SACHSEN-ANHALT

Landesarchiv

Bestandserhaltungsplanung von digitalem Archivgut mittels Verknüpfung von semantischen Technologien und KI

Björn Steffenhagen M. A. mult.
Landesarchiv Sachsen-Anhalt, Abt. 1 Zentrale Dienste

29. Archivwissenschaftliches Kolloquium
20.05.2025 in Marburg

Hintergrund

SACHSEN-ANHALT

Landesarchiv

1. Herausforderung dig. Bestandserhaltung in Archiven
2. KI: Erwartungen und Probleme
3. Semantische Technologien
4. Projekt PROMETHEUS
5. Schlussfolgerungen

Ziele der digitalen Bestandserhaltung im Archivwesen

SACHSEN-ANHALT

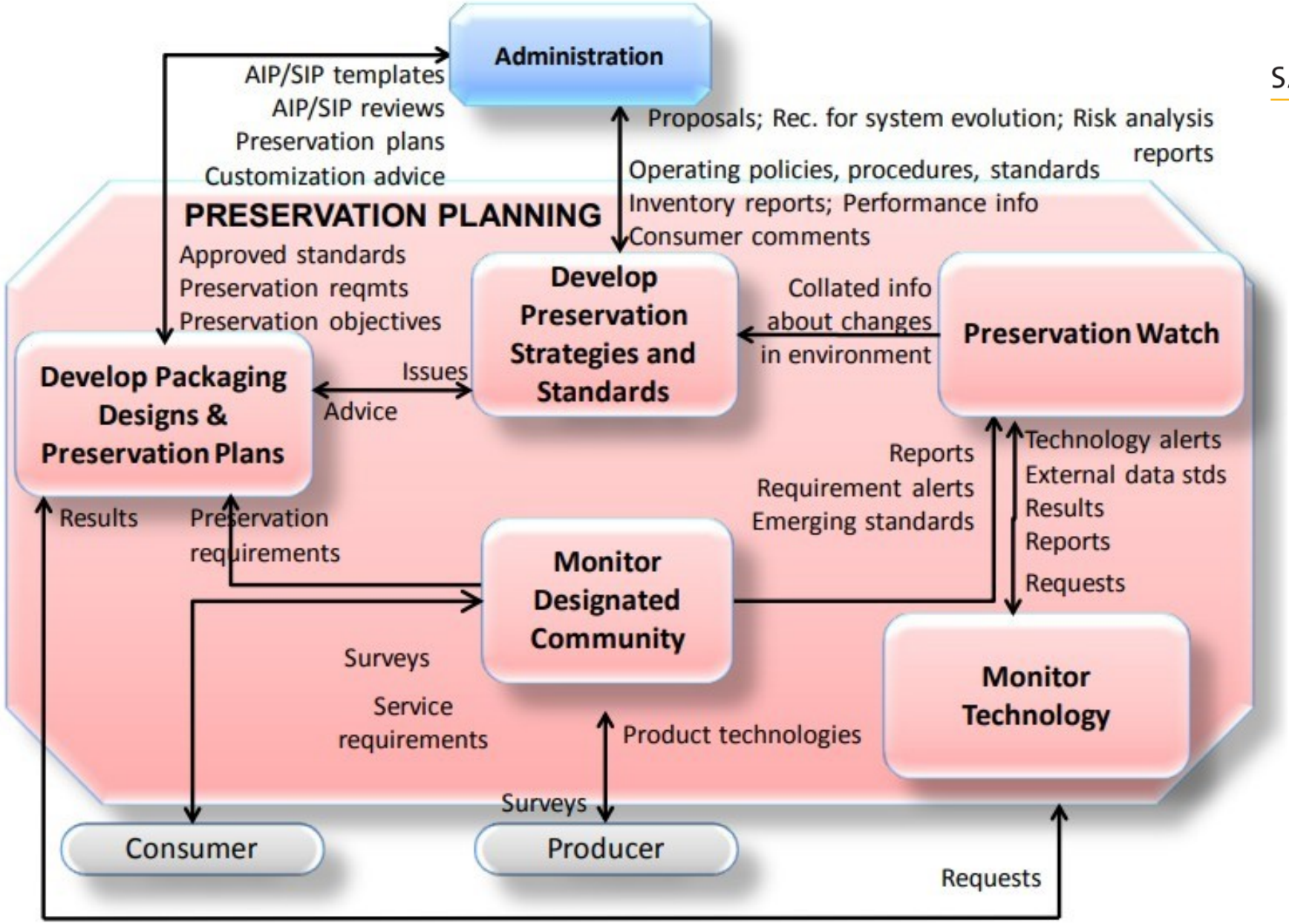
Landesarchiv

Sind u. a.:

- Erhalt der Integrität und Authentizität
- Nutz-, Verfüg-/und Verstehbarkeit erhalten
- Dokumentation des Verwaltungshandelns (Provenienz und Archiv)

->Vertrauenswürdigkeit herstellen

->evidenzbasierte Auswertungen ermöglichen



Quelle: Consultative Committee for Space Data Systems (CCSDS): Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-M-3, Magenta Book, Washington DC 2024, S. 4-15.

Herausforderungen der digitalen Bestandserhaltung

SACHSEN-ANHALT

Landesarchiv

- Masse an Primär-/und Metadaten nur noch (teil-)automatisiert verarbeitbar
