UNIVERSITÄT ZU KÖLN

INSTITUT FÜR DIGITAL HUMANITIES UNIVERSITÄT ZU KÖLN

Vertrauen in die Wirklichkeit

AI, Trust und Reliability in den DH

Susanne Kurz - Tagungsreihe Digitale Bibliothek 10 29.11.2024, Graz

Inhalt

Sogenannte KI

- Begrifflichkeit ‚Künstliche Intelligenz' erzeugt Unbehagen
- gefühlt ein gewaltiges Naturereignis, aber kein natürliches Phänomen
- proaktive Gestaltung vertrauenswürdiger und zuverlässiger Anwendungen

"We can only see a short distance ahead,

but we can see plenty there that needs to be done."

Alan Turing

Essentials I

Aufgaben der geisteswissenschaftlichen Forschung

- Beschreibung, Interpretation und Kontextualisierung von Kulturerbe-Objekten

Fälschungen – gab es schon immer

- Produktion von realen Objekten mit realitätsfremden Aussagegehalt

Vertrauen in Institutionen

- Vertrauen in Bibliotheken und Archive, keine Fälschungen zu verteilen

Fälschungen

- Kunstfälschungen von W. Beltracchi (überführt 2010)
- Protokolle der Weisen von Zion (um 1903)
- Vinland-Karte (um 1440)
- Konstantinische Schenkung (um 8. Jahrhundert)
- gefälschter Brief des Kaisers Mithridates an die Römer (um 1. Jahrhundert v. Chr.)

Protokolle der Weisen von Zion

- Historische Ungenauigkeiten
- Widersprüche
- sprachliche/stilistische Unterschiede
- Fälschung der Quellen
- Plagiatsnachweis

Titelblatt Sergej Nilus' Handexemplar seiner Ausgabe der „Protokolle der Weisen von Zion", 1917,
Sammlung M. Hagemeister, Foto: privat/Ansgar, Hoffmann, www.hoffmannfoto.de

Vinland-Karte

- Anomalien in der Tinte
- ungewöhnliche Alterungsmuster
- Stilistische Merkmale
- Ähnlichkeit mit anderen Fälschungen

Identifikation Fälschungen

typische Indikatoren

- Stil-, Material-, Handschriften-, Fotografie-Analysen und Vergleiche
- Überprüfung historische Genauigkeit
- Provenienzforschung

Verlässlichkeitsanzeichen

- Befragungen von Zeugen und mündliche Überlieferungen
- Offizielle Bestätigungen von führenden Persönlichkeiten oder Institutionen (kirchliche oder staatliche Autoritäten).

Essentials II

Digitales Kulturerbe als Forschungsgrundlage

- "born digital data" und retrodigitalisierte digitale Objekte

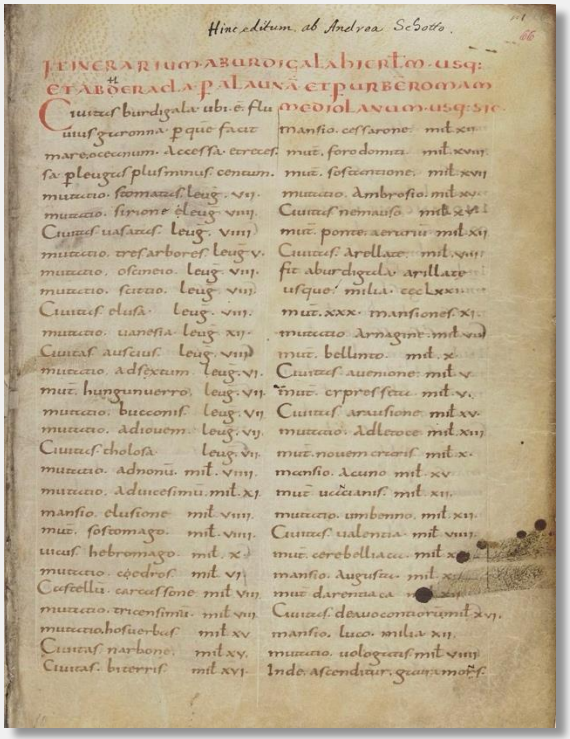
Digitale Verfälschungen

- Pixelmanipulation und vergleichbare Verfahren

Vertrauen in digitale Objekte

- Vertrauen in Bibliotheken und Archive, die für Digitalisierung verantwortlich sind
- Akzeptanz gewissen Fehlerrate bei digitalen Transformationsverfahren

