschiedene Interessengruppen vertreten. Es gibt die Empfängerländer, die Steuerzahler, die öffentlichen Organisationen und die Medienschaffenden, die alle unterschiedliche Vorteile in der Veröffentlichung von Entwicklungshilfedaten sehen. Auf der einen Seite stehen die Entwicklungs.hilfeempfänger, die Informationen darüber brauchen, in welchem Umfang sie finanzielle Mittel bekommen (IATI, 2016a). Bei einigen Ländern macht die Entwicklungshilfe einen grossen Teil des Bruttoinlandprodukts aus (Ngueira-Budny, 2015). Gelder fliessen dabei oftmals von verschiedenen Stellen rein. Für eine möglichst gute Budgetierung ist es daher wichtig, aktuelle und genaue Daten zu haben. Auf der anderen Seite muss die Gesellschaft wissen, wie Steuergelder eingesetzt werden, um deren Verwendung zu kontrollieren und Regierungen zur Verantwortung ziehen zu können. Korruption ist in einigen Ländern nach wie vor ein aktuelles Thema. Organisationen, die Projekte lancieren, sind auf Informationen über Aktivitäten von anderen Institutionen angewiesen, um untereinander Synergien zu knüpfen und Doppelspurigkeiten zu vermeiden. IATI unterscheidet dabei zwischen verschiedenen Organisationstypen wie internationale, nationale und regionale NGOs, private Spender, Stiftungen, Regierungen, Forschungsstellen, multilaterale Spender oder

andere Organisationen aus dem öffentlichen Sektor und öffentliche-private Partnerschaften. Mittels eines zentralen Systems sollen den Organisationen den Zugriff auf Budgetinformationen der Empfängerländer gewährt werden und Dokumente über Konditionen und Resultate von Projekten zur Verfügung stehen. Als letztes helfen Daten den Medien, Journalisten und Forschern die Verwendung der Ressourcen und die Effektivität zu beurteilen (IATI, 2012, 2016a). Damit soll begründet werden, welche Anreize für die unterschiedlichen Parteien in der Teilnahme an IATI bestehen.

2.2.1.4 Governance Struktur

Im folgenden Abschnitt soll die Governance Struktur von IATI aufgezeigt werden. Zuoberst steht die Mitgliederversammlung, welche sich aus allen IATI-Mitgliedern zusammensetzt. Ihre Aufgabe besteht darin, die vom Verwaltungsrat erhaltenen Empfehlungen zu Budgets, strategischen Entscheidungen und Arbeitsplänen zu begutachten und abzusegnen. Aktuell sind 72 Parteien in der Mitgliederversammlung repräsentiert, wovon 27 Partnerländer sind. Als Voraussetzung für die Mitgliedschaft müssen die Organisationen und Länder das IATI Accra Statement und die Vereinbarung für den Umsetzungsrahmen unterzeichnen. Damit sprechen sie ihr Commitment zu den Zielen der Initiative aus. Organisationen können jedoch unabhängig von der Mitgliedschaft ihre Daten über IATI publizieren. Der Verwaltungsrat als nächste Instanz besteht aus sieben Vertretern der Organisationen, wobei einer dieser Sitze durch die Technical Advisory Group (TAG) besetzt ist. Der Verwaltungsrat ist dazu beauftragt, Empfehlungen gegenüber der Mitgliederversammlung zu formulieren, welche die Performance von IATI in Bezug auf die Mission, Vision und die Wertvorstellungen betreffen. Zudem wählen die Repräsentanten den Vorsitzenden und Vizevorsitzenden der Mitgliederversammlung, welche auch im Verwaltungsrat Vorsitz nehmen. Die TAG ist eine Beratungsgruppe und für den technischen Support zuständig. Sie besteht aus einer Reihe von Entwicklern, Datennutzern, Herausgebern und Anwälten. Das Sekretariat als unterste Instanz ist die administrative Unterstützung von IATI und hält das tägliche Business am Laufen (IATI, 2016b).

2.2.1.5 Finanzierung

IATI deckt die finanziellen Anforderungen durch jährliche Mitgliedergebühren und freiwillige Beiträge ab. 70% des Einkommens wird aus Mitgliederbeiträgen von

Spendenländern (Regierungen), multilateralen Institutionen, Anbietern von Süd-Süd Kooperationen und philanthropischen Stiftungen generiert. Die restlichen 30% setzen sich zusammen aus freiwilligen Beitragsleistungen und Mitgliederbeiträgen von Partnerländern und zivilgesellschaftlichen Organisationen oder anderen Instituten, wie Forschungsstellen und technischen Organisationen. Alle finanziellen Mitteln fliessen in einen Topf, worüber die Mitgliederversammlung wacht und über deren Verwendung entscheidet (IATI, 2016b).

2.2.2 Datenqualität von IATI

Das folgende Kapitel beschreibt, wie IATI die Qualität und Validität der Daten handhabt und welche Tools sie dafür verwendet. IATI gibt vor, was zu publizieren ist, welche Definitionen zu verwenden sind und bestimmt zusätzlich einen allgemeinen Rahmen für den Datenaustausch (IATI, 2016a). Dabei beeinflusst IATI mehrheitlich die technische Datenqualität durch die Zurverfügungstellung der Infrastruktur und diverser Tools. Da die Daten nicht direkt bei ihr gelagert sind, sondern nur im Registry verlinkt, übernehmen die jeweiligen Organisationen die Verantwortung für den Inhalt und die Qualität der Daten. Die Aufgabe von IATI besteht unter anderem darin Commitment für Transparenz zu schaffen und die involvierten Organisationen über den Umgang mit den Daten zu informieren (Gemäss Mailauskunft von Rory Scott, IATI Support).

2.2.2.1 Infrastruktur und Ökosystem

Wie im Unterkapitel Open Data Initiativen beschrieben, ist es für den Erfolg einer Initiative wichtig eine E-Infrastruktur aufzubauen. Diese bildet die organisationale und physikalische Struktur ab und zeigt die Ausrüstung, die für den Betrieb benötigt wird. Die Abbildung 5 stellt die einzelnen Komponenten der Infrastruktur von IATI vor und visualisiert das Zusammenspiel mit dem Ökosystem (Davies, 2011, S. 3).

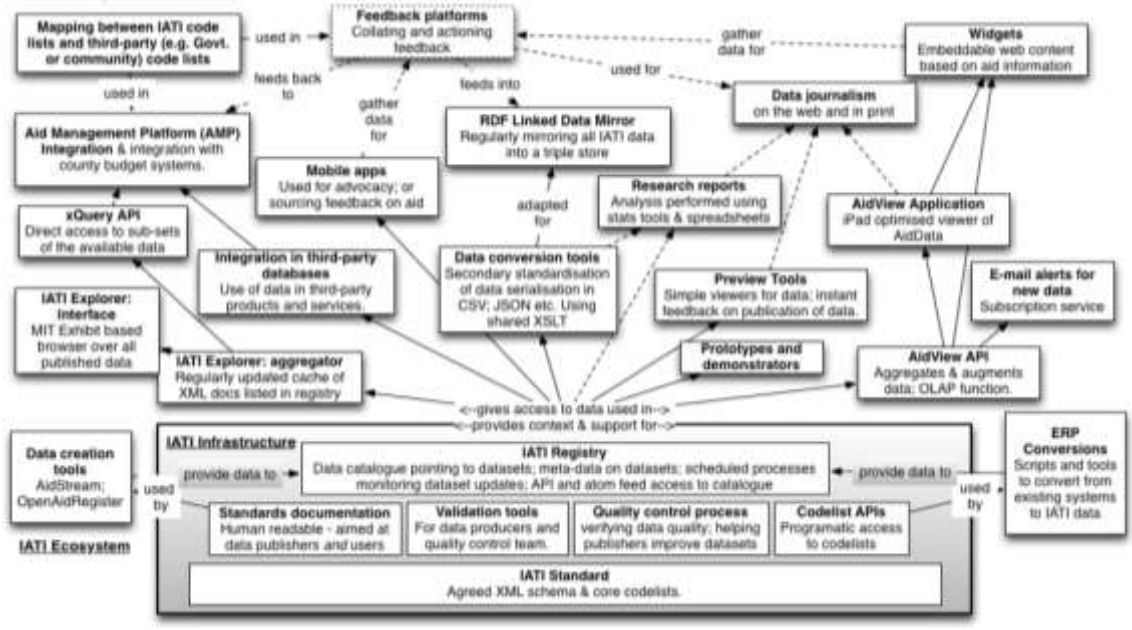

Abbildung 5: Infrastruktur und Ökosystem von IATI nach Davies (2012, S. 3).

Das Ökosystem stellt eine Übersicht dar, welche weiteren Anwendungen auf IATI zugreifen. Diese Komponenten sind direkt oder über Intermediäre mit der Infrastruktur verbunden. Lokale und globale Feedback Plattformen führen die Elemente zusammen. Wie die Theorie zeigt, sind solche Feedback Loops entscheidende Qualitätsindikatoren und helfen die Anwendungen kontinuierlich zu verbessern (Davies, 2011, S. 3ff.). Der IATI-Standard ermöglicht eine breite Palette an Tools zu nutzen. Als Beispiel könnte das d-Portal genannt werden, welches Informationen über Projekte und Finanzflüsse visuell darstellt (Bartlett, 2014). Ein anderes Hilfsmittel ist die online Plattform Aid Stream, die bei der Aufbereitung der Daten im gewünschten IATI-Standard unterstützt (YoungInnovations, 2016).

Die Übersicht zeigt, dass es sich lohnen kann, Investitionen in einen Standard zu stecken. IATI hat dadurch ein grosses Netzwerk geschaffen, wovon sie bereits in ihrer Anfangsphase profitieren konnte (Davies, 2011). Im Detail wird auf die einzelnen Komponenten im Ökosystem nicht eingegangen. Die Abbildung 5 soll jedoch das Ausmass zeigen, welches eine Initiative erreichen kann. Zudem wird ersichtlich, dass IATI unterschiedliche Anwendungen benutzt, um die Datenqualität und die Validität zu verbessern, wie auch den Publikationsprozess zu vereinfachen.

2.2.2.1.1 Registry

Vom Registry, welches ein wichtiger Teil der Infrastruktur ist, können ohne Registrierung Informationen zu Entwicklungshilfe der einzelnen Organisationen bezogen werden. Dabei handelt es sich nicht um eine Datenbank, sondern um einen Aufbewahrungsort für Links. Die Daten bleiben bei den jeweiligen Herausgebern. Das Registry verlinkt zu allen Rohdaten, die von den Organisationen gemäss dem IATI-Standard im XML-Format auf ihrer Webseite oder im eigenen Aid Management System publiziert werden. Damit wird die Thematik von Linked Open Data in diesem Kontext präsent. Betrieben wird das Registry vom Sekretariat, wobei die Herausgeber der Daten für deren Inhalt verantwortlich sind (IATI, 2016a). Das Registry stellt Datensets und Files zur Verfügung und liefert Informationen zu den Herausgebern. Die Herausgeber sind Organisationen, die ihre Daten nach dem IATI-Standard publizieren. Damit sie ihre Daten über IATI veröffentlichen können, müssen sie sich jedoch zuerst registrieren. Die Datensets werden durch die Verlinkung der Rohdaten gebildet. Eine API greift dabei auf die Metadaten des Registry zurück (IATI, 2016f). Diese Metadaten sind wichtige Bestandteile für die Identifizierung von Verwandtschaften zwischen den Daten, weshalb sie ebenfalls veröffentlicht werden (Davies, 2011, S. 3). Die Files wiederum entsprechen entweder dem Aktivitäten- oder Organisation-Standard (IATI, 2016f).

2.2.2.1.2 Datastore

Neben dem Registry verfügt IATI noch über einen weiteren online Service, von dem Daten bezogen werden können. Der Datastore speist die Daten direkt aus dem Registry ein. Änderungen im Registry werden innerhalb von 24 Stunden in den Datastore geladen. Von dort können die Daten dann in JavaScript Object Notation (JSON), in XML oder im CSV-Format bezogen werden. Der Datastore Query Builder hilft insbesondere bei der Generierung von CSV-Dateien. Im Vergleich zum Registry können über den Datastore komplexere Abfragemöglichkeiten ausgeführt werden. Ein Beispiel dazu sind Filterungen nach Sektoren und Zeitperioden (IATI, 2016i).

2.2.2.2 IATI-Standard

IATI hat einen Standard entwickelt, der Voraussetzungen definiert, wie Daten zu publizieren sind. Dabei geht es um die Veröffentlichung von zeitgerechten, umfas-

senden, vorausschauenden, offenen, strukturierten und vergleichbaren Daten (IATI, 2016q). Damit die Daten dann auch optimal genutzt werden können, sollte das XML-Format verwendet werden. Dieses ermöglicht die Transformation in andere Formate, wie zum Beispiel in ein CSV-Dokument. Daten können dadurch für eine grössere Anzahl von Applikationen gebraucht werden. Zudem können sie von unterschiedlichen Quellen miteinander kombiniert werden (IATI, 2012, 2016a). Zur Unterstützung des Publikationsprozesses stellt IATI ein Tool zur Verfügung, welches CSV-Dateien in das verlangte XML-Format umwandelt (IATI, 2016d).
Der IATI-Standard wird in den Organisation- und den Aktivitäten-Standard unterteilt (IATI, 2015b). Die Abbildung 6 zeigt einen Auszug aus dem Organisation-Standard. Dieser definiert diejenigen Informationen, die von allen Organisationen offenzulegen sind. Neben allgemeinen Angaben wie dem Organisationsnamen, sollen Informationen zu jährlichen Reportings und Länderplänen, aber auch zu aktuellen und vorausschauenden Budgets für jedes Land oder Region erfasst werden (IATI, 2016a).

```
<!--total-budget starts-->
<total-budget>
 <period-start iso-date="2014-01-01" />
 <period-end iso-date="2014-12-31" />
 <value currency="USD" value-date="2014-01-01">250000000</value>
 <budget-line ref="1234">
      <value currency="USD" value-date="2014-01-01">200000000</value>
      <narrative>Budget Line</narrative>
 </budget-line>
</total-budget>
<!--total-budget ends-->
<!--recipient-org-budget starts-->
<recipient-org-budget>
 <recipient-org ref="AA-ABC-1234567">
  <narrative>Org Name</narrative>
 </recipient-org>
 <period-start iso-date="2014-01-01" />
 <period-end iso-date="2014-12-31" />
 <value currency="USD" value-date="2014-01-01">2500000</value>
 <budget-line ref="1234">
      <value currency="USD" value-date="2014-01-01">2000000</value>
      <narrative>Budget Line</narrative>
 </budget-line>
</recipient-org-budget>
<!--recipient-org-budget ends-->
```

Abbildung 6: Organisation-Standard nach IATI (2015b).

Einen Auszug aus dem Aktivitäten-Standard ist in der Abbildung 7 dargestellt. Im Zentrum steht der Austausch von Informationen zu laufenden Projekten und Aktivi-

täten. Darin sind Beschreibungen zu vorausschauenden Budgets von einzelnen Aktivitäten, betroffenen Sektoren und Klassifikationen, sowie Konditionen und Resultate von Projekten enthalten. Ebenso werden Transaktionshistorien über Ausgaben, Aufwände und eingehende Mittel abgebildet und Informationen über vor Ort herrschende sub-nationale geografische Codes geliefert (IATI, 2016a).

```
<!--title starts-->
<title>
 <narrative>Activity title</narrative>
 <narrative xml:lang="fr">Titre de l'activité</narrative>
 <narrative xml:lang="es">Título de la actividad</narrative>
</title>
<!--title ends-->
<!--description starts-->
<description type="1">
 <narrative>General activity description text.  Long description of the
activity with no particular structure.</narrative>
 <narrative xml:lang="fr">Activité générale du texte de description.
Longue description de l'activité sans structure particulière.</narrative>
</description>
<description type="2">
 <narrative>Objectives for the activity, for example from a logical
framework.</narrative>
 <narrative xml:lang="fr">Objectifs de l'activité, par exemple à partir
d'un cadre logique.</narrative>
</description>
```

Abbildung 7: Aktivitäten-Standard nach IATI (2015b).

Neben der Einhaltung von formalen und inhaltlichen Vorschriften, sollen die Daten zudem in regelmässigen Abständen aktualisiert werden, um ihre Gültigkeit zu bestätigen. Die von IATI vorgeschriebene Zeitsequenz sollte sich zwischen monatlicher, viertel- oder mindestens halbjährlicher Aktualisierung bewegen. Zusätzlich legt IATI Wert darauf, bereits bestehende Vereinbarungen in den Standard zu integrieren, damit diese weiterhin verwendet werden können. Ein Beispiel dazu ist das Creditor Reporting System (CRS). Es ist eine Datenbank der Organisation for Economic Cooperation and Development (OECD) des Development Assistance Commitee (DAC), das jährliche Statistiken zu Entwicklungshilfe von international führenden Organisationen wie der Weltbank, der Vereinten Nationen (UN) oder der Europäischen Union (EU) erfasst (IATI, 2012, 2016a).

2.2.2.3 Datenvalidität

IATI bietet nicht nur eine Plattform an, von der Daten bezogen werden können, ebenso ist sie bemüht, Tools für die Qualitätsmessung in die Infrastruktur zu integrieren. Für die Überprüfung der Validität stellt IATI ein öffentliches Validierungstool zur Verfügung, welches verifiziert, ob das XML-Dokument den von IATI vorgegebenen XML-Schemas entspricht. Dazu kann entweder eine Datei hochgeladen, ein Uniform Resource Locator (URL) einer Webseite ergänzt oder direkt ein ganzer XML-Code eingefügt werden. Je nach Stand der Aktualisierung können verschiedene Schemas ausgewählt werden (IATI, 2016g).

IATI ist bemüht die interne Validität zu erfassen. Gemäss IATI erfüllen die Daten dafür drei Kriterien. Sie müssen zugänglich sein, indem sie direkt vom Registry bezogen werden können. Des Weiteren müssen die Dokumente die allgemein gültige XML-Syntax benutzen und gleichzeitig mit dem von IATI definierten XML-Schema übereinstimmen. Mittels dieser Art von Validität kann IATI inhaltliche Aspekte, wie konkurrierende Sektoren innerhalb von Organisationen auf dem relevanten Aktivitätenlevel kontrollieren. Eine externe Validität kann durch IATI jedoch nicht gegeben werden (Gemäss Mailauskunft von Rory Scott, IATI Support).

Da sich Umweltbedingungen laufend verändern und sich die Systeme an diesen Änderungen orientieren, ist es wichtig die technischen Bedingungen in regelmässigen Frequenzen anzupassen (Guerrini, Mesiti, & Sorrenti, 2007). Seit der Einführung und der Version 01.01 wurde der IATI Standard bereits mehrfach aktualisiert. Die Version 01.02 erschien im Dezember 2012 und Version 01.03 bereits im Frühling 2013. Noch im selben Jahr wurde im Winter die Version 01.04 eingeführt. Das nächste Update 01.05 erschien im Oktober 2014, welches anfangs Januar 2015 jedoch bereits wieder durch eine Neuversion 02.01 abgelöst wurde. Seit Dezember 2015 befindet sich der Standard in der Version 02.02. Bei jedem Update werden unterschiedliche Änderungen gemacht. Oftmals werden Elemente, besonders Kinderelemente, oder auch Attribute in den Schemas ergänzt oder abgeändert, oder die Möglichkeit von Freitext aus Elementen entfernt oder angefügt. Zusätzlich werden Codelisten oder Dokumentationen angepasst (IATI, 2016j). In den für die empirische Analyse verwendeten Datensätzen konnten keine Übereinstimmungen der Wertschwankungen des Anteils valider Dokumente mit der Terminierung der Aktua-

lisierungen festgestellt werden. Aus diesem Grund wurde dieser Aspekt für die statistische Auswertung nicht genauer betrachtet.