- Eingriffe durch automatisierte Erhaltungsmaßnahmen bergen Fehlerrisiko
- fragmentierte und heterogene Informationen erschweren konsistente Erhaltungsentscheidungen
- fehlende archivische Bewertungskriterien für Erhaltungsbedarf und Maßnahmenpriorisierung ->technischer **und** fachlicher Kontext eines Informationsobjekts

Herausforderungen der digitalen Bestandserhaltung - Beispiele

SACHSEN-ANHALT

Landesarchiv

- Qualität der PRONOM-Datenbank der National Archives UK zur Dateiformaterkennung
- fragmentiertes Wissen: PRONOM, Wikidata for DP, COPTR, LoC, KOST, Empfehlungen staatlicher Archive etc.
- Umgang mit Ergebnissen der Dateiformatvalidierung
- Migration im Kontext: Ist eine JPEG-Datei in einer E-Akte gleich wie in einer Webseite?
- Priorisierung: Fehlende Bewertungskriterien bei Bestandserhaltungsmaßnahmen

Probleme der KI bzw. LLM

SACHSEN-ANHALT

Landesarchiv

Automatisierung ist verlockend, Probleme aber u.a.:

- Halluzinationen
- unzuverlässige Entscheidungen
- Black-Box-Problematik – mangelnde Transparenz
- fehlendes, domänenspezifisches Wissen

➔Archive brauchen jederzeit reproduzierbare, verlässliche und steuerbare Methoden