Digitales Abbild



Itinerarium Burdigalense, 333-334

Bibliothèque nationale de France

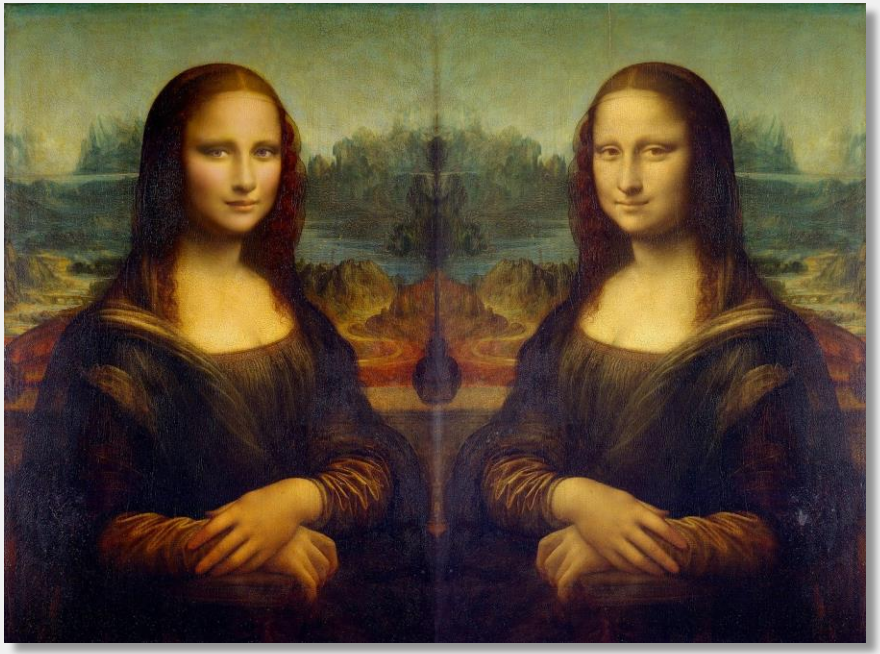
Digitales Abbild

da Vinci, Mona Lisa, um 1503

Paris, Musée du Louvre

Verfälschungen

Pixelmanipulation



Quelle: Pixabay

Verfälschungen

Pixelmanipulation

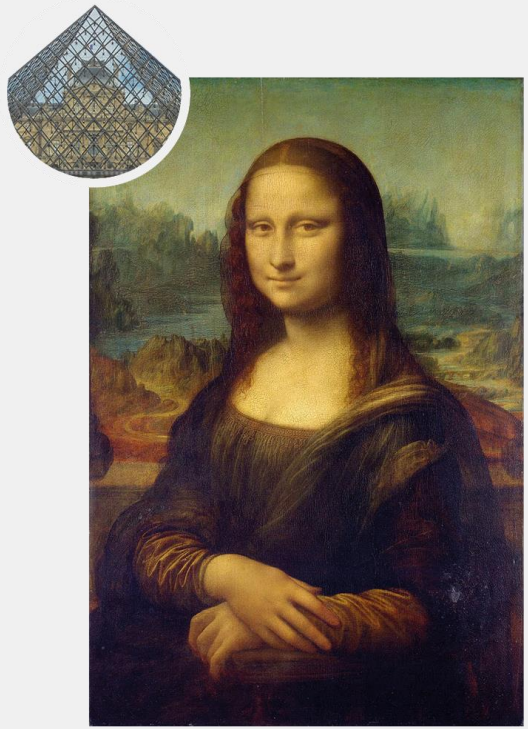
Quelle: Pixabay

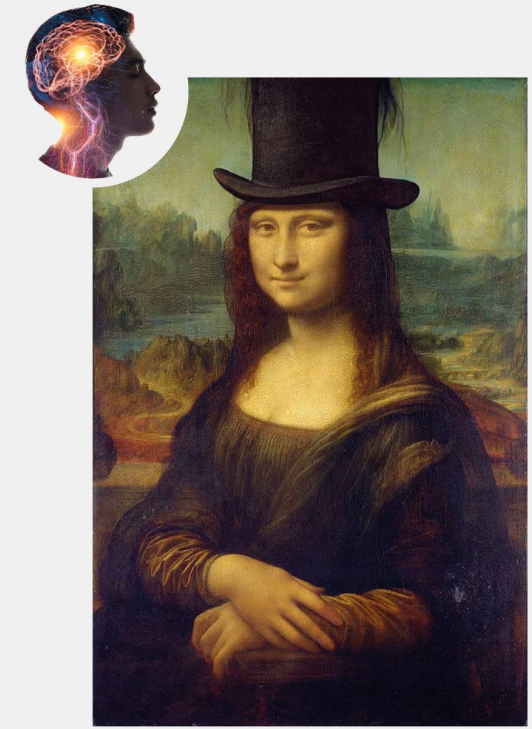
Essentials III

Generierende Software und KI-Algorithmen

- Möglichkeit des Generierens digitaler Objekte mit imaginären Inhalten
- Entstehung artifizieller digitaler Objekte als Abbild einer Vorstellung, die nicht außerhalb des Geistes existiert
- Problem der Unterscheidung → Ungewissheit

