Vertrauen in die Wirklichkeit

AI, Trust und Reliability in den Digital Humanities

Kurz, Susanne
susanne.kurz@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0002-2824-1485

Eide, Øyvind
oeide@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0002-7766-6287

Digitale Objekte

Eine der wesentlichen Aufgaben der geisteswissenschaftlichen Forschung ist die Bewahrung des Kulturerbes und damit die Beschreibung, Interpretation und Kontextualisierung von Objekten des Kulturerbes. Kulturerbe wird hier im Sinne der UNESCO als Summe von materiellen und immateriellen Kulturgütern verstanden, die Zeugnisse der menschlichen Schaffens- und Schöpfungsfähigkeit zu einer bestimmten Zeit an einem bestimmten Ort darstellen. Objekte des Kulturerbes bilden somit einen wesentlichen Teil des Fundaments der Forschung über die unterschiedlichen methodologischen Ansätze, Verfahrensweisen und Fragestellungen in den verschiedenen Disziplinen der Geisteswissenschaften hinweg.

Moderne Objekte des Kulturerbes liegen häufig originär in digitaler Form vor (born digital objects)[1], während historische Kulturerbe-Objekte als physische Objekte und zunehmend zusätzlich als deren digitalisiertes Abbild (Digitalisat) verfügbar sind. Ist die Digitalisierung der Original-Objekte nach geeigneten Qualitätskriterien durchgeführt worden, können, je nach Forschungsfrage, auch die Digitalisate die Basis für Forschung sein. Zudem stehen Original-Objekte Forschenden häufig aus verschiedensten Gründen nicht zur Verfügung und in der Regel weisen die digitalen Objekte bei tiefer Erschließung auch einen Mehrwert auf.

Dieses Paper konzentriert sich auf die digitalen Objekte, die ein digitales Abbild eines ehemals oder aktuell vorhandenen Original-Objekts[2] darstellen.

Stellen solche Objekte die Grundlage für Forschung dar, vertrauen die Forschenden den für die Digitalisierung verantwortlichen Institutionen, dass das digitale Objekt ein unverfälschtes Abbild auf der Grundlage der aktuellen technischen Möglichkeiten eines tatsächlich existierenden Original-Objektes darstellt. Durch den Einsatz verschiedener technischer und workflowbasierter Verfahren kann eine Institution in der Regel die für unterschiedliche Forschungsziele erforderlichen Qualitäten erzeugen und gleichzeitig besteht im Allgemeinen ein vorbehaltloses Vertrauen seitens der Forschenden in die Authentizität des digitalen Objektes.

Nichtsdestotrotz muss und wird eine gewisse Fehlerrate akzeptiert. Keine Institution kann zu 100% sicherstellen, dass es bei den angewendeten Transformationsverfahren zum digitalen Objekt zu keinen relevanten Fehlern und ggf. problematischen Veränderungen gekommen ist.

Neben solchen technischen Fehlern stellen bewusst manipulierte digitale Objekte eine weitere mögliche Fehlerquelle dar. Manipulationen können von so hoher Güte sein, dass diese nur mit aufwendigen Verfahren zu erkennen sind. Das Erstellen von solchen hochwertigen Manipulationen ist aber gleichermaßen aufwendig und somit ist die Anzahl der zu erwartenden manipulierten digitalen Objekte eher gering und fällt in den Bereich der akzeptierten Fehlerrate.

Artifizielle digitale Objekte

Mit der breiten Verfügbarkeit von unterschiedlichen Softwaresystemen, die auf der Basis von KI-Algorithmen digitale Objekte mit scheinbar authentischen, tatsächlich aber fiktiven[3] Inhalten generieren können, entsteht ein neuartiges Problem für die moderne geisteswissenschaftliche Forschung. Forschende müssen die Möglichkeit in Erwägung ziehen, dass vorliegende digitale Objekte nicht aus vertrauenswürdigen Digitalisierungsprozessen hervorgegangen sind, sondern von entsprechenden Softwaresystemen mit artifiziellen Inhalten generiert wurden. Diese generierten digitalen Objekte weisen keinerlei Manipulationen auf, da es sich nicht um ein manipuliertes Abbild eines realen Original-Objektes handelt, sondern um ein generiertes Abbild eines „fiktiven Objektes“[4]. Da es keinen Transformationsprozess analog-digital gab, sind sie als Ganzes Fälschungen und können alle derzeit anerkannten Sicherheitskriterien erfüllen. Einzig die Zuordnung der digitalen Objekte zu Realien ist nicht möglich. Diese Objekte werden im Folgenden als artifizielle digitale Objekte bezeichnet[5].