Für die empirische Analyse wird die Anzahl nicht valider Dokumente verwendet, welche von IATI erfasst wird. Ein File wird dann als invalide angezeigt, wenn das XML-Dokument nicht mit dem definierten XML-Schema übereinstimmt. Bei der Validierung treten dabei regelmässig dieselben Fehler auf. Ein Problem vor allem der Version 02.01 ist, dass Elemente im XML-Dokument oftmals falsch platziert werden und dadurch die definierte Reihenfolge des Schemas nicht einhalten. Des Weiteren müssen die URLs einer bestimmten maschinenlesbaren Codierung entsprechen, damit sie von den Systemen interpretiert werden können. Demnach sind auch Zeit- und Datumsvorgaben relevant. Das Datumsformat wird als YYYY-MM-DD definiert und muss durch ein T von der Zeitangabe HH:MM:SS abgetrennt werden. Die Nichteinhaltung solcher Vorgaben ist dafür verantwortlich, dass ein XML-Dokument nicht als valide gilt (IATI, 2016n).

2.2.2.4 Umgang der Organisationen mit der Datenqualität

Nachdem einige technische Spezifikationen der Datenqualität im oberen Teil erläutert wurden, soll das folgende Unterkapitel anhand eines Beispiels den Umgang von Organisationen mit der Datenqualität und dem Publikationsprozess von IATI aufzuzeigen.

2.2.2.4.1 DEZA

Als Beispiel für eine Institution, die ihre Daten über IATI publiziert, kann die Direktion für Entwicklung und Zusammenarbeit (DEZA) genannt werden. Sie zeigt welche Massnahmen ergriffen werden, um Transparenz in der Schweiz zu fördern. Als Teil des Eidgenössischen Departements für auswertige Angelegenheiten (EDA) ist die DEZA für die humanitäre Hilfe, sowie Entwicklungs- und Ostzusammenarbeit des Bundes zuständig. Die DEZA verlangt einen transparenten Austausch von Informationen über Projekte und Finanzplanungen zwischen den betroffenen Akteuren. Über ihre eigene Projektdatenbank können alle nötigen Informationen zu den Projekten bezogen werden. Dabei werden die Ergebnisse mittels Evaluationen, Studien und Jahresberichte überprüft, offengelegt und kontinuierlich verbessert. Bisher gemachte Erfahrungen sollen in den Lernprozess integriert werden, da institu-

tionelles Lernen Teil der Kultur des DEZAs ist. An regelmässig stattfindenden öffentlichen Veranstaltungen wird über den Status Quo und weitere Entwicklungen informiert, um die Bevölkerung in den Entscheidungsprozess zu integrieren (Eidgenössisches Departement für auswertige Angelegenheiten [EDA], 2016a, b). Die DEZA betont damit, die im Vorfeld erwähnte Wichtigkeit von Feedbackmechanismen, die Teil des Lernprozesses sind und eine Verbesserung der Datenqualität über die Zeit begründen.

Die DEZA ist seit November 2013 Mitglied von IATI und dafür zuständig ihre Daten im IATI-Standard zu publizieren. Gleichzeitig muss sie die Indikatoren des Aid Transparency Index für die Schweiz erfüllen (EDA, 2016a; Publish What You Fund, 2016d). Bei der Veröffentlichung der Daten ist die DEZA keinen Richtlinien des Bundes unterworfen, sondern orientiert sich am jeweiligen Standard des Reportingsystems (Gemäss Mailauskunft von Flavien Breitenmoser, Mitarbeiter der DEZA).

2.2.2.4.2 Herausforderungen für Organisationen

Der Publikationsprozess von IATI ist mit einem gewissen administrativen Aufwand verbunden. Plattformen wie Aid Stream können jedoch bei der Veröffentlichung der Daten unterstützen (YoungInnovations, 2016). Für einen längerfristigen Erfolg wird grundsätzlich empfohlen IATI als festen Bestandteil in die Organisation zu integrieren, da für die Offenlegung der Daten unterschiedliche Ressourcen in Anspruch genommen werden, die verfügbar sein sollten. Zusätzlich zum eigentlichen Publikationsprozess müssen im Vorfeld einige Vorbereitungen getroffen werden. Oftmals arbeiten die Organisationen mit Externen zusammen. Damit ihre Daten über IATI publiziert werden können, ist es wichtig, die Partner darüber zu informiert und ihre Zustimmung einzuholen. Des Weiteren sollte sich jede Organisation bewusst sein, dass der Einsatz von finanziellen Mitteln besonders genau geprüft wird und die Aufbereitung dieser Daten eine gewisse Herausforderung darstellt. Zusätzlich sollte das Sicherheitspotential von sensiblen Informationen von spezifischen Sektoren und Orten beurteilt werden, um sicherzustellen, dass eine Publikation dieser Daten nicht nachteilig wäre (IATI, 2014).

2.2.2.5 Fazit von IATI

Im oberen Abschnitt zu Open Data Initiativen wird gezeigt, dass die Vorteile von Open Data unter bestimmten Bedingungen auftreten. IATI erfüllt viele dieser

Voraussetzungen, weshalb sie als gutes Beispiel für eine erfolgreiche Open Data Initiative dient. Sie nutzt verschiedene Koordinationsmechanismen, wie Standardisierungen, offene Formate und nicht restriktive Lizenzen. IATI hat eine umfassende E-Infrastruktur aufgebaut, die den Nutzer ermöglicht von vielen Tools zu profitieren. Das Registry stellt Daten zur Verfügung und hilft dadurch den Informationsaustausch zu erleichtern. Eine API greift dabei auf die Metadaten zu. Mittels Statistiken betreibt IATI Monitoring, um Feedbacks für eine stetig Weiterentwicklung zu erhalten. Zusätzlich versucht IATI, wie das in der Literatur vorgeschlagen wird, den Usern ein digitales Umfeld zu schaffen, in dem sie gerne agieren. Dazu verwendet sie Videos, Tweets und informiert über aktuelle Events. Das Logo von IATI ist zudem eindeutig erkennbar und vermittelt das Gefühl einer Gemeinschaft. Zuletzt sorgt die Governance Struktur von IATI für die Mitbeteiligung der Mitglieder (IATI, 2016a-n).

2.3 Empirische Datenanalyse

Die Theorie zeigt, dass Open Data Initiativen zu einer höheren Datenvalidität führen. Bei verlinkten offenen Daten, zu welchen die Daten von IATI gehören, geht es spezifisch um die Validität von Dokumenten. Diese müssen syntaktisch richtig sein und ein einheitliches Vokabular verwenden (Zaveri et al., 2012, S. 13ff.). Die folgende empirische Analyse soll daher belegen, dass sich die Validität der XML-Dokumente, d.h. die Übereinstimmung eines XML-Dokuments mit einem XML-Schema, seit der Einführung von IATI verbessert hat. Um ein Mass für die Validität zu bekommen, wird der Anteil valider Dokumente überprüft. Es werden unterschiedliche Datensätze hinzugezogen, um verschiedene Variablen zu extrahieren, die einen Einfluss auf die validen Dokumente haben könnten. Zum Schluss soll das Modell gefunden werden, welches die Varianz der abhängigen Variable, d.h. des Anteils valider Dokumente, in diesem Kontext am besten erklärt.

Es werden die folgenden Hypothesen untersucht:

- *Hypothese 1: Die Zunahme des Alters von IATI führt zu einer höheren Validität von Dokumenten*
- *Hypothese 2: Die Zunahme der Anzahl hochgeladener Dokumente führt zu einer höheren Validität von Dokumenten*
- *Hypothese 3: Die Zunahme der Anzahl teilnehmender Organisationen führt zu einer höheren Validität von Dokumenten*
- *Hypothese 4: Eine längere Teilnahme an IATI führt zu einer höheren Validität von Dokumenten*

2.3.1 Methodisches Vorgehen

Die Analyse ist in drei Teile gegliedert. Zuerst wird die übergreifende Ebene betrachtet und die Auswertung über alle Organisationen und sämtliche Dokumente gemacht. Damit sollen die Hypothesen 1 bis 3 überprüft werden, um anschliessend Rückschlüsse auf das beste Modell zuzulassen. Im zweiten Teil werden die einzelnen Organisationen untersucht und Sonderfälle evaluiert. Zusätzlich soll die Hypothese 4 getestet werden. Im letzten Schritt wird eine Auswertung der statistischen Signifi-

kanz der Sonderfälle gemacht, um die zuvor erhaltenen Ergebnisse zu kontrollieren und Annahmen über Abweichungen zu treffen.

Im Laufe der Zeit sind zunehmend mehr Organisationen IATI beigetreten. Das hat zur Folge, dass die Anzahl hochgeladener Dokumente ebenfalls gestiegen ist. Um jedoch eine Aussage darüber zu machen, wie sich die Validität der Dokumente aufgrund diverser Einflüsse entwickelt hat, ist es sinnvoll die Anzahl valider Dokumente in Relation mit der Gesamtanzahl hochgeladener Dokumente zu setzen. Die abhängige Variable wird daher definiert als der Anteil valider Dokumente. Wären alle Dokumente valide, würde ein Wert von 1 erreicht werde. Der Anteil valider Dokumente läge damit bei 100%.

Um die Theorie der steigenden Validität weiter zu diskutieren, soll im ersten Teil zusätzlich geprüft werden, wie sich der Anteil Organisationen, die nur valide Dokumente haben, verändert hat. Diese Variable berechnet sich als Anzahl Organisationen mit nur validen Dokumenten im Vergleich zur Gesamtanzahl Organisationen. Würden alle Organisationen nur valide Files hochladen, wäre der Anteil Organisationen mit nur validen Dokumenten 100%. Das Hinzuziehen dieser Variable soll einen kritischen Blick auf die Resultate zulassen. Des Weiteren werden alle Modelle mittels Breusch-Pagan-Test auf Heteroskedastizität untersucht. Um diesen unerwünschten Effekt zu neutralisieren wird mit robusten Standardfehlern gerechnet.

Zur Vereinfachung der Schreibweise wird in diesem Kontext die Variable Anteil valider Dokumente mit dem Namen „Validität_Files“ betitelt. Gleichzeitig wird die Variable Anteil Organisationen mit nur validen Dokumenten als „Validität_Organisation“ bezeichnet.

Als unabhängige Variable werden die Einflussfaktoren Alter von IATI, Anzahl Dokumente, Anzahl teilnehmender Organisationen und Länge der Teilnahme an der Open Data Initiative definiert. Die Modelle werden alle auf dem Signifikanzniveau von 1% überprüft.

Für die Auswertung werden die Daten vom IATI Dashboard bezogen und im Statistikprogramm R empirisch analysiert. IATI stellt die Datensätze im JSON-Format zur Verfügung. Diese lassen sich mittels Konverter in ein CSV-Dokument umwandeln.

2.3.2 Organisationsübergreifende Analyse: Modelle 1 bis 9

Im folgenden Abschnitt sollen 8 Modelle aufgestellt werden, welche die Validität_Files und die Validität_Organisation mittels drei unabhängigen Variablen Alter IATI, Anzahl Dokumente und Anzahl teilnehmender Organisationen prüfen. Die Analyse wird organisationsübergreifend gemacht, d.h. unter Einbezug der Gesamtanzahl Dokumente und aller Organisationen. Zusätzlich soll im Modell 9 der Zusammenhang zwischen der unabhängigen Variable Zeit und der abhängigen Variable Anzahl Organisationen empirisch gemessen werden.

2.3.2.1 Beschreibung und Bereinigung der Datensätze

Die organisationsübergreifende Analyse verwendet den Datensatz, der die Anzahl invalider Files aller Organisationen pro Messzeitpunkt enthält. Die erste Messung wurde am 19.09.2013 vorgenommen. Für die Analyse wurde der Datensatz am 15.05.2016 bezogen. Alle späteren Messungen werden in dieser Arbeit nicht mehr berücksichtigt. Die Messungen wurden meist täglich, in Ausnahmefällen mit einem Zeitintervall von maximal 39 Tagen getätigt. Der Datensatz enthält 756 Beobachtungen.

Damit das CSV für die Analyse verwendet werden kann, muss es zuerst bereinigt und umgeformt werden. Im Folgenden werden die zentralen Schritte der Bereinigung und Formatierung aufgezeigt, welche auch für weitere Datensätze verwendet werden. Zuerst werden die Spalten und Zeilen für eine bessere Übersichtlichkeit getauscht und als Data Frame abgespeichert. Da für die Analyse nur die Tage, Monate und Jahre relevant sind, können die Zeitangaben aus der Datum-Variable entfernt werden. Danach wird das Datum in das Format YYYY-MM-DD gebracht. In einem nächsten Schritt wird überprüft, ob sich Duplikationen im Datensatz befinden. Da jedes Datum einmal für die Anzahl valide und einmal für die invaliden Files vorkommt, muss nach Datumsangaben gesucht werden, die mindestens drei Mal erwähnt werden. Heraus kommt, dass am 25.09.2013 drei, am 29.11.2013 und 20.12.2013 je zwei Messungen vorgenommen wurden. Für die weitere Analyse

reicht eine Messung der Anzahl validen und eine der nicht validen Dokumente pro Tag aus. Aus diesem Grund können alle Werte dieser drei Tage, ausser der letzten Tagesmessung, eliminiert werden.

Da die Beobachtungen der Anzahl invalider und der validen Files nicht in je einer separaten Spalte erfasst sind, müssen diese in eine andere Anordnung gebracht werden. Das Ziel ist je eine Variable mit der Anzahl aller validen und eine mit allen nicht validen Dokumenten zu generieren. Um dies zu erreichen, wird eine zusätzliche Spalte kreiert, die aus der Differenz des Messzeitpunkts besteht. Diese Werte werden dann in zwei Kategorien angeordnet, d.h. gleich oder grösser als 0, und in je eine separate Variable abgespeichert. Der Wert gleich 0 qualifiziert alle nicht validen Files und alle Werte grösser als 0 die validen Dokumente.

Diese zwei Variablen können dann in Files_pass und Files_fail unbenannt werden. Durch das Summieren dieser Variablen kann nun die Gesamtzahl aller hochgeladener Dokumente berechnet werden. Zum Schluss wird die Variable Validität_Files, definiert als der Anteil valider Dokumente an der Gesamtzahl, dem Datensatz hinzugefügt.

Um zusätzliche Variablen für das beste Modell zu finden, wird ein weiterer Datensatz vom IATI Dashboard hinzugezogen. Dieser zeigt die Anzahl Organisationen, die mindestens ein invalides Dokument zu einem bestimmten Zeitpunkt hochgeladen haben. Der Zeitrahmen und die Messpunkte, mit einer Ausnahme, sind identisch mit dem vorher beschriebenen Datensatz. Nach der Elimination des zusätzlichen Messwerts besteht dieser Datensatz ebenfalls aus 756 Beobachtungen.

Für die Bereinigung wurde analog wie beim vorherigen Datensatz vorgegangen. Eine Ausnahme besteht darin, dass in diesem Datensatz eine Messung am 06.09.2015 gemacht wurde, die im vorherigen Datensatz nicht vorhanden ist. Die Abweichung wurde mittels Zeitdifferenzmessung in Excel ermittelt. Da die Datensätze zusammengenommen werden, müssen die Messzeitpunkte übereinstimmen. Der zusätzliche Messwert wurde aus diesem Grund aus dem Datensatz entfernt.

Um die Variable Validität_Organisation zu berechnen, muss die Gesamtanzahl Organisationen für jeden einzelnen Messpunkt kalkuliert werden. Dies wird gemacht, indem die Anzahl Organisationen mit nur validen und diejenigen mit mindestens einem invaliden Dokument zusammengezählt werden. Wird anschliessend die Anzahl Organisationen mit nur validen Dokumenten durch die Anzahl aller Organisationen geteilt, ergibt sich daraus die Variable Validität_Organisation.

Die beiden beschriebenen Datensätzen können nun durch die gemeinsame Variable „Date“ miteinander verknüpft werden. Im Vorfeld müssen sie dazu als separate CSV-Dokumente „Validation_Publisher“ für die Berechnung der Validität nach den Organisationen und als „Validation_Files“ für die Variante mit der Anzahl Files abgespeichert werden.

Der zusammengefasste Datensatz kann nun für die Kalkulation der Modelle 1 bis 9 genutzt werden. Der zugehörige R-Code „Organisationsübergreifende Analyse“ kann dem Anhang entnommen werden.

2.3.2.2 Modell 1 & 4: Validität und Alter IATI

Um die Entwicklung der Validität der Dokumente in Abhängigkeit mit dem Alter der Open Data Initiative zu vergleichen wird eine Age-Variable generiert. Diese berechnet sich aus dem Einführungstermin von IATI, welcher auf den 01.12.2011 datiert wird, und der Differenz zum Messzeitpunkt. Für die Umrechnung in Jahre wird der Wert noch durch 365 Tage geteilt (Schaltjahre werden ausgeblendet).

Für den Vergleich der Variablen Validität_Files und der Validität_Organisation werden 2 Modelle definiert. Das Modell 1 betrachtet den Einfluss der x-Variable Alter IATI auf die y-Variable Validität_Files. Das Modell 4 zeigt die Beziehung zwischen der x-Variable Alter IATI und der y-Variable Validität_Organisation. Damit soll die Hypothese 1 überprüft werden.

2.3.2.2.1 Statistische Ergebnisse Modell 1 & 4

Die Abbildung 8 stellt visuell den Zusammenhang zwischen dem Modell 1 und 4 dar. Der Tabelle 3 können die dazugehörigen statistischen Resultate entnommen werden.

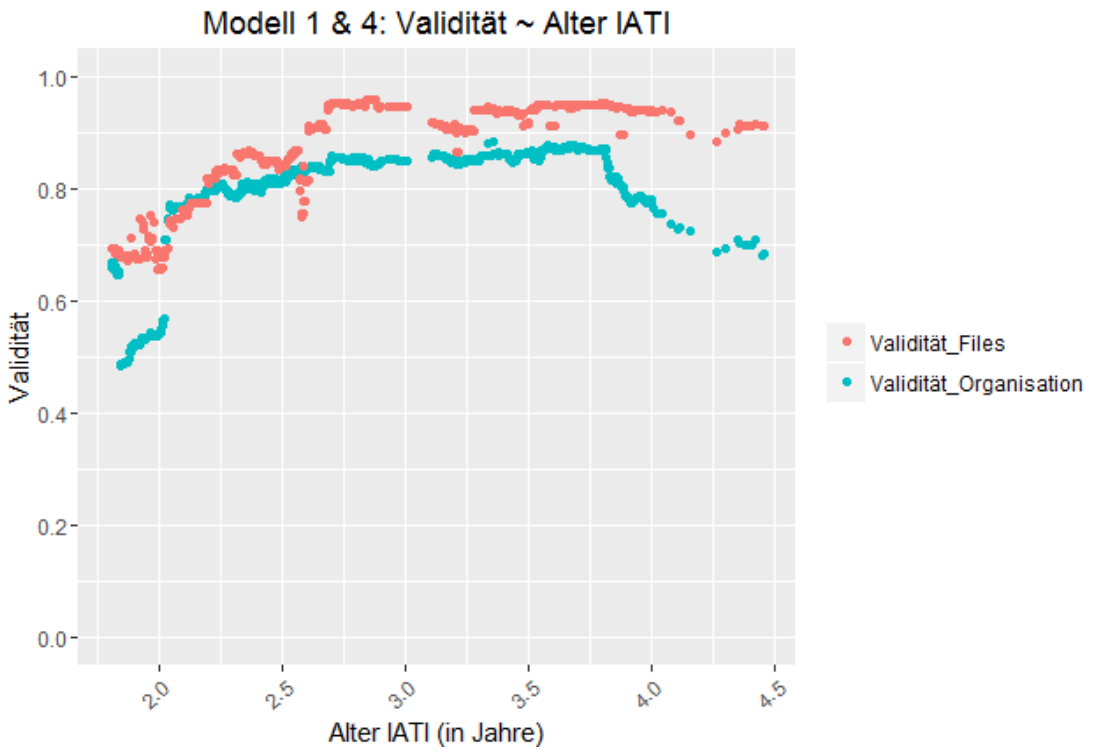

Abbildung 8: Validität und Alter IATI - Modelle 1 & 4

Modell 1

	Validity_Files	
	OLS SE	Robust SE
Constant	0.578***	0.578***
	(0.008)	(0.010)
Age	0.102***	0.102***
	(0.003)	(0.003)
Observations	756	756
Adjusted R2	0.641	0.641
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel1
BP = 47.283, df = 1, p-value = 6.145e-12

Modell 4

	Validity_Files	Validity_Publisher
	OLS SE	Robust SE
Constant	0.578***	0.586***
	(0.013)	(0.019)
Age	0.102***	0.073***
	(0.004)	(0.006)
Observations	756	756
Adjusted R2	0.641	0.284
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel4
BP = 47.352, df = 1, p-value = 5.932e-12

Tabelle 3: Resultate Modelle 1 & 4

Der Breusch-Pagan-Test bestätigt die Heteroskedastizität beider Modelle. Der p-Wert liegt unter 0.01, weshalb die Nullhypothese verworfen wird, dass die Residuen homoskedastisch sind. Aus diesem Grund werden für die Analyse robuste Standardfehler verwendet. Die Tabelle 3 zeigt die jeweilige Signifikanz der Werte bei der Verwendung der kleinsten Quadrate (OLS) und robusten Standardfehlern. Die Auswertung ergibt, dass die Variablen Validität_Files und Validi-

tät_Organisation mit dem Alter von IATI zunehmen. Es besteht in beiden Modellen ein positiver empirischer Zusammenhang zwischen der abhängigen und der unabhängigen Variable auf dem 1%-Signifikanzniveau. Das Modell 1 hat ein adjustiertes R^2 von 0.641. Die Variable Alter IATI erklärt damit 64.13% der Varianz der Variable Validität_Files. Das Modell 4 erreicht im Vergleich zum Modell 1 einen tieferen Wert. Das Alter von IATI erklärt die Varianz der Variable Validität_Organisation zu 28.4%, d.h. 71.6% sind auf andere Einflussfaktoren zurückzuführen.