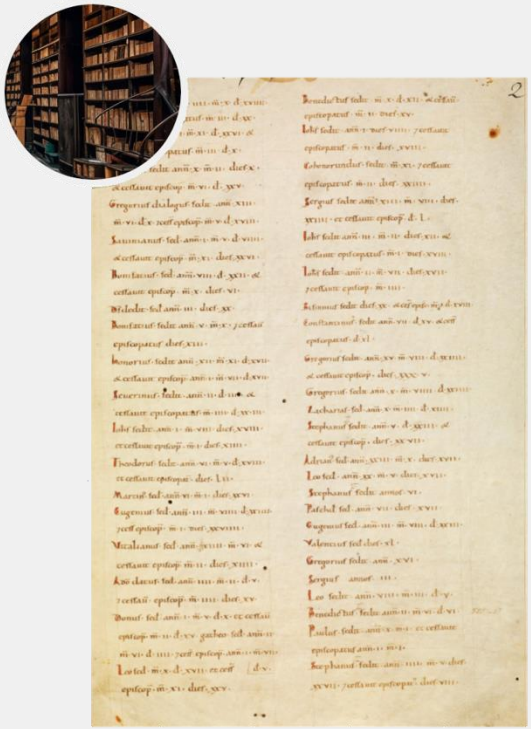
Fälschung oder alternatives Original?

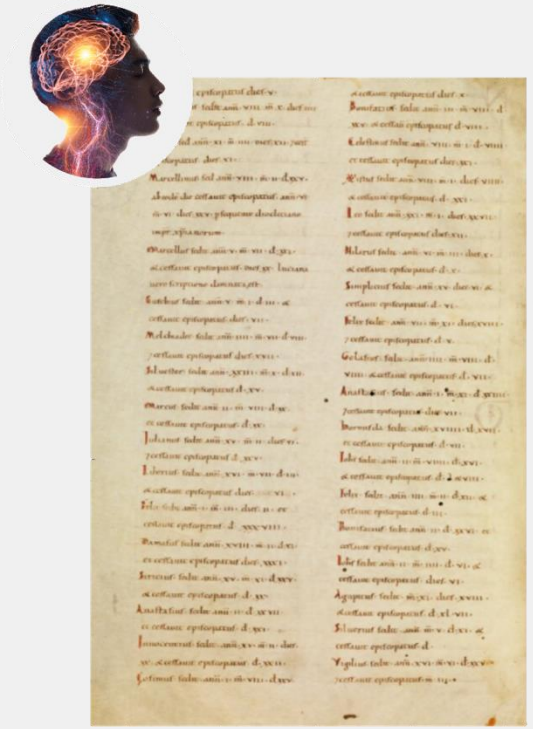




Digitales Abbild wovon?