➔Diese durch den alleinigen Einsatz von KI (noch) nicht erreichbar

Semantische Technologien

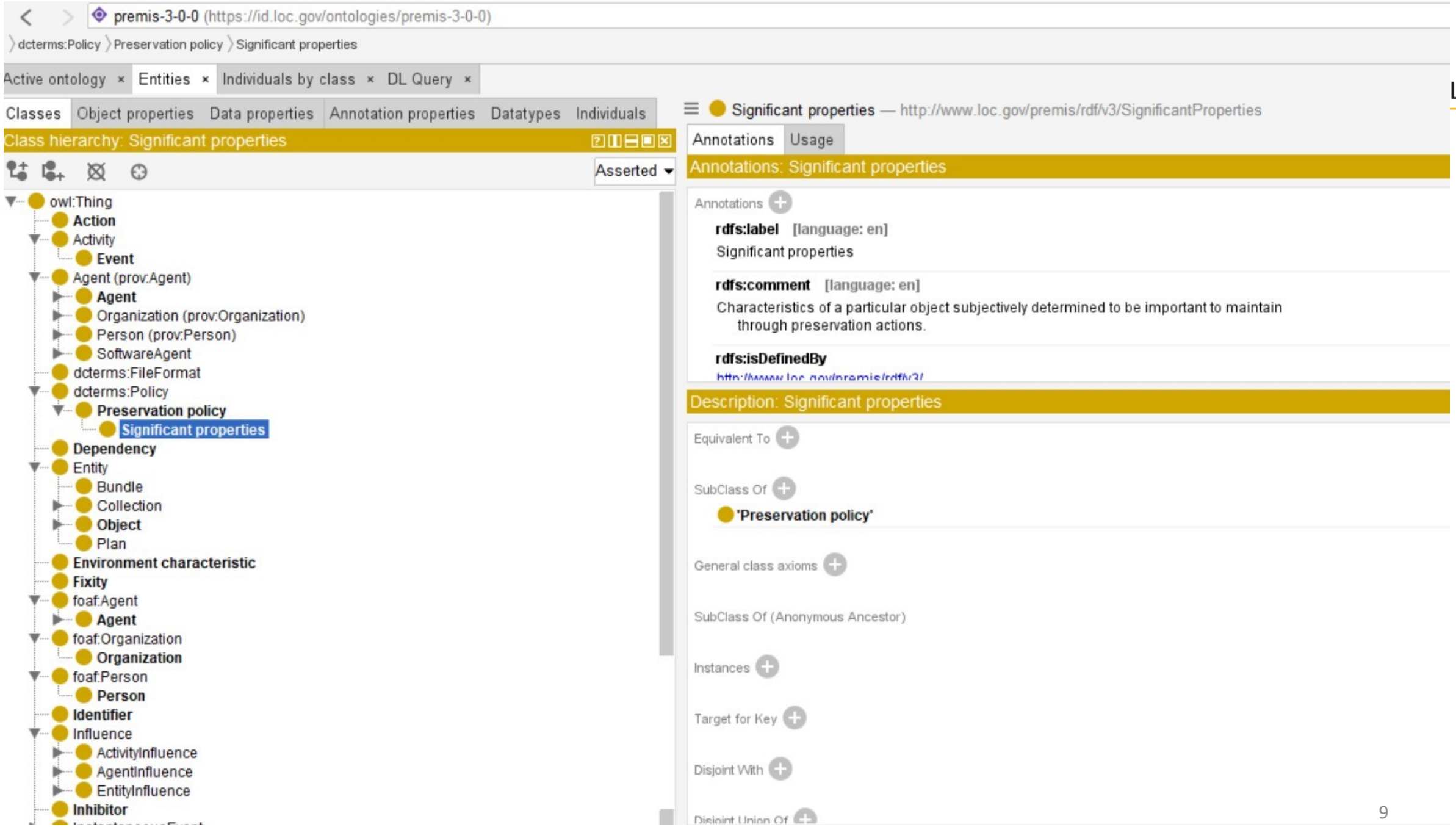
SACHSEN-ANHALT

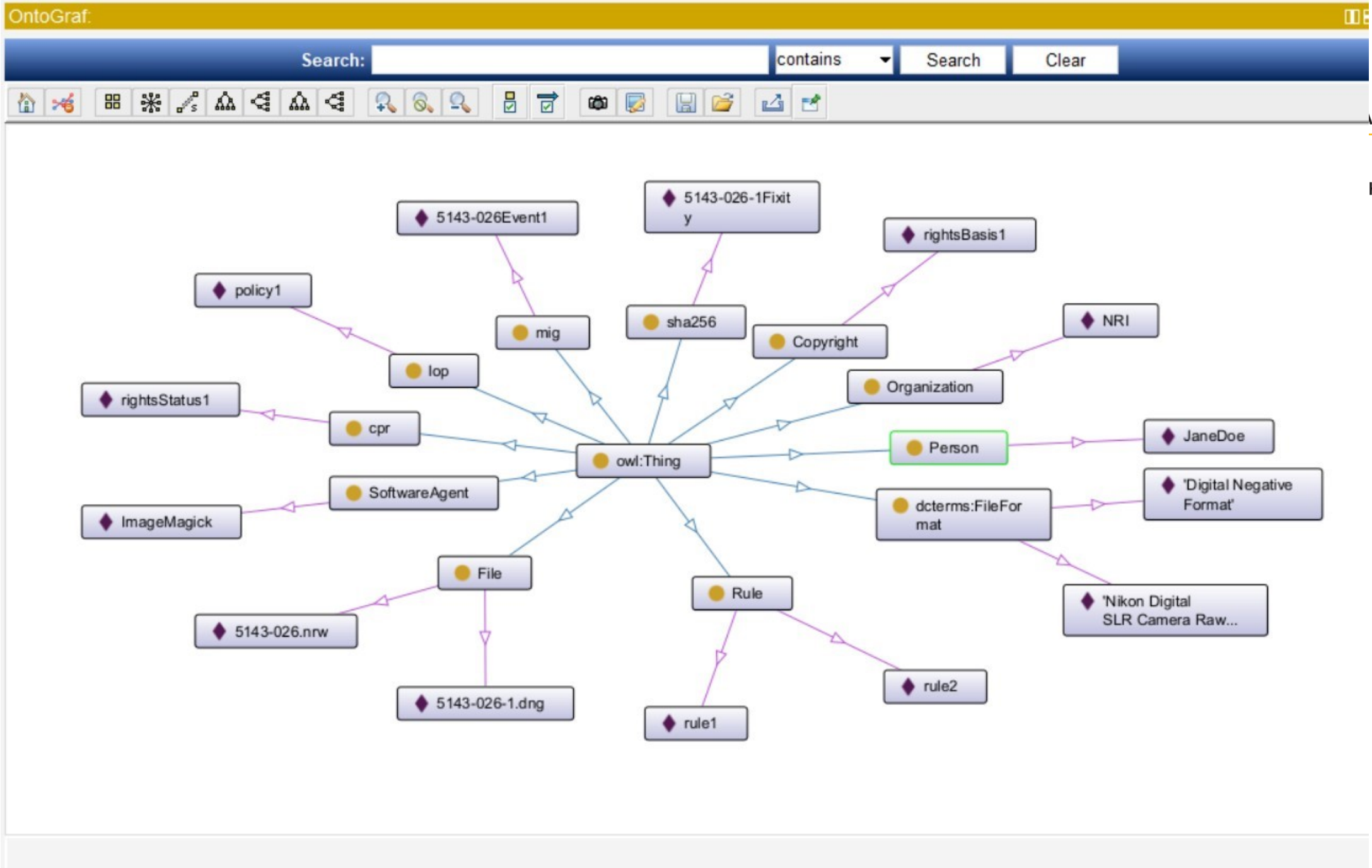
Landesarchiv

- Sollen maschinenlesbare Modelle mit Bedeutung schaffen, um Daten automatisiert zu verknüpfen, anzureichern und qualitativ zu bewerten.
- Explizite, formale Repräsentation von Konzepten + Relationen

-> Web of (meaningful) Data

- 3 wesentliche Technologien:
  - Ontologie
  - Wissensgraph
  - SPARQL


premis-3-0-0 (https://id.loc.gov/ontologies/premis-3-0-0)
dcterms:Policy 〉Preservation policy 〉Significant properties
Active ontology × Entities × Individuals by class × DL Query ×
Classes Object properties Data properties Annotation properties Datatypes Individuals
Class hierarchy: Significant properties
Asserted
owl:Thing
Action
Activity
Event
Agent (prov:Agent)
Agent
Organization (prov:Organization)
Person (prov:Person)
SoftwareAgent
dcterms:FileFormat
dcterms:Policy
Preservation policy
Significant properties
Dependency
Entity
Bundle
Collection
Object
Plan
Environment characteristic
Fixity
foaf:Agent
Agent
foaf:Organization
Organization
foaf:Person
Person
Identifier
Influence
ActivityInfluence
AgentInfluence
EntityInfluence
Inhibitor
Significant properties — http://www.loc.gov/premis/rdf/v3/SignificantProperties
Annotations Usage
Annotations: Significant properties
Annotations
rdfs:label [language: en]
Significant properties
rdfs:comment [language: en]
Characteristics of a particular object subjectively determined to be important to maintain through preservation actions.
rdfs:isDefinedBy
Description: Significant properties
Equivalent To
SubClass Of
'Preservation policy'
General class axioms
SubClass Of (Anonymous Ancestor)
Instances
Target for Key
Disjoint With
Disjoint Union Of