2.3.2.2.2 Zwischendiskussion Modell 1 & 4

Die beiden Modelle bestätigen die Hypothese 1, dass die Zunahme des Alters von IATI zu einer höheren Validität von Dokumenten führt. Der positive Zusammenhang könnte dadurch erklärt werden, dass sich über die Jahre ein Lernprozess einstellt. Durch die Publikation der Daten über eine Plattform und mittels integrierter Feedbackmechanismen, bekommen die User die Möglichkeit sich untereinander austauschen und Anregungen für weitere Verbesserungen abzugeben (Janssen, Charalabidis & Zuiderwijk, 2012; Juran, 1998). Dieser Feedback Loop hilft die Validität über die Zeit zu erhöhen. Ein weiterer Grund könnte das zunehmende Bewusstsein für die Wichtigkeit von transparenten und guten Daten sein. Das steigende Commitment gegenüber der Transparenz, welches sich in den letzten Jahren entwickelt hat, könnte die Motivation zur Bereitstellung von validen Daten positiv beeinflusst haben. Zusätzlich könnte die steigende Relevanz von IATI als Open Data Initiative den Druck auf die Zurverfügungstellung von validen Daten erhöht haben. Ab einem bestimmten Alter von IATI wird der Einfluss eines weiteren Jahres jedoch limitiert und die Validität der Dokumente kann nicht mehr weiter erhöht werden.

2.3.2.3 Modell 2 & 5: Validität und Anzahl Files

Die folgenden zwei Modelle untersuchen den Einfluss der Anzahl Files auf die Validität von Dokumenten. Die Gesamtanzahl Files pro Messwert wurde bereits für die Variable Validität_Files berechnet.

Das Modell 2 definiert Validität_Files als y-Variable und die Anzahl Dokumente als x-Komponente. Das Modell 5 verwendet als unabhängige Variable ebenfalls die Anzahl Files und betrachtet deren Einfluss auf die abhängige Variable Validität_Organisation.

2.3.2.3.1 Statistische Ergebnisse Modell 2 & 5

Die Abbildung 9 visualisiert die Modelle 2 und 5, um den Zusammenhang zwischen den Variablen zu zeigen. In der Tabelle 4 werden die Resultate der beiden Modelle einander gegenübergestellt.

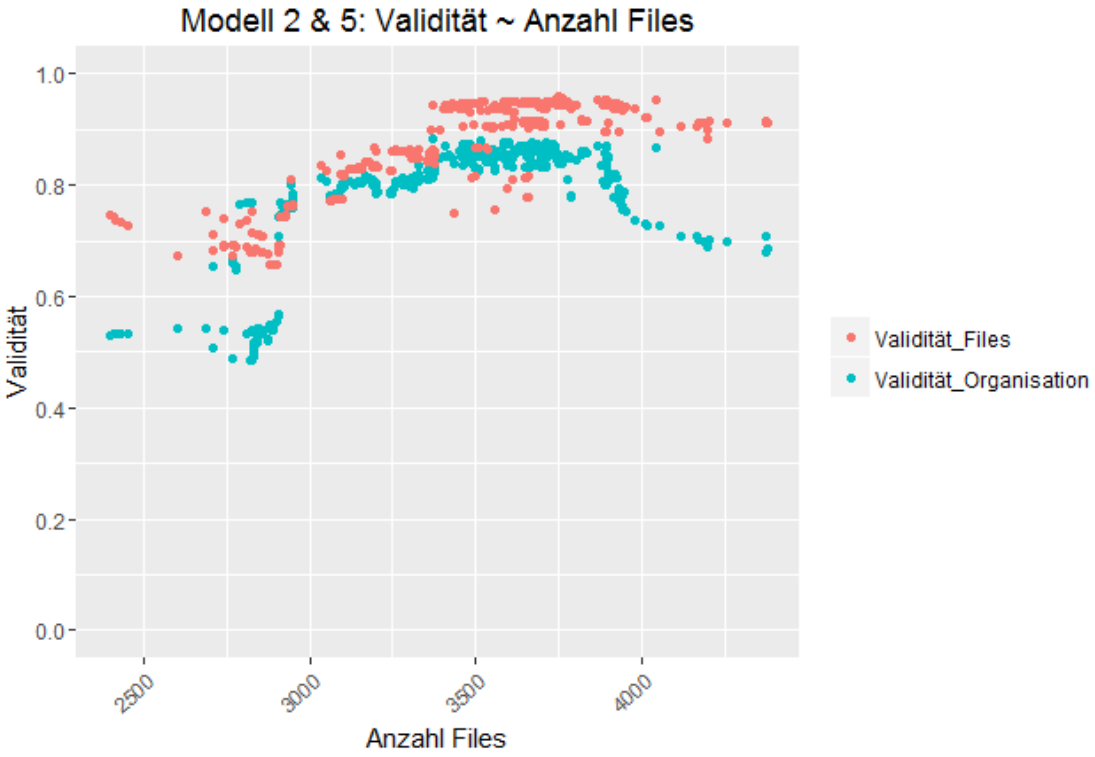

Abbildung 9: Validität und Anzahl Files - Modelle 2 & 5.

Modell 2

	Validity_Files	
	OLS SE	Robust SE
Constant	0.139***	0.139***
	(0.014)	(0.020)
Sum_Files	0.0002***	0.0002***
	(0.00000)	(0.00001)
Observations	756	756
Adjusted R2	0.781	0.781
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel2
BP = 0.68117, df = 1, p-value = 0.4092

Modell 5

	Validity_Publisher	
	OLS SE	Robust SE
Constant	0.218***	0.218***
	(0.025)	(0.041)
Sum_Files	0.0002***	0.0002***
	(0.00001)	(0.00001)
Observations	756	756
Adjusted R2	0.423	0.423
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel5
BP = 26.323, df = 1, p-value = 2.889e-07

Tabelle 4: Resultate Modelle 2 & 5

Die Abbildung 9 zeigt in beiden Modellen einen positiven Zusammenhang zwischen der abhängigen Variable und der unabhängigen Variable Anzahl Files. Der Tabelle 4 ist zu entnehmen, dass das Resultat signifikant auf dem Niveau von 1% ist. Das Modell 2 weist keine Heteroskedastizität auf, da der p-Wert grösser als 0.01 ist. Deshalb können für die Berechnung die OLS-Standardfehler benutzt werden. Hingegen ist das Modell 5 heteroskedastisch. Die Anzahl Files erklärt die Varianz der Variable Validität_Files zu 78.08%. Das Modell 2 erreicht damit einen höheren Wert als das Modell 5. Im Modell 5 wird 42.3% der Varianz der Variable Validität_Organisation durch die Anzahl Files begründet.

2.3.2.3.2 Zwischendiskussion Modell 2 & 5

Die Hypothese 2, dass die Anzahl Files einen positiven signifikanten Einfluss auf die Validität von Dokumenten hat, wird durch die Modelle 2 und 5 bestätigt. Die Argumentation für diesen Zusammenhang könnte sein, dass die Verpflichtung valide Dokumente zu liefern mit zunehmendem Datenvolumen und daraus resultierendem Einfluss wichtiger wird. Die Validität ermöglicht die Daten untereinander vergleichbar zu machen, was für die Open Data Thematik relevant ist. Des Weiteren könnte diskutiert werden, dass eine grössere Anzahl Dokumente zu einem höheren Erfahrungswert mit dem Publikationsprozess führt, was wiederum die Validität der Dokumente positiv beeinflusst. Gemäss der Abbildung 9 kann die Validität der Dokumente jedoch ab einer bestimmten Menge von Files nicht mehr verbessert werden, sondern stagniert oder ist sogar leicht rückläufig.

2.3.2.4 Modell 3 & 6: Validität und Anzahl Organisationen

Um die Variable Validität_Organisation zu bestimmen, wurde zuvor die Gesamtanzahl Organisationen kalkuliert. Diese Variable kann nun im Folgenden dafür verwendet werden, ihren Einfluss auf die Validität der Dokumente zu prüfen.

Das Modell 3 betrachtet die abhängige Variable Validität_Files im Zusammenhang mit der unabhängigen Variable Anzahl Organisationen. Das Modell 6 prüft ebenfalls den Einfluss der Anzahl Organisationen als x-Variable, spezifiziert als y-Variable jedoch die Validität_Organisation.

2.3.2.4.1 Statistische Ergebnisse Modell 3 & 6

In der Abbildung 10 wird der Zusammenhang zwischen der Validität der Dokumente und der Anzahl Organisationen dargestellt und die beiden Modelle 3 und 6 in den Vergleich gesetzt. Die Tabelle 5 zeigt die Resultate der Modelle gemäss der Berechnung mit den OLS und den robusten Standardfehlern.

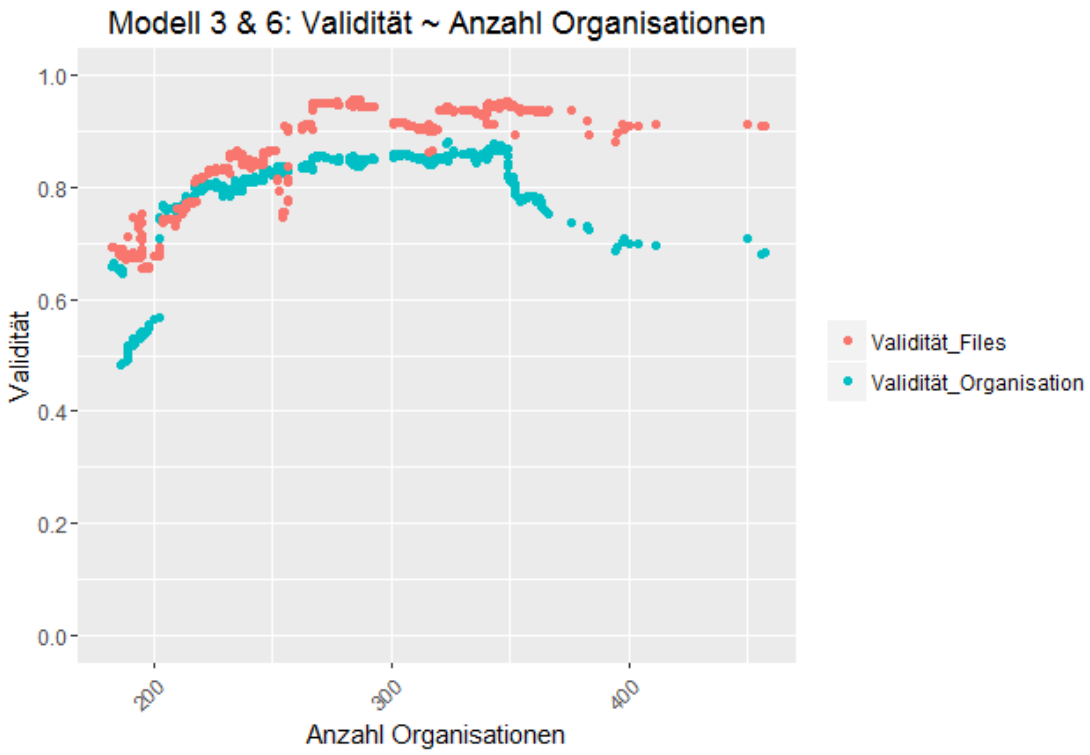

Abbildung 10: Validität und Anzahl Organisationen - Modelle 3 & 6.

Modell 3

	Validity_Files	
	OLS SE	Robust SE
Constant	0.527***	0.527***
	(0.008)	(0.011)
Sum_Publishers	0.001	0.001***
		(0.00004)
Observations	756	756
Adjusted R2	0.689	0.689
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel3
BP = 16.26, df = 1, p-value = 5.521e-05

Modell 6

	Validity_Publisher	
	OLS SE	Robust SE
Constant	0.540***	0.540***
	(0.014)	(0.022)
Sum_Publishers	0.001***	0.001***
	(0.00005)	(0.0001)
Observations	756	756
Adjusted R2	0.330	0.330
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel6
BP = 42.661, df = 1, p-value = 6.509e-11

Tabelle 5: Resultate Modelle 3 & 6

Die Abbildung 10 und die Tabelle 5 zeigen, dass eine Zunahme der Anzahl Organisationen zu einer besseren Validität der Dokumente führt. Das Resultat ist mittels Verwendung von robusten Standardfehlern, aufgrund gegebener Heteroskedastizität, signifikant auf dem 1%-Level. Die Varianz der im Modell 3 definierten Variable Validität_Files wird nur zu 31.09% auf externe Einflussfaktoren zurückgeführt, d.h. 68.91% wird durch die Anzahl Organisationen erklärt. Das Modell 6 erreicht mit 33.03% erklärter Varianz einen tieferen Wert als das Modell 3.

2.3.2.4.2 Zwischendiskussion Modell 3 & 6

Die Hypothese 3, dass die Erhöhung der Validität von Dokumenten mit der Zunahme der Anzahl Organisationen begründet werden kann, wird verifiziert. Es kann argumentiert werden, dass mehr Organisationen einen grösseren Druck auf die anderen Organisationen ausüben, valide Dokumente aufzuweisen. Gleichzeitig kann auch die Theorie der kollektiven Intelligenz zum Tragen kommen, welche in einem früheren Abschnitt bereits thematisiert wurde. Gemäss diesem Prinzip kann eine Gruppe zu einer besseren Entscheidungsfindung als Einzelpersonen neigen, da unterschiedlichen Fähigkeiten und Sichtweisen aufeinander treffen (Leimeister, 2010, S. 239). Dadurch könnte erklärt werden, wieso eine Vielzahl von unterschiedlichen Organisationen die Datenvalidität positiv beeinflusst. Die Abbildung 10 zeigt jedoch, dass irgendwann das Limit erreicht ist und die Validität der Dokumente durch die Teilnahme von zusätzlichen Organisationen nicht mehr weiter verbessert werden kann.

2.3.2.5 Modell 7 & 8: Suche nach dem besten Modell

Die vorhergehenden Modelle zeigen, dass alle drei unabhängigen Variablen Alter IATI, Anzahl Files und Anzahl Organisationen einen Teil der Varianz der Variablen Validität_Files und Validität_Organisation erklären. Im Modell 7 und 8 soll nun untersucht werden, ob sich das adjustierte R^2 durch das Zusammenführen der unabhängigen Variablen zusätzlich verbessert.

Das Modell 7 untersucht die Variable Validität_Files als abhängige Variable mit den unabhängigen Variablen Alter IATI, Anzahl Files und Anzahl Organisationen. Das Modell 8 definiert hingegen neben den erwähnten drei unabhängigen Variablen, die Validität_Organisation als abhängige Komponente.

2.3.2.5.1 Statistische Ergebnisse Modell 7 & 8

In der Tabelle 6 werden die beiden Modelle einander gegenübergestellt.

Modell 7		
	Validity_Files	
	OLS SE	Robust SE
Constant	0.169***	0.169***
	(0.017)	(0.027)
Age	-0.180***	-0.180***
	(0.017)	(0.047)
Sum_Files	0.0002***	0.0002***
	(0.00001)	(0.00001)
Sum_Publishers	0.002***	0.002***
	(0.0002)	(0.001)
Observations	756	756
Adjusted R2	0.821	0.821
Note:	*p<0.1; **p<0.05; ***p<0.01	

Modell 8		
	Validity_Publisher	
	OLS SE	Robust SE
Constant	0.140***	0.140***
	(0.030)	(0.048)
Age	-0.338***	-0.338***
	(0.030)	(0.078)
Sum_Files	0.0001***	0.0001***
	(0.00001)	(0.00001)
Sum_Publishers	0.004***	0.004***
	(0.0004)	(0.001)
Observations	756	756
Adjusted R2	0.505	0.505
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test	studentized Breusch-Pagan test
data: mymodel7	data: mymodel8
BP = 73.738, df = 3, p-value = 6.755e-16	BP = 37.134, df = 3, p-value = 4.312e-08

Tabelle 6: Resultate Modelle 7 & 8

Die Modelle 7 und 8 sind signifikant auf dem 1%-Niveau. Das Modell 7 erklärt 82.09% der Varianz der Variable Validität_Files. Im Modell 8 sind 50% der Varianz der Variable Validität_Organisation auf Alter IATI, Anzahl Files und Anzahl Organisationen zurückzuführen. In beiden Modellen wurde Heteroskedastizität nachgewiesen und daher mit robusten Standardfehlern gerechnet.

2.3.2.5.2 Zwischendiskussion Modell 7 & 8

Im Modell 7 wird die y-Variable von allen untersuchten Modellen am besten durch die überprüften unabhängigen Variablen erklärt. In Bezug auf dieses Kriterium kann das Modell 7 in diesem Kontext als das beste Modell bezeichnet werden. Das Modell 8 bestätigt ebenfalls, dass durch das Zusammenführen der drei unabhängigen Variablen das adjustierte R^2 nochmals verbessert werden kann. Damit wird gezeigt, dass die Validität der Dokumente von unterschiedlichen Faktoren abhängt.

2.3.2.6 Modell 9: Anzahl Organisationen über die Zeit

Seit der Einführung von IATI im Jahr 2011 sind laufend neue Organisationen der Open Data Initiative beigetreten. Das Modell 9 soll nun überprüfen, ob ein signifi-

kanter Zusammenhang zwischen der unabhängigen Variable Zeit und der abhängigen Variable Anzahl Organisationen besteht.

2.3.2.6.1 Statistische Ergebnisse Modell 9

Die Abbildung 11 zeigt die Entwicklung der Anzahl Organisationen über die Zeit. Der Tabelle 7 können die Resultate dazu entnommen werden.

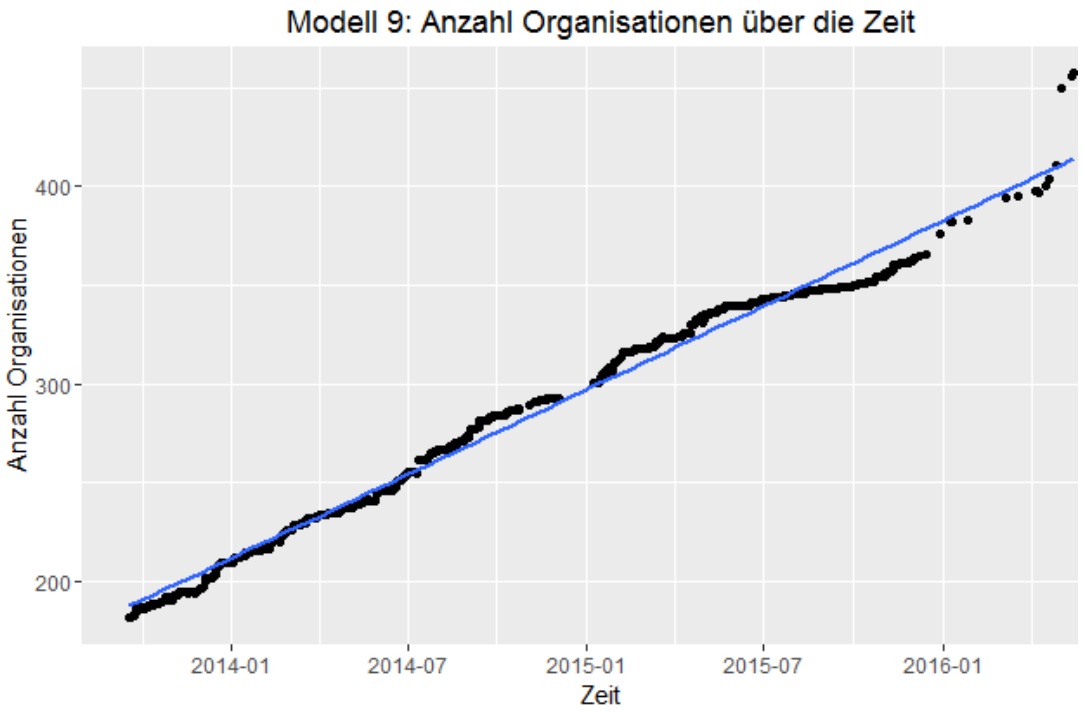

Abbildung 11: Entwicklung Anzahl Organisationen über die Zeit.

