Association for Information Systems
AIS Electronic Library (AISeL)

Wirtschaftsinformatik 2022 Proceedings | Special Track: Student Track

Jan 17th, 12:00 AM

Systematische Literaturanalyse und Harmonisierung von Datenlebenszyklen

Fabian Sauer
Hochschule Heilbronn, Germany, fsauer@stud.hs-heilbronn.de

Helmut Beckmann
Hochschule Heilbronn, Germany, helmut.beckmann@hs-heilbronn.de

Follow this and additional works at: https://aisel.aisnet.org/wi2022

Recommended Citation

Sauer, Fabian and Beckmann, Helmut, "Systematische Literaturanalyse und Harmonisierung von Datenlebenszyklen" (2022). *Wirtschaftsinformatik 2022 Proceedings*. 1.
https://aisel.aisnet.org/wi2022/student_track/student_track/1

This material is brought to you by the Wirtschaftsinformatik at AIS Electronic Library (AISeL). It has been accepted for inclusion in Wirtschaftsinformatik 2022 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Systematische Literaturanalyse und Harmonisierung von Datenlebenszyklen

Fabian Sauer[1] und Helmut Beckmann[2]

[1] Hochschule Heilbronn, Studiengang M. Sc. Wirtschaftsinformatik, Heilbronn, Deutschland
fsauer@stud.hs-heilbronn.de
[2] Hochschule Heilbronn, Studiengang M. Sc. Wirtschaftsinformatik, Heilbronn, Deutschland
helmut.beckmann@hs-heilbronn.de

Abstract. In einer sich immer schneller wandelnden digitalisierten Wirtschaft, welche hohe Anforderungen an Dynamik und Datenorientierung stellt, ist der strukturierte Umgang mit unternehmensinternen, aber auch -externen Daten von besonderer Relevanz. Trends wie Big Data und Internet of Things stellen hierbei große Herausforderungen dar, welche es zu adressieren gilt. Um ein einheitliches, möglichst übergreifend einsetzbares Bild der Datenlebenszyklen zu ermöglichen, fokussiert diese Arbeit ein systematisches Literaturreview, angeknüpft an bereits vorhandene Arbeiten sowie die Harmonisierung und Generalisierung der dort identifizierten Datenlebenszyklen. Die daraus resultierenden Phasen und Prozessschritte werden zudem basierend auf wissenschaftlichen Arbeiten näher beschrieben, machen den Lebenszyklus somit greifbar und erzeugen ein einheitliches Verständnis. Der Datenlebenszyklus fokussiert sich auf keine spezielle Art von Daten oder Informationen und ist demnach als Grundlage für verschiedenste Bereiche der Wirtschaft, aber auch Wissenschaft zu verstehen.

Keywords: Datenlebenszyklus, Informationslebenszyklus, Harmonisierung, State-of-the-Art

1 Einleitung

Durch die zunehmende Digitalisierung von Unternehmen müssen sich diese immer stärker mit dem Problem befassen, vorhandene Datenmengen strukturiert zu verwalten und zu beherrschen [1, 2]. Das hierbei zu verfolgende Ziel stellt die Sicherstellung einer hohen Datenqualität dar, welche für den Unternehmenserfolg von fundamentaler Bedeutung ist. Diese Relevanz ist besonders im Bereich der Stammdaten zu sehen, welche einen deutlichen Unternehmenswert aufweisen [3]. Hieraus resultiert die Notwendigkeit für Unternehmen, die Datenlebenszyklen der Unternehmensdaten zu fokussieren. Aufgrund der hohen Vielfalt spezifischer Datenlebenszyklen in der Literatur [1] gilt es diese Lebenszyklen miteinander zu vergleichen und mögliche Harmonisierungsmöglichkeiten zu identifizieren.

17th International Conference on Wirtschaftsinformatik,
February 2022, Nürnberg, Germany

1.1 Problemstellung und Zielsetzung

Bereits in Arbeiten wie denen von El Arass et al. [1], Elmekki et al. [4] oder auch Maindze et al. [5] wird ersichtlich, wie viele unterschiedliche, leicht abgewandelte und heterogene Datenlebenszyklen aktuell in der Literatur existieren. Dabei variieren die Lebenszyklen neben den inhaltlich fokussierten Phasen auch in der grundsätzlichen Anzahl der betrachteten Schritte. Demnach gibt es Modelle, welche aus fünf Phasen bestehen, aber auch jene, welche acht oder mehr Phasen betrachten [5]. Da die Auswahl des „richtigen" Lebenszyklus für einen speziellen Einsatzzweck bereits eine große Herausforderung darstellt und von El Arass et al. [1] durch eine kriterienbasierte Betrachtung bestehender Lebenszyklen unterstützt wurde, fokussiert sich diese Arbeit einerseits mit der systematischen Darstellung des aktuellen Forschungsstands im Bereich Datenlebenszyklen, andererseits aber auch mit der Harmonisierung dieser.

Auf Basis der identifizierten Problemstellung werden in dieser Arbeit folgende zwei Forschungsfragen thematisiert: Welche Datenlebenszyklen wurden im Rahmen der letzten vier Jahre (ab dem Jahr 2017) in der Literatur thematisiert? Wie lassen sich die Phasen der individuellen Datenlebenszyklen miteinander kombinieren und harmonisieren?

Das Ziel ist demnach ein State-of-the-Art Beitrag, welcher Arbeiten wie die von El Arass et al. [1] oder Maindze et al. [5] ergänzt und somit einen umfassenden Überblick über relevante sowie aktuelle Datenlebenszyklen ermöglicht. Besonders die Arbeit von El Arass et al. [1] stellt hierbei einen umfangreichen Überblick aus dem Jahr 2017 dar, welchen es fortzuführen gilt. Durch die Vielfalt der individuellen Datenlebenszyklen und der heterogenen Betrachtung dieser Modelle wird anschließend das Ziel einer Harmonisierung der identifizierten Lebenszyklen verfolgt. Hierbei werden die Modelle miteinander verglichen, die fokussierten Phasen zusammengeführt und die Lebenszyklen final in einen allgemeingültigen Datenlebenszyklus kombiniert. Dieses Vorgehen orientiert sich dabei an anderen Arbeiten wie der von Gupta und Müller-Birn [6] oder auch El Arass et al. [1] sowie an Arbeiten der Harmonisierung von Referenzmodellen [7]. Durch diese Zielsetzung werden die zuvor definierten Fragestellungen fokussiert und beantwortet.