OntoGraf:
Search:
contains
Search
Clear
5143-026Event1
5143-026-1Fixit y
rightsBasis1
policy1
mig
sha256
NRI
Copyright
lop
Organization
rightsStatus1
cpr
Person
JaneDoe
owl:Thing
SoftwareAgent
'Digital Negative Format'
ImageMagick
dcterms:FileFor mat
File
Rule
'Nikon Digital SLR Camera Raw...
5143-026.nrw
rule2
5143-026-1.dng
rule1


PANIC
DURAAK
CIDOC CRM
TOTEM
ENSURE
PERICLES
Europeana Data Model
Integrated Preservation Suite
Planets
XCL-Ontology
PREMIS OWL
InterPARES
PROV-O

Probleme der semantischen Technologien

SACHSEN-ANHALT

Landesarchiv

- streng formeller Modellierungsprozess
- Pflegeaufwand
- Expertensysteme für Fachdomänen
- Abfragemöglichkeiten via SPARQL unkomfortabel
- in Kultureinrichtungen (noch) ungenügend etabliert
- Beispiel: Archivführer Deutsche Kolonialgeschichte: https://archivfuehrer-kolonialzeit.de/index.php/

Lösungsansatz: Kombination aus semantischen Technologien und KI

SACHSEN-ANHALT

Landesarchiv

- Ontologie und Graphen als Wissensbasis
- KI-Modelle als
  - Abfragemöglichkeit
  - Möglichkeit der Entscheidungsfindung
  - Erweiterung des Gesamtmodells

- Ergebnis bildet ein RAG-System (Retrieval-Augmented Generation)