Modell 9

	Sum_Publishers	
	OLS SE	Robust SE
Constant	34.006***	34.006***
	(1.105)	(1.312)
Age	85.263***	85.263***
	(0.367)	(0.497)
Observations	756	756
Adjusted R2	0.986	0.986
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel9
BP = 90.052, df = 1, p-value < 2.2e-16

Tabelle 7: Resultate Modell 9

Die Tabelle 7 zeigt einen positiven empirischen Zusammenhang zwischen den beiden Variablen auf dem 1%-Signifikanzniveau. Das Resultat wird durch die Berechnung der robusten Standardfehler aufgrund Heteroskedastizität bestätigt. Die Varianz der Variable Anzahl Organisationen wird zu 98.62% durch die unabhängige Variable Zeit erklärt.

2.3.3.3.2 Zwischendiskussion Modell 9

Die Anzahl Organisationen hat über die Zeit signifikant zugenommen. Seit der Einführung von IATI gab es eine Zunahme um 145% von 187 auf 457 Organisationen. Diese Entwicklung zeigt, dass IATI im Laufe der Zeit stark an Bedeutung zugenommen hat.

2.3.2.7 Schlussdiskussion Modelle 1 bis 9

Die herausbekommenen Ergebnisse der Modelle 1 bis 8 sollen mit Vorsicht genossen werden. Obwohl die Signifikanz aller Modelle gegeben ist, kann es dennoch sein, dass die erklärte Varianz der abhängigen Variable überschätzt wird. Die y-Variablen Validität_Files und Validität_Organisation zeigen teilweise deutliche Unterschiede, wie gut ihre Varianz durch die x-Variablen erklärt wird. Die Variable Validität_Files erreicht insgesamt höhere Werte als die Variable Validität_Organisation. Im Durchschnitt beträgt der Anteil valider Dokumente über alle Messwerte 87.74%, mit einer Standardabweichung von 0.0866. Der durchschnittliche Anteil Organisationen mit nur validen Dokumenten ist mit 80.03% tiefer und hat eine Standardabweichung von 0.0928. Die Mittelwertdifferenzen sind signifikant verschieden auf dem Konfidenzintervall von 99%. Dies kann der Tabelle 8 entnommen werden.

```
        Welch Two Sample t-test

data:  m$Validity_Publisher and m$Validity_Files
t = -16.693, df = 1502.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -0.08900901 -0.06518505
sample estimates:
mean of x mean of y
0.8003378 0.8774348
```

Tabelle 8: Vergleich Validität_Files vs. Validität_Organisation

Die abweichenden Mittelwerte können ein Indiz dafür sein, dass die erklärte Varianz in den Modellen 1, 2, 3 und 7, welche die abhängige Variable als Anteil valider

Dokumente definieren, überschätzt werden und wichtige Faktoren nicht berücksichtigen. Die tieferen Werte der Variable Validität_Organisation könnten damit begründet werden, dass die unterschiedlichen Organisationstypen miteinfliessen und den Effekt dadurch abschwächen. Diese könnten einen relevanten Einfluss auf die Validität der Dokumente haben.

Ein weiterer Kritikpunkt, der auf eine Überschätzung der erklärten Varianz der Modelle schliessen lässt, ist die Abhängigkeit, die zwischen den unabhängigen Variablen besteht. Die Zunahme des Alters von IATI, bzw. die Zeit generell, wie im Modell 9 gezeigt wird, führt zu einer höheren Anzahl Organisationen. Gleichzeitig nimmt auch die Anzahl Files über die Zeit und mit der Anzahl Teilnehmer zu. Aus diesem Grund kann nicht klar abgegrenzt werden, wie sich die Variablen untereinander beeinflussen. Es kann jedoch davon ausgegangen werden, dass sie sich gegenseitig positiv verstärken.

2.3.3 Analyse auf Stufe Organisationen: Modelle 10 & 11

Nachdem im vorherigen Abschnitt organisationsübergreifend die Entwicklung des Anteils valider Dokumente über alle hochgeladenen Files und Teilnehmer geprüft wurde, sollen die einzelnen Organisationen im Folgenden etwas genauer betrachtet werden. Die Analyse hat zum Ziel, Sonderfälle zu evaluieren, um diese Organisationen in einem späteren Schritt als Kontrollvariablen zu verwenden. Des Weiteren soll die Hypothese 4 überprüft werden. Die abhängige Variable definiert den Anteil valider Dokumente.

2.3.3.1 Beschreibung und Bereinigung des Datensatzes

Der hier verwendete Datensatz betrachtet den Anteil valider Dokumente auf Stufe der einzelnen Organisationen. Er enthält die Anzahl valider und nicht valider Files pro Organisation, die zum Zeitpunkt des 09.06.2016 hochgeladen wurden.

Damit der Datensatz genutzt werden kann, muss er zuerst bereinigt werden. Wie bisher werden die Zeilen und Spalten des CSV vertauscht und als Data Frame abgespeichert. Das Problem in diesem Datensatz ist, dass zuerst die Anzahl aller nicht validen und anschliessend alle validen Dokumente pro Organisation aufgelistet werden. Für die weitere Anordnung der Beobachtungen wird daher eine ID kreiert, die es ermöglicht den Datensatz, in valide und nicht valide Dokumente zu unter-

teilen, indem ab einer bestimmten ID-Nummer abgetrennt wird. Die entstandenen Spalten können dann weiter in zwei verschiedene Data Frames abgespeichert werden. Somit wird ein Data Frame kreiert, der alle validen Files pro Organisation enthält und ein Weiterer, der die nicht validen Dokumente abdeckt. Beide Data Frames sind jedoch nicht vollständig, da kein Eintrag anzeigt wird, wenn der Wert 0 ist. Für die weitere Analyse ist dies relevant. Aus diesem Grund wird eine Sequenz in beiden Data Frames generiert, die den Wert 0 enthält und damit die Gegenseite abdeckt. Die beiden Data Frames mit den validen oder nicht validen Dokumenten können dann anhand der gemeinsamen Variable „Organisationen“ miteinander vereint werden. Die Schwierigkeit, die hier auftaucht ist, dass nur Organisationen angezeigt werden, die sowohl valide als auch nicht valide Dokumente haben. Alle Organisationen mit nur Entweder/Oder-Werten werden vernachlässigt. In einem nächsten Schritt muss diese Lücke gefüllt werden, indem ein Data Frame generiert wird, welches die zusammengefassten und die einzelnen Werte auflistet. Da es nun Überschneidungen gibt werden alle Organisationen, die bereits im zusammengefassten Data Frame vorkommen, herausgenommen. Der fertige Datensatz enthält nun alle Organisationen genau einmal. Dieser kann als „Final“ abgespeichert und für die Analyse verwendet werden. Der R-Code „Analyse auf Stufe Organisation“ kann dem Anhang entnommen werden.

2.3.3.2 Modell 10: Verteilung Files pro Organisation

Die Anzahl Files pro Organisation berechnet sich wie bis anhin. Diese Variable ist notwendig, um den Anteil valider Dokumente pro Organisation zu ermitteln. Das Modell 10 betrachtet die Verteilung der validen Dokumente auf der x-Achse und die nicht validen Files auf der y-Achse.

2.3.3.2.1 Statistische Ergebnisse Modell 10

Die Abbildung 12 zeigt die Verteilung der Files pro Organisation. Sie soll dabei helfen, einzelne Organisationen zu identifizieren, die besonders stark voneinander abweichen. Entscheidend ist jedoch nicht der Anteil valider Dokumente, da viele Organisationen einen Wert von 0% oder 100% erreichen. Vielmehr sollen Organisationen betrachtet werden, die besonders viele valide und/oder invalide Files haben.

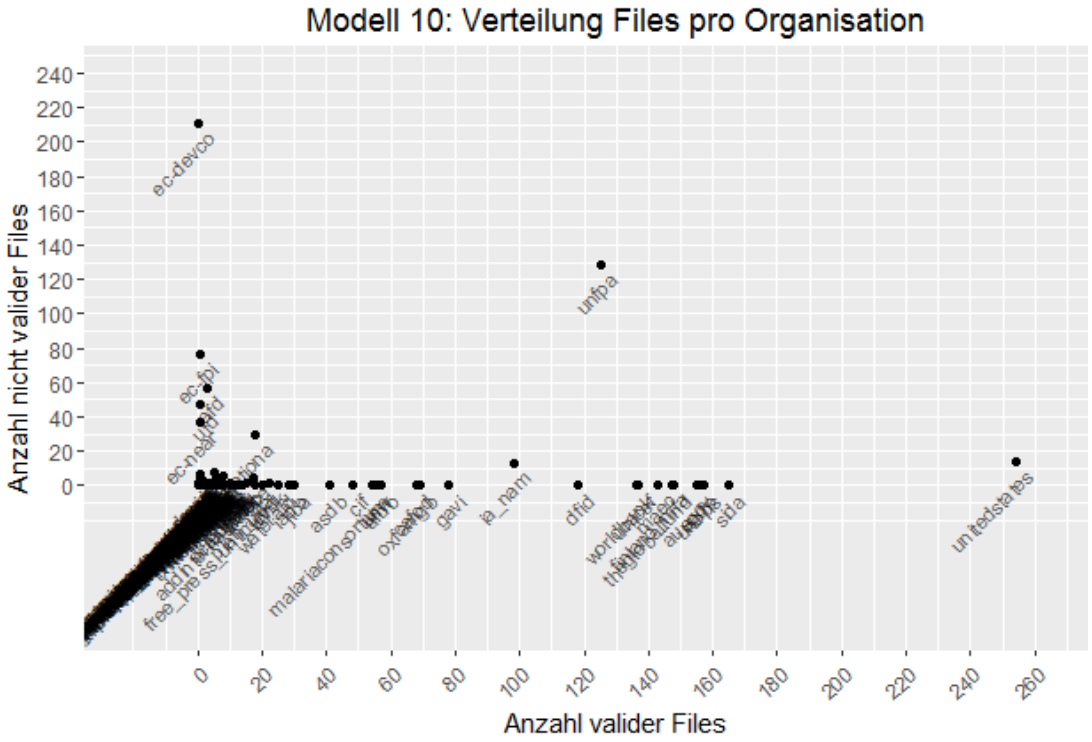

Abbildung 12: Verteilung valide vs. invalide Dokumente.

In der Abbildung 12 ist zu sehen, dass sich die Mehrheit der Organisationen entlang der x-Achse verteilen und damit mehr valide als invalide Dokumente aufweisen. Diese Erkenntnis kann auch der Tabelle 9 entnommen werden. Die Mittelwertsdifferenz zwischen der Anzahl valider und nicht valider Dokumente ist signifikant verschieden auf dem Konfidenzintervall von 99%. Die Organisationen haben im Durchschnitt eine signifikant höhere Anzahl valider als nicht valider Files. Der Durchschnittswert der Anzahl valider Dokumente beträgt 7.63 pro Organisation mit einer Standardabweichung von 26.18 Dokumenten. Im Vergleich dazu liegt der Wert der Anzahl invalider Files bei 1.76 mit einer Standardabweichung von 12.70. Im Durchschnitt hat jede Organisation 9.4 Dokumente hochgeladen.

```
        Welch Two Sample t-test

data:  Final$Files_Fail and Final$Files_Pass
t = -4.3332, df = 665.31, p-value = 1.697e-05
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -9.372640 -2.371395
sample estimates:
mean of x mean of y
 1.759219  7.631236
```

Tabelle 9: Vergleich Anzahl valider vs. invalider Dokumente

Besonders stechen die drei Organisationen European Comission – Development and Cooperation (ec-devco), United Nations Population Fund (unfpa) und Untied States (unitedstates) heraus. Sie vertreten die drei Formen von Extremen. Die unitedstates hat zum gemessenen Zeitpunkt 14 nicht valide und 254 valide Dokumente hochgeladen und ist damit die Organisation mit der höchsten Anzahl an validen Files. Ihr Anteil valider Dokumente beträgt 94.78%. Im Gegensatz dazu hat die ec-devco zum gemessenen Zeitpunkt 211 nicht valide und kein einziges valides Dokument publiziert und erreicht somit ein Anteil valider Dokumente von 0%. Die unfpa steht in der Mitte der beiden Extremwerte und hat 125 valide und 129 nicht valide Dokumente und einen Anteil valider Dokumente von 49.21%.

2.3.3.2.2 Zwischendiskussion Modell 10

Im Folgenden sollen die drei Organisationen etwas genauer betrachtet werden, um mögliche Unterschiede zu erkennen.

Die ec-devco hat zum ersten Mal am 15.09.2015 ihre Dokumente auf IATI hochgeladen und gehört damit zu einer eher jüngeren Teilnehmergruppe. Ihre Daten publiziert sie quartalsweise und verwendet dafür die Version 01.04 des IATI-Standards. Der Organisationstyp der ec-devco wird in die Kategorie andere öffentliche Sektoren eingeteilt (IATI, 2016o). Neben der Publikation der Daten über IATI, besitzt sie eine eigene Webseite, die ebenfalls Daten zu Entwicklungshilfe veröffentlicht (European Comission – Development and Cooperation-EuropeAid, 2016).

Am 19.09.2013 als die erste Messung der Anzahl valider und nicht valider Files gemacht wurde, hat die unfpa ebenfalls ihre Daten auf IATI hochgeladen. Aktuell benutzt sie die Version 02.01 des IATI Updates. Die unfpa ist eine multilaterale Organisation, welche ihre Daten wie die ec-devco quartalsweise veröffentlicht (IATI, 2016p). Ihre Strategie zur Verbesserung der Datenqualität richtet sie generell nach IATI und dem Busan Agreement aus (United Nations Population Fund, 2016).

Die unitedstates hat wie die unfpa ihre Daten bereits ab dem 19.09.2013 publiziert und verwendet ebenfalls den IATI Standard 02.01. Als Organisationstyp handelt es sich hierbei um die US Regierung. Sie legt starken Wert auf Offenheit und eine gute Datenqualität. Zum Ende eines Kalenderjahres werden Entwicklungshilfedaten im

IATI-Format in verifizierter, kompletter und statistischer Form nach Abschluss des OECD/DAC-Reportings veröffentlicht. Regelmässig werden in einem iterativen Prozess die Abweichungen zwischen vorab hochgeladenen Daten und zum Kalenderjahresabschluss verifizierten Daten überprüft, um eine bessere Qualität auf Quartalsebene zu erreichen (IATI, 2016l).

Die Organisationen unterscheiden sich vor allem im Organisationstyp und im Zeitpunkt des ersten Uploads der Daten über IATI. Die ec-devco hat im Vergleich zu den anderen Organisationen ihre Dokumente erst später hochgeladen. Dieser Aspekt soll im nächsten Modell untersucht werden, um die Hypothese 4 zu überprüfen.

2.3.3.3 Bereinigung und Beschreibung des Datensatzes

Da es keine JSON-Tabelle gibt, welche die Information über das erste Upload der Dokumente automatisch liefert, wurden die Daten einzeln vom IATI Dashboard bezogen und manuell über Excel dem vorhin abgespeicherten CSV File “Final“ als Variable First_Publication hinzugefügt. Das neue CSV „FinalIATI“ mit der zusätzlichen Variable First_Publication wird in R eingelesen. Da Excel das Datum in der Form dd.mm.yyyy erfasst, muss es zuerst in die von R verwendete Formatierung von YYYY-MM-DD gebracht werden.

Im nächsten Schritt kann geprüft werden, ob die Länge der Teilnahme an IATI, die durch das Datum des ersten Uploads ermittelt wird, einen Einfluss auf den Anteil valider Dokumente hat. Zu beachten ist, dass es viele Organisationen mit nur einem hochgeladenen Dokument gibt. Das hat zur Folge, dass in diesen Fällen der Anteil valider Dokumente nur 0% oder 100% sein kann. Da diese Extremwerte nicht repräsentativ erscheinen, werden sie nicht in die Auswertung einbezogen. Der Datensatz reduziert sich dadurch von 461 auf 341 Beobachtungen.

2.3.3.4 Modell 11: Validität und Datum erstes Upload

Das Modell 11 vergleicht die abhängige Variable Validität_Files, d.h. den Anteil valider Dokumente pro einzelne Organisation, mit der unabhängigen Variable Datum des ersten Uploads.

2.3.3.4.1 Statistische Ergebnisse Modell 10

In der Abbildung 13 wird der Zusammenhang der Variablen visuell dargestellt. Der Tabelle 10 können die statistischen Kennzahlen entnommen werden.

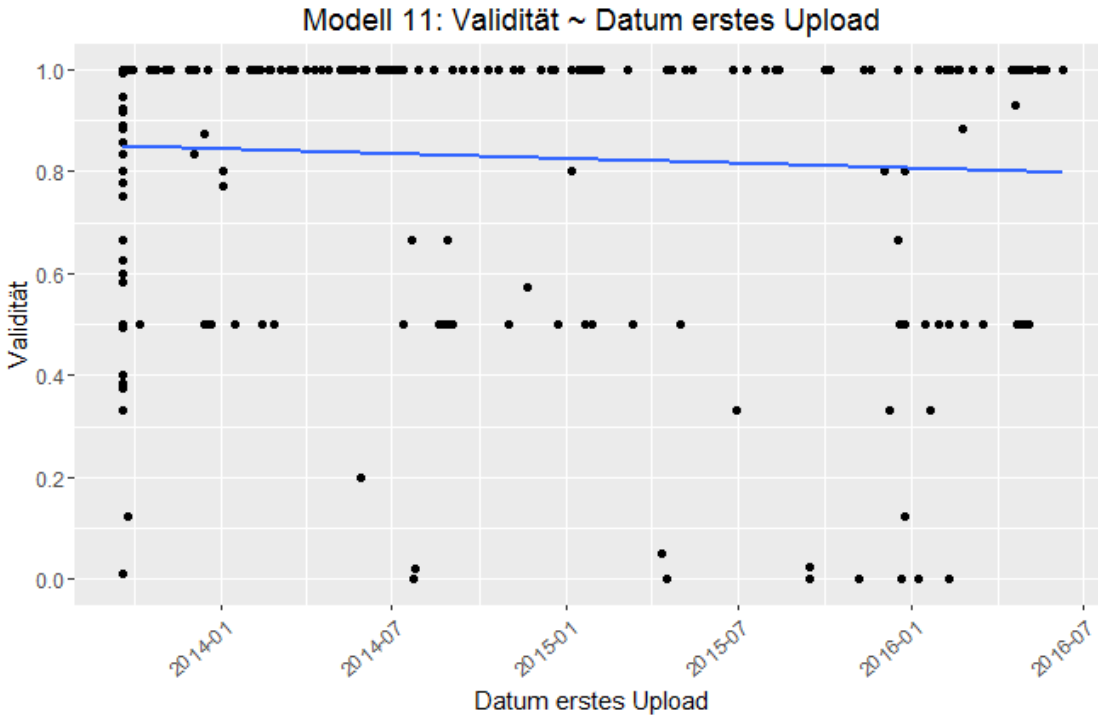

Abbildung 13: Validität und Datum erstes Upload.