1.2 Aufbau der Arbeit

Nachdem bereits in dem ersten Abschnitt dieser Arbeit die Problemstellung und Zielsetzung thematisiert wurden, ist die Arbeit im Anschluss wie folgt gegliedert: Im nachfolgenden zweiten Abschnitt werden Grundlagen und Begriffe der Arbeit thematisiert, welche sich speziell mit Lebenszyklen, insbesondere Daten- und Informationslebenszyklen befassen. Diese stellen die Basis für diese Arbeit dar und sind für das einheitliche Verständnis von hoher Relevanz. Danach folgt die Methodik zur Literaturanalyse im dritten Abschnitt, welche das Vorgehen der Literatursuche und -analyse beschreibt. Die hierbei gefundene Literatur sowie die Selektion der relevanten Arbeiten wird anschließend übersichtlich dargestellt. Das daran anknüpfende Kapitel fünf geht auf die identifizierte Literatur ein und stellt die dort identifizierte Menge an individuellen Datenlebenszyklen dar. Auf Basis der Lebenszyklen aus der Literatur

wird anschließend eine Harmonisierung durchgeführt, welche die relevanten Phasen der unterschiedlichen Lebenszyklen miteinander kombiniert und so einen allgemeingültigen Lebenszyklus ermöglicht. Den Abschluss der Arbeit bilden die Diskussion und der Ausblick, in welchem die Arbeit zusammengefasst wird und eine kritische Würdigung stattfindet. Aus der Arbeit resultierende Forschungsbedarfe werden hierbei ebenfalls aufgezeigt.

2 Grundlagen und Begriffe

Für ein einheitliches Verständnis der in dieser Arbeit verwendeten Begriffe werden nachfolgend die Begriffe „Data Lifecycle“ und „Data Lifecycle Management“ näher beschrieben. Entsprechend der Evolution von Daten über Informationen hin zu Wissen, welche auch von Ku et al. [8] thematisiert wurde, werden Informationen hierbei als verarbeitete und in einen Kontext gebrachte Daten verstanden. Somit wird die Transformation von Daten im Rahmen eines Datenlebenszyklus analog zu Informationslebenszyklen verstanden, sodass beide Lebenszyklen synonym zueinander betrachtet werden. Ein Unterschied zwischen diesen Lebenszyklen kann zwar in Form der Daten- / Informationskomplexität vorliegen, jedoch ist der Einfluss auf den zugrunde liegenden Lebenszyklus der entsprechenden Daten in dieser Arbeit zu vernachlässigen.

2.1 Data Lifecycle

Mit der Transformation von Daten aus der analogen in die digitale Form erhalten diese nach Blazquez und Domenech [9] sowie nach Lynch [10] die Eigenschaft, einfach ausgetauscht, vervielfältigt und wiederverwendet zu werden. Hierfür ist es jedoch notwendig, diese Daten zu sammeln, zu verarbeiten und zu nutzen. Diese Phasen, welche hierfür von hoher Relevanz sind, werden als Data Lifecycle verstanden. Dabei wird der Lebenszyklus beginnend bei der Erstellung bis hin zur Löschung aufgrund von Irrelevanz betrachtet [1]. Das hiermit zu verfolgende Ziel ist die kontrollierte Betrachtung und Verarbeitung sowie die damit einhergehende Reduktion des Risikos des Datenverlusts, der Inkonsistenz und der Compliance Verstöße [1, 10]. Ein Datenlebenszyklus, welcher nach El Arass et al. [1] die klassischen Bestandteile eines Lebenszyklus für Daten beinhaltet, ist der Lebenszyklus nach Yu und Wen [11], welche angelehnt an die CRUD Operatoren aus dem Umfeld der Datenbanken die Phasen „Create“, „Store“, „Use and share“, „Archive“ sowie „Destruct“ identifiziert haben. In Abbildung 1 wird dieser Lebenszyklus dargestellt und verdeutlicht, dass die beiden Phasen „Use and share“ sowie „Archive“ optionale Phasen darstellen.

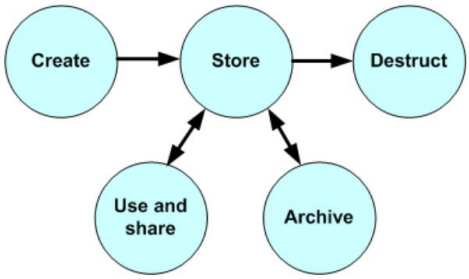

Abbildung 1: Data Life Cycle Model nach [11]

2.2 Data Lifecycle Management

Das Data Lifecycle Management steht in direktem Zusammenhang mit dem bereits definierten Data Lifecycle und beschreibt das Management der dort beinhalteten Prozesse [1, 12]. Das Data Lifecycle Management kann demnach als vorgelagerter sowie kontinuierlich fortlaufender Bestandteil des Datenlebenszyklus beziehungsweise als Metaebene des Zyklus verstanden werden, welche das Management der dort enthaltenen Daten und Prozesse steuert und kontrolliert. Dies ist auch in dem Literaturreview von El Arass et al. [1] zu sehen, da hier bereits einige Datenlebenszyklen (bspw. [12, 13]) um eine Planungsphase oder um Kontrollphasen erweitert wurden und einen Datenmanagementplan als Output aufweisen.

3 Methodik zur Literaturanalyse

In diesem Abschnitt der Arbeit wird die zugrunde liegende Methodik zur Literaturanalyse beschrieben. Die Literatursuche dieser Arbeit folgt dem fünfstufigen Forschungsprozess nach Fettke [14], bestehend aus der Problemformulierung, der Literatursuche, der Literaturauswertung, der Analyse und Interpretation sowie der Präsentation. Ebenso wird das strukturierte Vorgehen nach Webster und Watson [15] angewandt, wodurch die Auswahl der Literatur systematisch erfolgt. Auf Basis dessen wird die Wissenschaftlichkeit aber auch Transparenz des Vorgehens dieser Arbeit gewahrt, wodurch die Arbeit für die Wissenschaft und Praxis nutzbar, erweiterbar und zitierfähig wird. Eine Vorwärts- und Rückwärtssuche wird dabei nur in begrenztem Umfang für ergänzende und beschreibende Zwecke durchgeführt, um den Betrachtungsrahmen dieser Arbeit einzuhalten.

Die ausgewählten Datenbanken belaufen sich auf sieben Stück, namentlich: AIS Electronic Library (AISel), Association for Computing Machinery Digital Library (ACM), EmeraldInsight (Emerald), Google Scholar (GS), IEEE Xplore Digital Library (IEEE), Science Direct (SD) sowie Springer Link (SL). Die darin enthaltenen Konferenzen wurden in der Literaturanalyse ebenfalls berücksichtigt.

