Bausteine Forschungsdatenmanagement
Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanager*innen

Formaterkennung: eine Voraussetzung und eine Notwendigkeit für das Forschungsdatenmanagement

Kai Naumann[i]

2025

Zitiervorschlag

Naumann, Kai. 2025. Formaterkennung: eine Voraussetzung und eine Notwendigkeit für das Forschungsdatenmanagement. *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanager*innen* Nr. 3/2025: S. 1-11. DOI: 10.17192/bfdm.2025.3.8859.

[i]Landesarchiv Baden-Württemberg. ORCID: 0000-0002-2799-1030

Abstract

Jemand schickt mir ein Filmchen. Ich spiele es ab und freue mich. Aber was hat das mit unserer Forschungsdateninfrastruktur zu tun? Nun, damit aus der digitalen Bitfolge, in der Datei x34f9g.mpeg abgegrenzt, auf meinem Bildschirm ein Film mit Ton wird, braucht mein Endgerät Aufschluss darüber, wie es diese Datei darstellen soll. Die Basis hierfür ist das Dateiformat. Unsere alltägliche digitale Umwelt geht wie selbstverständlich mit Formaten um, weil die Hersteller entsprechende Vorkehrungen treffen. Doch was spielt sich genau in den Betriebssystemen ab und wie können Formate langfristig, also über die Lebensdauer von Hard- und Software hinweg, genau wiedererkannt und mit der passenden Wiedergabesoftware zusammengebracht werden? Dieser Artikel beleuchtet insbesondere, auf welchen Prozessen eine Formaterkennung beruht, was bei der langfristigen Aufbewahrung von Daten gegenüber dem Standardfall anders ist, ob die bisherigen Register für Formate unseren FAIR-Prinzipien entsprechen und schließlich, wie wir die erforderlichen Verbesserungen erreichen und damit zu einer souveränen Forschungsdateninfrastruktur beitragen.

1 Methoden der Formaterkennung

Um eine Datei einem Format zuzuweisen, werden Merkmale der Datei gegen eine Referenz abgeglichen. Hierfür kommen die Dateiendung oder bestimmte Bitfolgen am Beginn und am Ende der Datei in Frage, ähnlich wie Gensequenzen in der biologisch-medizinischen Forensik. Android-, Apple- und Windows-Systeme verlassen sich in der Regel auf Dateiendungen, während Linux-Desktopsysteme das Dateiformat anhand der Bytesequenzen am Beginn und am Ende der Datei erkennen. Die Bitfolgen der entsprechenden Referenz haben den Spitznamen „Magic Numbers” – weiter unten dazu mehr.

Eine weitere, besonders präzise Methode der Erkennung ist die Validierung, bei der einige oder sämtliche Merkmale eines Formats gegen eine definierte Spezifikation abgeglichen werden. Validierung ist aber aufwändiger als andere Methoden und sollte daher nur Anwendung finden, wenn andere Erkennungsmethoden nicht ausreichen. Nicht ohne Grund ist Validierung von Dateien sonst ein separater Schritt. Sie setzt eine verfügbare Spezifikation und eine saubere Analyse der Einzelteile einer Datei voraus, die auch als Parsing bezeichnet wird.

Eine recht neue Möglichkeit besteht darin, in die Referenz nicht etwa jeweils pro Format definierte Werte oder Prüfschritte aufzunehmen, sondern massenhaft Beispieldateien als Trainingskorpus für eine KI zu verwenden. So macht es das Tool Magika (o. D.), das also eher Ähnlichkeitswerte berechnet als dass es präzise Zuordnungen vornimmt. Im Prinzip kann eine Datei mehreren Merkmalen entsprechen und je nach Referenz auch unterschiedlich erkannt werden. So ist jede Datei vom Format DOCX zugleich eine Datei vom Format ZIP, wenn man einmal von der Dateiendung absieht.

Diese Problematik ist, nebenbei bemerkt, auch für die IT-Sicherheit relevant: Cyberkriminelle erschaffen gern Dateien, die sowohl als Format X, als auch als Format Y gelten. So lassen sich Schwachstellen eines Computersystems für dessen Eroberung ausnutzen.[1] Dies ist der Grund, weshalb Magika (o. D.) geschaffen wurde: es spürt potenzielle Hochstapler-Files auf, auch wenn der Wirkmechanismus noch geheim ist und damit noch kein Virenscanner anschlägt.