Modell 11

	Validity	
	OLS SE	Robust SE
Constant	1.678***	1.678**
	(0.639)	(0.653)
First_Publication	-0.0001	-0.0001
	(0.00004)	(0.00004)
Observations	341	341
Adjusted R2	0.002	0.002
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel11
BP = 6.5097, df = 1, p-value = 0.01073

Tabelle 10: Resultate Modell 11

Es ist keine Heteroskedastizität auf den Niveau von 1% gegeben, weshalb mit den OLS-Standardfehlern gerechnet werden darf. Die Variablen weisen jedoch keinen signifikanten Zusammenhang auf.

2.3.3.4.2 Zwischendiskussion Modell 11

Die Aussage des Modells ändert sich selbst dann nicht, wenn die Ausschlusskriterien weiter gesenkt und ausschliesslich Organisationen berücksichtig werden, die mindestens sechs Dokumente zur Verfügung stellen. Das Gleiche wäre bei der Eingrenzung der y-Variable. Würden alle Beobachtungen eliminiert, die einen Anteil valider Dokumente von 0% oder 100% haben, dann würde sich das Modell zwar verbessern, die Signifikanz wäre trotzdem nicht erfüllt. Die Hypothese 4, dass sich die Länge der Teilnahme an IATI positiv mit dem Anteil valider Dokumente verhält, kann nicht bestätigt werden. Der Anteil valider Dokumente wird demnach durch andere Faktoren beeinflusst.

2.3.4 Analyse Kontrollvariablen: Modelle 12 & 13

Um die Ergebnisse aus der organisationsübergreifenden Analyse zu kontrollieren, sollen einzelne Organisationen überprüft werden. Zu diesem Zweck werden die drei evaluierten Organisationen ec-devco, unfpa und unitedstates hinzugezogen. Die drei Vertreter weichen in Bezug auf die Anzahl valider und invalider Dokumente stark voneinander ab. Dadurch kann ein kritischer Blick auf die vorher erhaltenen Resultate geworfen werden. Die Bereinigung des Datensatzes läuft analog wie bei der Analyse auf der übergreifenden Ebene ab. Da das jeweilige CSV der untersuchten Organisationen einige Werte nicht in der gewünschten Formatierung übernimmt, müssen diese Werte manuell ergänzt werden. Der R-Code „Analyse Kontrollvariablen“ ist im Anhang hinterlegt.

2.3.4.1 Modell 12: Validität und Alter IATI, Vergleich Organisationen

Das Modell 12 betrachtet den Einfluss der unabhängigen Variable Alter IATI auf die abhängige Variable Anteil valider Dokumente der drei Organisationen ec-devco, unfpa und unitedstates.

2.3.4.1.1 Statistische Ergebnisse Modell 12

Die Abbildung 14 zeigt den Zusammenhang zwischen dem Alter von IATI und dem Anteil valider Dokumente der drei Organisationen im Vergleich. Den Tabellen 11 bis 13 können die Resultate entnommen werden.

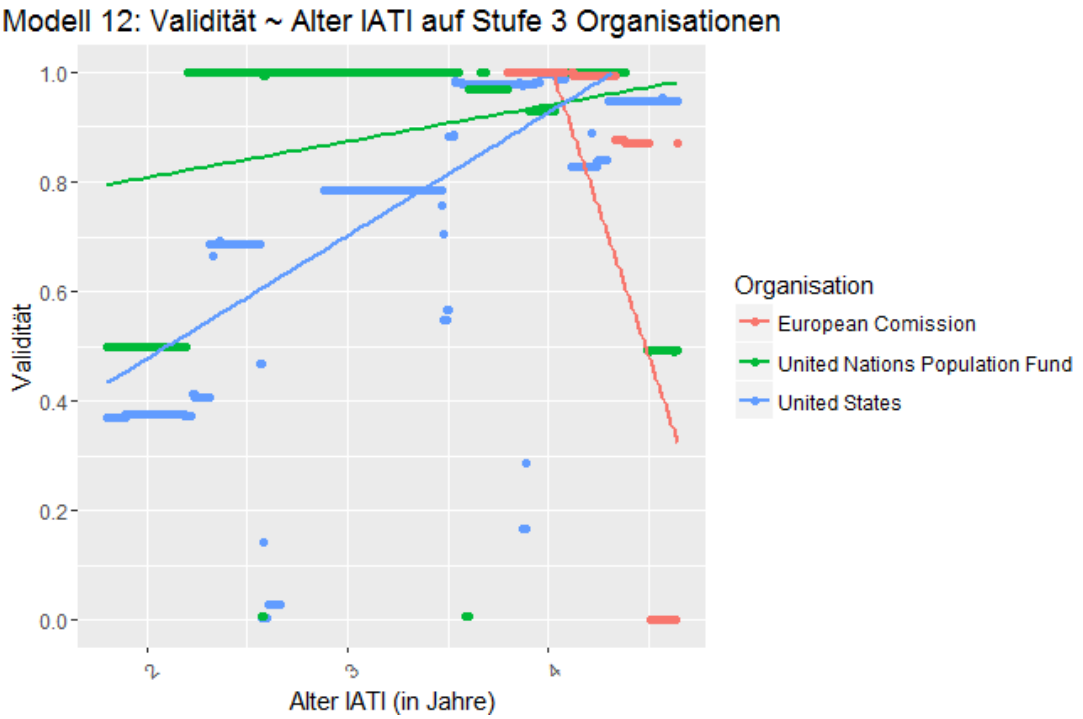

Abbildung 14: Validität und Alter IATI, Vergleich drei Organisationen.

Modell 12a

	validity_files	
	OLS SE	Robust SE
Constant	5.353***	5.353***
	(0.257)	(0.304)
Age	-1.083***	-1.083***
	(0.061)	(0.075)
Observations	270	270
Adjusted R2	0.538	0.538
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel12a
BP = 179.23, df = 1, p-value < 2.2e-16

Tabelle 11: Resultate Modell 12a – ec-devco

Modell 12b

	validity_files	
	OLS SE	Robust SE
Constant	0.676***	0.676***
	(0.028)	(0.036)
Age	0.066***	0.066***
	(0.009)	(0.011)
Observations	975	975
Adjusted R2	0.057	0.057
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel12b
BP = 0.072516, df = 1, p-value = 0.7877

Tabelle 12: Resultate Modell 12b - unfpa

Modell 12c

	Validity_Files	
	OLS SE	Robust SE
Constant	0.028	0.028
	(0.025)	(0.022)
Age	0.225***	0.225***
	(0.008)	(0.006)
Observations	964	964
Adjusted R2	0.481	0.481
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel12c
BP = 18.407, df = 1, p-value = 1.784e-05

Tabelle 13: Resultate Modell 12c - unitedstates

Der Abbildung 14 und der Tabelle 11 ist zu entnehmen, dass die ec-devco auf dem 1%-Signifikanzniveau einen negativen Zusammenhang zwischen der unabhängigen und der abhängigen Variable aufweist. Mit zunehmendem Alter sinkt der Anteil an validen Dokumenten. Die Varianz der Variable Anteil valider Dokumente wird zu 53.8% durch das Alter von IATI erklärt. Die unfpa und die unitedstates zeigen wiederum auf demselben Signifikanzniveau einen positiven Zusammenhang. Das Modell 12b der unfpa erreicht eine erklärte Varianz von 5.7%, wobei das Modell 12c der unitedstates auf einen Wert von 48.1% kommt. Dieser Wert liegt zwischen denjenigen von Modell 1 und 4 aus der vorherigen Analyse. Die Residuen des Modells 12a der ec-devco und 12c der unitedstates sind heteroskedastisch, weshalb mit robusten Standardfehlern gerechnet wird. Im Gegensatz dazu weist das Modell 12b der unfpa keine Heteroskedastizität auf.

2.3.4.1.2 Zwischendiskussion Modell 11

Die Modelle der unfpa und der unitedstates bestätigen die Hypothese 1, dass mit zunehmendem Alter von IATI der Anteil valider Dokumente steigt. Die erklärte Varianz erreicht in beiden Modellen hingegen unterschiedlich hohe Werte. Das Modell 12a der ec-devco widerspricht der aufgestellten Hypothese 1 und weist eine Abnahme des Anteils valider Dokumente mit zunehmendem Alter von IATI auf. Da die Organisation jedoch von einem Anteil valider Dokumente von 100% auf 0% gefallen ist, können gewisse Zweifel über den effektiven Zusammenhang der Variablen des Modells 12a aufkommen.

2.3.4.2 Modell 13: Validität und Anzahl Files Vergleich Organisationen

Das Modell 13 prüft die Beziehung der abhängigen Variable Anteil valider Dokumente mit der unabhängigen Variable Anzahl Files auf Stufe der drei evaluierten Organisationen ec-devco, unfpa und unitedstates.

2.3.4.2.1 Statistische Ergebnisse Modell 13

Die Abbildung 15 visualisiert die Zusammenhänge der Modelle 13a bis c und stellt die Resultate in den Tabellen 14 bis 16 einander gegenüber.

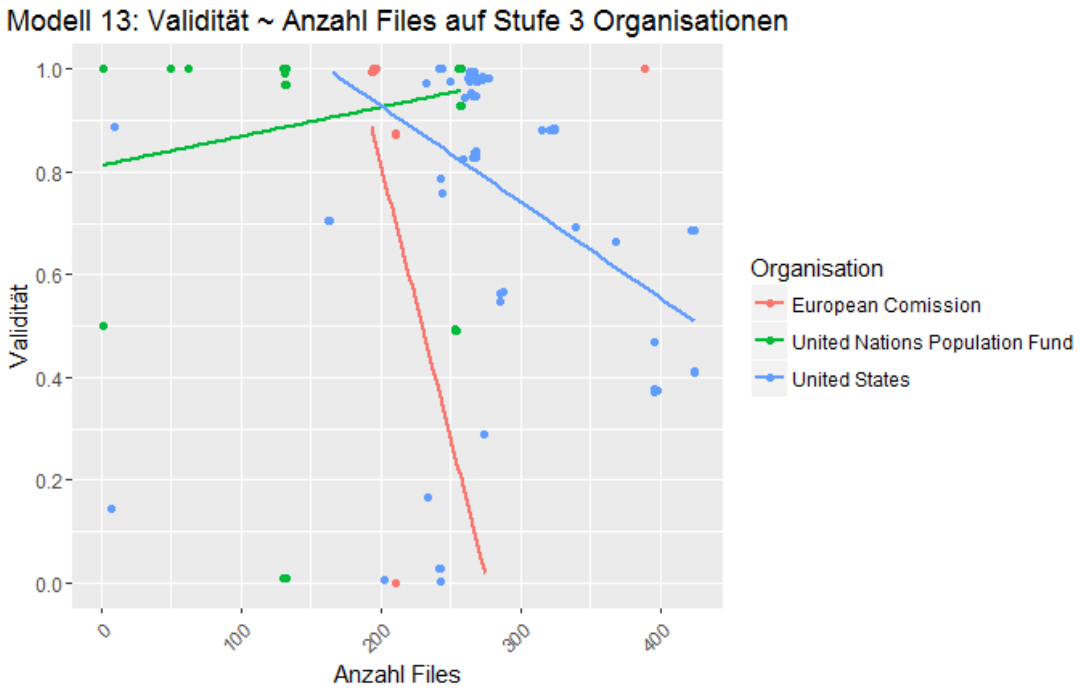

Abbildung 15: Validität und Anzahl Files, Vergleich drei Organisationen.

```
Modell 13a
=========================================
                    Validity_Files
                  OLS SE       Robust SE
-----------------------------------------
Constant         2.942***        2.942
                 (0.308)        (5.371)
Sum_Files       -0.011***       -0.011
                 (0.002)        (0.027)
-----------------------------------------
Observations       270             270
Adjusted R2       0.150           0.150
=========================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

```
        studentized Breusch-Pagan test

data:  mymodel13a
BP = 225.59, df = 1, p-value < 2.2e-16
```

Tabelle 14: Resultate Modell 13a – ec-devco

Modell 13b

	Validity_Files	
	OLS SE	Robust SE
Constant	0.813***	0.813***
	(0.011)	(0.013)
Sum_Files	0.001***	0.001***
	(0.0001)	(0.0001)
Observations	975	975
Adjusted R2	0.061	0.061
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel13b
BP = 10.01, df = 1, p-value = 0.001557

Tabelle 15: Resultate Modell 13b - unfpa

Modell 13c

	Validity_Files	
	OLS SE	Robust SE
Constant	1.304***	1.304***
	(0.031)	(0.040)
Sum_Files	-0.002***	-0.002***
	(0.0001)	(0.0001)
Observations	964	964
Adjusted R2	0.260	0.260
Note:	*p<0.1; **p<0.05; ***p<0.01	

studentized Breusch-Pagan test

data: mymodel13c
BP = 27.632, df = 1, p-value = 1.467e-07

Tabelle 16: Resultate Modell 13c - unitedstates

Den Tabellen 14 bis 16 ist zu entnehmen, dass alle drei Modelle auf dem Niveau von 1% Heteroskedastizität aufweisen, weshalb die Verwendung von robusten Standardfehlern notwendig ist. Das Modell 13a der ec-devco und 13c der unitedstates zeigen beide eine negative Beziehung zwischen dem Anteil valider Dokumente und der Anzahl Files. Die unitedstates bestätigt im Modell 13c diesen negativen Zusammenhang auf dem 1%-Signifikanzniveau und weist eine erklärte Varianz von 26% auf. Das Ergebnis des Modells 13a der ec-devco ist jedoch nicht signifikant. Das Modell 13b der unfpa verifiziert wiederum die erwartete signifikant positive Beziehung auf dem Niveau von 1%. Der Anteil Varianz der durch die Variable Anzahl Files erklärt wird erreicht im Modell 13b einen Wert von 6.1%.

2.3.4.2.2 Zwischendiskussion Modell 13

Das Modell 13b der unfpa bestätigt die Hypothese 2, dass mit zunehmender Anzahl Files sich die Validität der Dokumente verbessert. Die Variable Anzahl Files erklärt jedoch nur ein kleiner Teil der Varianz des Anteils valider Dokumente. Das Modell 13c der unitedstates widerlegt hingegen die Hypothese 2 und zeigt eine negative Beziehung. Diese Entwicklung wird aufgrund fehlender Signifikanz vom Modell 13a nicht noch zusätzlich gestützt. Ein Grund für die negative Beziehung des Modells 13c könnte unter anderem sein, dass sich die Anzahl Files rückläufig zur Variable Zeit verhält.

2.3.4.3 Schlussdiskussion Modelle 12 und 13

Der Vergleich dieser Modelle zeigt, dass trotz übergreifenden Zusammenhangs, sich die einzelnen Organisationen untereinander differenzieren können. Es lässt sich kein eindeutiges Muster in den Ergebnissen auf Stufe der einzelnen Organisationen erkennen. Die Hypothesen 1 und 2 konnten dennoch bestätigt werden. Für die Analyse der Modelle 12 und 13 wurden jedoch bewusst Organisationen gewählt, die stark voneinander abweichen. Die Resultate zeigen vor allem Unterschiede, wie stark die Varianz der y-Variable erklärt wird. Daraus lässt sich schliessen, dass neben den erwähnten Variablen noch andere Faktoren die Validität der Dokumente beeinflussen. Die drei untersuchten Organisationen werden unterschiedlichen Organisationstypen zugeordnet. Dies könnte ein Hinweis darauf sein, dass die Organisationsstruktur ein Einflussfaktor ist. Hinzu kommt, dass die Organisationen selbst für die Datenqualität verantwortlich sind. IATI versucht den Teilnehmern Commitment zu übermitteln und sie über allgemeine Normen der Datenqualität zu informieren. Wie hoch jedoch das Commitment gegenüber der Transparenz in den Einzelfällen ist, könnte ein weiterer relevanter Aspekt sein, der die Validität der Dokumente beeinflusst. Ob sich zudem nationale Guidelines bemerkbar machen, müsste weiter überprüft werden.

2.3.5 Übersicht Resultate

Die Tabelle 17 stellt eine Übersicht aller untersuchten Modelle dar und zeigt, welche x-Variablen die Varianz der y-Variable wie stark erklärt. Es wird ersichtlich, dass das Modell 7 den höchsten adjustierten R^2-Wert aufweist. Aufgrund dieses Kriteriums könnte das Modell 7 als bestes Modell definiert werden. Weitere Unter-

suchungen haben jedoch ergeben, dass davon ausgegangen werden muss, dass die erklärte Varianz der Modelle 1, 2, 3 und 7 überschätzt wird. Demzufolge kann keine eindeutige Aussage darüber gemacht werden, welches Modell die Varianz der y-Variable am besten erklärt.

Modell	**Organisationen**	**Variable**		**Adjustiertes R^2**
		y-Wert	**x-Wert**	
Modell 1	alle	Validität_Files	Alter IATI	0.641
Modell 2	alle	Validität_Files	Anzahl Files	0.781
Modell 3	alle	Validität_Files	Anzahl Organisationen	0.689
Modell 4	alle	Validität_Organisation	Alter IATI	0.284
Modell 5	alle	Validität_Organisation	Anzahl Organisationen	0.423
Modell 6	alle	Validität_Organisation	Anzahl Files	0.330
Modell 7	alle	Validität_Files	Alter IATI, Anzahl Organisationen, Anzahl Files	0.821
Modell 8	alle	Validität_Organisation	Alter IATI, Anzahl Organisationen, Anzahl Files	0.505
Modell 9	alle	Anzahl Organisationen	Zeit	0.986
Modell 10	Einzelne Organisationen	Anzahl invalide Files	Anzahl valide Files	-
Modell 11	Einzelne Organisationen	Validität_Files	Datum erstes Upload	0.002
Modell 12	unitedstates	Validität_Files	Alter IATI	0.481
	unfpa			0.057
	ec-devco			0.538
Modell 13	unitedstates	Validität_Files	Anzahl Files	0.260
	unfpa			0.061
	ec-devco			0.105

Tabelle 17: Übersicht Resultate

3 Zusammenfassung und Ausblick

3.1 Zusammenfassung

Auf theoretischer Ebene wird gezeigt, dass Open Data Initiativen zu einer besseren Datenqualität über die Zeit führen. Integrierte Koordinationsmechanismen wie Standardisierungen und Feedbacksysteme helfen den Open Data Prozess aktiv zu steuern und dadurch die Datenqualität zu verbessern (Zuiderwijk & Janssen, 2013). Der Aufbau einer E-Infrastruktur bietet dabei die nötigen Rahmenbedingungen, um den Informationsaustausch und das Feedbackverhalten zu fördern (De Cindio, 2012). Zudem begründet die Theorie der Usefulness und der Stewardship Ansatz die Motivation der Regierungen, qualitativ gute Daten zur Verfügung zu stellen (Dawes, 2010). Das Konstrukt der Datenqualität beschreibt die Datenvalidität als wichtiges Subkriterium. Daraus kann gefolgert werden, dass mit zunehmender Qualität gleichzeitig die Validität steigt. Die theoretische Hypothese, dass Open Data Initiativen die Datenvalidität über die Zeit erhöhen, kann damit verifiziert werden.

In der empirischen Analyse wurde die Validität von Dokumenten untersucht, welche als wichtiges intrinsisches Qualitätskriterium von verlinkten Daten gilt (Zaveri et al., 2012, S. 6). Die Auswertung zeigt, dass sich die Validität der Dokumente, welche als Anteil valider Dokumente operationalisiert ist, über die Zeit von 2013 bis 2016 erhöht hat. Diese Steigerung kann durch die Zunahme des Alters von IATI, der Anzahl valider Dokumente und der Anzahl teilnehmender Organisationen begründet werden. Die Resultate sind signifikant auf dem 1%-Niveau. Damit können die Hypothesen 1 bis 3 verifiziert werden. Gründe für die Zusammenhänge könnte die kollektive Intelligenz, der Lernprozess oder auch die Erfahrung mit dem Publikationsprozess sein (Juran, 1998; Leimeister, 2010). Weitere Auswertungen zeigen jedoch, dass die Varianz des Anteils valider Dokumente, die durch die drei unabhängigen Variablen erklärt wird, überschätzt ist. Eine Aussage darüber, welches der untersuchten Modelle die Varianz der y-Variable am besten erklärt und damit das beste Modell ist, kann daher nicht gemacht werden. Dies zeigt, dass noch weitere Faktoren die Validität der Dokumente beeinflussen, die in dieser Analyse nicht berücksichtigt wurden. Die Hypothese 4, dass die Länge der Teilnahme der Organisationen an IATI zu einer höheren Validität von Dokumenten führt, kann hingegen nicht bestätigt werden.

Diese Masterarbeit soll insgesamt zeigen, dass es für Organisationen und Institutionen lohnenswert ist, einer Open Data Initiative wie IATI beizutreten. Neben der Möglichkeit der Benutzung verschiedener Tools und dem Netzwerk, welches sich durch den Zusammenschluss aufbauen lässt, helfen Standardisierungen den Herausgebern eine bessere Datenvalidität zu erreichen. Diese ist relevant, damit es nicht zu Fehlinterpretationen der Daten kommt und sie somit zweckmässig verwendet werden können. Dadurch können die Vorteile von Open Data wie Transparenz, Wirtschaftswachstum und Innovationen auftreten.

3.2 Ausblick

In dieser Arbeit wird mehrmals die kollektive Intelligenz erwähnt. Dabei geht es insgesamt um die Frage, wie sich der Staat intelligent ausrichten kann. Da die Thematik sehr umfangreich ist und nur ein Randthema in diesem Kontext darstellt, wurde nicht genauer darauf eingegangen. Für weitere Forschungen, auch in Bezug auf Open Data, wäre es jedoch interessant diesen Ansatz vertiefter zu betrachten. Weiterführende Informationen könnten den Publikationen von Bonabeau (2009) und Leimeister (2010) entnommen werden.

Im Unterkapitel zu Open Data Initiativen werden einige Voraussetzungen beschrieben, welche für den Erfolg einer Open Data Initiative relevant sind. Diese nicht abschliessende Auflistung dient in dieser Arbeit als Erklärungsgrundlage. Mittels einer empirischen Analyse könnte jedoch überprüft werden, welche dieser Bedingungen einen signifikanten Einfluss auf den Erfolg einer Open Data Initiative haben.

Die Auswertung der IATI-Daten hat ergeben, dass neben den erwähnten Variablen noch andere Faktoren einen Einfluss auf die Validität der Dokumente haben. Ein zentraler Aspekt, der weiter untersucht werden könnte, wären die unterschiedlichen Organisationstypen, die verschiedene Strategien verfolgen und unterschiedliche Kulturen aufweisen. Festgelegte Strukturen oder auch die Einstellung innerhalb der Organisationen gegenüber der Transparenz könnten die Validität der Dokumente beeinflussen. Des Weiteren könnten nationale Guidelines Abweichungen zwischen den einzelnen Organisationen erklären. Da dies ein sehr komplexes Thema ist, wurde im Rahmen dieser Arbeit nicht genauer darauf eingegangen.

Zum Schluss wäre es interessant zu prüfen, wie sich die Validität der Dokumente in anderen Open Data Initiativen innerhalb der Entwicklungshilfe und im Vergleich zu anderen Branchen über die Zeit entwickelt hat.

3.3 Anhang A

3.3.1 R-Code Organisationsübergreifende Analyse