Da diese Arbeit bereits bestehenden Arbeiten, insbesondere der von El Arass et al. [1] folgt sowie die Suchergebnisse bereits möglichst relevante Ergebnisse enthalten sollen, werden nachfolgende Filter- und Ausschlusskriterien angewandt:

- **Zeitraum der Suche:** 2017 (inklusive) bis 2021 (soweit vorhanden inklusive)
- **Zu verwendende Suchfelder:** Title, Abstract und Keywords
- **Ausnahme:** In den Literaturdatenbanken Google Scholar und Springer Link werden die Suchen ausschließlich auf den Titel begrenzt.

Auf Basis dieser Kriterien werden Arbeiten fokussiert, welche nach 2017 publiziert wurden und sich möglichst detailliert mit dem Lebenszyklus von Daten und Informationen befassen. Zur vollumfänglichen und nachvollziehbaren Identifikation von Literatur wurden die Datenbanken auf Basis nachfolgender Begriffe in deutscher und englischer Sprache durchsucht: Datenlebenszyklus, Informationslebenszyklus, Datenmanagement Lebenszyklus und Informationsmanagement Lebenszyklus. Die Begriffe wurden dabei in Suchtermen (bspw.: „data lifecycle“ OR „information lifecycle“) kombiniert. Die Suchterme wurden explizit kompakt gehalten, um einerseits andere Formen von Lebenszyklen wie die der Produkte oder Prozesse auszuschließen. Andererseits wird jedoch versucht, durch möglichst geringe Einschränkungen (beispielsweise auf spezielle Arten von Daten wie Stammdaten oder Forschungsdaten) ein vollumfängliches Bild der Datenlebenszyklen zu erreichen.

Auf Basis der bereits thematisierten Suchterme sowie Filter- und Ausschlusskriterien wurden, die in Tabelle 1 dargestellten Suchergebnisse erzielt. Im Anschluss an die Suche wurden die Arbeiten auf Basis des Titels und Abstracts betrachtet und gefiltert. Dabei wurden von den 542 Arbeiten 85 potenziell relevante Arbeiten identifiziert. Nach Entfernung von Duplikaten sowie Betrachtung des Volltextes beläuft sich die Summe an relevanter Literatur auf 68 Arbeiten.

Tabelle 1: Identifizierte Literatur nach Datenbank

#	Datenbank	Suchfelder	Suchergebnisse	Relevant
L1	AISel	Title/Abstract	5	1
L2	ACM	Title/Abstract/Keyword	3	2
L3	Emerald	Title/Abstract	8	4
L4	GS	Title	200	22
L5	IEEE	Title/Abstract	87	24
L6	SD	Title/Abstract	43	10
L7	SL	Title	196	22
Summe			542	85
Ohne Duplikate				78
Nach Volltextbetrachtung				68

4 Literaturanalyse

Die auf Basis der Datenbanken identifizierte Literatur wird in nachfolgenden Abschnitten systematisch aufbereitet dargestellt sowie selektierte Arbeiten kurz beschrieben. Bereits vor der Analyse kann die Erkenntnis getroffen werden, dass die Literatur in Verbindung mit den Datenlebenszyklen vorrangig der englischen Sprache angehört und somit die englischsprachigen Suchterme einen deutlich stärkeren Erfolg

aufweisen konnten. Für die Analyse aller relevanten Arbeiten wurde in Tabelle 2 eine Übersicht über die Datenlebenszyklen erstellt, in welcher die Arbeiten einerseits nach ihrem Einsatzbereich, andererseits nach Eigen- und Fremdleistung aufgegliedert werden. Die Einsatzbereiche wurden hierbei nach markanten Ausprägungen der jeweiligen Arbeit ausgewählt. Wurde keine direkte Ausprägung im dargestellten Lebenszyklus erkannt, wird der entsprechende Lebenszyklus der Kategorie „Allgemeingültig“ zugeordnet. Innerhalb dieses Einsatzbereichs werden zusätzlich Arbeiten, welche einen besonderen Fokus auf das Datenmanagement oder die Data Governance legen, unter „Management“ aufgeführt. Arbeiten wie die von Gavalas et al. [16], Elmekki et al. [4] oder auch Maindze et al. [5], welche sich mit Traffic, Government oder Vehicle Health Daten befasst haben, wurden zudem aufgrund der zugrunde liegenden allgemeingültigen Datenlebenszyklen in den allgemeinen, nicht spezifischen Einsatzbereich eingeordnet. Die Unterscheidung nach Eigen- und Fremdleistung wurde anhand des Kriteriums bestimmt, ob der dargestellte Lebenszyklus ohne Hilfe fremder Literatur (zum Beispiel durch die Beschreibung des vorhandenen Lebenszyklus in der Arbeit) erstellt wurde, auf Basis fremder Literatur eigenständig abgeleitet wurde oder ob es sich um einen referenzierten Lebenszyklus einer fremden Arbeit handelt. Bei Letzterem ist es zudem möglich, dass es sich um Lebenszyklen handelt, welche vor 2017 erstellt, jedoch nach 2017 in der betrachteten Arbeit referenziert wurden (wie bei [1, 4, 17]).
Bei der Darstellung der Lebenszyklen gilt es anzumerken, dass die Summe aller Lebenszyklen nicht mit der identifizierten Literatur verglichen werden kann. Grund hierfür sind Arbeiten, welche über ein Literaturreview mehrere Lebenszyklen aufzeigen und eventuell darauf basierend eigene Lebenszyklen abgeleitet haben. Zudem führen Lebenszyklen wie der von Khaloufi et al. [18] dazu, dass eine Mehrfachnennung in der abgebildeten Tabelle möglich ist. Die Gesamtsumme von 92 ist demnach nicht die Gesamtanzahl an Lebenszyklen. Die Anzahl der betrachteten Lebenszyklen beläuft sich tatsächlich auf 82 Stück.

