RESEARCH

Otmane Azeroual

Untersuchungen zur Datenqualität und Nutzerakzeptanz von Forschungs-informationssystemen

Framework zur Überwachung und Verbesserung der Qualität von Forschungsinformationen

MOREMEDIA

Springer Vieweg

Untersuchungen zur Datenqualität und Nutzerakzeptanz von Forschungsinformationssystemen

Otmane Azeroual

Untersuchungen zur Datenqualität und Nutzerakzeptanz von Forschungsinformationssystemen

Framework zur Überwachung und Verbesserung der Qualität von Forschungsinformationen

Springer Vieweg

Otmane Azeroual
Deutsches Zentrum für Hochschul- und
Wissenschaftsforschung (DZHW)
Berlin, Deutschland

ISBN 978-3-658-36701-5 ISBN 978-3-658-36702-2 (eBook)
https://doi.org/10.1007/978-3-658-36702-2

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Planung/Lektorat: Stefanie Eggert
Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Vorwort

Die vorliegende Dissertation ist im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Deutschen Zentrum für Hochschul- und Wissenschaftsforschung GmbH (DZHW)[1] zwischen 2016 und 2021 entstanden. Maßgeblich für diese Arbeit war das Projekt „Helpdesk für die Einführung des Kerndatensatzes Forschung (KDSF)“[2], das vom Bundesministerium für Bildung und Forschung (BMBF) und von allen Bundesländern gefördert wurde. Im KDSF-Projekt wurde ein freiwilliger Standard für die Erhebung, Vorhaltung und den Austausch von Forschungsinformationen entwickelt und sowohl für Dateneigner (z. B. Hochschulen und außeruniversitäre Forschungseinrichtungen (AUFs)) als auch für Datenabfrager (z. B. Forschungsförderer, Behörden usw.) im deutschen Wissenschaftsraum bereitgestellt. Der KDSF beschreibt die Datengruppierung nach Inhalten und die Formate zur Implementierung in Forschungsinformationssysteme (FIS), weil heutzutage die Vorhaltung und der Austausch von Forschungsinformationen an einer zunehmenden Anzahl an Hochschulen und außeruniversitären Forschungseinrichtungen über ein FIS erfolgen. Ein FIS als Datenbanksystem soll mittel- und langfristig interne sowie externe Berichtspflichten und -legungen für die Hochschulverwaltung und politische Entscheidungsträger erleichtern. Es existieren nationale und internationale Standards (z. B. KDSF-Datenmodell und Common European Research Information Format (CERIF)) zur Unterstützung von Forschungsinformationssystemen, und diese ermöglichen Kompatibilität und Interoperabilität zwischen verschiedenen Systemen, um die Forschungsbereiche zu repräsentieren. Um diese systematisch aufbauenden Bereiche effizient nutzen zu können, wird eine gute Datenqualität verlangt, da ein wichtiger Erfolgsfaktor

[1] https://www.dzhw.eu

[2] https://kerndatensatz-forschung.de

für die Implementierung des FIS in Hochschulen und AUFs die Sicherstellung der Datenqualität darstellt. Allerdings wird dieser Punkt bisher kaum in der Wissenschaft untersucht, obwohl das Thema für Hochschulen, AUFs und deren Forscher relevant und durchaus aufschlussreich ist und in internationalen wissenschaftlichen Publikationen erwähnt wurde. Bislang fehlen die erforderliche Beachtung der Bedeutung der Datenqualität, die Dimensionen zur Messung und Prüfung der Datenqualität in einem FIS, neue Methoden bzw. Techniken zur Verhinderung und Verbesserung von Ursachen mangelnder Qualität in einem FIS und die Verstärkung des Vertrauens bzw. der Akzeptanz in das FIS. Ohne eine ausreichende Datenqualität kann das FIS sein Nutzenpotenzial als Lieferant entscheidungsrelevanter Daten allerdings nicht ausschöpfen. Grundsätzlich gilt, je höher und sicherer die Qualität der Daten in Forschungsinformationssystemen sind, desto größer ist die Nutzerakzeptanz.

Die Entstehung der Dissertation haben zahlreiche Personen ermöglicht und unterstützt. Bei ihnen möchte ich mich an dieser Stelle herzlichst bedanken.

