Bausteine Forschungsdatenmanagement
Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern

Wir Machen Daten FAIR

Die Konzeption von Datenservices im GESIS - Datenarchiv für Sozialwissenschaften

Sebastian Netscher[i] Oliver Watteler[ii] Anja Perry[iii]

2020

Zitiervorschlag

Netscher, Sebastian, Watteler, Oliver und Anja Perry. 2020. Wir Machen Daten FAIR. Die Konzeption von Datenservices im GESIS - Datenarchiv für Sozialwissenschaften. *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern* Nr. 1/2020: S. 45-52. DOI: 10.17192/bfdm.2020.1.8159.

[i]GESIS - Datenarchiv für Sozialwissenschaften. ORCID: 0000-0002-2784-6968
[ii]GESIS - Datenarchiv für Sozialwissenschaften. ORCID: 0000-0002-1634-9229
[iii]GESIS - Datenarchiv für Sozialwissenschaften. ORCID: 0000-0003-0574-9275

1 Abstract

Das *Datenarchiv für Sozialwissenschaften*, eine Abteilung von *GESIS – Leibniz-Institut für Sozialwissenschaften*, besitzt eine über Jahrzehnte hinweg aufgebaute Expertise in der Aufbereitung, Dokumentation und Kuratierung von Daten großer (internationaler) Umfrageprogramme ebenso wie in der Entwicklung und Anwendung internationaler Standards in diesem Rahmen. Steigenden Anforderungen zur Generierung FAIRer (Forschungs-)Daten, etwa durch Forschungsförderer, begegnet das Datenarchiv mit der Bereitstellung der *GESIS Datenservices* rund um die Archivierung quantitativer sozialwissenschaftlicher Umfragedaten. In einem internen Projekt wurden ab 2016 Dienstleistungen des Datenarchivs systematisiert und professionalisiert. So entstand ein Angebotsportfolio (größtenteils) kostenpflichtiger Datenservices, die seit 2019 von allen, die entsprechende Daten generieren, aufbereiten, dokumentieren, aufwerten und archivieren möchten, in Anspruch genommen werden können. Der vorliegende Beitrag skizziert die Konzeption und Bepreisung dieser Datenservices und erörtert die Vorteile derartiger Angebote für Forschende, Förderer und die Forschungsgemeinschaft in der Praxis.

2 Einleitung

In den letzten Jahren gewinnt die Forderung nach nachnutzbaren (Forschungs-)Daten zunehmend an Bedeutung. Damit verbunden ist die Erwartung, die Transparenz in der Forschung zu erhöhen und sowohl Forschungsergebnisse als auch zugehörige Daten replizier- bzw. reproduzierbar zu machen. Die Bereitstellung nachnutzbarer Daten fördert ferner Forschung und Innovation und dient einem effizienten Einsatz von Geldern in der Forschungsförderung. In diesem Kontext sind auch Auflagen, z. B. von Forschungsfördern, zur Generierung von Daten nach den sogenannten *FAIR Data Principles*[1] zu verstehen, die für Dritte auffindbar, zugänglich, interoperabel und (analytisch) nachnutzbar sein sollen. Das Aufgreifen dieser Prinzipien, etwa durch die Europäische Kommission im Rahmen des Förderprogramms *Horizon 2020*[2], hat den Themenkomplex Forschungsdatenmanagement nochmals neu in eine breite interdisziplinäre Diskussion gebracht. Im Unterschied zu allgemeinen Forderungen, wie beispielsweise der *Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen*[3], bietet

[1] Weitere Informationen zu den *FAIR Data Principles* finden sich bei Wilkinson, Mark D. et al. (2016): *The FAIR Guiding Principles for Scientific Data Management and Stewardship*. Science Data No. 3. DOI: 10.1038/sdata.2016.18 sowie bei FORCE11 unter https://www.force11.org/group/fairgroup/fairprinciples [04.12.2019].

[2] Siehe hierzu Artikel 29.3 des *Annotated Model Grant Agreement* des *Horizon 2020 Framework Programme* der EU, unter http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf [04.12.2019].