Problem: Ungewissheit



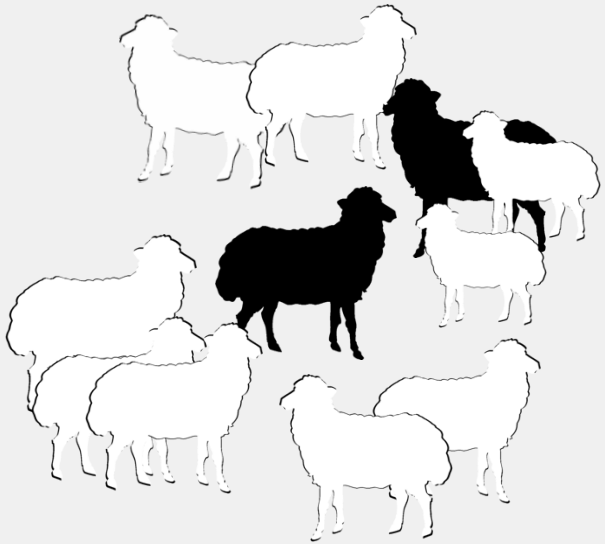


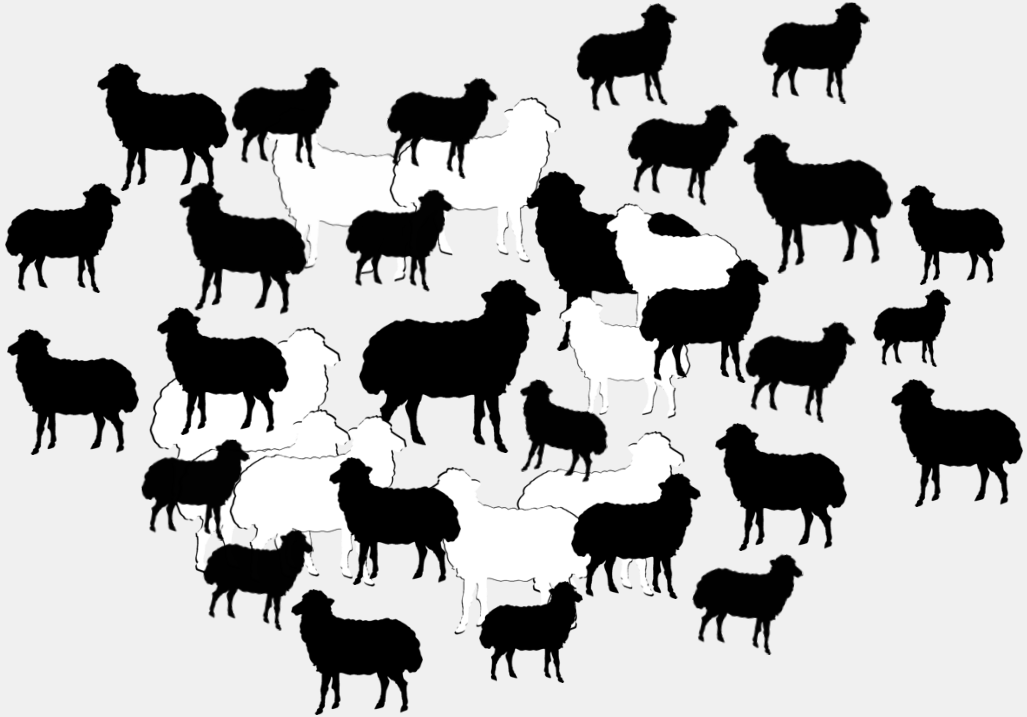
Artifizielle digitale Objekte

Herausforderungen

1. hohe **Skalierbarkeit** generierender Verfahren
2. Problematik **Kontextanalyse** durch imaginären Kontext
3. **Scheinsicherheit**: CIA-Triade & friends können gleichermaßen auf alle digitalen Objekte angewendet werden

1. Skalierbarkeit





2. Kontextanalyse

Generieren von Deepfake

- Erzeugung plausibler Texte, Bilder, Audios und Videos
- Algorithmus erzeugt oder manipuliert Videos, Gesichter und Stimmen

Was ist »Fake Kontext«?

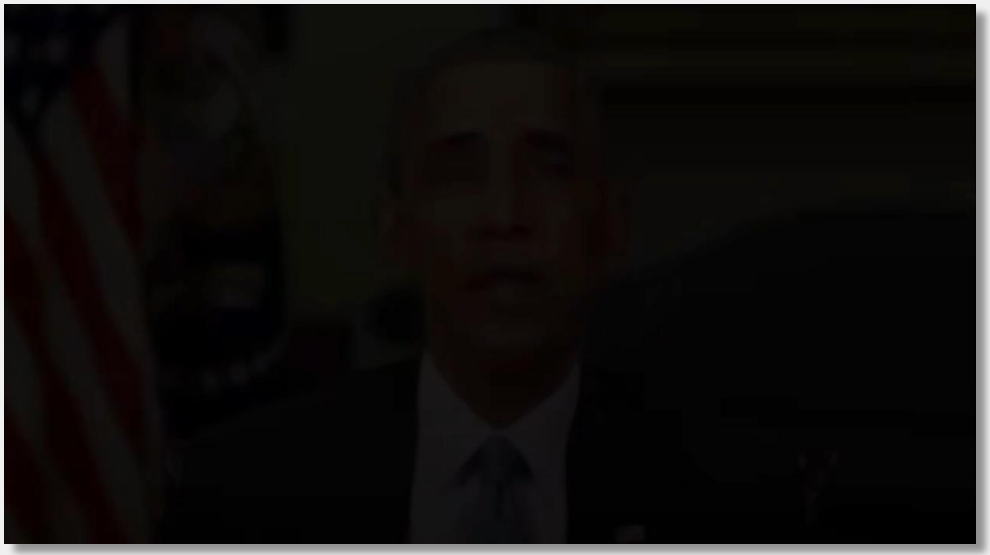
- Identifizierungsproblem aufgrund generativer Verfahren
- Deepfake + Fake News untergräbt Vertrauen in digitalen Content »Fake Reality«

Objekte kulturellen Erbes

- Kontextualisierung von artifiziellen Objekten
- Schaffung von Glaubwürdigkeit für artifizielle Objekte

Keine Identifikationsoption

Deepfake Video



https://youtu.be/gLoI9hAX9dw?si=7Bd6MDKSSBgrf7Ta

Sora generiertes Video

Prompt: Reflections in the window of a train traveling through the Tokyo suburbs

https://cdn.openai.com/sora/videos/train-window.mp4

Speech to Speech (ElevenLabs)

- Generieren von gesprochenen Audioaufnahmen in vielen Sprachen
- Klonen von Stimmen in Echtzeit – Basis: 1 Minute Audiorecording
- Problem der Unterscheidung / Identifizierung

3. Scheinsicherheit

Sicherheitsmaßnahmen

- Elektronische Signaturen, Siegel, Wasserzeichen, Blockchain-Lösungen

CIA-Triade und Identitätsmanagement

- Vertraulichkeit, Integrität und Verfügbarkeit als Schlüsselziele
- + Authentizität und Verbindlichkeit – Vertrauenssignale

Anwendungsbereich

- Schutzziele finden Anwendung bei jeder Art von digitalen Objekten
- Erfolgt **nach** einer Transformation in digitale Objekte