An erster Stelle gilt dies ganz besonders meinem Doktorvater, Herrn Prof. Dr. Gunter Saake. Durch seine persönliche, fortwährende Betreuung, die freundliche Hilfe, die Unterstützung während des Entstehungsprozesses dieser Dissertation, durch die Schaffung von hervorragenden Rahmenbedingungen für ein praxisorientiertes Forschen im Bereich der Datenbanken und Informationssysteme und durch die Freiheiten, die er mir für die Dissertation gewährte, hat er maßgeblich zum erfolgreichen Abschluss des Forschungsvorhabens beigetragen.

Ein besonderer Dank gilt auch Herrn Dr.-Ing. Eike Schallehn für seine hilfreiche und aufmerksame Unterstützung, Ermutigung und stete Motivation.

Zudem habe ich Herrn Prof. Dr.-Ing. Mohammad Abuosba für die Übernahme des Zweitgutachtens sowie für seine ständige, motivierende Beratung, Unterstützung und sein wertvolles Feedback zu diversen Fragen rund um die Dissertation zu danken. Die zahlreichen Gespräche an der Otto-von-Guericke-Universität Magdeburg werden mir stets eine kostbare und gute Erinnerung sein.

Dem Deutschen Zentrum für Hochschul- und Wissenschaftsforschung (DZHW) danke ich vielmals für die finanzielle Unterstützung dieser Dissertation, insbesondere in Form von Förderungen von Konferenz- und Workshopbesuchen, die das Vorhaben wesentlich bereichert haben. Ebenfalls danke ich meinem KDSF-Team am DZHW Berlin für die kollegiale Arbeit und interessante Zeit, die mir in guter Erinnerung bleiben wird.

Verbunden bin ich auch Herrn Dr. Joachim Schöpfel für seine Unterstützung, die konstruktiven Anregungen und die ebenso außerordentliche wie freundschaftliche Zusammenarbeit. Ein zusätzlicher Dank geht an meinem weiteren Gutachter Herrn Prof. Dr. Sören Auer für die zuverlässige und hilfreiche Unterstützung.

Ebenfalls gilt auch ein großer Dank allen Teilnehmenden meiner Befragung für ihre Auskunftsbereitschaft, interessanten Beiträge und Antworten auf meine Fragen.

Außerdem möchte ich Herrn Dr. Volker Manz für das Korrekturlesen meiner Dissertation danken.

Abschließend möchte ich mich von ganzem Herzen bei meinen lieben Eltern Jamila Morjane und Mustapha Azeroual bedanken, die mir das Studium überhaupt erst ermöglicht haben und auf deren Unterstützung ich immer zählen kann und konnte. Sie standen mir während meiner gesamten Ausbildungszeit äußerst liebevoll und mit vollen Kräften zur Seite. Ein besonderer und persönlicher Dank gilt meiner Frau Virginia Azeroual für ihre ganze Bandbreite an Geduld, Motivation und unglaublich hilfreicher Unterstützung während der Anfertigung meiner Dissertation.

Meinen allerherzlichsten Dank an Allah!

Otmane Azeroual

Inhaltsverzeichnis

Abkürzungsverzeichnis

AUFs	Außeruniversitäre Forschungseinrichtungen
BI	Business Intelligence
BMBF	Bundesministerium für Bildung und Forschung
CASRAI	Consortia Advancing Standards in Research Administration Information
CERIF	Common European Research Information Format
CMS	Content Management System
CMS	Campus Management System
CORDIS	Community Research and Development Information Service
CRF	Conditional Random Field
CRIS	Current Research Information Systems
CRM	Customer Relationship Management
DBMS	Datenbankmanagementsystem
DCMI	Dublin Core Metadata Initiative
DFG	Deutsche Forschungsgemeinschaft
DINI	Deutsche Initiative für Netzwerkinformation e. V. – Arbeitsgruppe
DNB	Deutsche Nationalbibliothek
DOI	Digital Object Identifier
DPMA	Deutsches Patent- und Markenamt
DQ	Datenqualität
DQD	Datenqualitätsdimensionen
DQM	Datenqualitätsmanagement
DWH	Data Warehouse
DZHW	Deutsches Zentrum für Hochschul- und Wissenschaftsforschung
eG	eingetragene Genossenschaft
EOSC	European Open Science Cloud