[3] *Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen* vom 22. Oktober 2003, unter https://openaccess.mpg.de/68053/Berliner_Erklaerung_dt_Version_07-2006.pdf [04.12.2019].

das FAIR-Konzept konkrete Anhaltspunkte, was für einen offenen Austausch von Forschungsdaten zu tun ist. Damit einhergehend wird derzeit auch wieder verstärkt über die Kosten des Forschungsdatenmanagement gesprochen.

Die Generierung FAIRer Daten gestaltet sich jedoch häufig problematisch. Zum einen fehlt den Forschenden oftmals das Verständnis für die Bedeutung entsprechend aufbereiteter Daten. Zum anderen entsteht ein Trade-off zwischen Forschungszeit und der Zeit zur Generierung qualitativ hochwertiger und nachnutzbarer Daten, der zumeist zu Ungunsten der Daten ausfällt. Daraus ergibt sich der Bedarf, Angebote zur Erstellung FAIRer Daten zu etablieren und die Auffindbarkeit, Zugänglichkeit etc. durch Expertinnen und Experten der Datenaufbereitung[4], -dokumentation und -archivierung sicherzustellen. Um Forschende bei der Generierung FAIRer Daten zu unterstützen, konzipierte das *Datenarchiv für Sozialwissenschaften* bei *GESIS – Leibniz-Institut für Sozialwissenschaften*[5] (im Folgenden kurz *Datenarchiv*) ein Angebotsportfolio von Dienstleistungen rund um die Archivierung sozialwissenschaftlicher Umfragedaten, die sogenannten *GESIS Datenservices*.

Der vorliegende Beitrag skizziert die Konzeption dieses Portfolios sowie die Berechnung der Preise für die Datenservices und erörtert die entsprechenden Dienstleistungen in der Praxis. Er trägt dabei auf zwei Arten zur Generierung FAIRer Daten bei: Zum einen wird ein Weg aufgezeigt, entsprechende Dienstleistungen zu konzipieren und zu bepreisen. Zum anderen werden die Datenservices als niedrigschwelliges Angebot vorgestellt, das von Forschenden genutzt werden kann, um eigene Daten verfügbar zu machen. Der Beitrag versucht so auch andere Akteure in die Lage zu versetzen, für die Erbringung entsprechender Leistungen die notwendigen Ressourcen bereitstellen bzw. einwerben zu können.

3 Die Konzeption der *GESIS Datenservices*

Gegründet als Zentralarchiv der Universität zu Köln, sichert das Datenarchiv seit 1960 qualitativ hochwertige sozialwissenschaftliche Daten. Seit den 1980er Jahren liegt ein Schwerpunkt auf der Archivierung und Bereitstellung großer (inter-)nationaler Umfrageprogramme, wie *ALLBUS*[6], *European Value Survey*[7] oder *International Social Survey*

[4] In Abgrenzung zur Generierung FAIRer Forschungsdaten, die den kompletten Forschungsdatenlebenszyklus von der Studienplanung über die Datenerhebung, Aufbereitung und Auswertung bis hin zur Archivierung umfasst, bezieht sich die Datenaufbereitung auf die Bearbeitung und Kontrolle der Rohdaten, z. B. auf inhaltliche Konsistenz, im Vorfeld der Datenanalyse.

[5] Weitere Informationen zu *GESIS – Leibniz-Institut für Sozialwissenschaften* finden sich unter https://www.gesis.org/home/ [04.12.2019].

[6] Weitere Informationen zum *ALLBUS* bei GESIS finden sich unter https://www.gesis.org/allbus/allbus/ [04.12.2019].

[7] Weitere Informationen zum *European Value Survey* bei GESIS finden sich unter https://www.gesis.org/angebot/daten-analysieren/internationale-umfragen/european-values-study/ [04.12.2019].

Programme[8]. Das Datenarchiv ist heute fester Bestandteil der nationalen und internationalen sozialwissenschaftlichen Infrastruktur und hat sowohl systematisch Expertise im Bereich der Datenkuratierung aufgebaut als auch (inter-)nationale Standards mit entwickelt, wie z. B. im Rahmen der *DDI Alliance*[9]. Damit steigert das Datenarchiv etwa die Auffindbarkeit von Daten aus Umfrageprogrammen durch eine standardisierte Dokumentation sowie durch die Vergabe von *Digital Object Identifiern*[10] (DOI) über die Registrierungsagentur *da|ra*[11]. Tabelle 1 gibt einen (unvollständigen) Überblick über unterschiedliche Dienstleistungen des Datenarchivs und deren Zusammenhang mit den FAIR-Prinzipien.