Problematisch wird dies durch die hohe Skalierbarkeit der generierenden Verfahren. Die geschickte Nutzung von entsprechenden Softwaresystemen ermöglicht die kurzfristige Produktion einer großen Anzahl von unterschiedlichen, dabei aber inhaltlich zusammenhängenden artifiziellen Objekten (Urkunden, Briefe, Bilder, Ton-/Videoobjekte, ...). Solchermaßen kontextbezogene und untereinander abgestimmte Objekte erzeugen trotz ihrer Fiktion eine hohe Plausibilität.

Wird eine große Anzahl von in sich abgestimmten artifiziellen Objekten erzeugt, erscheinen die ggf. wenigen ordnungsgemäßen Objekte als manipuliert. So sind Kontextanalysen und weitere bisher zielführenden Verfahren zur Offenbarung von Falsifikaten (Barata, 2004) nur bedingt

geeignet, um konstruierte, alternative Realitätsdarstellungen aufzudecken.

Deepfake[6] kann im Kontext der digitalen Bewahrung von Kulturgütern insbesondere zur Kontextualisierung von artifiziellen Objekten eingesetzt werden. Nicht nur in gefälschten generierten Texten kann plausibler Kontext geschaffen werden, sondern Personen oder Institutionen, denen wir Vertrauen schenken, belegen, erläutern oder bekräftigen angeblich in Videos oder Podcasts Sachverhalte, die erst durch artifizielle Objekte geschaffen wurden. Dies stärkt einerseits die Glaubwürdigkeit artifizieller Objekte und ist andererseits als Fälschung nicht identifizierbar, da es sich, genau wie bei jedem artifiziellen Objekt, um ein generiertes Objekt handelt, das keine technischen Fälschungsparameter aufweist.

Existenzfrage

Es kann die Frage gestellt werden, ob artifizielle digitale Objekte bereits in unseren Sammlungen und Archiven vorhanden sind. Grundsätzlich existiert die Option artifizielle Objekte in unsere Speichersysteme einzuschleusen, weil nicht alle Softwaresysteme über entsprechende Sicherheitsmaßnahmen zum Schutz vor Injection-Angriffe[7] verfügen.

Letztlich kann aber keine Antwort auf die Existenzfrage gegeben werden. Auch wenn die angenommene Wahrscheinlichkeit tendenziell eher niedrig ist, muss bedacht werden, dass nur die Angriffe auf Softwaresystem wirklich erfolgreich sind, die nicht bemerkt werden. Die Realisation und Offenlegung von erfolgten Angriffen bestimmt jedoch unsere Einschätzung für deren Auftrittswahrscheinlichkeit. Erschwerend kommt hinzu, dass der Erfolg von Angriffen daran zu bemessen ist, dass niemand darauf aufmerksam wird.

Dies, gepaart mit der fehlenden Detektionsoptionen für artifizielle Objekte, führt zu der Feststellung, dass es sich unserer Kenntnis entzieht, ob und wenn ja wie viele artifizielle digitale Objekte bereits verbreitet sind.

Vertrauen in die Institutionen

Das Fälschen von Berichten, Bildern und Objekten ist kein neues Phänomen einer digitalen Gesellschaft. Zu allen Zeiten wurden Informationen absichtlich oder aus Versehen falsch weitergegeben. Das Vertrauen in verantwortliche Institutionen war und ist eine der wesentlichen Komponenten für die Entscheidung, ob Menschen Inhalte als wahrheitsgemäß oder falsch bewerten. Durch die Etablierung der digitalen Informationsverbreitung und -bewahrung hat sich nur die Art des Veränderns von Inhalten geändert, nicht aber die Tatsache selbst.

Generell konnten in der Prä-KI-Zeit die meisten Fälschungen durch Kontext- und Plausibilitätsprüfungen aufgespürt werden. Erst mit Einführung der generativen KI ist es möglich geworden, eine so große Anzahl an untereinander abgestimmten Fälschungen zu erstellen, dass diese in sich geschlossen und plausibel sind und die wahrhaft authentischen Objekte als Fälschungen erscheinen zu lassen.