Keine Identifikationsoption

Verantwortung

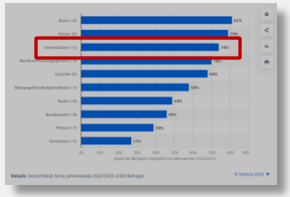
Vertrauen als soziale Kulturtechnik -Luhmann-

- alltäglicher Umgang mit Unsicherheiten und Komplexitäten
- Erwartung an Verlässlichkeit

Zuverlässigkeit der Bibliotheken und Archive

- Verantworten korrekte Transformationsprozesse ✅
- Gewährleisten, dass digitale Objekte authentische Abbilder sind ✅
- Sicherstellen der Schutzziele (CIA-Triade) ☝

Vertrauen



Urvertrauen in Kulturerbe – Institutionen

- Forschenden haben derzeit Urvertrauen

Vertrauenserhalt

- Vertrauen basiert auf individuelle Wahrnehmung von Kompetenz und Ehrlichkeit
- ebenso Integrität, Authentizität, Verlässlichkeit → klass. Schutzziele

NIST: CIA-Triade & Friends

US **N**ational **I**nstitute of **S**tandards and **T**echnology

Informationssicherheit

- **C**onfidentiality – Vertraulichkeit (Schutz vor unbefugtem Zugriff)
- **I**ntegrity – Integrität (Nachweis für keine Datenmanipulation)
- **A**vailability – Verfügbarkeit (Daten sind jederzeit verfügbar und nutzbar)

+ Schutzziele Identitätsmanagement

- Authenticity - Echtheit im Sinne von Ursprünglichkeit (Bestätigung der Identität)
- Non-Repudiation – Nicht Abstreitbarkeit

  Erstellung des Objekts ist unbestreitbar.

Vertrauenssignale

Schutzziele

- Umsetzung der NIST Schutzziele für digitale Daten

Labeling

- Ausdruck der Verantwortung und Zuverlässigkeit
- anerkanntes Label für ordnungsgemäße digitale Objekte
- Kennzeichnungs-Selbstverpflichtung für Forschende wenig hilfreich

ausreichend?

mehr Vertrauenssignale

weitere Vertrauenssignale

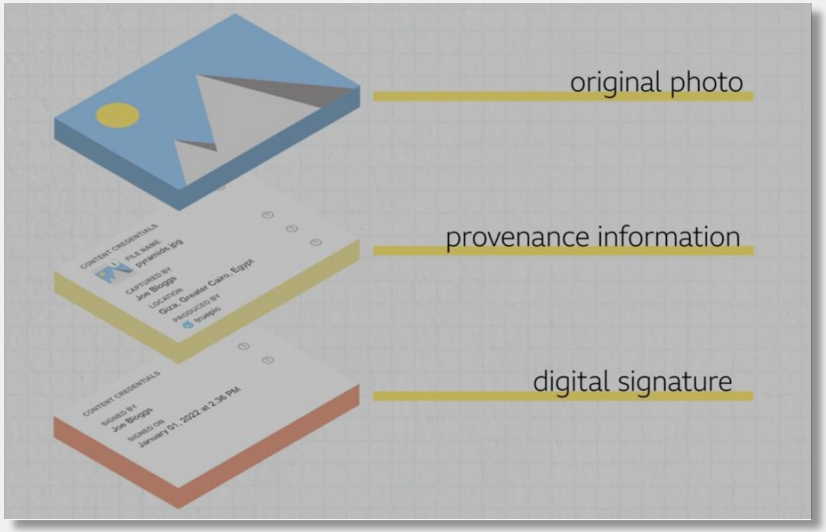
- Provenienz
- Transparenz über Veränderung

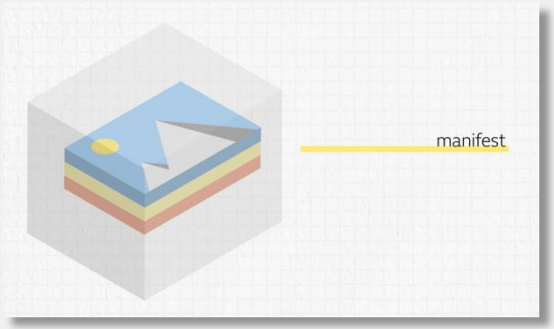
Coalition for Content Provenance and Authenticity

- C2PA
- offener Metadatenstandard von 2021
- Ziele
  - Herkunft und Authentizität digitaler Inhalte nachverfolgen
  - Desinformation bekämpfen