Projekt PROMETHEUS

SACHSEN-ANHALT

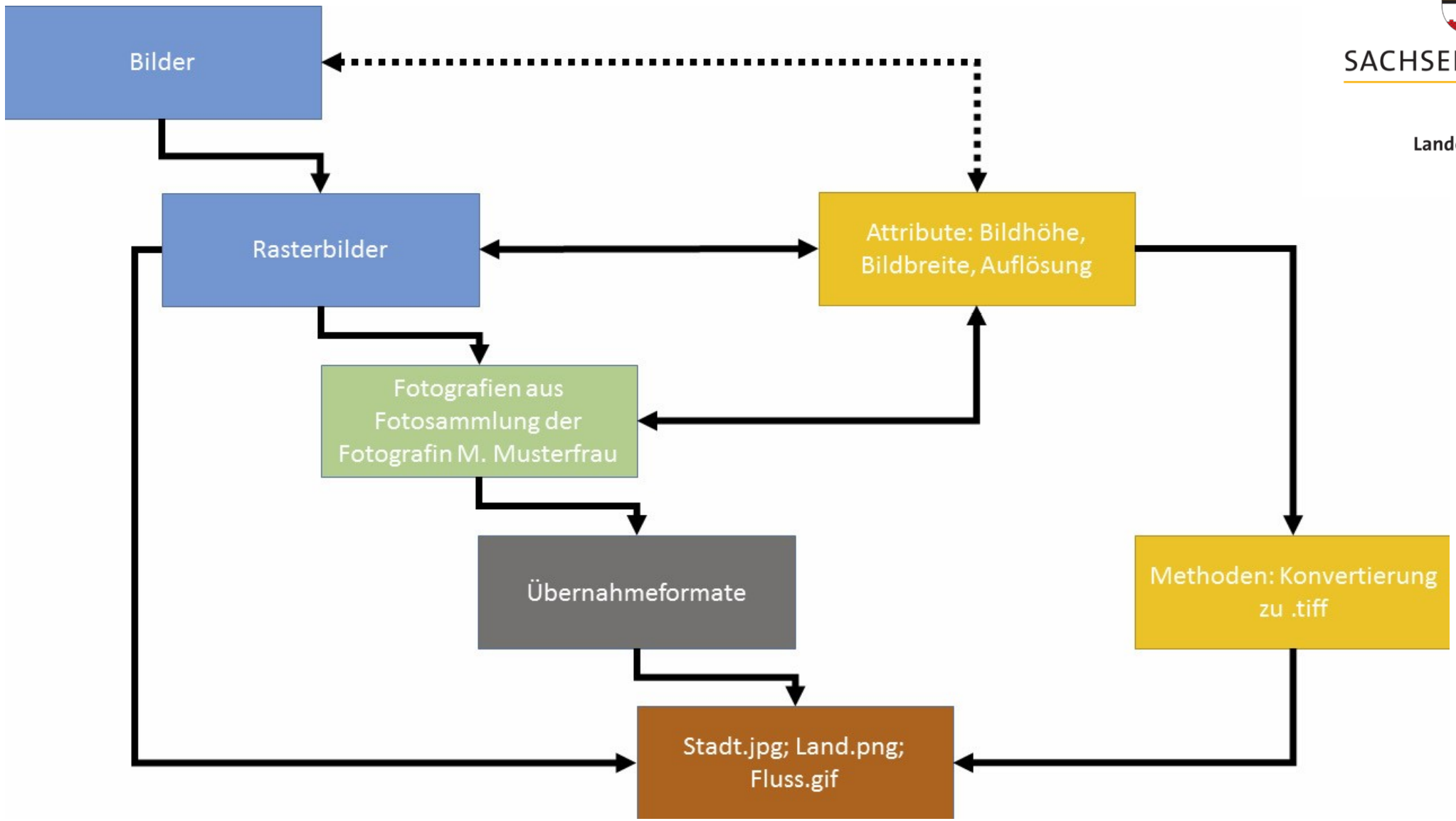
Landesarchiv

PROMETHEUS – **Pr**eservation of digital **o**bjects and **me**tadata: encoding and decoding **th**e digital **e**nvironment under the **u**se of ontologie**s**

- Kernbereich Ontologie
- darauf aufbauend hybrid-reasoning-basierte Entscheidungslogik für risikoorientierte Priorisierung
- Klassifikation und Priorisierung von Erhaltungsmaßnahmen anhand von technischen und archivischen Merkmalen

->Einsatz von KI für die Risikoplanung und Preservation Watch

->Bildung eines einheitlichen Modells zur Bestandserhaltungsplanung



Quelle: Steffenhagen, Björn: Die Modellierung des Kontexts: ein objektorientierter Ansatz zur automatisierten Verarbeitung digitaler Archivalien mittels signifikanter Eigenschaften, in: ABI-Technik 43(2023)1, S. 46-54, hier S. 51.

PROMETHEUS – Stand und Ausblick

SACHSEN-ANHALT

Landesarchiv

- Modellierung der Ontologie
- Auswahl von KI-Modellen zur Anbindung an Ontologie
- Fachkonzept

- dauerhafte Verknüpfung mit externen Informationsquellen oder eigenes Vorhalten von Informationen?
- Erweiterung zur interaktiven Ontologieabfrage über einen Chatbot?
- GUI?

Schlussfolgerungen

SACHSEN-ANHALT

Landesarchiv

1. Digitale Bestandserhaltung ist ein Big-Data-Problem
2. KI im Archiv benötigt eine qualitativ hochwertige Wissensbasis
3. Ontologien und Wissensgraphen eignen sich dafür
4. Die Verknüpfung von KI und Ontologien fördert die Stärken beider Ansätze

->Kein entweder-oder der Technologien, sondern ein sowohl-als-auch

5. Kooperationen zwischen Archiven werden noch relevanter

Kontakt

SACHSEN-ANHALT

Landesarchiv

Björn Steffenhagen

Abteilung 1 - Zentrale Dienste

Landesarchiv Sachsen-Anhalt

bjoern.steffenhagen@sachsen-anhalt.de

https://orcid.org/0000-0002-5849-7123