Tabelle 1: FAIRe Dienstleistungen des Datenarchivs (Auswahl)

FAIR-Prinzipien	Dienstleistungen
Findable	- Studienbeschreibung nach DDI-Standard auf Deutsch & Englisch - DOI-Registrierung über *da|ra* - internet-optimierte Datenkataloge (*GESIS Datenbestandskatalog*)
Accessible	- internet-optimierte Datenkataloge - unterschiedlichen Zugangsklassen (on-site & off-site)
Interoperable	- einheitliche technische Formate (*SPSS, Stata*) - Metadatenstandards (*DDI*)
Re-Usable	- Qualitätssicherung und Datendokumentation (*DDI*) - Lizenzvorlagen und Klärung urheberrechtlicher Fragen - *Bitstream Preservation* für mindestens 25 Jahre - Langzeitarchivierung

Die Generierung FAIRer Daten ist jedoch nicht nur für große Umfrageprogramme relevant. Zunehmende Auflagen, z. B. von Förderern, führen auch zu einer zunehmenden Nachfrage nach Angeboten zur Aufbereitung, Dokumentationen etc. kleinerer Studien nach den FAIR-Prinzipien. Um dieser gesteigerten Nachfrage systematisch zu begegnen, startete das Datenarchiv im Jahr 2016 ein internes Projekt zur Konzeption der *GESIS Datenservices* rund um die Archivierung quantitativer sozialwissenschaftlicher Umfragedaten. Ein Datenservice ist dabei definiert als eine (theoretisch) unabhängige Dienstleistung, die nach außen anbietbar ist und die nicht weiter in untergeordnete Leistungen zerlegt werden kann. So beinhaltet beispielsweise der Datenservice *Prüfung auf inhaltliche Konsistenz* eine Qualitätskontrolle der Angaben in einzelnen Variablen. Dabei wird u. a. geprüft, ob Angaben fehlerhaft sind oder ob zum Beispiel eine Altersangabe außerhalb der anvisierten Stichprobe (z. B. Erwachsene zwischen 18 und

[8] Weitere Informationen zum *International Social Survey Programme* bei GESIS finden sich unter https://www.gesis.org/issp/home/ [04.12.2019].
[9] Weitere Informationen zur *DDI Alliance* finden sich unter https://www.ddialliance.org/ [04.12.2019].
[10] Weitere Informationen zu *Digital Object Identifiern* finden sich unter https://www.doi.org/ [04.12.2019].
[11] Weitere Informationen zu *da|ra* finden sich unter https://www.da-ra.de/home/ [04.12.2019]

65 Jahren) liegt. Derartige Inkonsistenzen werden im Rahmen des genannten Datenservices durch einen Abgleich von Messinstrument und erhobenen Daten identifiziert, dokumentiert und (ggf.) korrigiert.

Zur Konzeption des Angebotsportfolios wurde zunächst eine Marktanalyse durchgeführt. Die Sichtung der Angebote von zehn nationalen und vier internationalen Umfrageinstituten und Datenarchiven führte dabei zu dem Ergebnis, dass bislang keine systematischen und bepreisten Dienstleistungen rund um die Generierung FAIRer Daten existieren. Parallel zu dieser Marktanalyse wurden alle am Datenarchiv erstellten Dienstleistungen systematisch erfasst und entschieden, welche dieser Leistungen in Datenservices überführt werden sollen. Diese Dienstleistungen wurden dann hinsichtlich ihrer weiteren Zerlegbarkeit in kleinere Einheiten überprüft und so eine vorläufige Sammlung potentieller Datenservices aufgebaut. Daran anschließend wurden alle Datenservices in Arbeitsschritte unterteilt und Qualitätsstandards definiert. Ein Arbeitsschritt bildet dabei die kleinste organisatorische Einheit im Workflow eines Datenservices, dem klar definierte zeitliche und personelle Ressourcen zugewiesen werden können. Schließlich wurden die zur Erbringung eines Datenservice notwendigen Aufwendungen ermittelt und entsprechende Kalkulationsvorlagen erstellt.