2 Aufgaben für die Formaterkennung bei der langfristigen Nutzung von Forschungsdaten

Formaterkennung findet, wie oben dargestellt, in unserem Alltag ständig statt und nicht nur die Wissenschaft braucht sie dringend, sondern alle Menschen. Wo genau ist Formaterkennung im Lebenszyklus von Forschungsdaten unerlässlich, wo ist sie zumindest relevant, und was ist anders als bei der alltäglichen Formaterkennung? Die Antwort ist: Sie wird von der Übernahme in ein Forschungsdatenzentrum über Erhaltungsschritte in dessen Archiv bis hin zur Nutzung gebraucht. Und anders als im Alltag greift sie weit in die Vergangenheit zurück, was den Support durch die Hersteller der Betriebssysteme verringert. Wie sagt es der Rat für Informationsinfrastrukturen in seinem jüngsten Positionspapier: „Auch digitale Datenbestände können unwiederbringlich sein" (Rat für Informationsinfrastrukturen (RfII), 2025). Um einem Verlust der Lesbarkeit vorzubeugen, ist schon im Vorfeldbereich Einiges zu tun:

- Ehe die Forschungsdaten in das Repositorium übernommen werden, stellt sich die Frage, ob die Daten eine Bedeutung und einen Wert haben, der den Aufwand für längere Aufbewahrung rechtfertigt. Anhaltspunkte dafür geben präzise Formaterkennungsergebnisse. Unter Umständen hat die abgebende Instanz falsche Vorstellungen über die Qualitätsstandards der Daten, die es gemeinsam mit dem Forschungsdatenzentrum richtigstellen kann.
- Wenn Formate im Vorfeld bereits als ungewöhnlich und schwer zugänglich gelten, sollte vor der längerfristigen Verwahrung eine Umwandlung in ein besseres Format erfolgen, um die langfristige Zugänglichkeit zu erhalten. Das Wandlungsergebnis ist zu überprüfen.

Während der Aufbewahrung auf beispielsweise 15, 40 oder 120 Jahre fallen weitere Aufgaben an:

- Nutzende wollen in 15, 40, 120 Jahren Informationen darüber, welches (womöglich schon etwas betagte Format) sie erhalten werden. Was es „kann" und was

[1] Der Freiberufler Ange Albertini macht regelmäßig auf die Möglichkeiten und Gefahren solcher sogenannter „Polyglots" aufmerksam. Vgl. den Vortrag „Fearsome File Formats", Chaos Communication Congress, 28.12.2024, https://fahrplan.events.ccc.de/congress/2024/fahrplan/talk/QS9AXX/. Homepages: https://github.com/corkami und https://github.com/angea.

nicht. Welche Software es öffnet. Diese Information muss an die Kataloge des Forschungsdatenzentrums weitergereicht werden.
- Auch während der Aufbewahrung könnte der Tag kommen, an dem Formate systematisch umgewandelt werden müssen, weil die Benutzer des FDZ eine neuere Form der gleichen Daten benötigen. Wieder empfehlen sich Kontrollmechanismen für das Ergebnis.
- Auch Details des Formats (z. B. Abspieldauer, Seitenzahl) werden für das Katalogsystem eventuell auf Zuruf der Nutzenden flächendeckend gebraucht und müssen extrahiert werden. Da Formaterkennung technisch gesehen eng mit Parsing verwandt ist, können entsprechende Instanzen auch solche Informationen erheben.

Selbst bei der Nutzung fallen darüber hinaus Erkennungsaufgaben an. Benutzerin X wünscht etwa im Jahr 2039 die Bereitstellung des Films x34f9g.mpeg in einem gängigen Streaming-Format. Wieder ein Prüfvorgang für die Formaterkennung, der in 14 Jahren etwa so vor sich gehen könnte: „Was ist denn MPEG?" – „Keine Ahnung. Aber die Referenz sagt, das ist ein Tonfilm und Tool Z kann es ohne Probleme rendern." – „Tool Z, nie gehört. Aber OK, bitte auf meine Smartbrille streamen."