Das Ergebnis jeder Forschungstätigkeit[8] unabhängig von der Forschungsdisziplin wäre wertlos, wenn diese auf artifiziellen digitalen Objekten mit fiktiven Inhalten beruhen würde. Aus diesem Grund benötigen Forschenden zukünftig eine neue Vertrauenskomponente in digitalen Objekten, die sicherstellt, dass es sich nicht um artifizielle, sondern um sorgfältig digitalisierte substanzielle Objekte handelt, deren Gegenstand das authentische digitale Abbild eines zum Zeitpunkt der Digitalisierung tatsächlich vorhandenen realen Original-Objektes ist.

Das wichtige Urvertrauen, das von Forschenden und der Gesellschaft in die Kulturerbe-Institutionen gesetzt wird, wird allein nicht mehr ausreichend sein und muss durch technische Komponenten unterstützt und gerechtfertigt werden.

Security Objectives

Lt. Simon (2020) ermöglicht Vertrauen auch in Situationen der Ungewissheit die Entwicklung eines Sicherheitsgefühls, wenn es möglich ist, sich auf das Handeln anderer zu verlassen.

Die Umsetzung von präventiven Schutzmaßnahmen ist ein bewährtes Mittel zur Vertrauensbildung. Elementare Kernpunkte für den Schutz von digitalen Objekten des Kulturerbes sind in den Schutzzielen der Informationssicherheit formuliert (NIST, 2004). Institutionen können das Vertrauen von Forschenden in die von ihnen zur Verfügung gestellten digitalisierten Objekte durch Beachten der international anerkannten Schutzziele für Informationssicherheit[9] rechtfertigen.

CIA-Triade

1. Confidentiality-Vertraulichkeit
2. Integrity-Integrität
3. Availability-Verfügbarkeit

Zwei Schutzziele des Identitätsmanagements ergänzen diese Triade:

1. Authenticity - Echtheit im Sinne von Ursprünglichkeit
2. Non-Repudiation - überprüfbare Beweise werden erstellt, dass das Erstellen des Objektes nicht in Abrede gestellt werden kann.

Vertraulichkeit bedeutet hier, dass nur berechtigte Personen digitale Objekte einsehen können (Autorisierung/Verschlüsselung).

Integrität hingegen zielt auf unbemerkte Modifikationen/Manipulationen und stellt die Korrektheit und Vollständigkeit sicher (elektronische Signatur/Siegel[10] oder Blockchain (Lo Duca et al., 2020)).

Die *Verfügbarkeit* beschreibt die Sicherstellung des Zugriffs auf die Daten und Vermeidung von Datenverlusten (BackUp-Strategien und gegebene Abrufbarkeit).

Ist ein Objekt mit geeigneten Verfahren und kryptografischen Algorithmen signiert, gesiegelt oder in einer Blockchain gesichert, kann sichergestellt werden, von wem und wann es erstellt und dass es anschließend nicht verändert wurde (BSI, 2020).

Bedacht werden muss aber unbedingt, dass elektronische Signaturen und Siegel sowie Blockchain Verfahren und digitale Wasserzeichen immer auf bestehende digitale Objekte angewendet werden und sie finden grundsätzlich erst nach der Transformation eines realen in ein digitales Objekt statt.

Für die Umsetzung der oben genannten Schutzziele gibt es konkrete Handlungsempfehlungen für die Institutionen des Kulturerbes, die auch im Forschungsdatenmanagement der DH zu finden sind. Jedoch bieten all diese keine unmittelbare Möglichkeit die oben beschriebene Gefährdung durch artifizielle Objekte zu verhindern, da sie erst dann ansetzen, wenn die Transformation von einem real existierenden in ein digitales Objekt erfolgt ist.

Es handelt sich um Schutzziele für alle digitale Objekte, zu denen auch die artifiziellen Objekte gehören. Sie stellen prinzipbedingt kein geeignetes Mittel zur Identifikation von artifiziellen digitalen Objekten dar.

Jedoch besteht die Möglichkeit, zertifikatsbasierte Signaturen/Siegel oder Blockchain-Lösungen sowie digitale Wasserzeichen von unterschiedlicher Ausprägung in digitalen Objekten zu verwenden, um sicherzustellen, dass eine bestimmte Institution (Zertifikatsinhaber) ein Objekt erzeugt hat und sich damit für die korrekte Transformation verantwortlich zeigt. Auf solchen Objekten basierende Forschungsergebnisse beruhen dann mit hoher Wahrscheinlichkeit nicht auf artifiziellen Objekten.