Tabelle 2: Lebenszyklen nach Einsatzbereich

	Einsatzbereich									**Summe**
								Allgemeingültig		
	Personal Data/Information	Internet of Things	Big Data	Scientific/Research Data	HealthCare	Machine Learning	Blockchain	Kein spezifischer Fokus	Management	
Eigen	1	2	10	10	3	3	1	21	6	57
Fremd	2	0	9	3	1	0	1	18	1	35
Summe	3	2	19	13	4	3	2	39	7	**92**

In der Tabelle ist sehr gut zu erkennen, dass die meisten Lebenszyklen keinen spezifischen Managementfokus aufweisen und allgemeingültig anwendbar sind. Dies ist darauf zurückzuführen, dass viele Arbeiten die Lebenszyklen lediglich als Grundlage definieren, auf welcher die Arbeit basiert. Hierbei steht der Datenlebenszyklus nicht im primären Fokus der Arbeit und lässt sich als „Basis"-Lebenszyklus bezeichnen, welcher basierend auf den verwendeten Daten innerhalb der Arbeit abgeleitet wurde. Beispielarbeiten sind die von Firdhous und Hussien [19], Li et al. [20] oder auch Sari und Frisila [21]. Neben den allgemeingültigen Lebenszyklen lassen sich aber auch vermehrt Lebenszyklen in den Bereichen Big Data (bspw. [9, 22, 23]) und Scientific/Research Data (bspw. [6, 24, 25]) erkennen. Im Bereich der wissenschaftlichen Datenlebenszyklen überwiegen zudem die eigenerstellten und/oder abgeleiteten Lebenszyklen gegenüber der referenzierten Lebenszyklen aus fremden Arbeiten deutlich.

Um eine Übersicht über die seit 2017 entstandenen Lebenszyklen zu erhalten, stellen die Arbeiten von Alshammari und Simpson [26], El Arass et al. [27] und Griffin et al. [25] Vertreter für die Bereiche der persönlichen Daten, der Big Data und der wissenschaftlichen Daten dar. Alshammari und Simpson [26] haben hierbei einen „Abstract Personal Data Lifecycle" (APDL) dargestellt, welcher sich auf die Besonderheiten und rechtlichen Rahmenbedingungen von persönlichen Daten fokussiert.

Einen etwas anderen Fokus weist hingegen die Arbeit von El Arass et al. [27] auf, welche sich mit der Transformation von Big Data hin zu Smart Data, also sinnvoll einsetzbaren Daten, befasst hat. Dabei wurde die NIST Big Data Referenzarchitektur als Grundlage genutzt, um den Smart Datalifecycle (Smart DLC) zu entwerfen. Hierbei ist der Fokus der Autoren auf die Management- und Supportprozesse sehr gut zu erkennen, welche übergreifend über den eigentlichen operationalen Datenlebenszyklus Anwendung finden. Die Sicherstellung der Sicherheit und Qualität der großen, unstrukturierten Datenmengen sowie die detaillierte Planung, das kontinuierliche Management und die Kontrolle des operativen Prozesses nehmen hierbei eine wichtige Rolle ein.

Der dritte Datenlebenszyklus, welcher einen Kontrast zu den vorherigen Arbeiten darstellt, ist der wissenschaftliche Lebenszyklus von Griffin et al. [25]. Die Autoren fokussieren dabei keine derart großen Datenmengen wie zuvor angesprochen und auch keine sensiblen persönlichen Daten, sondern wissenschaftliche Daten aus dem Bereich der Biologie, wozu neben wissenschaftlichen Publikationen ebenso DNA- oder Protein-Datenbanken zählen. Dabei liegt der Fokus auf der Wiederverwendung der Erkenntnisse und dem Publizieren von Arbeiten/Daten, um andere Interessenten an den Ergebnissen teilhaben zu lassen. Der Prozessschritt des Löschens von Daten wird hingegen nicht direkt thematisiert.

Trotz dieser drei Arbeiten ist jedoch anzumerken, dass bereits etablierte Lebenszyklen wie zum Beispiel der DataOne, DDI, USGS oder auch PII Lebenszyklus, aber auch Grundlagen wie die Informationspyramide oder der „CRUD" Lifecycle noch von hoher Relevanz sind und einen Einfluss auf die jüngeren Arbeiten haben [1, 4]. Der Einfluss bereits existierender Datenlebenszyklen reicht inzwischen weit in die Unternehmen hinein und spielt dort eine fundamentale Rolle. ETL (Extract, Transform,

Load) Prozesse im Kontext der Industrie 4.0 [28] oder auch klassische Anlage-, Nutzungs- und Pflege- sowie Löschprozesse von Stammdaten oder Bewegungsdaten in ERP Systemen wie von SAP [29] sind Beispiele für die grundlegende Präsenz der Datenlebenszyklen im Unternehmensalltag.

5 Harmonisierter Datenlebenszyklus

Nachdem im vorherigen Kapitel die vorhandenen Datenlebenszyklen aus der Literatur kategorisiert wurden, befasst sich dieses Kapitel mit einem harmonisierten Lebenszyklus basierend auf der Literatur. Auf Grundlage der betrachten Arbeiten und den darin enthaltenen Datenlebenszyklen wurden folgende sechs Phasen als relevant identifiziert: Planning, Collecting, Processing, Storage/Maintenance, Usage und Destruction. Die Auswahl folgte dabei den Arbeiten von El Arass et al. [1], welche auf Basis der Literatur die relevantesten Phasen identifizierten konnten; von Gupta und Müller-Birn [6], welche einen Vergleich der Lebenszyklen vorgenommen haben als auch von Pardo et al. [7], welche die Möglichkeiten der Vergleiche, Zuordnungen, Kombinationen und Harmonisierungen aufgezeigt haben.
Die Phasen werden im Laufe dieses Abschnitts gemeinsam mit den darin enthaltenen Prozessschritten näher beschrieben. Auf Basis dieser Phasen und den Prozessschritten der Lebenszyklen aus der Literatur wurde der harmonisierte Datenlebenszyklus aus Abbildung 2 abgeleitet. Neben der Betrachtung des eigentlichen Lebenszyklus der Daten wurden ebenso die relevanten Managementaufgaben, welche für das Data Lifecycle Management von hoher Relevanz sind, als parallelen und übergreifenden Prozess mitaufgenommen. Hierzu zählt neben der Dokumentation von unter anderem Metainformationen [4, 5, 30] auch die kontinuierliche (Zwischen-) Speicherung [27], das Qualitätsmanagement [4, 5, 22, 30] sowie das Sicherheitsmanagement [5, 30, 31].
In nachfolgenden Absätzen werden die bereits genannten Phasen inklusive der darin enthaltenen Prozessschritte des Datenlebenszyklus beschrieben.