3 Gängige Vokabulare für Formate . . .

Vokabulare oder Register für Formate gibt es schon eine ganze Weile. Schon in den 1970er Jahren wurde für UNIX-artige Betriebssysteme das Programm „File" erstellt. Das Programm basiert seit 2003 auf einer ausgelagerten Bibliothek namens `libmagic`[2], die nicht nur in unixoiden Betriebssystemen interaktiv über das „file"-Kommando benutzt werden kann, sondern auch für eigene Programme in verschiedenen Programmiersprachen zur Verfügung steht (beispielsweise für C, Go und Python). Ob `libmagic` als Vokabular anzusehen ist, sei dahingestellt. Wieviele Magic Numbers und zugehörige Formate verwaltet werden, ist nicht sofort zu sehen, denn diese sind Stand von Mai 2025 im Code-Repository auf über 347 Codedateien verteilt, die die Magic Numbers thematisch gruppieren.[3] In einer schönen Illustration von Ange Albertini ist die Magic Number einer JPEG-Datei in einer anatomischen Darstellung unter den Bezeichnungen „identifier" und „version" zu sehen (Abbildung 1). Zu vielen, aber nicht allen Einträgen sind die MIME-Types (s. u.) angegeben, was eine gewisse Interoperabilität ermöglicht. Andere Betriebssysteme verlassen sich, wie bereits geschildert, auf Dateiendungen. Das Windows-System, auf dem der Verfasser diese Zeilen verfasst hat, listet 1.153 verschiedene Dateiendungen auf, wie eine Abfrage der Windows Registry ergab (Girnus, 2024). Zu vielen Endungen gibt das Windows-Register auch einen MIME-Type an.

[2] https://github.com/file/file/, https://web.archive.org/web/20161228123411/https://mx.gw.com/pipermail/file/2003/000043.html.
[3] https://github.com/file/file/blob/master/magic/Magdir/.

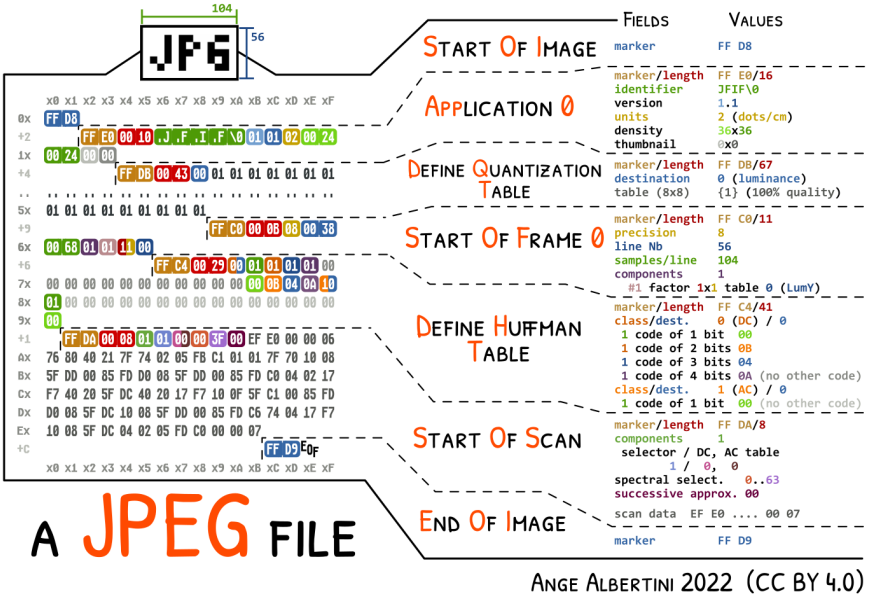

Abbildung 1: Anatomie eines Formats, hier JPEG, auf https://github.com/corkami/pics/blob/master/binary/, gestaltet von Ange Albertini. Unter den Bezeichnungen identifier und version ist hier die Magic Number aufgeführt.

Mit dem Aufstieg des Internets setzte sich 1996 die Internet Assigned Numbers Authority (IANA) das Ziel, alle durch das Netz verbreiteten Formate in einem einheitlichen Register (IANA Media Types, 2025) zusammenzufassen. Es entstand eine Liste der Media Types (ungebrochen beliebt ist die Altbezeichnung MIME-Types[4]), die heute fast 2300 Einträge umfasst. Die Liste ist wenig anschlussfähig, denn sie verweist nur auf Autor:innen oder zugehörige Spezifikationen, nicht aber auf Dateiendungen oder Magic Numbers. Durch den Registrierungsprozess sind die Werte für MIME-Type immerhin persistent und damit über die Zeit hinweg vergleichbar.