C2PA Manifeste

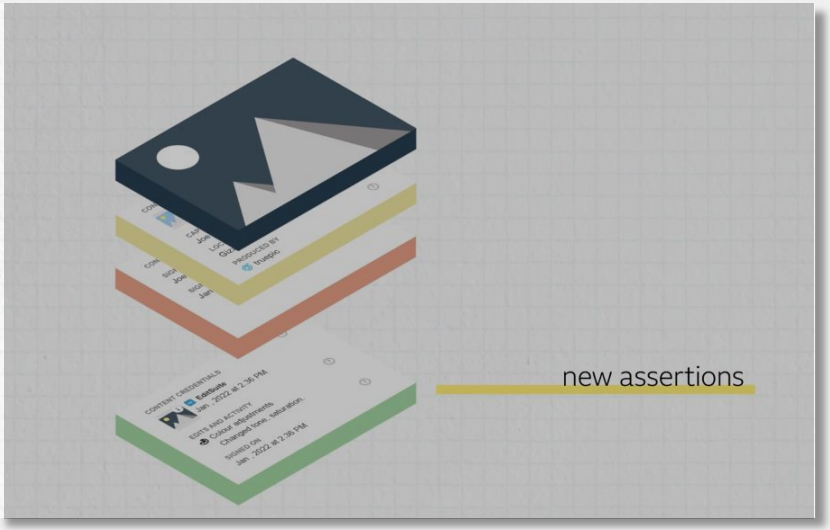
C2PA Metadaten in Manifesten

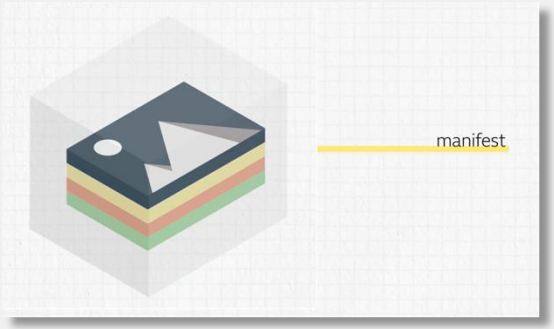




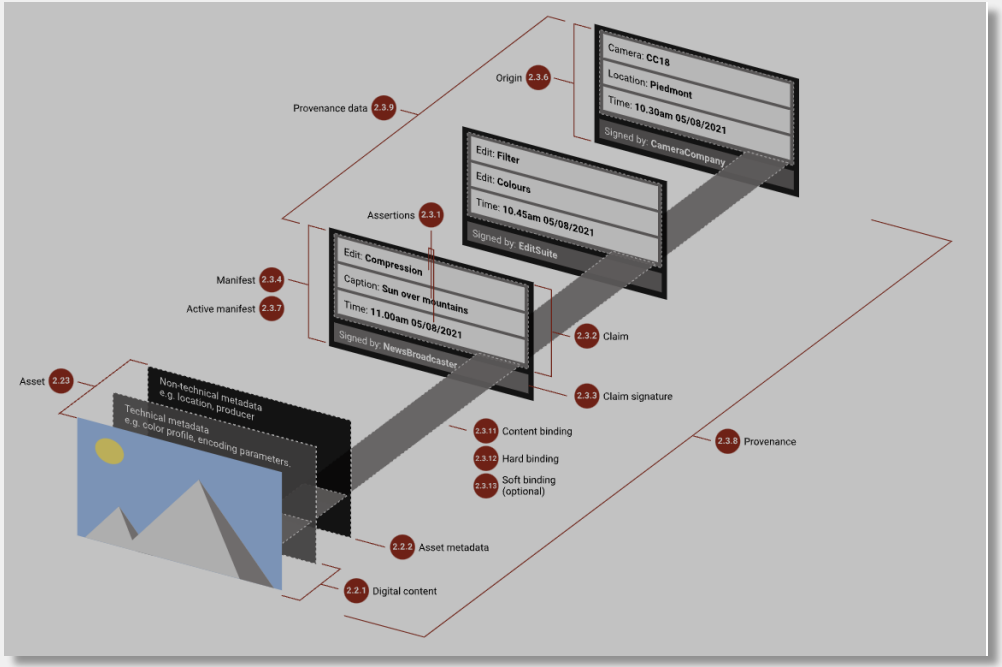
C2PA Assertations

Manifeste stellen **Vertrauensignale** bereit.





C2PA Provenance Information



“It is important to highlight that

C2PA specifications do not provide value judgments

about whether a given set of provenance data is 'true', but instead merely whether the provenance information can be verified as associated with the underlying asset, correctly formed, and free from tampering.”

“In the C2PA Specifications,

trust decisions are made by the consumer

of the asset based on the identity of the actor(s) who signed the provenance data along with the information in the assertions contained in the provenance.”

Vertrauen in digitalen Räumen

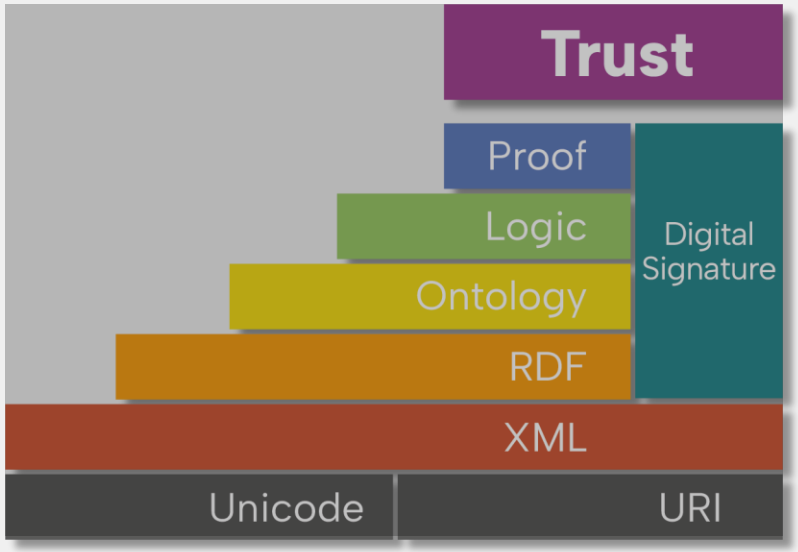
Vertrauen modellieren?