EPO	European Patent Office
ERM	Entity Relationship Modell
ERP	Enterprise Resource Planning
ETL	Extraktion, Transformation, Laden
EU	Europäische Union
EUNIS	European University Information Systems
e. V.	eingetragene Verein
FDM	Full Data Model
FhG	Fraunhofer-Gesellschaft
FI	Forschungsinformationen
FIS	Forschungsinformationssysteme
GEPRIS	Geförderte Projekte Informationssystem
GLM	General Linear Model
GmbH	Gesellschaft mit beschränkter Haftung
HAC	Hierarchical Agglomerative Clustering
HGF	Helmholtz-Gemeinschaft Deutscher Forschungszentren
HTML	Hypertext Markup Language
IDM	Identity Management
Ids	Identifikationen
IE	Informationsextraktion
iFQ	Institut für Forschungsinformation und Qualitätssicherung e. V.
IT	Informationstechnologie
KDSF	Kerndatensatz Forschung
KMO	Kaiser-Meyer-Olkin-Kriterium
MDM	Master Data Management
MPG	Max-Planck-Gesellschaft
NER	Named Entity Recognition
NLP	Natural Language Processing
OCLC	Online Computer Library Center
ODI	Oracle Data Integrator
OPAC	Online Public Access Catalogue
ORCID	Open Researcher and Contributor ID
PAISY	Personal-Abrechnungs-und-Informations-System
PCA	Principal Components Analysis
PDI	Pentaho Data Integration
PLS	Partial Least Squares
POS	Part-of-Speech
RCD	Research Core Dataset
RDBMS	Relational Database Management System

RDF	Resource Description Framework
RIS	Research Information Systems
RMS	Root Mean Square
SAP	Systemanalyse Programmentwicklung
SEM	Structural Equation Modeling
SSOAR	Social Science Open Access Repository
TAM	Technology Acceptance Model
TDM	Text Data Mining
WGL	Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz
WR	Wissenschaftsrat
XML	Extensible Markup Language

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

Listingsverzeichnis

Einleitung

1

Mit zunehmendem Bedarf der Forschungsaktivitäten, den steigenden Anforderungen an die Forschungsberichterstattung im deutschen Wissenschafts- und Hochschulsystem und den wachsenden Datenmengen bei gleichzeitigen Veränderungen der Informationsbedürfnisse von Forschern und seitens der Öffentlichkeit werden alle Organisationen vor neue Herausforderungen gestellt. Langfristig werden sich jene Hochschulen und außeruniversitären Forschungseinrichtungen (AUFs) durchsetzen, die sich auf diese Bedingungen einstellen, die also in der Lage sein werden, flexibel und schnell auf Veränderungen zu reagieren, und gleichzeitig ihre Kosten im Griff haben und auch reduzieren. Hierfür ist jedoch eine genaue Kenntnis der aktuellen Organisationssituation unverzichtbar. Um dies zu gewährleisten und das Management von Forschungsinformationen bei den eigenen Planungs- und Entscheidungstätigkeiten mit den benötigten Daten zu versorgen, werden hoch entwickelte Forschungsinformationssysteme eingesetzt. Diese wurden von einer europäischen Community namens euroCRIS[1] (*Current Research Information Systems (CRIS)*) entwickelt und unterstützt. Mittlerweile hat sich in der Forschung der Begriff Forschungsinformationssystem (FIS) etabliert. Dabei beschreibt ein FIS Ansätze wie das Sammeln, Integrieren, Speichern und Analysieren von Forschungsinformationen einer Organisation und dient als föderierte Datenbank [ASA18b].

In den letzten Jahren hat sich das Thema Forschungsinformationssysteme in den deutschen und internationalen Organisationen stark verbreitet und ist zu einem festen Bestandteil der universitären IT-Landschaften geworden [ASW18], [ASA18a],

[1] https://www.eurocris.org

Ergänzende Information Die elektronische Version dieses Kapitels enthält Zusatzmaterial, auf das über folgenden Link zugegriffen werden kann https://doi.org/10.1007/978-3-658-36702-2_1.