Das in der LZA-Community bekannteste Register PRONOM (o. D.) wurde 2002 zu dem Zweck geschaffen, eine umfassende Referenz für die Formaterkennung und eine Verbindung zwischen all diesen Verzeichnissen durch persistente Identifier für Formate zu ermöglichen. Jede einzelne Version eines Formats trägt eine Pronom Unique ID (PUID), eine Dateiendung und soweit vorhanden den MIME-Type und eine Bytesequenz (die aber nicht komplett mit der Magic Number in `libmagic` übereinstimmt). Eine stattliche Zahl von über 2400 Einträgen wurde über die Jahre von Freiwilligen weltweit in dieses Register eingebracht. In einer idealen Welt hätte PRONOM umgehend die anderen Register abgelöst. Es war präziser, es war quelloffen und transparent. Es war aber mit den gängigen Methoden nicht vollständig kompatibel und für manche Verwen-

[4] MIME stand für Multipurpose Internet Mail Extensions, weil zunächst nur an E-Mail-Anhänge gedacht war. Vermutlich auch aus diesem Grund wurde das Akronym offiziell zurückgezogen. Wegen des Sprachwitzes (ein mime ist auf Englisch ein Bühnendarsteller) bleibt es aber sehr beliebt.

dungen zu präzise. So kennt es (Stand 2025) 42 verschiedene Varianten von Dateien mit der Endung PDF.[5] Das ist für einige Anwender:innen sehr wichtig, für andere viel zu kleinteilig. PRONOM bleibt deshalb immer noch einem Nischenpublikum vorbehalten, während `libmagic`, die Windows-Registry-Einträgen und die Liste der MIME-Types weitergeführt wurden. In den Jahren 2005 bis 2012 gab es Bestrebungen, PRONOM in ein amerikanisch-britisches Gemeinschaftsprojekt einzubinden, was vielleicht eine größere Verbreitung ermöglicht hätte, aber nicht gelang.

Zu Beginn der 2010er Jahre betrat Apache Tika[6] die Szene. Als Programmpaket zur Identifikation und Analyse von Merkmalen in Dateien wurde Tika für die Welt der Webdienste hochattraktiv. Jeder auf Apache beruhende Webserver, auf dem etwas hochgeladen wird, macht sich heute Tika zunutze. Tika kann derzeit über 1000 Dateitypen anhand von Magic Number oder Dateiendung unterscheiden.[7] Die Erkennungsmethode ist in einer XML-Datenstruktur übersichtlich dargestellt. Mit dem Jahr 2012 kam eine weitere Innovation in die Welt, die zunächst wenig mit Formaterkennung zu tun hatte. Wikidata als neuer Zweig des Wikipedia-Projekts setzte sich zum Ziel, einen maschinenlesbar strukturierten Wissensraum zu schaffen. Entitäten sind in Wikidata über Eigenschaften mit anderen Entitäten verknüpft. Am 1. Dezember 2012 wurde die Entität Dateiformat geschaffen, und nach und nach gesellten sich viele Instanzen der Entität Dateiformat hinzu. Heute sind es 7000 Einträge – das größte Formatregister der Welt.

Es gibt noch weitere Formatregister, die hier wegen ihrer geringeren Verbreitung nicht erwähnt sind. Wer alle Formatregister „in Aktion" sehen will, dem sei die durch Andrew Jackson von der Digital Preservation Coalition aufgesetzte DigiPres Workbench (o. D.) ans Herz gelegt. Auf einer interaktiven Webseite können Sie schauen, welches Register wie viele und welche Formate enthält. Als Bindeglied dienen hier aber ausschließlich Dateiendungen.

4 . . . und ihre Übereinstimmung mit den FAIR-Prinzipien?