5.1 Planning

Die erste Phase des harmonisierten Datenlebenszyklus stellt die Planning-Phase dar. Diese wurde in Arbeiten wie den von El Arass et al. [1, 27] aber auch Gupta und Müller-Birn [6] oder Maindze et al. [5] beschrieben. Dabei wird das Ziel verfolgt, die zu verfolgende Aufgabe inklusive der notwendigen Ressourcen und Regeln je Prozessschritt zu definieren und zu planen. Dafür werden Anforderungen unter anderem an den Verwendungserfolg, die Qualität, die Integrität und die Sicherheit gestellt. Ebenso werden Pläne für das bereits angesprochene übergreifende Management erarbeiten. Neben den in der Abbildung exemplarisch dargestellten Managementprozessen zählen ebenso die Data Governance Prozesse, welche unter anderem von El-Zoghby und Azer [32] thematisiert wurden, zu den übergreifenden Managementaufgaben.

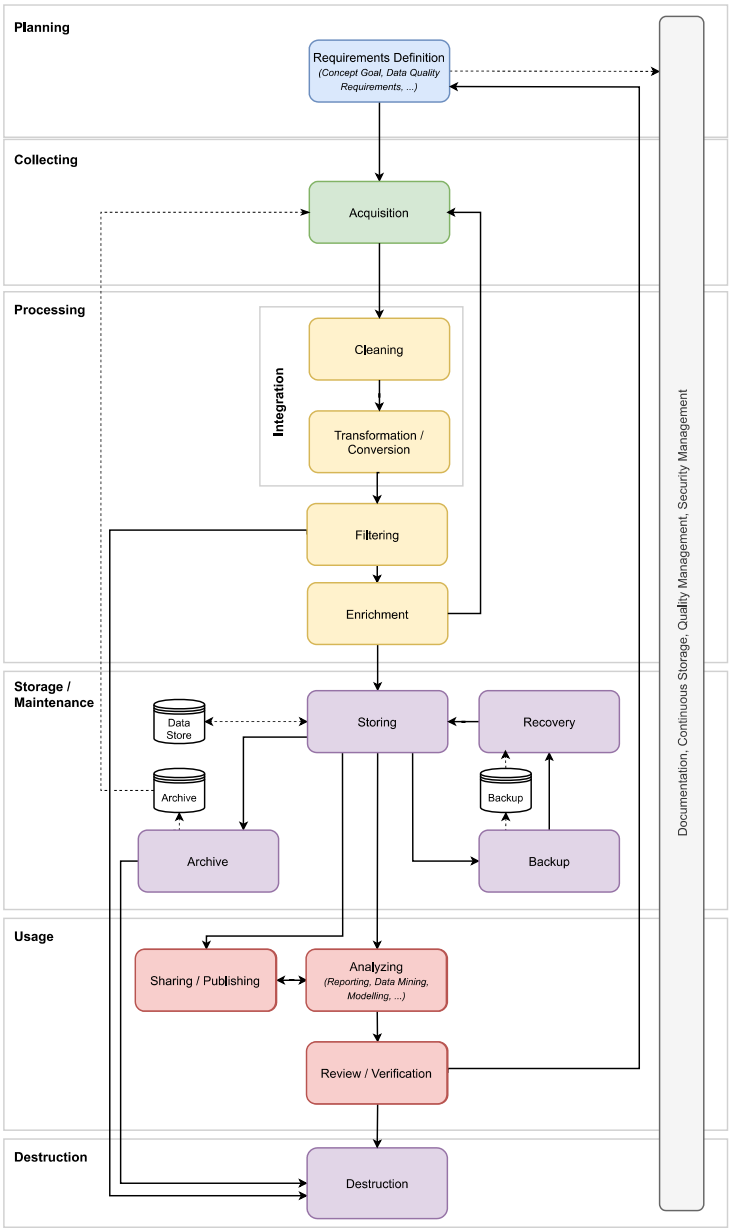

Abbildung 2: Harmonisierter Datenlebenszyklus

5.2 Collecting

Auf Basis der gestellten Anforderungen aus der vorherigen Phase folgt nun die wohl wichtigste Phase für einen Datenlebenszyklus – die Akquisition der Daten. Diese Phase

ist in jedem klassischen Datenlebenszyklus vorhanden und fokussiert das Generieren und/oder Abfragen von Daten. Dabei kann es sich sowohl um Primär- oder Sekundärquellen handeln, aus welchen beispielsweise persönliche Daten gewonnen werden [26], als auch um interne oder externe Quellen, aus welchen (relevante) Daten abgefragt werden [9, 23]. Die hieraus resultierenden Daten liegen anschließend in ihrer Rohform vor und wurden noch in keiner Form transformiert.

5.3 Processing

Aus den nun vorliegenden unverarbeiteten Daten folgt anschließend die Phase des Processings. Diese Phase umfasst die Integration, bestehend aus dem Cleaning und der Transformation/Conversion, dem Filtering und dem Enrichment. Der Prozess der Integration verfolgt dabei das Ziel, heterogene Daten aus unterschiedlichen Datenquellen miteinander zu kombinieren [27]. Dabei werden die Daten bereinigt [23, 33, 34], um die vorausgesetzte Datenqualität zu erzielen, aber auch transformiert (bspw. Normalisierung) und konvertiert, um eine Harmonisierung der unterschiedlichen Daten(-formate) zu ermöglichen [33]. Das hieraus entstandene Resultat, eine einheitliche gemeinsame Datenbasis wird anschließend gefiltert [35]. Beim Filtern der Daten wird das Ziel (besonders im Zusammenhang mit Big Data) verfolgt, relevante Daten von irrelevanten Daten zu trennen und so die Menge an Daten zu reduzieren [1]. Auch hier werden Anforderungen aus der Planungsphase in Bezug auf die Definition von relevanten und irrelevanten Daten berücksichtigt [2, 27]. Qualitativ eventuell hochwertige Daten, welche für die Erreichung des definierten Ziels keine Relevanz haben, werden hier entfernt und enden somit in der Destruction. Der letzte Prozessschritt der Processing Phase ist das Enrichment. Hierbei werden die bereits aufbereiteten Daten durch weitere Datenakquisitionen angereichert [8]. Dadurch kann die Qualität der Datenbasis weiter erhöht werden, wodurch darauf basierende Entscheidungen eine höhere Qualität aufweisen können [27, 36].