```
#Organisationsübergreifende Ebene
#Packages
install.packages("stargazer")
library(stargazer)
install.packages ("sandwich")
library(sandwich)
install.packages("lmtest")
library(lmtest)
install.packages("ggplot2")
library(ggplot2)
install.packages("data.table")
library(data.table)
install.packages ("zoo")
library(zoo)
-----------------------------------------------------------------------------------------------
#Bearbeitung Datensatz Anzahl invalider Dokumente
#File "convertcsv.csv" einlesen (Import Dataset)
read.csv("convertcsv.csv", header=T)
head(convertcsv)
Validation <- t(convertcsv) #Spalten und Zeilen wechseln
head(Validation)
Validation <- as.data.frame(substr(Validation, 1, 10)) #Datum aus Text Beobachtungen rausfiltern
names(Validation) <- c("Date", "Files") #Header umbenennen
Validation$Date <- as.character(Validation$Date)
Validation$Date <- as.Date(Validation$Date, "%Y-%m-%d")
head(Validation)

tt <- with(Validation, table(Date))
data.frame(count = tt[tt > 2]) #Prüft an welchen Tagen mehrere Messungen vorgenommen wurden
#am 25.09.2013 wurden 3 Messungen gemacht, am 29.11.2013 und 20.12.2013 je 2

Validation <- Validation[-c(13:16), ] #2 Messungen vom 25.09.2013 eliminieren
Validation <- Validation[-c(133:134), ] #1 Messung vom 29.11.2013 eliminieren
Validation <- Validation[-c(175:176), ] #1 Messung vom 20.12.2013 eliminieren

#Age-Variable kreieren
Validation$December2011 <- c("2011-12-01") #Start IATI
Validation$December2011 <- as.Date(Validation$December2011, "%Y-%m-%d")
Validation$Age <- ((Validation$Date - Validation$December2011)/365) #Age in Jahre anzeigen

Validation <- data.table(Validation)

Validation <- Validation[, list(Date, Age, Files,Diff=diff(Date))] #Reihenfolge
```

```
Variablen ändern, Differenz Datum berechnen & Variable December2011 rausnehmen
ValidationNew <- subset(Validation, Diff>0) #Variable bilden bei der Diff > 0
ValidationNew2 <- subset(Validation, Diff==0) #Variable bilden bei der Diff = 0

Validation <- cbind(ValidationNew, ValidationNew2$Files)#Spalte mit invaliden Files dem Datensatz hinzufügen
Validation$Diff <- NULL #Spalte Differenz rausnehmen
names(Validation) <- c("Date","Age", "Files_pass", "Files_fail") #Header umbenennen
Validation$Files_pass <-
as.numeric(levels(Validation$Files_pass))[Validation$Files_pass] #Spalte numerisch machen
Validation$Files_fail <-
as.numeric(levels(Validation$Files_fail))[Validation$Files_fail] #Spalte numerisch machen
head(Validation)

Validation <- Validation[-c(757:758), ] #Zwei letzten Zeilen rausnehmen

Validation$Sum_Files <- Validation$Files_pass + Validation$Files_fail #Anzahl Files berechnen
Validation$Validity_Files <- Validation$Files_pass / Validation$Sum_Files #Anteil valider Dokumente berechnen

write.csv(Validation, file = "Validation_Files.csv")
-----------------------------------------------------------------------------------------------------
#Bearbeitung Datensatz Anzahl Organisationen mit nur validen Dokumenten -> Ablauf analog oben

#File "convertcsv_publishers" einlesen (import Dataset)
read.csv("convertcsv_publishers.csv", header=T)
head(convertcsv_publishers)
Validation_publishers <- t(convertcsv_publishers)
Validation_publishers <- as.data.frame(substr(Validation_publishers, 1, 10))
names(Validation_publishers) <- c("Date", "Publishers")
Validation_publishers$Date <- as.character(Validation_publishers$Date)
Validation_publishers$Date <- as.Date(Validation_publishers$Date, "%Y-%m-%d")
head(Validation_publishers)

tq <- with(Validation_publishers, table(Date))
data.frame(count = tq[tq > 2])
#Prüft an welchen Tagen mehrere Messungen vorgenommen wurden
# am 25.09.2013 wurden 3 Messungen gemacht, am 29.11.2013 und 20.12.2013 2.

#2 Messungen vom 25.09.2013 eliminieren
Validation_publishers <- Validation_publishers[-c(13:16), ]
#1 Messung vom 29.11.2013 eliminieren
Validation_publishers <- Validation_publishers[-c(133:134), ]
#1 Messung vom 20.12.2013 eliminieren
Validation_publishers <- Validation_publishers[-c(175:176), ]
```

```
Validation_publishers$Diff #sieht in welchem Tagesabstand die Daten erhoben wur-
den.

Validation_publishers <- data.table(Validation_publishers)
Validation_publishers <- Validation_publishers[,
list(Date,Publishers,Diff=diff(Date))]
Validation_publishersNew <- subset(Validation_publishers, Diff>0)
Validation_publishersNew2 <- subset(Validation_publishers, Diff==0)
Validation_publishers <- cbind(Validation_publishersNew, Valida-
tion_publishersNew2$Publishers)


Validation_publishers$Diff <- NULL
names(Validation_publishers) <- c("Date", "Publisher_pass", "Publisher_fail")


Validation_publishers$Publisher_pass <-
as.numeric(levels(Validation_publishers$Publisher_pass))[Validation_publishers$P
ublisher_pass]
Validation_publishers$Publisher_fail <-
as.numeric(levels(Validation_publishers$Publisher_fail))[Validation_publishers$Pu
blisher_fail]
head(Validation_publishers)

#Anzahl Organisationen berechnen
Validation_publishers$Sum_Publishers <- Validation_publishers$Publisher_pass +
Validation_publishers$Publisher_fail

#Anteil Organisationen mit nur validen Dokumenten berechnen
Validation_publishers$Validity_Publisher <- Validation_publishers$Publisher_pass /
Validation_publishers$Sum_Publishers

Validation_publishers <- Validation_publishers[-c(652), ]
Validation_publishers <- Validation_publishers[-c(757:765), ]

Validation_publishers$Diff_Sum_pub <-
diff(Validation_publishers$Sum_Publishers)

write.csv(Validation_publishers , file = "Validation_Publisher.csv")

-----------------------------------------------------------------------------------------------
#CSV-Files zusammenführen
#Files "Validation_Publisher.csv" & "Validation_Files.csv" einlesen
read.csv("Validation_Publisher.csv", header=T)
read.csv("Validation_Files.csv", header=T)
m <- merge(Validation_Publisher, Validation_Files, by="Date")
#Datensätze zusammenfügen
m <- m [,-c(2, 8) ]

write.csv(m , file = "Zusammengefasster_Datensatz.csv")
-----------------------------------------------------------------------------------------------
#Statistische Ergebnisse
```

```
#Graphik Validität & Alter IATI
ggplot(m, aes(x=Age)) +
  geom_point(aes(y=Validity_Publisher, color="Validität_Organisation")) +
  geom_point(aes(y=Validity_Files, color="Validität_Files")) +
  geom_smooth(se=F,method="lm") +
  labs(x = "Alter IATI (in Jahre)") +
  labs(y = "Validität") +
  labs(title = "Modell 1 & 4: Validität ~ Alter IATI") +
  theme(axis.text.x=element_text(angle = 40, hjust = 1),
        legend.title=element_blank()) +
  scale_y_continuous(limits=c(0, 1), breaks = c(0, 0.2, 0.4, 0.6, 0.8, 1))

#Graphik Validität & Anzahl Files
ggplot(m, aes(x=Sum_Files)) +
  geom_point(aes(y=Validity_Publisher, color="Validität_Organisation")) +
  geom_point(aes(y=Validity_Files, color="Validität_Files")) +
  geom_smooth(se=F, method="lm") +
  labs(x = "Anzahl Files") +
  labs(y = "Validität") +
  labs(title = "Modell 2 & 5: Validität ~ Anzahl Files") +
  theme(axis.text.x=element_text(angle = 40, hjust = 1),
        legend.title=element_blank()) +
  scale_y_continuous(limits=c(0, 1), breaks = c(0, 0.2, 0.4, 0.6, 0.8, 1))

#Graphik Validität & Anzahl Organisationen
ggplot(m, aes(x=Sum_Publishers)) +
  geom_point(aes(y=Validity_Publisher, color= "Validität_Organisation")) +
  geom_point(aes(y=Validity_Files, color="Validität_Files")) +
  geom_smooth(se=F,method="lm") +
  labs(x = "Anzahl Organisationen") +
  labs(y = "Validität") +
  labs(title = "Modell 3 & 6: Validität ~ Anzahl Organisationen") +
  theme(axis.text.x=element_text(angle = 40, hjust = 1),
        legend.title=element_blank()) +
  scale_y_continuous(limits=c(0, 1), breaks = c(0, 0.2, 0.4, 0.6, 0.8, 1))

#Graphik Anzahl Organisationen über die Zeit
ggplot(Validation_publishers, aes(x=Date, y=Sum_Publishers)) +
  geom_point() +
  geom_smooth(se=F,method="lm") +
  labs(x = "Zeit") +
  labs(y = "Anzahl Organisationen") +
  labs(title = "Modell 9: Anzahl Organisationen über die Zeit")
theme(axis.text.x=element_text(angle = 40, hjust = 1))

-----------------------------------------------------------------------------------------
#Modelle testen:
mymodel1 <- lm(Validity_Files ~ Age, data = m, na.action = na.exclude)
summary(mymodel1)
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel1)
```

```
# Robuste Standardfehler
rob.se1 <- sqrt(diag(vcovHC(mymodel1)))
# OLS Standardfehler
OLS.se1 <- sqrt(diag(vcov(mymodel1)))
stargazer(mymodel1,mymodel1, se=list(OLS.se1, rob.se1),
        title="Modell 1",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)


mymodel2 <- lm(Validity_Files ~ Sum_Files, data = m, na.action = na.exclude)
summary(mymodel2)
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel2)
# Robuste Standardfehler
rob.se2 <- sqrt(diag(vcovHC(mymodel2)))
# OLS Standardfehler
OLS.se2 <- sqrt(diag(vcov(mymodel2)))
stargazer(mymodel2,mymodel2, se=list(OLS.se2, rob.se2),
        title="Modell 2",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)


mymodel3 <- lm(Validity_Files ~ Sum_Publishers, data = m, na.action = na.exclude)
summary(mymodel3)
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel3)
# Robuste Standardfehler
rob.se3 <- sqrt(diag(vcovHC(mymodel3)))
# OLS Standardfehler
OLS.se3 <- sqrt(diag(vcov(mymodel1)))
stargazer(mymodel3,mymodel3, se=list(OLS.se3, rob.se3),
        title="Modell 3",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)
```

```
mymodel4 <- lm(Validity_Publisher ~ Age, data = m, na.action = na.exclude)
summary(mymodel4)
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel4)
# Robuste Standardfehler
rob.se4 <- sqrt(diag(vcovHC(mymodel4)))
# OLS Standardfehler
OLS.se4 <- sqrt(diag(vcov(mymodel4)))
stargazer(mymodel1,mymodel4, se=list(OLS.se4, rob.se4),
        title="Modell 4",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)


mymodel5 <- lm(Validity_Publisher ~ Sum_Files, data = m, na.action = na.exclude)
summary(mymodel5)
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel5)
# Robuste Standardfehler
rob.se5 <- sqrt(diag(vcovHC(mymodel5)))
# OLS Standardfehler
OLS.se5 <- sqrt(diag(vcov(mymodel5)))
stargazer(mymodel5,mymodel5, se=list(OLS.se5, rob.se5),
        title="Modell 5",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)


mymodel6 <- lm(Validity_Publisher ~ Sum_Publishers, data = m, na.action =
na.exclude)
summary(mymodel6)
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel6)
# Robuste Standardfehler
rob.se6 <- sqrt(diag(vcovHC(mymodel6)))
# OLS Standardfehler
OLS.se6 <- sqrt(diag(vcov(mymodel6)))
stargazer(mymodel6,mymodel6, se=list(OLS.se6, rob.se6),
        title="Modell 6",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
```

```
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)


mymodel7 <- lm(Validity_Files ~ Age + Sum_Files + Sum_Publishers, data = m,
na.action = na.exclude)
summary(mymodel7)
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel7)
# Robuste Standardfehler
rob.se7 <- sqrt(diag(vcovHC(mymodel7)))
# OLS Standardfehler
OLS.se7 <- sqrt(diag(vcov(mymodel7)))

stargazer(mymodel7,mymodel7, se=list(OLS.se7, rob.se7),
        title="Modell 7",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)

mymodel8 <- lm(Validity_Publisher ~ Age + Sum_Files + Sum_Publishers, data =
m, na.action = na.exclude)
summary(mymodel8)
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel8)
# Robuste Standardfehler
rob.se8 <- sqrt(diag(vcovHC(mymodel8)))
# OLS Standardfehler
OLS.se8 <- sqrt(diag(vcov(mymodel8)))
stargazer(mymodel8,mymodel8, se=list(OLS.se8, rob.se8),
        title="Modell 8",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)


mymodel9 <- lm(Sum_Publishers ~ Age, data = m, na.action = na.exclude)
summary(mymodel9)
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel9)
```

```
# Robuste Standardfehler
rob.se9 <- sqrt(diag(vcovHC(mymodel9)))
# OLS Standardfehler
OLS.se9 <- sqrt(diag(vcov(mymodel9)))
stargazer(mymodel9,mymodel9, se=list(OLS.se9, rob.se9),
        title="Modell 9",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)


#Vergleich der Mittelwerte
t.test(m$Validity_Publisher, m$Validity_Files, conf.level = 0.99)
#Standardabweichung Anteil Organisationen mit validen Dokumenten
sd(m$Validity_Publisher)
#Standardabweichung Anteil valider Dokumente
sd(m$Validity_Files)
```

3.3.2 R-Code: Analyse auf Stufe Organisation

```
#Auswertung Verteilung Anzahl valider und invalider Dokumente pro Organisation
#Packages
install.packages("stargazer")
library(stargazer)
install.packages ("sandwich")
library(sandwich)
install.packages("lmtest")
library(lmtest)
install.packages("ggplot2")
library(ggplot2)
install.packages("data.table")
library(data.table)
install.packages ("zoo")
library(zoo)
-----------------------------------------------------------------------------------------------
#Datensatz einlesen und bereinigen

#CSV File "convertcsv_files.per.organisation.csv" einlesen (Import Dataset)
read.csv("convertcsv_files.per.organisation.csv", header=T)
head(convertcsv_files.per.organisation) #File anzeigen
x <- t(convertcsv_files.per.organisation) #Zeilen und Spalten vertauschen
x <- as.data.frame(x) #Als Dataframe abspeichern
names(x) <- c("Organisation", "Files") #Header umbenennen
head(x)
x$Id <- seq(1,574) # Zahlen von 1-574 fÜr Id
x <- x[,c(3,1:2)]  #Id-Variable soll am Anfang stehen zur besseren Übersicht
```

```
xFail <- subset(x, Id < 144) #Variable mit ID kleiner als 144
xPass <- subset(x, Id > 143) #Variable mit ID grösser als 143

#Beschriftung Beobachtungen auf Organisationsnamen reduzieren
xFail$Organisation <- substr(xFail$Organisation, 6, 30)

#Beschriftung Beobachtungen auf Organisationsnamen reduzieren
xPass$Organisation <- substr(xPass$Organisation, 6, 30)


xFail$Files <- as.numeric(as.character(xFail$Files))
xPass$Files <- as.numeric(as.character(xPass$Files))

names(xFail) <- c("Id", "Organisation", "Files_Fail")
names(xPass) <- c("Id", "Organisation", "Files_Pass")

write.csv(xFail, file = "Files_Fail_per_Organisation.csv")
write.csv(xPass, file = "Files_Pass_per_Organisation.csv")

xFail <- as.data.frame(xFail)
xPass <- as.data.frame(xPass)

#die beiden Dataframes mittels der gemeinsamen Variable Oraganisation verknüp-
fen
#es werden nur die Organisationen angezeigt, die valide und nicht valide Dokumente
haben
xm=merge(xPass, xFail, by="Organisation")

xm <- xm [,-c(2, 4)] #unerwünschte Spalten rausnehmen

xPassNew <- xPass [,-c(1)] #unerwünschte Spalte rausnehmen
xPassNew$Files_Fail <- seq(0, 0) #Spalte mit Nullwerten hinzufügen

xFailNew <- xFail [,-c(1)] #unerwünschte Spalte rausnehmen
xFailNew$Files_Pass <- seq(0, 0) #Spalte mit Nullwerten hinzufügen

#y weist nun einige Organisationen doppelt auf (diejenigen die in xm und xFail-
New/xPassNew vorkommen)
y <- rbind(xm, xFailNew, xPassNew)


#eliminiert diejenigen Organisationen die bereits in xm enthalten sind
q<- y[ !(y$Organisation %in% xm$Organisation), ]


Final <- rbind(xm,q) #fertiger Datensatz

#Variable Summe Files hinzufügen
Final$Sum <- Final$Files_Pass + Final$Files_Fail

#Variable Anteil valider Dokumente hinzufügen
Final$Validity <- Final$Files_Pass / Final$Sum
```

```
write.csv(Final , file = "Final.csv") #bereinigter Datensatz als CSV abspeichern
-----------------------------------------------------------------------------------------------------
#Statistische Analyse

#Berechnung der Mittelwerte und Standardabweichungen
mean(Final$Files_Fail) #1.759
sd(Final$Files_Fail) #12.704
mean(Final$Files_Pass) #7.6312
sd(Final$Files_Pass) #26.176