Digitales Vertrauen

Ein ungelöstes Kernproblem in der digitalen Welt ist das Abbilden von Vertrauen. Unterschieden werden muss unbedingt zwischen Proof und Trust. Nachweise stellen ein der Grundlagen für Vertrauen dar, können selbst aber nur bedingt Vertrauen schaffen. Vertrauen entsteht jenseits des Zweifels und ist wie bei Luhmann beschrieben eine soziale Kulturtechnik. Es stellt sich die Frage,

1. wann kann ein digitaler Inhalt als vertrauenswürdig gelten?
2. kann Vertrauen hergestellt werden, indem eine Vertrauenskomponente in digitale Objekte verankert wird?
3. was bedeutet der Verlust von Vertrauen in jegliche digitalen Objekte/Inhalte?

Content Authenticity Initiative

Die Idee der *Content Authenticity Initiative* "CAI"[11] ist, durch Anwendung kryptografischer Verfahren Inhalte und Metadaten gegen unbemerkte Manipulationen mit zertifizierten digitalen Signaturen zu schützen und Transparenz zu erzeugen, so dass Nutzende entscheiden können, ob sie den Inhalten Vertrauen schenken oder nicht. Basierend auf den C2PA Spezifikationen[12] werden Herkunfts-, Veränderungs- und Urhebernachweise festgehalten, die über ein Info-Icon in der Repräsentation des Objekts aufgerufen werden kann.

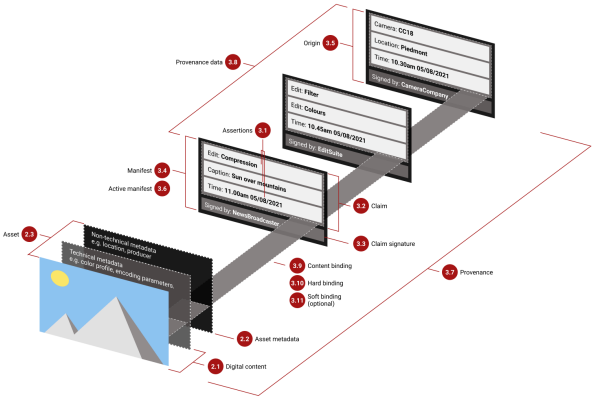

Abbildung 1: Bildnachweis: Coalition for Content Provenance and Authenticity, CC BY-SA 4.0, via Wikimedia Commons

So entsteht die Möglichkeit, selbsterzeugte Objekte CAI-konform zu kennzeichnen und festzuhalten, ob und wenn ja in welchem Umfang KI-Methoden Anwendung gefunden haben. Es erfolgt aber an keiner Stelle eine Überprüfung auf Wahrheitsgehalt. Es liegt ausschließlich bei den Usern zu entscheiden, ob sie dem System vertrauen oder nicht.

Bei dieser Vorgehensweise kann, genau wie bei allen anderen freiwilligen Selbstverpflichtungs-Maßnahmen bezüglich einer KI-Verwendungskennzeichnung[13], nicht sichergestellt werden, dass ein Objekt ohne Kennzeichnung, trotzdem KI generierte Inhalte enthält. Jede Art der freiwilligen Kennzeichnung darf nicht zu dem Umkehrschluss führen, dass nicht gekennzeichnete Objekte garantiert ohne KI entstanden sind.

Implikationen

Es kann durch unentdeckte artifizielle Objekte eine Bedrohungslage für eine authentische wissenschaftliche Forschung entstehen und es wird möglich, gezielt die Realität und die Historie verzerrt darzustellen. Die Manipulationsstrategie bei Verwendung artifizieller Objekte setzt nicht wie bisher auf der Ebene einzelner Objekte, sondern auf der Ebene des Gesamtverständnisses eines Sachverhaltes durch den Menschen an. Dies stellt eine deutlich umfangreichere Bedrohungslage dar.

Insbesondere eine Kombination aus Verbreitung artifizieller Objekte über soziale Netzwerke und deren Verfügbarkeit bei Portalen vertrauenswürdiger Institutionen stellt eine große Gefahr dar. Über die sozialen Netzwerke wird eine sehr hohe Reichweite erzielt und über Portale vertrauenswürdiger Institutionen großes Vertrauen in die Objekte hergestellt.