O. Azeroual, *Untersuchungen zur Datenqualität und Nutzerakzeptanz von Forschungsinformationssystemen*, https://doi.org/10.1007/978-3-658-36702-2_1

[ASA18b]. Gleichzeitig arbeiten viele Hochschulen und AUFs derzeit noch an der Implementierung solcher Informationssysteme. Durch den Einsatz von Forschungsinformationssystemen können die Organisationen und die Forscher darin unterstützt werden, ihre Forschungsaktivitäten und Forschungsergebnisse transparent und intelligent zu gestalten. Mit der Hilfe eines FIS können Forschungsinformationen wie *Projekte, Drittmittel, Patente, Kooperationspartner, Preise, Publikationen usw.* gespeichert und verwaltet sowie in den jeweiligen Webauftritt eingebunden werden. Für die Hochschulen und AUFs bilden Forschungsinformationen die Grundlage für die interne und externe Bewertung ihrer Leistungen als selbstständige wissenschaftliche Organisation. Nebenbei haben sie die Möglichkeit, das eigene Forschungsprofil übersichtlich darzustellen; ein FIS dient somit als ein effektives Werkzeug des Informationsmanagements. Die Vereinheitlichung und Erleichterung der Berichterstattung schafft einen Mehrwert sowohl für die Verwaltung und das Management als auch für die Forscher. Forschende und andere Interessengruppen bekommen mit dem FIS ein geeignetes Mittel, um Informationen über ihre Forschungsaktivitäten und aktuelle Trends, das Erstellen von Lebensläufen, die Ergebnisse und Kooperationen in ihren Forschungsfeldern, existierende Projekte, Publikationen und Förderungen sowie den Kontakt zu den Interessenten aus der Wirtschaft zu managen und zu begleiten bzw. den Anforderungen der Berichterstattung gerecht zu werden. Durch all diese Prozesse soll eine Mehrarbeit für die FIS-Nutzer vermieden werden, was z. B. die Reduktion des Zeitaufwandes bei der Erstellung von Berichten oder bei der Außendarstellung ihrer Forschungsleistung und wissenschaftlichen Expertise bedeutet.

1.1 Motivation des Forschungsvorhabens

Obwohl das FIS schon seit Längerem existiert, hat es erst vor wenigen Jahren besonders durch viele verschiedene Konferenzen etwa im Rahmen der europäischen Organisation euroCRIS und der Deutschen Initiative für Netzwerkinformation e. V. (DINI AG-FIS)[2] mit Experten und Entwicklern an Bekanntheit gewonnen. Hier wurde das Potenzial von FIS entdeckt, und einige deutsche Hochschulen und AUFs begannen damit, solche Systeme zu implementieren. Ein FIS einzuführen, bedeutet, die erforderlichen Informationen über Forschungsaktivitäten und -ergebnisse in gesicherter Qualität zur Verfügung zu stellen. Da die Einführung und der Betrieb von Forschungsinformationssystemen mit erheblichen Kosten und einem hohen Ressourcenbedarf verbunden sind, gewinnen die Qualität der Daten und die Nutzerakzep-