Wie oben schon erwähnt: Forschungsdaten können obsolet werden. Wenn sie ein bestimmtes Alter überschritten haben und die ursprünglichen Darstellungsumgebungen verfallen, sind sie nur dann noch verwendbar, wenn ihr Format präzise bestimmt werden kann. Es liegt also nahe, auch die Formaterkennung im Lichte der FAIR-Prinzipien (2016) zu betrachten. In Tabelle 1 ist das aktuelle Verhalten der hier geschilderten Register dargestellt. Hinsichtlich der Nachnutzbarkeit liegen dem Verfasser nicht immer Angaben darüber vor, inwieweit die Ergebnisse reproduzierbar sind, d. h. ob zum Beispiel der Stand der Register im Jahr 2009 wiederhergestellt werden kann. Bei `libmagic`

[5] Der aktuelle Stand kann hier abgefragt werden.
[6] https://tika.apache.org/.
[7] Die Angaben schwanken zwischen 1400 auf https://en.wikipedia.org/wiki/Apache_Tika und 1174 auf https://www.digipres.org/workbench/registries/compare (abgerufen 27.5.2025).

ist es mit der Interoperabilität nicht weit her, denn die Entscheidungsbasis, auf der die Formaterkennung stattfindet, ist in separaten Dateien gemäß einem eigenen, textbasierten Format eingebettet, was auch die Findbarkeit beeinträchtigt. Die Einträge lassen sich nur mit Programmierkenntnissen als umfassende Liste darstellen. Oftmals finden sich keine persistenten, opaken Identifier, sondern sprechende Bezeichnungen, wie bei den MIME-Types. Diese Unzulänglichkeiten betreffen vor allem die klassischen Formatregister.

Tabelle 1: FAIR-Kompatibilität ausgewählter Formatregister in der Reihenfolge der Darstellung in diesem Aufsatz

	Findable	**Accessible**	**Interoperable**	**Reusable**
Windows	ja, aber keine PI	als Liste	bedingt, via MIME	ja
`libmagic`	bedingt, keine PI	als proprietär strukturierter Text	bedingt, via MIME	ja
IANA MIME Type Registry	ja, aber keine PI	als XML, HTML, TXT	bedingt, via MIME	ja
PRONOM	ja, PUID	als XML, HTML, Abfrage-GUI	bedingt, via MIME, PUID	ja
Tika	ja, aber keine PI	als XML	bedingt, via MIME	ja
Wikidata	ja, WD item ID	als Linked Data, via SparQL	via MIME, PUID, Wikidata item ID	ja

Bei den Registern PRONOM und Wikidata finden wir immerhin persistente Identifier. Entsprechend könnte in den späten 2020er Jahren ein erneuter Anlauf gelingen, die Verbindungen zwischen den Wissensbeständen auf eine zeitgemäße Grundlage zu stellen. Dank der Linked-Data-Technologie gibt es inzwischen Wege, unterschiedliche Datenbestände durch Verknüpfungen mit URLs aufeinander zu beziehen. Das Portal Wikidata (o. D.) kennt Formatfamilien, die Formatversionen umschließen, welche jeweils einer Wikidata item ID und (sofern vorhanden) einer PUID, einem MIME-Type, einer Dateiendung zugeordnet sind und nicht nur auf die hier dargestellten, sondern auch weitere Register verweisen können.

5 Anwendungsgebiete und Herausforderungen bei langer Aufbewahrung

Wie dargestellt, können aktuelle Systeme dank einer emsigen Community meist alle gängigen Formate erkennen und öffnen. Doch wie steht es mit obsoleten Formaten? Anwendende von Archivsystemen stehen vor folgenden Problemen:

- Es gibt eine Vielzahl enthusiastischer Menschen, die aufgrund der vorhandenen Vokabulare Erkennungstools erstellen (COPTR, o. D.). Doch sobald deren Arbeitsleistung versiegt, beginnt der Code zu altern und IT-Sicherheitsprobleme kommen auf, so dass diese Tools aus dem Prozess wieder herausgenommen werden müssen.
- Die verschiedenen Werkzeuge geben unterschiedliche Ergebnisse aus und widersprechen einander im Detail. Ein Werkzeug kann sich auch selbst widersprechen, weil seine Referenz sich verändert und es folglich die Datei X im Jahr 2011 als Format P erkannt hat, im Jahr 2022 hingegen als Format Q. Meist ist dann Q als präzisere Variante von P identifiziert worden oder P war ein Fehlgriff. Wie oben gesagt, können Dateien mehreren Referenzvorgaben entsprechen und folglich ist es anspruchsvoll, die vom Ersteller gewünschte Option zu treffen.
- Die Detailtiefe der Register ist unterschiedlich. Was als MIME-Type application/pdf ist, kann bei PRONOM eine von 42 Varianten sein.
- Manche Tools entscheiden je nach Formatfamilie, welches Vokabular das andere verdrängt und welche Ergebnisse wie zu Klassen zusammengefasst werden. Als Beispiel lassen sich PDF/A-2a, PDF/A-2b und PDF/A-2u zur Klasse PDF/A-2 zusammenfassen, aber auch zur übergeordneten Klasse PDF/A. Diese Entscheidungen sind nicht immer transparent dokumentiert.
- Wenn Widersprüche bestehen, sei es zwischen Tool A im Jahr 2011 und Tool A im Jahr 2022 oder sei es zwischen Tool A und Tool B am gleichen Tag, dann fehlt eine Systematik für den Abgleich und für die verzwickte Entscheidung, welches Ergebnis das bessere ist. Die Anwendung von KI, um eine Ergebnisgewichtung zu erzeugen, wie es Magika tut, bringt zwar der IT-Sicherheit Vorteile, denn polyglotte Formate lassen sich auf diese Weise sehr effizient aufspüren und die riskanten Dateien werden aussortiert. Für langfristige Erhaltungszwecke aber geht es nicht um verkappte Trojaner, sondern um Forschungsdaten, deren Format nach einem reproduzierbaren Regelwerk erkannt werden muss. Ob ein Format nun als Format A oder B gedacht war, muss nachvollziehbar sein.
- Die Entscheidung, welches Format die besten Aussichten auf langfristige Verfügbarkeit hat, muss jede verantwortliche Stelle für sich selbst treffen. Die dafür notwendige Gewichtung sollte aber vergleichbar sein. Insofern bedarf es eines Standards, der über Gattungen und Familien von Formaten hinweg diese Gewichtung transparent ermöglicht.

Die geschilderten Probleme sind im wissenschaftlichen Sinne kein Problem, sondern eine Aufgabe. Und es gibt Hoffnung. Einrichtungen können ihren Kenntnisstand als

Linked Open Data herausgeben, so dass er in Wikidata einfließen kann oder wir ihn uns zu eigen machen können. Die Library of Congress (Library of Congress, o. D.) und die National Archives and Records Administration (National Archives and Records Administration, o. D.) in Washington machen es vor.

6 Ein Ausblick, speziell zum deutschsprachigen Raum

Diese Zeitschrift richtet sich vor allem an die deutschsprachige Community, weshalb die aktuelle Situation kurz dargestellt sei. Im Kompetenznetzwerk nestor war von 2014 bis 2024 eine AG Formaterkennung (o. D.) tätig. Die damals Beteiligten sind zu großen Teilen immer noch mit dem Problem der Formaterkennung betraut. Manche melden regelmäßig neue Formate in PRONOM an, um zur globalen Lösung des Problems beizutragen. Andere tüfteln gemeinsam an Verbesserungen ihrer Archivierungssoftware, um eine zweckmäßige Umsetzung im Einklang mit den globalen Entwicklungen zu erzielen. Wieder andere beteiligen sich an entsprechenden Gesprächsrunden der Digital Preservation Coalition (Digital Preservation Coalition, o. D.). Außerhalb der AG, im Kontext der Entwicklung von Archivsoftware, sind Gedanken darüber entstanden, wie Formaterkennungsergebnisse klassifiziert und mit Erhaltungsmaßnahmen in Verbindung gestellt werden können (Steffenhagen, 2023).

Aus Sicht des Verfassers kommt sowohl nestor als auch der Nationalen Forschungsdateninfrastruktur (NFDI) im Kontext dieser Entwicklungen in Deutschland eine wichtige Rolle zu. Die verschiedenen Konsortien für Forschungsdaten kennen am besten die unterschiedlichen Formate, in denen sich ihre Forschungsdaten präsentieren. Sie kennen am besten die Spezifikationen und den Bedarf an Granularität, den eine Formaterkennung erreichen muss. Sie sind den FAIR-Prinzipien verhaftet. Es ist unabdingbar, dass alle Konsortien, die Daten für länger als 10 Jahre aufbewahren, im Rahmen ihrer künftigen Arbeit gemeinsam die globalen Formatregister, sei es Wikidata, sei es PRONOM, anreichern, um die Wissensallmende der Forschungsdaten nachhaltiger zu machen. Deshalb muss die NFDI als Ganzes im Schulterschluss mit der European Open Science Cloud (EOSC) Mechanismen vorantreiben, die die Formaterkennung weltweit besser und sicherer machen. Der zu schaffende europäischen Knoten des globalen Registers ist einer der Datendienste, die laut den jüngsten Empfehlungen des Wissenschaftsrats (2025, S. 54f) und des RfII (2025, S. 55) für „verlässliche Verfügbarkeit und Integrität" dauernd finanziert werden müssen.