Die Kontrolle der Vergangenheit kann als eine Methode der Machtausübung eingesetzt werden, wie bereits George Orwell in seinem Roman 1984[14] schreibt: " Wer die Vergangenheit kontrolliert, der kontrolliert die Zukunft. ... Wer die Gegenwart kontrolliert, der kontrolliert die Vergangenheit." Falsche Aussagen von Machtausübenden können durch manipulierte historische Dokumente gestützt werden. Artifizielle Objekte unterstützen derartige Prozesse optimal.

Trustmanagement

Eine eindeutige Identifizierung von KI-generierten Material und ein berechtigtes Vertrauen in digitale nicht automatisch generierte Inhalte wird nicht nur in den Digital Humanities, sondern in vielen Disziplinen aus verschiedensten Gründen dringend benötigt.[15]

Zur Absicherung des prinzipiell vorhandenen Urvertrauens, das den Institutionen von Forschenden entgegengebracht wird, sollten Institutionen und Einrichtungen des Kulturerbes Maßnahmen (vor allem technischer Natur) ergreifen, um dieses Vertrauen in ihre digitalen Inhalte zu rechtfertigen.

Vertrauen ist auch im Forschungsprozess ein zentraler Aspekt.

Leider wird der Begriff Vertrauen oder Trust häufig nicht für das verwendet, was hier unter Vertrauen verstanden wird. Golbecks, Parsias, und Hendlers ‚Web of Trust' (2003) beschreibt ein Netz gegenseitiger Bestätigungen für digitale Schlüssel und Rahimzadeh Holagh und Mohebbi (2019) beschreiben für ihre im Semantic Web of Things enthaltenen Trust-Layer einen Blockchainansatz als Nachweiskette. Beides wird hier als Proof, nicht als Trust verstanden.

Ein interessanter Ansatz ist das ‚TrustNet' aus der angewandten KI-Forschung von Schillo und Funk (2000), das Agenten in einem Multiagenten System in die Lage versetzt, das Vertrauen in andere zu bewerten.

Im Kontext des Modells des Semantic Webs wurde 2001 die Idee eines Trust-Layer vorgestellt[16], aber es finden sich keine technischen Realisationen dazu.

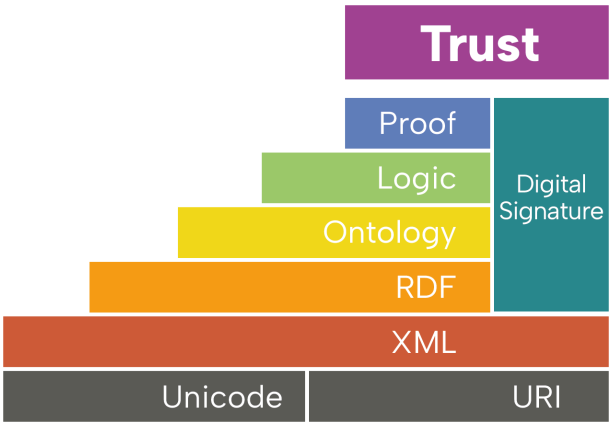

Abbildung 2: Semantic Web Layer Cake

Die Frage, die sich stellt, ist, wie ein erweitertes Modell der nicht neuen Idee des Trustmanagements aussehen könnte, bei dem eine Vertrauenskomponente explizit auf die Authentizität zur Sicherstellung von Non-Artifizialität im Objekt angewendet wird und somit eine Zero-Trust Lösung darstellt, die im Objekt selbst verankert ist und nicht auf einer Trust Chain beruht.