- verschiedene Arten von Vertrauen im digitalen Raum unterscheiden
- Modellierung von Vertrauen in digitalen Räumen

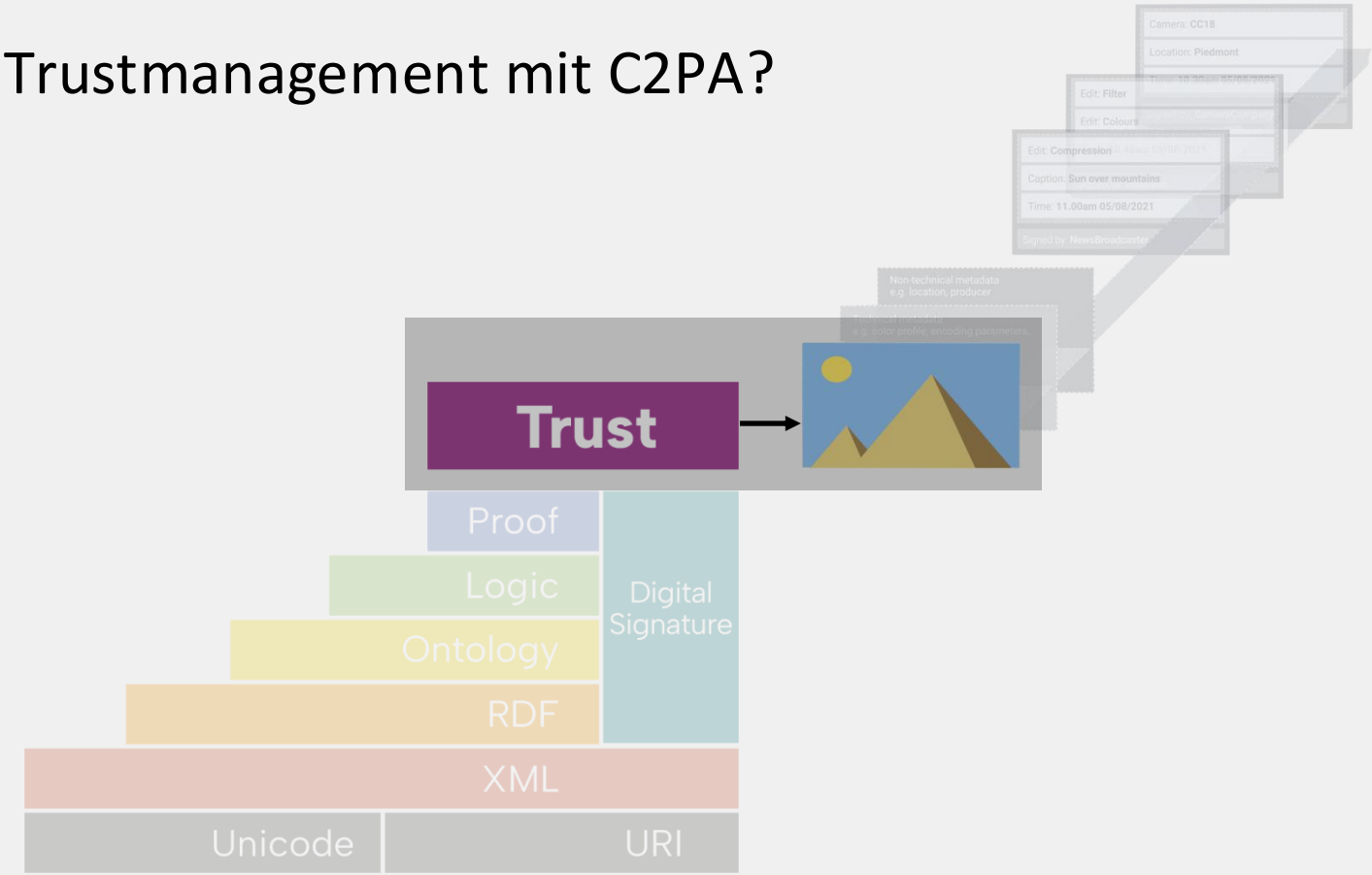
Trust-Leveling-Engine TLE

- technische Lösung zur Trust Level Bestimmung
- basierend auf vorliegenden Trust-Informationen
- gewichtet über Vertrauensmodelle
- Visualisierungsoptionen von Rating bis vollständige Metadaten