[2] https://dini.de/ag/fis/

tanz zunehmend an Bedeutung. Die Datenqualität gehört neben der Datensicherheit, der Bedienerfreundlichkeit und anderen Variablen zu den wesentlichen Bedingungen der Nutzerakzeptanz eines FIS. Dabei geht es in erster Linie um Vertrauen – Vertrauen in das System, in dessen Anbieter und in die Administration. Einem System, das Datenprobleme nicht zuverlässig identifiziert oder korrigiert oder das selbst eine Quelle für Datenqualitätsmängel ist, kann (und wird) kein Vertrauen entgegengebracht werden. Wahrgenommene oder empfundene Qualitätsprobleme beeinträchtigen die subjektive Leistungserwartung gegenüber dem System. Dieser Sachverhalt ist nicht neu [Dav93], [WT05]; im Fall eines FIS ist eine mangelhafte Datenqualität aber umso problematischer, als es um strategische und zum Teil hochsensible Informationen und Entscheidungshilfen geht, wie personenbezogene oder finanzielle Daten. Für die Nutzbarkeit und Interpretation einrichtungsspezifischer Forschungsinformationen spielt die Datenqualität eine wichtige Rolle; sie entscheidet über Erfolg oder Misserfolg bei Hochschulen und AUFs. Nur mit einer hohen Qualität können die FIS-Nutzer zuverlässige und nutzbringende Forschungsergebnisse liefern und eine fundierte Entscheidungsfindung mit einer guten Präsentation von Forschung ermöglichen. Daher sollte das Thema Datenqualität Gegenstand ständiger Sorgfalt und Aufmerksamkeit sein; ein „Qualitätsmanagement“ ist erforderlich, um die Systemleistung und die Zufriedenheit und Akzeptanz der Benutzer zu gewährleisten. Die vorliegende Dissertation adressiert die Problemstellung der Datenqualität bei FIS nutzenden Hochschulen und AUFs in Deutschland und verschiedenen europäischen Ländern und stellt entsprechende Lösungen bereit, um die Akzeptanz des FIS seitens dieser Einrichtungen zu erhöhen.

1.2 Problemstellung

Das Management von Forschungsinformationen wird zunehmend zu einer wichtigen Aufgabe für die Hochschulen und AUFs [AA18]. Wachsende Datenmengen und die größer werdende Anzahl an internen und externen Datenquellen, so z. B. durch Informationssysteme für Humanressourcen, Finanzhaushalte und Bibliotheken, führen in einem FIS zu mehr Fehlern, z. B. Rechtschreibfehlern, Dubletten, zu einer fehlerhaften Formatierung oder zu fehlenden, inkorrekten und uneinheitlichen Daten, die zunehmend zu einer Herausforderung in Hochschulen und AUFs werden und sich zu einem ernsthaften Problem entwickeln können. Diese unterschiedlichen Datenfehler können bei der Erfassung, der Übertragung sowie der Integration von Forschungsinformationen in das FIS entstehen und sich über verschiedene Bereiche erstrecken sowie schwer auffindbar sein [ASA18b]. Außerdem kann die Qualität der externen Datenquellen (wie z. B. unterschiedliche Publikationsdatenbanken,

Identifiers, Textdateien oder XML-Files usw.) einen ungünstigen Einfluss auf die Qualität der FIS haben. Heutzutage gehört der Umgang mit großen Datenquellen für die Hochschulen und AUFs zum täglichen Betrieb. Dies liegt hauptsächlich an der Tatsache, dass viele Schnittstellen im Forschungsmanagementprozess entwickelt werden müssen, um einen Informationsaustausch zwischen den Quellsystemen zu ermöglichen, da das FIS als integraler Bestandteil einer kompletten Verwaltungssoftware für die Administration von Einrichtungen konzipiert werden kann und eine Plattform für die Datenintegration darstellt [BEL12]. Während der Datenintegration können die Hochschulen und AUFs jederzeit über den Zustand ihrer Forschungsinformationen informiert und ihre Qualitätsfehler schnell identifiziert und korrigiert werden. Daher muss bei der Integration von Forschungsinformationen auf die Datenheterogenität und Datenverteilung geachtet werden. Hierbei ist eine einmalige Bereinigung in den internen und externen heterogenen und verteilten Quellsystemen nicht ausreichend, denn Forschungsinformationen müssen kontinuierlich gepflegt und ihre Qualität muss immer wieder optimiert werden [ASA18a], [ASA18b]. Solange die Nutzer nicht in der Lage sind, auf die am dringendsten benötigten Informationen zuzugreifen und schnelle Entscheidungen zu treffen, sinkt der Wert der verwendeten Forschungsinformationen und das Vertrauen in das FIS und dessen Akzeptanz [ASA18b]. Die Datenqualität bei den Hochschulen und AUFs sollte daher mit hoher Priorität behandelt werden; dies bedingt die Einführung einer Strategie, die als Frühindikator dient, um die Probleme der Datenqualität im FIS anzugehen.