5.4 Storage/Maintencance

Nachdem in den vorherigen Phasen eine einheitliche Datenbasis erstellt wurde, folgt in der Phase Storage/Maintenance die grundlegende Speicherung der Daten. Wie bereits bei der parallel ablaufenden Managementphase beschrieben, handelt es sich bei der hier angesprochenen Speicherung nicht um die einzig vorhandene Speicherung. Bereits in den vorherigen, aber auch in den nachfolgenden Prozessschritten sind (Zwischen-) Speicherungen möglich und notwendig [37]. Die hier angesprochene Speicherung fokussiert sich jedoch auf die strukturierte Datenspeicherung in Datenbanken oder File-Systemen [38], welche auf eine lange Zeitspanne ausgerichtet ist [23]. Aufgaben wie das Migrieren von Datenbasen zwischen unterschiedlichen Systemen zählen ebenso zu den Aufgaben der Speicherung [9]. Dabei können besonders bei sensiblen Daten unterschiedliche Anforderungen an die Speicherung gestellt worden sein, welche es zu adressieren gilt [17]. In engem Zusammenhang zu der Datenspeicherung steht ebenso die Datensicherung und die Wiederherstellung einer solchen Sicherung. Um sich gegen Datenverlust jeglicher Art zu sichern und den Datenmanagementanforderungen zu

entsprechen, ist eine solche Sicherung sowie die Möglichkeit der Wiederherstellung von hoher Relevanz [1, 24]. Der letzte Prozessschritt, welchen es in der Phase Storage/Maintenance anzusprechen gilt, ist die Archivierung von Daten. Hierbei werden Daten, welche in naher Zukunft nicht benötigt werden, archiviert und somit aus dem aktiven Lebenszyklus genommen [5, 35]. Die Archivierung sollte hierbei vordefinierten Regeln folgen, wodurch Daten beispielsweise anhand des Alters archiviert werden [1]. Das Archiv kann für künftige Datenlebenszyklen zudem als eine mögliche Datenressource dienen, aus welcher Daten abgefragt werden. Eine Wiederverwendung der Daten ist daher ebenfalls Fokus einer angemessenen Datenspeicherung und -archivierung [9].

5.5 Usage

Im Anschluss an die Speicherung, Sicherung und Archivierung der betrachteten Daten folgt die Verwendung dieser. Die Verwendung wird häufig als „Access“ bezeichnet, da hier eine notwendige Schnittstelle für die Konsumenten bereitgestellt werden muss [27]. Die Datenverwendung kann dabei auf unterschiedlichsten Weisen erfolgen. Unter anderem können hierfür einfache Abfragen, statistische Vorgehen sowie komplexe Tools aus dem maschinellen Lernen eingesetzt werden [34]. Die geeignete Vorgehensweise der Analyse ist dabei von den entsprechenden Daten und Anforderungen aus der Planungsphase abhängig [9, 25]. Das übergreifend fokussierte Ziel in der Datenverwendung ist es, neue Erkenntnisse durch beschreibende oder vorhersagende Analysen zu erzielen [6, 23]. Nachdem die Daten analysiert wurden, folgt der Prozessschritt des Reviews und der Verification. Dieser dient dazu, die Ergebnisse der Analyse sowie den vorangegangen Lebenszyklus anhand der zu Beginn definierten Anforderungen an das Ergebnis und die Managementdisziplinen zu evaluieren [26]. Die hieraus resultierenden Erkenntnisse können dazu genutzt werden, die Anforderungen der Planungsphase anzupassen und den Datenlebenszyklus nachhaltig zu verbessern [34].
Um auch anderen Interessenten den Zugang zu den Daten zu ermöglichen, wird ebenso der Prozessschritt Sharing/Publishing in die Phase der Verwendung eingeordnet. Hierbei können neben den zugrunde liegenden Daten auch die gewonnenen Erkenntnisse geteilt werden [1, 9]. Bei der Veröffentlichung von Daten und Ergebnissen ist es jedoch von hoher Relevanz, dass die Qualität der veröffentlichten Gegenstände möglichst hoch ist, um darauf basierende Entscheidungen oder Forschungen (im Kontext von wissenschaftlichen Daten [6, 39]) nicht negativ zu beeinflussen [1]. Die Veröffentlichung kann dabei über direkte Abfragen, Subskriptionsmodelle [40], aber auch in Form von Arbeiten (speziell in der Wissenschaft) oder Erkenntnispräsentationen erfolgen.

5.6 Destruction

Das Ende des Datenlebenszyklus bildet die Destruction Phase inklusive des gleichnamigen Prozessschritts. Das dort enthaltene Löschen von Daten kann durch rechtliche Rahmenbedingungen (bspw. bei persönlichen Daten [17, 26]) aber auch

aufgrund der Irrelevanz der Daten selbst notwendig sein. Das Löschen von Daten ist besonders dann von hoher Relevanz, wenn diese bereits erfolgreich verwendet wurden, keinen Mehrwert mehr liefern und somit nutzlos geworden sind [27]. Auch hier sind die festgelegten Vorgaben aus der Planungsphase die Basis für den Löschprozess.

6 Diskussion und Ausblick

Durch die zunehmende Relevanz von Daten in der Wissenschaft und Praxis durch Thematiken wie Big Data, Internet of Things oder auch Machine Learning hat sich diese Arbeit mit den Fragen befasst, welche Datenlebenszyklen angrenzend an bereits vorhandene Literaturreviews aus 2017 existieren und ob sich diese miteinander harmonisieren und kombinieren lassen. Für ein systematisches Literaturreview wurden hierfür entsprechende Suchterme gebildet und eine Auswahl an Literaturdatenbanken durchsucht. Dadurch konnten insgesamt 542 Arbeiten identifiziert werden, welche durch eine Betrachtung des Titels, des Abstracts, aber auch des Inhalts sowie der Entfernung von Duplikaten auf 68 Arbeiten reduziert werden konnten. Diese Arbeiten stellten die Grundlage für eine Auswertung dar, in welcher die dort enthaltenen Lebenszyklen einerseits nach deren Einsatzbereich, andererseits nach Eigen- und Fremdleistung kategorisiert wurden. Dabei belief sich die Summe der eingeordneten Datenlebenszyklen auf 82 Stück. Ein Ergebnis dieser Betrachtung war der identifizierte starke Fokus auf die Bereiche Big Data und Scientific/Research Data.
Nachdem die Unterschiede der vorhandenen Datenlebenszyklen an drei Lebenszyklen aufgezeigt wurden, wurde ein harmonisierter Datenlebenszyklus basierend auf der Literatur präsentiert. Die sechs Phasen mit insgesamt 14 Kernprozessschritten, in welche der Prozess unterteilt wurde, wurden im Detail und mit Referenz auf die bereits vorhandenen Lebenszyklen aus den unterschiedlichen Bereichen der Literatur beschrieben. Der harmonisierte Datenlebenszyklus fokussiert mit seinen Phasen keine spezifischen Daten, ist demnach allgemeingültig einsetzbar, ermöglicht ein grundsätzlich einheitliches Verständnis und dient als Basis für Anpassungen in speziellen Anwendungsbereichen.
Es gilt jedoch anzumerken, dass Lebenszyklen aus den Jahren vor 2017 durch daraus abgeleitete jüngere Datenlebenszyklen und Nennungen nur indirekten Einfluss auf den harmonisierten Lebenszyklus hatten. Ebenso musste durch die Harmonisierung eine Generalisierung durchgeführt werden, wodurch einzelne Prozessschritte aus spezifischen Anwendungsbereichen nicht in die harmonisierte Form übernommen wurden. Der hier dargestellte Datenlebenszyklus muss aufgrund dieses Informationsverlusts je nach Einsatzbereich möglicherweise angepasst werden und sollte nicht ohne nähere kritische Betrachtung eingesetzt werden. Eine Validierung und Überprüfung der praktischen Realisierbarkeit dieses Lebenszyklus gilt es ebenso in zukünftigen Arbeiten vorzunehmen. Zudem wäre eine Einbeziehung spezifischer Kernarbeiten aus den Jahren vor 2017 denkbar sowie eine direkte Gegenüberstellung des hier entwickelten Datenlebenszyklus mit den bereits etablierten Lebenszyklen, um spezifische Anpassungsbedarfe für die jeweiligen Einsatzbereiche herauszuarbeiten.