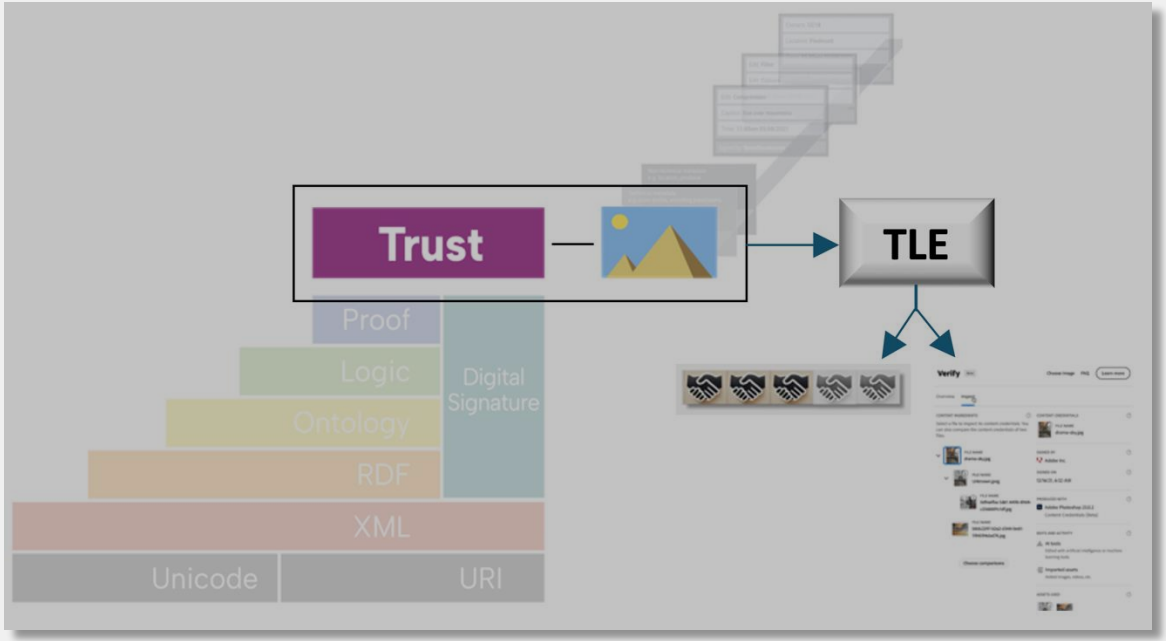
Semantic Web 2001



Trustmanagement mit C2PA?



Trust Leveling Engine TLE



Zusammenfassung

- neue Qualität von ‚Fälschungen' durch KI
- bewährte Detektionsmechanismen greifen nicht mehr

- derzeit gibt es keine zuverlässige Identifikationsmöglichkeit
- Verzerrung der Realität durch artifizielle Objekte »Fake Reality«

- Forschende haben Urvertrauen in Bibliotheken und Archive
- Verantwortung übernehmen und Zuverlässigkeit visualisieren
- Provenienzdaten und Transparenz – Trustsignale für User

Quellen

- Goals and Non-Goals of C2PA Specifications https://c2pa.org/specifications/specifications/1.0/explainer/Explainer.html#_how_is_trust_in_digital_assets_established
- https://www.heise.de/download/product/deepfakes-fakeapp ; 26.10.2023
- https://doi.org/10.1001/jamaophthalmol.2023.5162 . 24.11.2023
- http://www.itref.ir/uploads/editor/42890b.pdf 12.07.2023
- 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
- https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/ 30.08.2023
- https://newsroom.tiktok.com/de-de/neue-hinweise-zur-kennzeichnung-von-ki-generierten-inhalten 12.11.2023
- https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/ 04.11.2023
- https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 04.11.2023
- https://contentauthenticity.org/ 20.11.2023
- https://c2pa.org/ 20.11.2023
- https://www.w3.org/2001/12/semweb-fin/w3csw 21.11.2023
- https://doi.org/10.1080/0037981042000199151
- https://doi.org/10.1007/s11623-021-1401-x
- https://doi.org/10.1007/s11623-021-1468-4
- https://doi.org/10.1007/978-3-540-45217-1_18
- https://www.vorausschau.de/SharedDocs/Downloads/vorausschau/de/Foresight_Vertraueensstudie_Langfassung.pdf?__blob=publicationFile&v=1 (24.11.2023)
- https://doi.org/10.1109/CCGRID.2017.8
- https://doi.org/10.6028/NIST.FIPS.199
- https://doi.org/10.1007/s42452-019-1598-6
- https://doi.org/10.1080/08839510050127579
- https://doi.org/10.48550/arXiv.2310.13828
- https://doi.org/10.4324/9781003005117
- https://www.heise.de/news/Deepfakes-Neuronale-Netzwerke-erschaffen-Fake-Porn-und-Hitler-Parodien-3951035.html; 26.10.2023

Danke!

Ich freue mich auf Diskussionen.