#Vergleich der Mittelwerte
t.test(Final$Files_Fail, Final$Files_Pass, conf.level = 0.99) #Die Mittelwerte sind
signifikant verschieden von 0 auf dem 99% Konfidenzintervall
#Die Anzahl valider Files ist signifikant höher als die Anzahl nicht valider Files

#Mittelwert Anteil valider Dokumente
mean(Final$Validity)

#Graphik Anzahl valider Files vs. nicht valider Files
ggplot(Final, aes(x=Files_Pass, y=Files_Fail)) +
  geom_point() +
  labs(x = "Anzahl valider Files") +
  labs(y = "Anzahl nicht valider Files") +
  labs(title = "Modell 10: Verteilung Files pro Organisation") +
  theme(axis.text.x=element_text(angle = 40, hjust = 1)) +
  geom_text(aes(label = Organisation), angle=45, alpha=0.6,
            hjust = 1, vjust = 1.5, size = 3) +
  scale_x_continuous(limits=c(-20,280), breaks = c(0, 20, 40, 60, 80, 100, 120, 140,
160, 180, 200, 220, 240, 260)) +
  scale_y_continuous(limits=c(-80,240), breaks = c(0, 20, 40, 60, 80, 100, 120, 140,
160, 180, 200, 220, 240))
-----------------------------------------------------------------------------------------------------
#Analyse Länge der Teilnahme an IATI:

#CSV "Final_Teilnahme_.IATI.csv" einlesen, manuell wurde das Datum der ersten
Publikation im Excel-File Final ergänzt

read.csv("Final_Teilnahme_.IATI.csv", header=T)
head(Final_Teilnahme_.IATI)
FinalIATI <- Final_Teilnahme_.IATI [,-c(1)]

#Datum konvertieren
FinalIATI$First_Publication <- gsub(".","/",FinalIATI$First_Publication,fixed=T)
FinalIATI$First_Publication <- strp-
time(as.character(FinalIATI$First_Publication), "%d/%m/%Y")
FinalIATI$First_Publication <- as.Date(FinalIATI$First_Publication, "%Y-%m-
%d")
 mean(FinalIATI$Sum) # Durchschnittlich hat jede Organisation 9.390456 Files
hochgeladen
```

```
FinalIATI <- subset(FinalIATI, Sum>1) #Organisationen rausnehmen, die nur 1 File hochgeladen haben

#Graphik Datum erstes Upload
library(ggplot2)
ggplot(FinalIATI, aes(x=First_Publication, y=Validity)) +
  geom_point() +
  geom_smooth(se=F,method="lm") +
  labs(x = "Datum erstes Upload") +
  labs(y = "Validität") +
  labs(title = "Modell 11: Validität ~ Datum erstes Upload") +
  theme(axis.text.x=element_text(angle = 40, hjust = 1)) +
  scale_y_continuous(limits=c(0, 1), breaks = c(0, 0.2, 0.4, 0.6, 0.8, 1))

# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel11)
# Robuste Standardfehler
rob.se11 <- sqrt(diag(vcovHC(mymodel11)))
# OLS Standardfehler
OLS.se11 <- sqrt(diag(vcov(mymodel11)))
stargazer(mymodel11,mymodel11, se=list(OLS.se11, rob.se11),
          title="Modell 11",
          no.space=TRUE,
          omit.stat=c("LL","ser","f", "rsq"),
          column.labels=c("OLS SE", "Robust SE"),
          dep.var.caption="",
          type="text",
          intercept.bottom=FALSE,
          model.numbers=FALSE)

#Untersuchung Signifikanz des Modells mittels Einschränkung der Bedingungen

#Anzahl Dokumente sollte grösser als 5 sein.
FinalIATI_Test <- subset(FinalIATI, Sum> 5)
mymodelx <- lm(Validity ~ First_Publication, data = FinalIATI_Test, na.action =
na.exclude)
summary(mymodelx) #zeigt keine Signifikanz

#Anteil valider Dokumente sollte grösser als 0 und kleiner als 1 sein.
FinalIATI_Test2 <- subset(FinalIATI, Validity < 1 & Validity > 0 )
mymodely <- lm(Validity ~ First_Publication, data = FinalIATI_Test, na.action =
na.exclude)
summary(mymodely) #zeigt keine Signifikanz
```

3.3.3 R-Code: Analyse Kontrollvariablen

```
#Auswertung der Kontrollvariablen
#Packages
install.packages("stargazer")
library(stargazer)
install.packages ("sandwich")
library(sandwich)
install.packages("lmtest")
library(lmtest)
install.packages("ggplot2")
library(ggplot2)
install.packages("data.table")
library(data.table)
install.packages ("zoo")
library(zoo)
-----------------------------------------------------------------------------------------------
#Bereinigung CSV-File

#EC_DEVCO
#CSV "ec-devco.csv" in R einlesen
read.csv("ec.devco.csv", header=T)
head(ec.devco)
ec.devco <- t(ec.devco) #Spalten und Zeilen wechseln
ec.devco <- as.data.frame(substr(ec.devco, 1, 10)) #Datum aus text Beobachtungen
rausfiltern
names(ec.devco) <- c("Date", "Files")
ec.devco$Date <- as.character(ec.devco$Date)
ec.devco$Date <- as.Date(ec.devco$Date, "%Y-%m-%d")
head(ec.devco)

#Prüft an welchen Tagen mehrere Messungen vorgenommen wurden
tt <- with(ec.devco, table(Date))
data.frame(count = tt[tt > 2]) #Keine Doppelmessungen gefunden

#Age-Variable kreieren
ec.devco$December2011 <- c("2011-12-01") #Start IATI
ec.devco$December2011 <- as.Date(ec.devco$December2011, "%Y-%m-%d")
ec.devco$Age <- ((ec.devco$Date - ec.devco$December2011)/365) #Age in Jahre
anzeigen

ec.devco <- data.table(ec.devco)
ec.devco <- ec.devco[, list(Date, Age, Files,Diff=diff(Date))] #Reihenfolge Variab-
len ändern, Differenz Datum berechnen & Variable December2011 rausnehmen
ec.devcoNew <- as.data.frame(subset(ec.devco, Diff>0)) #Variable bilden bei der
Diff > 0
ec.devcoNew2 <- as.data.frame(subset(ec.devco, Diff==0)) #Variable bilden bei
der Diff = 0

#Zwei Data Frames über Variable Date verbinden
ECDEVCO=merge(ec.devcoNew, ec.devcoNew2, by="Date")
```

```
ECDEVCO$Diff.x <- NULL #Unnötige Spalte entfernen
ECDEVCO$Diff.y <- NULL #Unnötige Spalte entfernen
ECDEVCO$Age.y <- NULL #Unnötige Spalte entfernen

names(ECDEVCO) <- c("Date","Age", "Files_pass", "Files_fail") #Header umbenennen

ECDEVCO$Files_pass <- as.numeric(levels(ECDEVCO$Files_pass))[ECDEVCO$Files_pass] #als nummerisch abspeichern
ECDEVCO$Files_fail <- as.numeric(levels(ECDEVCO$Files_fail))[ECDEVCO$Files_fail] #als nummerisch abspeichern

#Anzahl Files berechnen
ECDEVCO$Sum_Files <- ECDEVCO$Files_pass + ECDEVCO$Files_fail
#Anteil valider Dokumente berechnen
ECDEVCO$Validity_Files <- ECDEVCO$Files_pass / ECDEVCO$Sum_Files


write.csv(ECDEVCO, file = "ECDEVCO.csv") #als CSV abspeichern

mean(UNFPA$Sum_Files) #111.4968 Dokumente im Durchschnitt
----------------------------------------------------------------------------------------------------

#Da das ECDEVCO.csv einige Werte nicht in der entsprechenden Formatierung übernommen hat, mussten diese manuell nachgetragen werden

#Neues CSV "ECDEVCO2.csv" mit Ergänzungen reinladen
read.csv("ECDEVCO2.csv", header=T)
head(ECDEVCO2)

ECDEVCO2$Organisation <- c("European Comission") #Organisationsnamen ergänzen
ECDEVCO2 <- ECDEVCO2[,-c(1)] #unnötige Spalte rausnehmen
ECDEVCO2 <- ECDEVCO2[,c(7,1:6)] #Organisationsname an die erste Stelle setzen

write.csv(ECDEVCO2, file = "ECDEVCO3.csv")
----------------------------------------------------------------------------------------------------
#Bereinigung CSV
#UNFPA - analoges Vorgehen wie bei der EC-DEVCO
#CSV "unfpa.csv" in R einlesen
read.csv("unfpa.csv", header=T)
head(unfpa)
unfpa <- t(unfpa) #Spalten und Zeilen ändern
unfpa <- as.data.frame(substr(unfpa, 1, 10)) #Datum aus Text Beobachtungen rausfiltern
names(unfpa) <- c("Date", "Files") #Header unbenennen
unfpa$Date <- as.character(unfpa$Date)
unfpa$Date <- as.Date(unfpa$Date, "%Y-%m-%d")
```

```
head(unfpa)

#Prüfen an welchen Tagen mehrere Messungen vorgenommen wurden
tt <- with(unfpa, table(Date))
data.frame(count = tt[tt > 2])
#am 25.09.2013 wurden 3 Messungen gemacht, am 29.11.2013 und 20.12.2013 je 2

unfpa <- unfpa[-c(13:16), ] #2 Messungen vom 25.09.2013 eliminieren
unfpa <- unfpa[-c(129:130), ] #1 Messung vom 29.11.2013 eliminieren
unfpa <- unfpa[-c(171:172), ] #1 Messung vom 20.12.2013 eliminieren

#Age-Variable kreieren
unfpa$December2011 <- c("2011-12-01") #Start IATI
unfpa$December2011 <- as.Date(unfpa$December2011, "%Y-%m-%d")
unfpa$Age <- ((unfpa$Date - unfpa$December2011)/365)

unfpa <- data.table(unfpa)
unfpa <- unfpa[, list(Date, Age, Files,Diff=diff(Date))] #Reihenfolge variablen ändern, Differenz Datum berechnen & Variable December2011 rausnehmen
unfpaNew <- as.data.frame(subset(unfpa, Diff>0))#Variable bilden bei Diff > 0
unfpaNew2 <- as.data.frame(subset(unfpa, Diff==0)) #Variable bilden bei Diff = 0

#Zwei Data Frames über Variable Date verbinden
UNFPA=merge(unfpaNew, unfpaNew2, by="Date")


UNFPA$Diff.x <- NULL #unnnötige Spalte rausnehmen
UNFPA$Diff.y <- NULL #unnnötige Spalte rausnehmen
UNFPA$Age.y <- NULL #unnnötige Spalte rausnehmen

names(UNFPA) <- c("Date","Age", "Files_pass", "Files_fail") #Header umbenennen

UNFPA$Files_pass <-
as.numeric(levels(UNFPA$Files_pass))[UNFPA$Files_pass] #als nummerisch abspeichern
UNFPA$Files_fail <- as.numeric(levels(UNFPA$Files_fail))[UNFPA$Files_fail]
#als nummerisch abspeichern

UNFPA$Sum_Files <- UNFPA$Files_pass + UNFPA$Files_fail #Variable Anzahl Files berechnen
UNFPA$Validity_Files <- UNFPA$Files_pass / UNFPA$Sum_Files #Variable Anteil valider Dokumente berechnen

write.csv(UNFPA, file = "UNFPA.csv")

mean(UNFPA$Sum_Files) #111.4968 Dokumente im Durchschnitt
-----------------------------------------------------------------------------------------------
#Da das UNFPA.csv einige Werte nicht in der entsprechenden Formatierung übernommen hat, mussten diese manuell nachgetragen werden
```

```
#Neues CSV "UNFPA2.csv" mit Ergänzungen reinladen
read.csv("UNFPA2.csv", header=T)
head(UNFPA2)

UNFPA2$Organisation <- c("United Nations Population Funds") #Organisations-
namen ergänzen
UNFPA2 <- UNFPA2[,-c(1)] #unnötige Spalte rausnehmen
UNFPA2 <- UNFPA2[,c(7,1:6)] #Organisation an erste Stelle setzen

write.csv(UNFPA2, file = "UNFPA3.csv")
-----------------------------------------------------------------------------------------------
#Bereinigung CSV
#United States - analoges Vorgehen wie bei den anderen beiden Organisationen
#CSV "unitedstates.csv" in R einlesen
read.csv("unitedstates.csv", header=T)
head(unitedstates)
UnitedStates <- t(unitedstates) #Spalten und Zeilen ändern
UnitedStates <- as.data.frame(substr(UnitedStates, 1, 10)) #Datum aus Text Be-
obachtungen rausfiltern
names(UnitedStates) <- c("Date", "Files") #Header umbenennen
UnitedStates$Date <- as.character(UnitedStates$Date)
UnitedStates$Date <- as.Date(UnitedStates$Date, "%Y-%m-%d")
head(UnitedStates)

#Prüft an welchen Tagen mehrere Messungen vorgenommen wurden
tt <- with(UnitedStates, table(Date))
data.frame(count = tt[tt > 2])
#am 25.09.2013 wurden 3 Messungen gemacht, am 29.11.2013 und 20.12.2013 je 2

UnitedStates <- UnitedStates[-c(13:16), ] #2 Messungen vom 25.09.2013 eliminieren
UnitedStates <- UnitedStates[-c(123:124), ] #1 Messung vom 29.11.2013 eliminieren
UnitedStates <- UnitedStates[-c(165:166), ] #1 Messung vom 20.12.2013 eliminieren

#Age-Variable kreieren
UnitedStates$December2011 <- c("2011-12-01") #Start IATI
UnitedStates$December2011 <- as.Date(UnitedStates$December2011, "%Y-%m-
%d")
UnitedStates$Age <- ((UnitedStates$Date - UnitedStates$December2011)/365) #Age
in Jahre anzeigen

UnitedStates <- data.table(UnitedStates)
UnitedStates <- UnitedStates[, list(Date, Age, Files,Diff=diff(Date))] #Reihenfolge
variablen ändern, Differenz Datum berechnen & Variable December2011 rausneh-
men
UnitedStatesNew <- as.data.frame(subset(UnitedStates, Diff>0)) #Variable bilden
bei der Diff > 0
UnitedStatesNew2 <- as.data.frame(subset(UnitedStates, Diff==0)) #Variable bil-
den bei der Diff = 0
```

```
#Zwei Data Frames über Variable Date verbinden
US=merge(UnitedStatesNew, UnitedStatesNew2, by="Date")
US$Diff.x <- NULL #unnötige Spalte rausnehmen
US$Diff.y <- NULL #unnötige Spalte rausnehmen
US$Age.y <- NULL #unnötige Spalte rausnehmen

names(US) <- c("Date","Age", "Files_pass", "Files_fail") #Header unbenennen

US$Files_pass <- as.numeric(levels(US$Files_pass))[US$Files_pass] #als nummerisch abspeichern
US$Files_fail <- as.numeric(levels(US$Files_fail))[US$Files_fail] #als nummerisch abspeichern

US$Sum_Files <- US$Files_pass + US$Files_fail #Variable Anzahl Files berechnen
US$Validity_Files <- US$Files_pass / US$Sum_Files #Variable Anteil valider Dokumente berechenen

write.csv(US, file = "UnitedStates_Val.csv")

mean(US$Sum_Files) #303.533 Dokumente im Durchschnitt
-----------------------------------------------------------------------------------------------
#Da das UnitedStates_Val.csv einige Werte nicht in der entsprechenden Formatierung übernommen hat, mussten diese manuell nachgetragen werden


#Neues CSV "UnitedStates_Val2.csv" mit Ergänzungen reinladen
read.csv("UnitedStates_Val2.csv", header=T)

US2 <- UnitedStates_Val2[,-c(8:10)] #unnötige Spalten rausnehmen
head(US2)

US2$Organisation <- c("United States") #Organisationsnamen ergänzen
US2 <- US2[,-c(1)] #unnötige Spalte rausnehmen
US2 <- US2[,c(7,1:6)] #Organisation an erste Stelle setzen

write.csv(US2, file = "US3.csv")
-----------------------------------------------------------------------------------------------
#In diesem Teil werden die einzelnen Organisationen zusammengefügt, um sie untereinander vergleichen zu können.
#CSV-Files einlesen: "ECDEVCO3.csv", "UNFPA3.csv", "US3.csv"
read.csv("ECDEVCO3.csv", header=T)
read.csv("UNFPA3.csv", header=T)
read.csv("US3.csv", header=T)

UNFPA3 <- UNFPA3[,-c(1)] #unnnötige Spalte löschen
ECDEVCO3 <- ECDEVCO3[,-c(1)] #unnnötige Spalte löschen
US3 <- US3[,-c(1)] #unnnötige Spalte löschen

USUNFEC <- rbind(US3, UNFPA3, ECDEVCO3) #Organisationen verbinden

#Graphik Organisationen und Alter IATI
```

```
ggplot(USUNFEC, aes(x=Age, y=Validity_Files, colour = Organisation)) +
  geom_point() +
  geom_smooth(se=F,method="lm") +
  labs(x = "Alter IATI (in Jahre)") +
  labs(y = "Validität") +
  labs(title = "Modell 12: Validität ~ Alter IATI auf Stufe 3 Organisationen") +
  theme(axis.text.x=element_text(angle = 40, hjust = 1)) +
  scale_y_continuous(limits=c(0, 1), breaks = c(0, 0.2, 0.4, 0.6, 0.8, 1))

#Modelle testen

mymodel12a <- lm(Validity_Files ~ Age, data = ECDEVCO3, na.action =
na.exclude)
summary(mymodel12a) #ec-devco
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel12a)
# Robuste Standardfehler
rob.se12a <- sqrt(diag(vcovHC(mymodel12a)))
# OLS Standardfehler
OLS.se12a <- sqrt(diag(vcov(mymodel12a)))
stargazer(mymodel12a,mymodel12a, se=list(OLS.se12a, rob.se12a),
        title="Modell 12a",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)

mymodel12b <- lm(Validity_Files ~ Age, data = UNFPA3, na.action = na.exclude)
summary(mymodel12b) #unfpa
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel12b)
# Robuste Standardfehler
rob.se12b <- sqrt(diag(vcovHC(mymodel12b)))
# OLS Standardfehler
OLS.se12b <- sqrt(diag(vcov(mymodel12b)))
stargazer(mymodel12b,mymodel12b, se=list(OLS.se12b, rob.se12b),
        title="Modell 12b",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)

mymodel12c <- lm(Validity_Files ~ Age, data = US3, na.action = na.exclude)
summary(mymodel12c) #unitedstates
```

```
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel12c)
# Robuste Standardfehler
rob.se12c <- sqrt(diag(vcovHC(mymodel12c)))
# OLS Standardfehler
OLS.se12c <- sqrt(diag(vcov(mymodel12c)))
stargazer(mymodel12c,mymodel12c, se=list(OLS.se12c, rob.se12c),
          title="Modell 12c",
          no.space=TRUE,
          omit.stat=c("LL","ser","f", "rsq"),
          column.labels=c("OLS SE", "Robust SE"),
          dep.var.caption="",
          type="text",
          intercept.bottom=FALSE,
          model.numbers=FALSE)

#Graphik Organisationen und Anzahl Files
ggplot(USUNFEC, aes(x=Sum_Files, y=Validity_Files, colour = Organisation)) +
  geom_point() +
  geom_smooth(se=F,method="lm") +
  labs(x = "Anzahl Files") +
  labs(y = "Validitätt") +
  labs(title = "Modell 13: Validität ~ Anzahl Files auf Stufe 3 Organisationen") +
  theme(axis.text.x=element_text(angle = 40, hjust = 1)) +
  scale_y_continuous(limits=c(0, 1), breaks = c(0, 0.2, 0.4, 0.6, 0.8, 1))

mymodel13a <- lm(Validity_Files ~ Sum_Files, data = ECDEVCO3, na.action =
na.exclude)
summary(mymodel13a) #ec-devco
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel13a)
# Robuste Standardfehler
rob.se13a <- sqrt(diag(vcovHC(mymodel13a)))
# OLS Standardfehler
OLS.se13a <- sqrt(diag(vcov(mymodel13a)))
stargazer(mymodel13a,mymodel13a, se=list(OLS.se13a, rob.se13a),
          title="Modell 13a",
          no.space=TRUE,
          omit.stat=c("LL","ser","f", "rsq"),
          column.labels=c("OLS SE", "Robust SE"),
          dep.var.caption="",
          type="text",
          intercept.bottom=FALSE,
          model.numbers=FALSE)

mymodel13b <- lm(Validity_Files ~ Sum_Files, data = UNFPA3, na.action =
na.exclude)
summary(mymodel13b) #unfpa
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel13b)
```