Literaturverzeichnis

1. El Arass, M., Tikito, I., Souissi, N.: Data lifecycles analysis: Towards intelligent cycle. 2017 Intell. Syst. Comput. Vision, ISCV 2017. (2017). https://doi.org/10.1109/ISACV.2017.8054938.
2. El Arass, M., Souissi, N.: Data Lifecycle: From Big Data to SmartData. In: 2018 IEEE 5th International Congress on Information Science and Technology (CiSt). pp. 80–87. IEEE (2018). https://doi.org/10.1109/CIST.2018.8596547.
3. Hubert Ofner, M., Straub, K., Otto, B., Oesterle, H.: Management of the master data lifecycle: a framework for analysis. J. Enterp. Inf. Manag. 26, 472–491 (2013). https://doi.org/10.1108/JEIM-05-2013-0026.
4. Elmekki, H., Chiadmi, D., Lamharhar, H.: Open Government Data. In: Proceedings of the ArabWIC 6th Annual International Conference Research Track on - ArabWIC 2019. pp. 1–6. ACM Press, New York, New York, USA (2019). https://doi.org/10.1145/3333165.3333180.
5. Maindze, A., Skaf, Z., Jennions, I.: Towards an Enhanced Data- and Knowledge Management Capability: A Data Life Cycle Model Proposition for Integrated Vehicle Health Management. Annu. Conf. PHM Soc. 11, 1–14 (2019). https://doi.org/10.36001/phmconf.2019.v11i1.842.
6. Gupta, S., Müller-Birn, C.: A study of e-Research and its relation with research data life cycle: a literature perspective. Benchmarking An Int. J. 25, 1656–1680 (2018). https://doi.org/10.1108/BIJ-02-2017-0030.
7. Pardo, C., Pino, F.J., García, F., Piattini Velthius, M., Baldassarre, M.T.: Trends in Harmonization of Multiple Reference Models. In: Communications in Computer and Information Science. pp. 61–73 (2011). https://doi.org/10.1007/978-3-642-23391-3_5.
8. Ku, T., Park, W., Choi, H.: Energy Big Data Life Cycle Mechanism for Renewable Energy System. In: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). pp. 1067–1068. IEEE (2019). https://doi.org/10.1109/INFCOMW.2019.8845036.
9. Blazquez, D., Domenech, J.: Big Data sources and methods for social and economic analyses. Technol. Forecast. Soc. Change. 130, 99–113 (2018). https://doi.org/10.1016/j.techfore.2017.07.027.
10. Lynch, C.: How do your data grow? Nature. 455, 28–29 (2008). https://doi.org/10.1038/455028a.
11. Yu, X., Wen, Q.: A View about Cloud Data Security from Data Life Cycle. In: 2010 International Conference on Computational Intelligence and Software Engineering. pp. 1–4. IEEE (2010). https://doi.org/10.1109/CISE.2010.5676895.
12. Faundeen, J.L., Burley, T.E., Carlino, J.A., Govoni, D.L., Henkel, H.S., Holl, S.L., Hutchison, V.B., Martin, E., Montgomery, E.T., Ladino, C.C., Tessler, S., Zolly, L.S.: The United States Geological Survey Science Data Lifecycle Model. U.S. Geol. Surv. Open-File Rep. 2013–1265. 1–4 (2013). https://doi.org/http://dx.doi.org/10.3133/ofr20131265.
13. Ma, X., Fox, P., Rozell, E., West, P., Zednik, S.: Ontology dynamics in a data life cycle: Challenges and recommendations from a Geoscience Perspective. J. Earth Sci. 25, 407–412 (2014). https://doi.org/10.1007/s12583-014-0408-8.
14. Fettke, P.: State-of-the-Art des State-of-the-Art. WIRTSCHAFTSINFORMATIK. 48, 257 (2006). https://doi.org/10.1007/s11576-006-0057-3.
15. Webster, J., Watson, R.T.: Analyzing Past To Prepare For Future: Writing Literature Review. MIS Q. 26, 13–23 (2002). https://doi.org/https://doi.org/10.2307/4132319.