Diese Ermittlung von Aufwendungen für die Kostenkalkulation erwies sich dabei als besonders komplex. Im Projekt wurde zunächst zwischen Gemein-, Sach- und Personalkosten unterschieden. Die Gemeinkosten, z. B. für Büros, Computer, Verwaltung etc., können analog zum Overhead in Drittmittelprojekten als Anteil am Gesamtbudget einer Institution ermittelt werden. Sachkosten, die zur Erstellung eines Datenservices über die in den Gemeinkosten verbuchten Aufwendungen hinausgehen, beziehen sich beispielsweise auf zusätzlich notwendige Softwarelizenzen, erhöhte Speicherkapazitäten usw. Auch derartige Kosten lassen sich im Vorfeld der Erstellung eines Datenservices gut erfassen, budgetieren und in die Kostenkalkulation integrieren.

Eine Herausforderung für die Kostenkalkulation stellte hingegen vor allem der personelle Aufwand dar, d. h. im Wesentlichen die Arbeitszeit zur Erstellung eines bestimmten Datenservices. Um diesen Aufwand zu ermitteln bezog das Projekt Mitarbeitende des Datenarchivs mit ein, die in ihrer täglichen Arbeit mit unterschiedlichsten Aspekten der Datenaufbereitung, -dokumentation und -archivierung betraut sind. Diese wurden gebeten ihre Aktivitäten, den damit verbundenen zeitlichen Aufwand ebenso wie zentrale Charakteristika der bearbeiteten Daten, z. B. deren Umfang, exakt zu protokollieren. Durch die Auswertung dieser Protokolle konnten zum einen Faktoren ermittelt werden, die den zeitlichen Arbeitsaufwand beeinflussen. Hierzu zählen neben der Datenqualität und der Kommunikation mit den Datengebenden vor allem der Datenumfang (Variablenanzahl) ebenso wie die Anzahl an Abzweigungen bzw. Filtern[12] im

[12] Derartige Abzweigungen bzw. Filtern im Messinstrument dienen der Aufteilung der Befragten in verschiedene Teilgruppen, denen dann im weiteren Verlauf der Befragung unterschiedliche Fragen gestellt werden. So kann beispielsweise die Frage nach der Erwerbstätigkeit dazu dienen, Erwerbstätige von Nicht-Erwerbstätigen zu trennen, um anschließend nur der Teilgruppe der Erwerbstätigen Fragen zu ihren Arbeitsbedingungen zu stellen.

jeweiligen Messinstrument.

Zum anderen dienten die Protokolle dazu, den durchschnittlich notwendigen zeitlichen Aufwand für jeden einzelnen Arbeitsschritt zu kalkulieren. So ergab etwa die Auswertung der personellen Aufwendungen zur *Prüfung auf inhaltliche Konsistenz*, dass pro Variable und Abzweigung bzw. Filter im Messinstrument etwa eine Minute Arbeitszeit veranschlagt werden muss. Diese Zeit umfasst anteilig das Einarbeiten in die jeweiligen Daten, den Abgleich der Umfragedaten mit dem Messinstrument (Fragebogen), die Dokumentation auftretender Inkonsistenzen sowie die Vorbereitung eventueller Korrekturen. Aus dem so ermittelten zeitlichen Aufwand in Stunden ergeben sich, multipliziert mit dem tatsächlichen Stundenlohn einer Person, die Netto-Personalkosten zur Konsistenzkontrolle. Der Bruttopreis eines Datenservices berechnet sich dann unter Aufschlag von zusätzlichen Sachkosten, den (anteiligen) Gemeinkosten und der fälligen Mehrwertsteuer.

4 Die Datenservices in der Praxis

Das erstellte Angebotsportfolio mit größtenteils kostenpflichtigen Datenservices wurde im Jahr 2018 zunächst in einer einjährigen Pilotphase anhand konkreter Aufträge getestet. Dabei kam es zu einigen kleineren Änderungen am Portfolio ebenso wie in den Kalkulationsvorlagen. Seit 2019 sind die GESIS Datenservices voll implementiert und können auf Anfrage erworben werden.