Danksagung

Besonders viel Anregung zu diesem Artikel verdankt der Verfasser der DigiPres Workbench, die Andrew Jackson von der Digital Preservation Coalition für den Workshops „Digital Preservation Registries: What We Have & What We Need" im Rahmen der

Konferenz iPRES2024 in Gent aufgesetzt hat. Aus diesem Workshop ist eine Gruppe hervorgegangen, die sich unter dem Dach der DPC regelmäßig weiter trifft. Ein weiterer wesentlicher Bestandteil dieser Ausarbeitung sind Diskussionen zur Formaterkennung mit Kolleginnen und Kollegen im DIMAG-Verbund. Auch danke ich Claire Röthlisberger-Jourdan und den freundlichen Personen, die meinen Peer Review übernommen haben, für ihre zahlreichen und wertvollen Anregungen.

Literaturverzeichnis

AG Formaterkennung. (o. D.). *Homepage der Nestor-AG*. Zugriff 6. August 2025 unter https://www.langzeitarchivierung.de/Webs/nestor/DE/Arbeitsgruppen/AG_Formaterkennung/ag_formaterkennung_node.html

COPTR. (o. D.). *Unterseite Formatidentifikation*. Zugriff 6. August 2025 unter https://coptr.digipres.org/index.php/File_Format_Identification

DigiPres Workbench. (o. D.). *Comparing Format Registries: Ways of exploring what formats are where*. Zugriff 6. August 2025 unter https://www.digipres.org/workbench/registries/compare

Digital Preservation Coalition. (o. D.). *Preservation Registries Special Interest Group*. Zugriff 6. August 2025 unter https://www.dpconline.org/events/eventdetail/446/-/preservation-registries-special-interest-group

Girnus, P. (2024). *Where to Find File Extension Associations in the Windows Registry* [Blogpost]. Zugriff 6. August 2025 unter https://www.petergirnus.com/blog/where-to-find-file-extension-associations-in-the-windows-registry

GO FAIR Initiative. (2016). *The FAIR Principles*. Zugriff 6. August 2025 unter https://www.go-fair.org/fair-principles/

Google Magika. (o. D.). *Magika*. Zugriff 6. August 2025 unter https://google.github.io/magika/

IANA Media Types. (2025). *IANA Media Types*. Zugriff 6. August 2025 unter https://www.iana.org/assignments/media-types/media-types.xhtml

Library of Congress. (o. D.). *Digital Formats Descriptions as XML*. Zugriff 6. August 2025 unter https://www.loc.gov/preservation/digital/formats/fdd/fdd_xml_info.shtml

National Archives and Records Administration. (o. D.). *Digital Preservation Framework Linked Open Data*. Zugriff 6. August 2025 unter https://www.archives.gov/preservation/digital-preservation/linked-data

PRONOM. (o. D.). *The Technical Registry*. Zugriff 6. August 2025 unter https://www.nationalarchives.gov.uk/PRONOM/

Rat für Informationsinfrastrukturen (RfII). (2025). *Leistung in Verantwortung: Zur Zukunft der wissenschaftlichen Informationsinfrastrukturen in Deutschland* (61 S.). Göttingen. Zugriff 6. August 2025 unter https://rfii.de/?p=12040

Steffenhagen, B. (2023). Die Modellierung des Kontexts. Ein objektorientierter Ansatz zur automatisierten Verarbeitung digitaler Archivalien mittels signifikanter Eigenschaften. *ABI Technik*, *43*(1), 46–54. https://doi.org/doi:10.1515/abitech-2023-0006

Wikidata. (o. D.). *Formate des WikiProject Informatics*. Zugriff 6. August 2025 unter https://www.wikidata.org/wiki/Wikidata:WikiProject_Informatics/Structures/File_formats

Wissenschaftsrat. (2025). *Strukturevaluation der Nationalen Forschungsdateninfrastruktur (NFDI)* [Drs. 2627-25]. https://doi.org/10.57674/wcdc-6d36