16. Gavalas, D., Giannakopoulou, K., Kasapakis, V., Kehagias, D., Konstantopoulos, C., Kontogiannis, S., Kypriadis, D., Pantziou, G., Paraskevopoulos, A., Zaroliagis, C.: Renewable Mobility in Smart Cities:TheMOVESMART Approach. In: EAI/Springer Innovations in Communication and Computing. pp. 135–157 (2020). https://doi.org/10.1007/978-3-030-39986-3_7.
17. Altorbaq, A., Blix, F., Sorman, S.: Data subject rights in the cloud: A grounded study on data protection assurance in the light of GDPR. In: 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST). pp. 305–310. IEEE (2017). https://doi.org/10.23919/ICITST.2017.8356406.
18. Khaloufi, H., Abouelmehdi, K., Beni-hssane, A., Saadi, M.: Security model for Big Healthcare Data Lifecycle. Procedia Comput. Sci. 141, 294–301 (2018). https://doi.org/10.1016/j.procs.2018.10.199.
19. Firdhous, M.F.., Hussien, N.A.: Data Security Implementations in Cloud Computing: A Critical Review. In: 2018 3rd International Conference on Information Technology Research (ICITR). pp. 1–5. IEEE (2018). https://doi.org/10.1109/ICITR.2018.8736153.
20. Li, F., Li, H., Niu, B., Chen, J.: Privacy Computing: Concept, Computing Framework, and Future Development Trends. Engineering. 5, 1179–1192 (2019). https://doi.org/10.1016/j.eng.2019.09.002.
21. Sari, I.M., Frisila, L.: Information System Management on Asset Management in PT. PLN West Java Transmission Regional. In: 2019 2nd International Conference on High Voltage Engineering and Power Systems (ICHVEPS). pp. 223–226. IEEE (2019). https://doi.org/10.1109/ICHVEPS47643.2019.9011033.
22. El Arass, M., Tikito, I., Souissi, N.: An Audit Framework for Data Lifecycles in a Big Data context. In: 2018 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT). pp. 1–5. IEEE (2018). https://doi.org/10.1109/MoWNet.2018.8428883.
23. Coyne, E.M., Coyne, J.G., Walker, K.B.: Big Data information governance by accountants. Int. J. Account. Inf. Manag. 26, 153–170 (2018). https://doi.org/10.1108/IJAIM-01-2017-0006.
24. Chen, X., Wu, M.: Survey on the Needs for Chemistry Research Data Management and Sharing. J. Acad. Librariansh. 43, 346–353 (2017). https://doi.org/10.1016/j.acalib.2017.06.006.
25. Griffin, P.C., Khadake, J., LeMay, K.S., Lewis, S.E., Orchard, S., Pask, A., Pope, B., Roessner, U., Russell, K., Seemann, T., Treloar, A., Tyagi, S., Christiansen, J.H., Dayalan, S., Gladman, S., Hangartner, S.B., Hayden, H.L., Ho, W.W.H., Keeble-Gagnère, G., Korhonen, P.K., Neish, P., Prestes, P.R., Richardson, M.F., Watson-Haigh, N.S., Wyres, K.L., Young, N.D., Schneider, M.V.: Best practice data life cycle approaches for the life sciences. F1000Research. 6, 1618 (2017). https://doi.org/10.12688/f1000research.12344.1.
26. Alshammari, M., Simpson, A.: Personal Data Management: An Abstract Personal Data Lifecycle Model. In: Lecture Notes in Business Information Processing. pp. 685–697 (2018). https://doi.org/10.1007/978-3-319-74030-0_55.
27. El Arass, M.: Data Life Cycle: Towards a Reference Architecture. Int. J. Adv. Trends Comput. Sci. Eng. 9, 5645–5653 (2020). https://doi.org/10.30534/ijatcse/2020/215942020.
28. Santos, M.Y., Oliveira e Sá, J., Andrade, C., Vale Lima, F., Costa, E., Costa, C., Martinho, B., Galvão, J.: A Big Data system supporting Bosch Braga Industry 4.0 strategy. Int. J. Inf. Manage. 37, 750–760 (2017). https://doi.org/10.1016/j.ijinfomgt.2017.07.012.
29. Hildebrand, K.: Master Data Life Cycle – Management der Materialstammdaten in SAP®. In: Daten- und Informationsqualität. pp. 299–310. Springer Fachmedien Wiesbaden, Wiesbaden (2018). https://doi.org/10.1007/978-3-658-21994-9_18.

30. Demestichas, K., Daskalakis, E.: Data Lifecycle Management in Precision Agriculture Supported by Information and Communication Technology. Agronomy. 10, 1648 (2020). https://doi.org/10.3390/agronomy10111648.
31. Peng, X.: Construction of University Scientific Data Management Frame. In: 2020 International Conference on Big Data and Informatization Education (ICBDIE). pp. 60–63. IEEE (2020). https://doi.org/10.1109/ICBDIE50010.2020.00021.
32. El-Zoghby, A.M., Azer, M.A.: Cloud computing privacy issues, challenges and solutions. In: 2017 12th International Conference on Computer Engineering and Systems (ICCES). pp. 154–160. IEEE (2017). https://doi.org/10.1109/ICCES.2017.8275295.
33. Henaien, A., Ben Elhadj, H., Chaari Fourati, L.: Combined Machine Learning and Semantic Modelling for Situation Awareness and Healthcare Decision Support. Presented at the (2020). https://doi.org/10.1007/978-3-030-51517-1_16.
34. Moulos, V., Chatzikyriakos, G., Kassouras, V., Doulamis, A., Doulamis, N., Leventakis, G., Florakis, T., Varvarigou, T., Mitsokapas, E., Kioumourtzis, G., Klirodetis, P., Psychas, A., Marinakis, A., Sfetsos, T., Koniaris, A., Liapis, D., Gatzioura, A.: A Robust Information Life Cycle Management Framework for Securing and Governing Critical Infrastructure Systems. Inventions. 3, 71 (2018). https://doi.org/10.3390/inventions3040071.
35. Sinaeepourfard, A., Petersen, S.A.: Distributed-to-Centralized Data Management Through Data LifeCycle Models for Zero Emission Neighborhoods. In: Communications in Computer and Information Science. pp. 132–142. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-33495-6_11.
36. Polyzotis, N., Roy, S., Whang, S.E., Zinkevich, M.: Data Lifecycle Challenges in Production Machine Learning. ACM SIGMOD Rec. 47, 17–28 (2018). https://doi.org/10.1145/3299887.3299891.
37. Gurcan, F., Berigel, M.: Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges. In: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). pp. 1–6. IEEE (2018). https://doi.org/10.1109/ISMSIT.2018.8567061.
38. Chatzipanagiotou, N.: Toward an Integrated Approach to Information Management: A Literature Review. Presented at the (2017). https://doi.org/10.1007/978-3-319-33865-1_81.
39. Bychkov, I., Demichev, A., Dubenskaya, J., Fedorov, O., Haungs, A., Heiss, A., Kang, D., Kazarina, Y., Korosteleva, E., Kostunin, D., Kryukov, A., Mikhailov, A., Nguyen, M.-D., Polyakov, S., Postnikov, E., Shigarov, A., Shipilov, D., Streit, A., Tokareva, V., Wochele, D., Wochele, J., Zhurov, D.: Russian–German Astroparticle Data Life Cycle Initiative. Data. 3, 56 (2018). https://doi.org/10.3390/data3040056.
40. Al-Shdifat, A., Emmanouilidis, C., Starr, A.: Context-Awareness in Internet of Things - Enabled Monitoring Services. In: Lecture Notes in Mechanical Engineering. pp. 889–896. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-48021-9_98.