Aus Platzgründen werden die GESIS Datenservices im Folgenden nur ansatzweise beschrieben, eine detaillierte Übersicht findet sich unter gesis.org/datenservices/home/. Das Angebotsportfolio gliedert sich zunächst in verschiedene Servicebereiche, wie in Abbildung 1 dargestellt. Diese umfassen unterschiedliche Archivierungs- und Bereitstellungsoptionen (Sicherung der Zugänglichkeit), die Registrierung und Bekanntmachung der Daten (Erhöhung der Auffindbarkeit) sowie deren Aufbereitung, Anreicherung und Dokumentation (Gewährleistung der technischen und analytischen Nachnutzbarkeit). Innerhalb dieser Bereiche sind dann verschiedene Datenservices gruppiert. So können Daten im Bereich *bereitstellen* z. B. vollkommen frei für jedwede Nachnutzung zur Verfügung stehen, erst auf Basis einer Nutzendenregistrierung zugänglich gemacht werden oder aber speziellen Zugangsvoraussetzungen, wie etwa der Nutzung in einem *Secure Data Center*[13] (on- und off-site), unterliegen. Analog können Daten im Bereich *dokumentieren* lediglich auf der Studienebene beschrieben oder in unterschiedlicher Granularität auch auf Variablenebene dokumentiert werden.

Daneben wurden verschiedene Datenservices auf Basis bisheriger Anfragen in Servicepaketen kombiniert. So beinhaltet etwa das kostenlose Servicepaket 1 (*Archivierung Basis*) die Sicherung und Bereitstellung der Daten für 25 Jahre ebenso wie deren

[13] Weitere Informationen zum *Secure Data Center* bei GESIS finden sich unter https://www.gesis.org/angebot/daten-analysieren/weitere-sekundaerdaten/secure-data-center-sdc/ [04.12.2019].

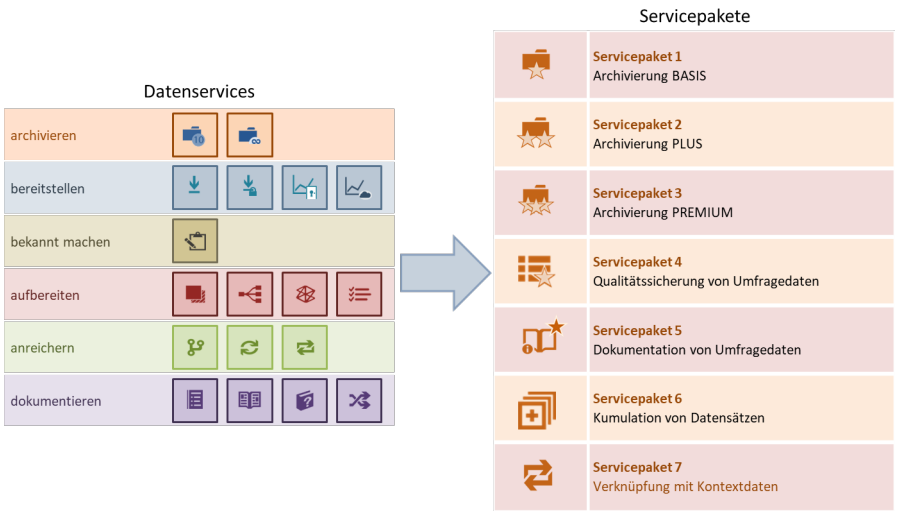

Abbildung 1: Die Datenservices und Servicepakete im Überblick

DOI-Registrierung und den Datennachweis im *GESIS Datenbestandskatalog*[14]. Darauf aufbauend bietet das kostenpflichtige Servicepaket 2 (*Archivierung Plus*) eine Langzeitarchivierung über 25 Jahre hinaus, die Basisprüfung der Daten(-qualität) ebenso wie die Beschreibung der Daten auf Studienebene im DDI-Standard. Im ebenfalls kostenpflichtigen Servicepaket 3 (*Archivierung Premium*) kommen noch die standardisierte Dokumentation der Daten auf Variablenebene ebenso wie weitere qualitätskontrollierende Maßnahmen hinzu.