Fußnoten

1. Born digital data bringen für Kulturerbe-Institutionen viele Probleme mit sich und es besteht ein großer Forschungsbedarf in diesem Sektor. Dies ist ausdrücklich nicht Thema dieses Papers.
2. Konkrete Beispiele für digitale Objekte ohne verfügbares, zugehöriges Original-Objekt sind Abbilder von mit der Zeit unbrauchbar gewordenen Materialien wie Zeitungen, bei denen die zunächst vorhandenen Original-Objekte aufgrund der Papierbeschaffenheit mit der Zeit zerfallen oder digitale Objekte der Ruinenstadt Palmyra, wo bewusste Zerstörung des Baal-Tempels 2015 dazu führte, dass kein Original-Objekt verfügbar ist.
3. Fiktiv wird hier nicht in literaturwissenschaftlichen Sinn verwendet, sondern im Sinn von angenommen, erdacht oder erdichtet.
4. „Fiktive Objekte" werden hier als imaginäre Objekte gesehen, die das Potential haben, dass es sie tatsächlich geben könnte. Tatsächlich existieren sie aber nur in der Vorstellungskraft.
5. Born digital Objects haben ebenfalls keine Existenz außerhalb ihrer digitalen Form. Dennoch sind diese keine artifiziellen Objekte in diesem Sinn, da das Konzept dieser Objekte keine Idee eines digitalen Abbildes aufweist.
6. Deepfake bezeichnet eine Technologie, die die Erstellung von täuschend echt gefälschten Bildern, Videos oder Audiodateien ermöglicht. Diese Inhalte werden mithilfe von KI erzeugt, um Personen in bestimmten Situationen oder Handlungen darzustellen, die in Wirklichkeit nicht stattgefunden haben. Deepfakes können verwendet werden, um Gesichter in Videos zu manipulieren, Stimmen zu fälschen oder Szenarien zu erstellen, die authentisch wirken, obwohl sie in Wahrheit konstruiert sind. Die Technologie hat weitreichende Auswirkungen auf die Medienlandschaft, die Privatsphäre und die Glaubwürdigkeit von Inhalten, da sie die Grenzen zwischen Realität und Fiktion verschwimmen lässt.
7. Als Injection-Angriff wird das Ausnutzen von Sicherheitslücken im Zusammenhang mit Datenbanken verstanden, wobei Angreifende Datenbankbefehle einschleusen, Daten einfügen, auslesen, verändern oder löschen oder die Kontrolle über den Datenbankserver erlangen.
8. Gemeint ist hier jede Forschung, die auf digitalen Objekten beruht und Forschende davon ausgehen, dass diese ein ordnungsgemäß transformiertes Abbild eines realen Originals darstellen. Denkbar ist eine zukünftige For-

schung, die artifizielle Objekte zum Gegenstand hat. Für diese gilt diese Aussage nicht.
9. Beschrieben in ISO/IEC 27001, http://www.itref.ir/uploads/editor/42890b.pdf 12.07.2023 und NIST (2004)
10. „Technisch sind diese vergleichbar mit den einer juristischen anstatt einer natürlichen Person.“. https://www.bsi.bund.de/DE/Themen/Oeffentliche-Verwaltung/eIDAS-Verordnung/Elektronische-Signaturen-Siegel-und-Zeitstempel/elektronische-signaturen-siegel-und-zeitstempel_node.html 13.07.2023
11. https://contentauthenticity.org/ 20.11.2023
12. https://c2pa.org/ 20.11.2023
13. Zum Beispiel 21.07.2023 Selbstverpflichtung der große Techunternehmen, 09.2023 Kennzeichnungsoption bei TikTok, 11.2023 Konzept AI Safty Institute und Bletchley Declaration und vieles mehr.
14. https://politik.brunner-architekt.ch/wp-content/uploads/orwell_george_1984.pdf S. 366; 11.07.2023
15. Wie reagieren Systeme auf die Verwendung von KI-generierten Texten für das Trainieren von KI-Softwaresystemen? Shumailov et al. (2023) in‚The curse of recursion: Training on generated data makes models forget‘ festgestellt, das dies zu einem Modellkollaps führt. Außerdem werden bei Data Poisoning Attacken die Vorhersagemodelle korrumpiert und die gesellschaftlichen Auswirkungen eines Vertrauensverlustes in die Kulturerbeinstitutionen sind weitreichend.
16. https://www.w3.org/2001/12/semweb-fin/w3csw 21.11.2023

Bibliographie

Barata, Kimberly. 2004. Archives in the Digital Age. Journal of the Society of Archivists 25 (1): 63–70. https://doi.org/10.1080/0037981042000199151

BSI, Bundesamt für Sicherheit in der Informationstechnik. 2020. Leitlinie für digitale Signatur-/ Siegel-, Zeitstempelformate sowie technische Beweisdaten (Evidence Record). https://www.bundesnetzagentur.de/EVD/DE/SharedDocuments/Downloads/Anbieter_Infothek/BSI_TR_03125.pdf?__blob=publicationFile&v=1 13.07.2023