```
# Robuste Standardfehler
rob.se13b <- sqrt(diag(vcovHC(mymodel13b)))
# OLS Standardfehler
OLS.se13b <- sqrt(diag(vcov(mymodel13b)))
stargazer(mymodel13b,mymodel13b, se=list(OLS.se13b, rob.se13b),
        title="Modell 13b",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)

mymodel13c <- lm(Validity_Files ~ Sum_Files, data = US3, na.action = na.exclude)
summary(mymodel13) #unitedstates
# Breusch-Pagan Test auf Heteroskedastizität
bptest(mymodel13c)
# Robuste Standardfehler
rob.se13c <- sqrt(diag(vcovHC(mymodel13c)))
# OLS Standardfehler
OLS.se13c <- sqrt(diag(vcov(mymodel13c)))
stargazer(mymodel13c,mymodel13c, se=list(OLS.se13c, rob.se13c),
        title="Modell 13c",
        no.space=TRUE,
        omit.stat=c("LL","ser","f", "rsq"),
        column.labels=c("OLS SE", "Robust SE"),
        dep.var.caption="",
        type="text",
        intercept.bottom=FALSE,
        model.numbers=FALSE)
```

Abbildungsverzeichnis

Abbildung 1: Open Data Prozess nach Zuiderwijk et al. (2012, S. 157). 12
Abbildung 2: Fünf Sterne Modell nach Bauer & Kaltenböck (2012). 14
Abbildung 3: Informationsidentifizierung nach Ren & Glissmann (2012, S. 96). 16
Abbildung 4: Qualität verlinkter Daten nach Zaveri et al. (2012, S. 6). 25
Abbildung 5: Infrastruktur und Ökosystem von IATI nach Davies (2012, S. 3). 32
Abbildung 6: Organisation-Standard nach IATI (2015b). 34
Abbildung 7: Aktivitäten-Standard nach IATI (2015b). 35
Abbildung 8: Validität und Alter IATI - Modelle 1 & 4 45
Abbildung 9: Validität und Anzahl Files - Modelle 2 & 5. 47
Abbildung 10: Validität und Anzahl Organisationen - Modelle 3 & 6. 49
Abbildung 11: Entwicklung Anzahl Organisationen über die Zeit. 52
Abbildung 12: Verteilung valide vs. invalide Dokumente. 56
Abbildung 13: Validität und Datum erstes Upload. 59
Abbildung 14: Validität und Alter IATI, Vergleich drei Organisationen. 61
Abbildung 15: Validität und Anzahl Files, Vergleich drei Organisationen. 63

Tabellenverzeichnis

Tabelle 1: Informationsqualität: Akademische Sicht nach Lee et al. (2002, S. 134). 21
Tabelle 2: Informationsqualität: Praktische Sicht nach Lee et al. (2002, S. 136). 21
Tabelle 3: Resultate Modelle 1 & 4 45
Tabelle 4: Resultate Modelle 2 & 5 47
Tabelle 5: Resultate Modelle 3 & 6 49
Tabelle 6: Resultate Modelle 7 & 8 51
Tabelle 7: Resultate Modell 9 52
Tabelle 8: Vergleich Validität_Files vs. Validität_Organisation 53
Tabelle 9: Vergleich Anzahl valider vs. invalider Dokumente 56
Tabelle 10: Resultate Modell 11 59
Tabelle 11: Resultate Modell 12a 61
Tabelle 12: Resultate Modell 12b 61
Tabelle 13: Resultate Modell 12c 62
Tabelle 14: Resultate Modell 13a 63
Tabelle 15: Resultate Modell 13b 64
Tabelle 16: Resultate Modell 13c 64
Tabelle 17: Übersicht Resultate 66

Abkürzungsverzeichnis

API	Application Programming Interface
CC-BY	Creative Common Attribution License
CRS	Creditor Reporting System
CSV	Comma-Separated Values
DAC	Development Assistance Commitee
DoD	Departements of Defense
DTD	Dokumenten Typ Definition
ec-devco	European Comission - Development and Cooperation
EDA	Eidgenössisches Departement für auswertige Angelegenheiten
EU	Europäische Union
GSMA	Groupe Speciale Mobile Association
IS	Informationssystem
JSON	JavaScript Object Notation
MDI	Mobile and Development Intelligence
NGO	Nichtregierungsorganisationen
OLS	Ordinary least squares
PDF	Portable Document Format
RDF	Resource Description Framework
TAG	Technical Advisory Group
UN	Vereinte Nationen
unfpa	United Nations Population Fund
unitedstates	Untied States
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USAID	United States Agency for International Development
USD	US-Dollar
XML	Extensible Markup Language

Literaturverzeichnis

Barnickel, N., & Klessmann, J. (2012). Open Data - Am Beispiel von Informationen des öffentlichen Sektors. In U. Herb (Hrsg.), *Open Initiatives: Offenheit in der digitalen Welt und Wissenschaft* (S. 127-158). Saarbrücken: universaar.

Bartlett, M. (2014). *Bringing IATI data to life: the d-portal generator*. Abgerufen auf: http://www.aidtransparency.net/news/bringing-iati-data-to-life-the-d-portal-generator [Abruf: 2016-09-08].

Bauer, F., & Kaltenböck, M. (2012). *Linked Open Data: The Essentials – A Quick Start Guide for Decision Makers* (1. Aufl.). Wien: edition mono/monochrom.

Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. (1984). The Problem of Informant Accuracy: The Validity of Retrospective Data. *Annual Review of Anthropology, 13*, 495-517.

Bonabeau, E. (2009). Decisions 2.0: The Power of Collective Intelligence. *MIT Sloan Management Review*. *50*(2), 45-53.

Bracht, U., Geckler, D., & Wenzel, S. (2011). Datenmanagement und Softwarewerkzeugklassen. In U. Bracht, D. Geckler und S. Wenzel (Hrsg.), *Digitale Fabrik – Methoden und Praxisbeispiele* (S. 164-217). Berlin: Springer Verlag.

Busan HL-4. (2011, Dezember). *Busan Partnership for Effective Development Co-Operation.* Fourth High Level Forum on Aid Effectiveness. Abgerufen auf: http://www.oecd.org/dac/effectiveness/49650173.pdf [Abruf: 2016-09-11].

Creative Commons. (2016). *Mehr über die Lizenzen*. Abgerufen auf: https://creativecommons.org/licenses/ [Abruf: 2016-09-07].

Dawes, S. S. (2010). Stewardship and Usefulness: Policy Principles for Information-based Transparency. *Government Information Quarterly*, *27*(4), 377-383.

Davies, T. (2011). Open Data: Infrastructure and Ecosystem. *Open Data Research,* 1-6.

De Cindio, F. (2012). Guidelines for Designing Deliberative Digital Habitats: Learning from e-Participation for Open Data Initiatives. *The Journal of Community Informatics, 8*(2).

DeLone, W. H., & McLean, E. R. (2003) .The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. *Journal of Management Information Systems, 19*(4), 9-30.

Eidgenössisches Departement für auswertige Angelegenheiten (EDA). (2016a). *Grundsätze der Transparenz*. Abgerufen auf: https://www.eda.admin.ch/deza/de/home/aktivitaeten_projekte/grundsaetze-transparenz.html [Abruf: 2016-09-07].

Eidgenössisches Departement für auswertige Angelegenheiten. (2016b). *Institutionelles Lernen*. Abgerufen auf: https://www.eda.admin.ch/deza/de/home/deza/strategie/grundsaetze_der_zusammenarbeit/institutionelleslernenundvernetzung.html [Abruf: 2016-06-04].

European Comission – Development and Cooperation-EuropeAid. (2016). *Aid Transparency*. Abgerufen auf: http://ec.europa.eu/europeaid/policies/eu-approach-aid-effectiveness/aid-transparency_en [Abruf: 2016-09-10].

Guerrini, G., Mesiti, M., & Sorrenti, M. A. (2007). XML Schema Evolution : Incremental Validation and Efficient Document Adaptation. In D. Barbosa, A. Bonifati, Z. Behllahsène, E. Hunt, & R. Unland (Hrsg.), *Database and XML Technologie: 5th International XML Database Symposium* (S. 92-106). Heidelberg: Springer.

GSMA. (2016). *Mobile for Development*. Abgerufen auf: http://www.gsma.com/mobilefordevelopment/ [Abruf: 2016-08-06].

Hartung, C., Lerer, A., Anokwa, Y., Tseng, C., Brunette, W., & Borriello, G. (2010, Dezember). *Open Data Kit: Tools to Build Information Services for Developing Regions*. Artikel präsentiert an der 4th ACM/IEEE International Conference on Information and Communication Technologies and Development. http://dl.acm.org/citation.cfm?doid=2369220.2369236

Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. *European Journal of ePractice*, (12), 1-13.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adaption Barriers and Mythos of Open Data and Open Government. *Information Systems Management (ISM), 29*(4), 258-268.

Juran, J. M. (1998). The Quality Improvement Process. In J. M. Juran,. & A. B. Godfrey (Hrsg.), *Juran's Qualtity Handbook* (S. 1-1730). New York: McGraw-Hill.

International Aid Transparency Initiative (IATI). (2012). *IATI and Aid Transparency*. Abgerufen auf: http://www.aidtransparency.net/wp-content/uploads/2012/11/IATI_4_PAGER_2012_FINAL.pdf [Abruf: 2016-09-08].

International Aid Transparency Initiative (IATI). (2014). *Preparing your Organisation*. Abgerufen auf: http://iatistandard.org/202/guidance/how-to-publish/prepare-your-org/ [Abruf: 2016-09-25].

International Aid Transparency Initiative (IATI). (2015a). *IATI Annual Report 2015*. Abgerufen auf: http://www.aidtransparency.net/annualreport2015/downloads/IATI_Annual_Report_2015.pdf [Abruf: 2016-09-08].

International Aid Transparency Initiative (IATI). (2015b). *Upgrades*. Abgerufen auf: http://iatistandard.org/202/upgrades/all-versions/ [Abruf: 2016-09-08].

International Aid Transparency Initiative (IATI). (2016a). *About.* Abgerufen auf: http://www.aidtransparency.net/about [Abruf: 2016-09-08].

International Aid Transparency Initiative (IATI). (2016b). *Aid Transparency*. Abgerufen auf: http://www.aidtransparency.net [Abruf: 2016-09-08].

International Aid Transparency Initiative (IATI). (2016c). *IATI Governance*. Abgerufen auf: http://www.aidtransparency.net/governance [Abruf: 2016-09-08].

International Aid Transparency Initiative (IATI). (2016d). *CSV2IATI*. Abgerufen auf: http://csv2iati.iatistandard.org/ [Abruf: 2016-06-23].

International Aid Transparency Initiative (IATI). (2016e). *Using IATI Data*. Abgerufen auf: https://iatiregistry.org/using-iati-data [Abruf: 2016-05-09].

International Aid Transparency Initiative (IATI). (2016f). *About.* Abgerufen auf: https://iatiregistry.org/about [Abruf: 2016-05-09].

International Aid Transparency Initiative (IATI). (2016g). *IATI Public Validator.* Abgerufen auf: http://validator.iatistandard.org/ [Abruf: 2016-06-23].

International Aid Transparency Initiative (IATI). (2016h). *Dashboard - Validation Against the Schema: Which files fail schema validation?* Abgerufen auf: http://dashboard.iatistandard.org/validation.html [Abruf: 2016-09-10].

International Aid Transparency Initiative (IATI). (2016i). *Datastore.* Abgerufen auf: http://datastore.iatistandard.org/docs/user-guide/what-is-the-iati-data-store/ [Abruf: 2016-08-17].

International Aid Transparency Initiative (IATI). (2016k). *Reference.* Abgerufen auf: http://iatistandard.org/202/reference/ [2016-09-08].

International Aid Transparency Initiative (IATI). (2016l). *United States.* Abgerufen auf: https://www.iatiregistry.org/publisher/about/unitedstates [Abruf: 2016-09-10].

International Aid Transparency Initiative (IATI). (2016m). *Licence Types.* Abgerufen auf: http://iatistandard.org/102/getting-started/licencing/licence-types/ [Abruf: 2016-09-11].

International Aid Transparency Initiative (IATI). (2016n). *Common Errors*. Abgerufen auf: http://validator.iatistandard.org/common_errors.php [Abruf: 2016-09-11].

International Aid Transparency Initiative (IATI). (2016o). *European Commission - Development and Cooperation- EuropeAid.* Abgerufen auf: https://www.iatiregistry.org/publisher/about/ec-devco [Abruf: 2016-09-18].

International Aid Transparency Initiative (IATI). (2016p). *United Nations Population Fund*. Abgerufen auf: https://www.iatiregistry.org/publisher/about/unfpa [Abruf: 2016-09-18].

International Aid Transparency Initiative (IATI). (2016q). *Introduction*. Abgerufen auf: http://iatistandard.org/202/introduction/ [2016-09-25].

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management, 40*, 133-146.
Leimeister, J. M. (2010). Kollektive Intelligenz. *Wirtschaftsinformatik, 52*(4), 239-242.

Linders, D. (2013). Towards Open Development: Leveraging Open Data to Improve the Planning and Coordination of International Aid. *Government Information Quarterly, 30*(4), 426-434.

Murray-Rust, P. (2008). Open Data in Science. *Serials Review, 34*(1), 52-64.

Nicolaou, A. I., & McKnight, H. (2006). Perceived Information Quality in Data Exchange: Effects on Risk, Trust, and Intention to Use. *Information Systems Research, 17*(4), 332-351.

Ngueira-Budny, D. (2015, Januar 29). The importance of open aid data to open governance. [Web log post]. Abgerufen auf: http://blogs.worldbank.org/governance/importance-open-aid-data-open-governance [Abruf: 2016-09-10].

OECD. (2016a). *Paris Declaration and Accra Agenda for Action*. Abrufbar auf: http://www.oecd.org/dac/effectiveness/parisdeclarationandaccraagendaforaction.htm [Abruf: 2016-09-07].

OECD. (2016b). *The Busan Partnership for Effective Development Cooperation*. Abrufbar auf: http://www.oecd.org/development/effectiveness/busanpartnership.htm [Abruf: 2016-09-07].

Open Knowledge. (2016). *Open Definition 2.1*. Abgerufen auf: http://opendefinition.org/od/2.1/en/ [Abruf: 2016-09-11].

Orr, K. (1998). Data Quality and Systems Theory. *Communications of the ACM, 41*(2), 66-71.

Publish What You Fund. (2016a). *Conclusions.* Abgerufen auf: http://ati.publishwhatyoufund.org/index-2016/conclusions-recommendations/ [Abruf: 2016-09-11].

Publish What You Fund. (2016b). *2016 Index.* Abgerufen auf: http://ati.publishwhatyoufund.org/wp-content/uploads/2016/02/ATI-2016_Report_Proof_DIGITAL.pdf [Abruf: 2016-09-07].

Publish What You Fund. (2016c). *Methodology*. Abgerufen auf: http://ati.publishwhatyoufund.org/approach/methodology/ [Abruf: 2016-09-11].

Publish What You Fund. (2016d). *Switzerland - Swiss Agency for Development and Cooperation*. Abgerufen auf: http://ati.publishwhatyoufund.org/donor/switzerland/ [2016-09-20].

Ren, G.-J., & Glissmann, S. (2012). Identifying Information Assets for Open Data. The Role of Business Architecture and Information Quality. In IEEE Computer Science (Hrsg.), *Proceedings of the 2012 IEEE 14th International Conference on Commerce and Enterprise Computing* (S. 94-100). Washington, DC: IEEE Computer Society Press.

Salkind, N. J. (2011). Internal and External Validity. In L. Moutinho, & G. D. Hutcheson (Hrsg.), *The SAGE Dictionary of Quantitative Management Research* (147-149). SAGE Publications.

Schwegmann, C. (2012, Februar). *Open Data in Developing Countries*. Arbeitsbericht Nr. 2013/02. European Public Information Platform.

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A Framework for Information Quality Assessment. *Journal of the American Society for Information Science and Technology*, *58*(12), 1720-1733.

Sunlight Foundation. (2010). *Ten Principles for Opening Up Government Information. Sunlight Foundation.* Abgerufen auf: https://sunlightfoundation.com/policy/documents/ten-open-data-principles/ [Abruf: 2016-09-24].

Tayi, G. K., Ballou, D. P., & Guest Editors. (1998). Examining Data Quality. *Communication of the ACM, 41*(2), 54-57.

Taylor, R. S. (1986). *Value-added processes in information systems*. Norwood, NJ: Ablex Publishing.

The World Bank Group. (2012). *Governance and Anti Corruption Strategy Update 2012 – Online Feedback Summary*. Abgerufen auf: http://siteresources.worldbank.org/PUBLICSECTORANDGOVERNANCE/Resources/285741-1326816182754/SummaryConsultations.pdf [Abruf: 2016-09-08].

The World Bank. (2016a). *Data Quality and Effectiveness*. Abgerufen auf: https://datahelpdesk.worldbank.org/knowledgebase/articles/906534-data-quality-and-effectiveness [Abruf: 2016-08-06].

The World Bank. (2016b). *What We Do.* Abgerufen auf: http://www.worldbank.org/en/about/what-we-do [Abruf: 2016-08-06].

Uhlir, P. F., & Schröder, P. (2007). Open Data for Global Science. *Data Science Journal, 6*(Open Data Issue), 36-53.

United Nations Population Fund. (2016). *Data Quality and Usability*. Abgerufen auf: http://www.unfpa.org/data-quality-and-usability [Abruf: 2016-09-10].

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5-33.

World Wide Web Consortium (W3C). (2004). *Resource Description Framework (RDF): Concepts and Abstract Syntax.* Abgerufen auf: https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ [Abruf: 2016-09-11].

World Wide Web Consortium (W3C). (2012). *W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures*. Abgerufen auf: https://www.w3.org/TR/xmlschema11-1/ [Abruf: 2016-09-07].

YoungInnovations. (2016). *Aid Stream - About us. Abgerufen auf: https://aidstream.org/about [Abruf: 2016-09-25]*.

Yu, L. (2011). Linked Open Data. In L. Yu (Hrsg.), *A Developer's Guide to the Semantic Web (S. 409-466)*. Heidelberg: Springer.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2012). Quality Assessment Methodologies for Linked Open Data: A Systematic Literature Review and Conceptual Framework. *Undefined,* 1-5.

Zuiderwijk, A., & Janssen, M. (2013). A Coordination Theory Perspective to Improve the Use of Open Data in Policy-Making. In M .A. Wimmer, M. Janssen, & H. J. Scholl (Hrsg.), *EGOV 2013, LNCS 8074* (S. 38-49). Koblenz: IFIP International Federation for Information Processing 2013.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-technical Impediments of Open Data. *Electronic Journal of e-Government. 10*(2), 156-172.

Selbständigkeitserklärung

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe o des Gesetzes vom 5. September 1996 über die Universität zum Entzug des aufgrund dieser Arbeit verliehenen Titels berechtigt ist.“

Handschriftliche Unterschrift

Bern, Datum Vorname Name

Veröffentlichung der Arbeit

(nur für Master- / Lizentiats- /Bachelorarbeit)

I.d.R. werden schriftliche Arbeiten in der Bibliothek des Instituts für Wirtschaftsinformatik öffentlich zugänglich gemacht.

- ☐ Hiermit erlaube ich, meine Arbeit in der Bibliothek des Instituts für Wirtschaftsinformatik zu veröffentlichen.
- ☐ Ich möchte auf eine Veröffentlichung meiner Arbeit verzichten.

Falls eine Vertraulichkeitserklärung unterschrieben wurde, ist es Sache des Studierenden, das Einverständnis des Praxispartners einzuholen. Es muss der Arbeit eine schriftliche Bestätigung des Praxispartners beigelegt werden.

Die Benotung der Arbeit erfolgt unabhängig davon, ob die Arbeit veröffentlicht werden darf oder nicht.

Handschriftliche Unterschrift

Bern, Datum Vorname Name