1.3 Forschungslücken und Forschungsmethodik

Je mehr Forschungsinformationssysteme es gibt und je mehr Forschungsinformationen dort gesammelt, gespeichert und verarbeitet werden, desto wichtiger wird die Betrachtung der Qualität, damit die Stakeholder qualitativ hochwertige Ergebnisse erhalten können und die Systemakzeptanz erhöht wird. Ohne eine akzeptable Datenqualität im jeweiligen FIS ist eine Akzeptanz der Nutzer nicht möglich. In der Literatur zu FIS beschäftigen sich nationale und internationale FIS-Experten in ihren wissenschaftlichen Publikationen im Allgemeinen nur mit dem Aufbau, dem Mehrwert und den Herausforderungen eines FIS in Form von Erfahrungsberichten. Mit Blick auf die Datenqualität des FIS ließen sich hingegen kaum Quellen in der Literatur finden, und es gibt bisher keine Methodik, die vorschreibt, welche Qualitätsschritte in welcher Reihenfolge durchgeführt werden müssen, um anschließend ein hohes Qualitätsniveau des FIS langfristig und dauerhaft zu garantieren. Insbesondere mangelt es an Untersuchungen über die Maßnahmen und deren praktischen

Einsatz zur Bewertung und Verbesserung der Datenqualität im FIS. Allerdings wurden diese Punkte in wissenschaftlichen Arbeiten zu anderen, ähnlichen Bereichen der Informationssysteme und des Informationsmanagements behandelt und publiziert.

Immerhin haben Hochschulen, AUFs und Forschende dem Thema Datenqualität in FIS in den ersten Jahren im Rahmen von Konferenzen und Workshops große Aufmerksamkeit geschenkt, und immer öfter ist die Rede von Forschungslücken im Hinblick auf die Sicherstellung und Verbesserung der Datenqualität bei der Integration von Forschungsinformationen mit unterschiedlichen Quellsystemen in das FIS sowie von der Notwendigkeit, die Nutzerakzeptanz von FIS zu steigern [Mün17], [SSC+19]. Vor diesem Hintergrund ist die vorliegende Dissertation im Forschungsbereich der Wirtschaftsinformatik (WI) zum Themengebiet Datenbanken und Software Engineering entstanden und soll dabei unterstützen, FIS effektiv und effizient zu nutzen. Das Ziel besteht darin, im Rahmen einer anwendungsbezogenen Forschung zum einen die untersuchten Datenqualitätsprobleme in FIS aus der Praxis zu lösen. Grundlage hierfür bilden Erfahrungen, die aus dem KDSF-Projekt und organisierten Workshops am DZHW Berlin stammen, sowie die Erkenntnisse aus einer quantitativen Untersuchung (Umfrage) mit FIS nutzenden Hochschulen und AUFs. Zum anderen geht es darum, ein Konzept bzw. Framework für einen kontrollierten Umgang mit der Datenqualität in FIS zu entwickeln.

Abbildung 1.1 stellt die Referenzarchitektur als Konzept- bzw. Framework-Lösung für die vorliegende Dissertation dar. Die einzelnen Komponenten werden in **Kapitel** 3 weiter vertieft. Das Thema ist von hoher Praxisrelevanz an praktisch allen

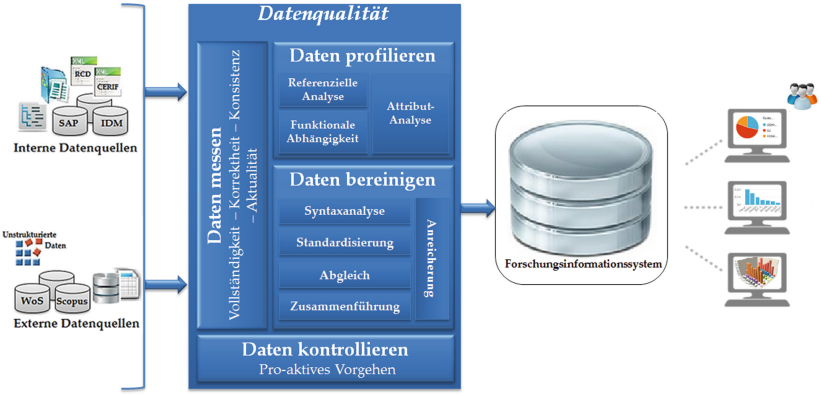

Abbildung 1.1 Architektur des Konzepts bzw. Frameworks