BSI, Bundesamt für Sicherheit in der Informationstechnik. 2021. Leitlinie für die beweiswerterhaltende Aufbewahrung gemäß BSI TR-03125 TR-ESOR. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/TechnischeRichtlinien/TR03125/BSI_TR-ESOR-LEIT.pdf;jsessionid=9F95D437F4BE4F66A6D623B5D3491164.internet481?1732. https://d__blob=publicationFile&v=5 13.07.2023

BSI, Bundesamt für Sicherheit in der Informationstechnik. 2023. IT-Grundschutz-Kompendium. 6. Edition 2023, Reguvis Fachmedien GmbH, Köln. https://d-nb.info/1282075888 und https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/IT-Grundschutz-Kompendium/it-grundschutz-kompendium_node.html 12.07.2023

Coester, Ulla, Norbert Pohlmann. 2021. Vertrauen – ein elementarer Aspekt der digitalen Zukunft. Datenschutz und Datensicherheit – DuD. Springer Nature. https://doi.org/10.1007/s11623-021-1401-x

Dittmann, Jana. 2000. Digitale Wasserzeichen: Grundlagen, Verfahren, Anwendungsgebiete., Berlin: Springer.

Fliehe, Marc, Brummel, Elisa. 2021. Eckpunkte eines sicheren Ökosystems für KI-Anwendungen. Datenschutz Datensicherheit - DuD 45, 444–447. https://doi.org/10.1007/s11623-021-1468-4

Golbeck, Jennifer, Bijan Parsia und James Hendler. 2003. Trust Networks on the Semantic Web. In: Klusch, Matthias, Andrea Omicini, Sascha Ossowski und Heimo Laamanen. 2003. Cooperative Information Agents VII. Lecture Notes in Computer Science, vol 2782. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-45217-1_18

Hornig, Anna, Christian Grünwald, Daniel Bonin, Jan Reichert, Marie-Kristin Komendzinski, Julian Sachs, Holger Glockner, Michael Astor. 2022. Studie: Die Zukunft des Vertrauens in digitalen Welten. [PDF]. Abgerufen von https://www.vorausschau.de/SharedDocs/Downloads/vorausschau/de/Foresight_Vertraueensstudie_Langfassung.pdf?__blob=publicationFile&v=1 (24.11.2023)

Liang, Xueping, Sachin Shetty, Deepack Tosh, Charles Kamhoua, Kevin Kwiat und Laurent Njilla. 2017. ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability. 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Madrid, Spain, 2017, pp. 468-477. https://doi.org/10.1109/CCGRID.2017.8

Lo Duca, Angelica, Clara Bacciu und Andrea Marchetti. 2020. The Use of Blockchain for Digital Archives: a comparison between Ethereum and Hyperledger (AIUCD 2019). Umanistica Digitale, 4(8). https://doi.org/10.6092/issn.2532-8816/9959

NIST National Institute of Standards and Technology. 2004. Standards for Security Categorization of Federal Information and Information Systems. Federal Information Processing Standards Publications FIPS PUB 199. Gaithersburg USA. https://doi.org/10.6028/NIST.FIPS.199

Rahimzadeh Holagh, Sam und Keyvan Mohebbi. 2019. A glimpse of Semantic Web trust. SN Appl. Sci. 1, oi.org/10.1007/s42452-019-1598-6

Simon, Judith. 2020. The Routledge Handbook of Trust and Philosophy. Routledge New York https://doi.org/10.4324/9781315542294

Schillo, Michael, Petra Funk, und Michael Rovatsos 2000. Using trust for detecting deceitful agents in artificial

societies. Applied Artificial Intelligence, 14:8, 825-848, https://doi.org/10.1080/08839510050127579

Shan, Shawn, Ding, Wenxin, Passananti, Josephine, Zheng Haitao, Zhao, Ben. Oktober 2023. Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. arXiv:2310.13828 [cs.CR]. https://doi.org/10.48550/arXiv.2310.13828

Shumailov, Ilia, Shumaylov, Zakhar, Zhao, Yiren, Gal, Yarin, Papernot, Nicolas, Anderson, Ross. Mai 2023. The curse of recursion: Training on generated data makes models forget. arXiv:2305.17493v2 [cs.LG]. https://doi.org/10.48550/arXiv.2305.17493

Stančić, Hrvoje. 2020. Trust and Records in an Open Digital Environment (1st ed.). Routledge. https://doi.org/10.4324/9781003005117