Mit den Datenservices und den zugehörigen Servicepaketen richtet sich das Datenarchiv an unterschiedliche Nutzendengruppen der akademischen und außerakademischen Forschung, wie z. B. universitäre und (inter-)nationale Forschungsprojekte, Infrastruktureinrichtungen, Ressortforschung, Stiftungen etc. Vor allem für die universitäre Forschung und für (inter-)nationale Forschungsprojekte stellt sich dabei die Frage nach der Finanzierung der Datenservices. In diesem Zusammenhang bleibt zunächst auf die eingangs erwähnten Förderauflagen zur Generierung FAIRer Daten hinzuweisen. Mit derartigen Auflagen gehen zumeist auch entsprechende Finanzierungszusagen einher, wenn die Kosten der Aufbereitung, Dokumentation, etc. im Rahmen des Förderantrags beziffert werden können.[15] Hier zeigen sich zwei großen Vorteile der *GESIS Datenservices*: Zum einen können die unterschiedlichen Maßnahmen rund um die Archivierung von Forschungsdaten definiert sowie die dadurch entstehenden Kos-

[14] Weitere Informationen zum *GESIS Datenbestandskatalog* finden sich unter https://dbk.gesis.org/dbksearch/index.asp?db=d [04.12.2019].

[15] So können Antragsstellende beispielsweise im Rahmen des *Horizon 2020 Framework Programme* der EU (https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm, 04.12.2019) oder der Förderung Digitalisierung im Bildungsbereich des Bundesministeriums für Bildung und Forschung (https://www.bmbf.de/foerderungen/bekanntmachung-1420.html, 04.12.2019) Mittel für ihr Forschungsdatenmanagement beantragen.

ten kalkuliert und geltend gemacht werden. Zum anderen werden entsprechende Maßnahmen - in Form von Datenservices - durch Expertinnen und Experten erbracht, deren tägliche Routine es ist, Daten nach internationalen Standards aufzubereiten, zu dokumentieren und zu archivieren. Auf diesem Weg wird die Generierung FAIRer Daten gewährleistet.

5 Fazit

Insgesamt kann die Konzeption und Implementierung der *GESIS Datenservices* als eine Win-win-Situation beschrieben werden. Erstens können Primärforschende die Kosten für Maßnahmen rund um die Datenarchivierung kalkulieren, den notwendigen Ressourceneinsatz planen und ggf. gegenüber Mittelgebenden geltend machen. Die Auslagerung der Generierung FAIRer (Forschungs-)Daten ermöglicht dabei die Fokussierung auf das eigentliche Forschungsvorhaben. Forschungsfördernde und Mittelgebende erhalten zweitens eine klare Kostenkalkulation und stellen gleichzeitig sicher, dass die geförderten Daten den FAIR-Prinzipien entsprechen und Dritten verfügbar gemacht werden. Schließlich profitiert auch die Forschungsgemeinschaft von qualitativ hochwertigen Daten, die auffindbar, zugänglich, interoperabel und (analytisch) nachnutzbar sind.

Die Konzeption der *GESIS Datenservices* am *Datenarchiv für Sozialwissenschaften* zeigt, dass die Implementierung kostenpflichtiger Dienstleistungen machbar ist und Mehrwerte nicht nur für die Dienstleistenden sondern vor allem für die Forschenden und die Forschungsgemeinschaft bietet. Das Konzept kann somit als Blaupause für andere Datenarchive ebenso wie für weitere Dienstleistungen bzw. andere Forschungsdisziplinen dienen. Es hilft auch, die Aufwendungen und Kosten des Forschungsdatenmanagements genauer zu identifizieren und zu beziffern. Zu bedenken bleibt, dass sich die am Datenarchiv ermittelten zeitlichen Aufwendungen nicht ohne weiteres auf Forschungsprojekte übertragen lassen. Die Beauftragung von Dienstleistenden wie GESIS, die über die notwendige Expertise und die entsprechende Erfahrung verfügen, geht mit Effizienzgewinnen bei der Aufbereitung, Dokumentation und Archivierung der Daten einher. Sie ermöglicht es Forschenden aber auch, ihren Aufwand für die Bearbeitung ihrer Daten zu reduzieren und so Zeit für die inhaltliche Forschung zu gewinnen.