Die Aussicht auf maschinelle Inhaltserschließung von Dateien allgemeinen Formats

Frank Tristram

Abstract

Die Datenberge in Wissenschaft, Industrie und Gesellschaft werden immer unübersichtlicher. Doch schon simple Formatgrenzen behindern Konsistenzprüfungen von Einzelergebnissen und das Erkennen größerer Zusammenhänge. Es besteht Einigkeit darin, dass wir verlässliche „Wissensautomaten" benötigen, damit Menschen öffentliche Informationen umfassend nachnutzen können. Die maschinelle Verarbeitungsqualität von Bildern und Sprache soll bald auch für Messdaten aus dem Labor erreicht werden. Ein Prototyp am Exzellenzcluster 3DMM2O zeigt, wie die maschinelle Inhaltserschließung von solchen Dateien gelingen kann.

The amount of data in science, industry and society is becoming more and more complex. But even simple format boundaries hinder consistency checks of individual results and the recognition of larger relations. There is a consensus that we need reliable „knowledge automatons" so that people can comprehensively reuse public information. The machine processing quality of images and speech should soon also be achieved for measurement data from the laboratory. A prototype at the Cluster of Excellence 3DMM2O demonstrates, how automated content mining of such files can succeed.

1 Der Stand der Technik – und Lücken darin

Wenn man das KI-Zeitalter daran festmacht, dass die breite Öffentlichkeit datengetriebene KI bewusst als Alltagserfahrung wahrnimmt und als solche akzeptiert, dann beginnt dieses spätestens jetzt, – denn in den letzten Jahren hat KI auch die breite Öffentlichkeit erreicht. Neben der Berichterstattung über herausragend gelöste Herausforderungen (z.B. Schach (1996), Jeopardy (2011), Go (2016)) findet die Interaktion mit der breiten Öffentlichkeit hauptsächlich über Sprache (z.B. Chatbots, Übersetzungsprogramme, Textanalyse) oder Bilder (Bildersuche, Captchas, generative AI) statt. Aktuell kann KI Bilder und Sprache gezielt verändern, erweitern oder ineinander übersetzen.

So beeindruckend diese Erfolge sind, so trivial ist das alles informationstheoretisch. Die Trivialität liegt darin, dass es für diese Problemkategorien schier endlos „vollständige" Information gibt, aus der man Ausschnitte als Trainingsmaterial erzeugen kann. Die KI hat dann die Aufgabe, von diesen Ausschnitten wieder auf das Vollständige zurückzukommen. So kann man beispielsweise einer KI beibringen, aus einzelnen Fotos kurze Videosequenzen zu generieren, indem man ein neuronales Netz auf Basis der Einzelbilder von vorhandenen Videos trainiert.

In der Wissenschaft liegt allerdings weder eine vergleichbar vollständige Informationsgrundlage (z.B. in Form perfekt annotierter Daten) vor, noch kann man KI sorglos Inhalte generieren lassen, wenn man wissenschaftliche Analysen darauf aufbauen will. Messdaten, aber auch Programminput wie -output, sind zudem mehrheitlich nicht Sprache oder Bilder, sondern Spezialformate mit wenig Verbreitung. Die dadurch entstehenden Problemklassen für datengetriebene Ansätze sollen ein paar **Leitbeispiele** verdeutlichen, die an den meisten naturwissenschaftlich arbeitenden Forschungsinstituten vorkommen:

Formatvielfalt: Hunderte Messgeräte bringen nicht selten eigene, einzigartige Formate mit, in denen sie Ergebnisse ausgeben.

Metadaten: Konfigurationen oder Bedingungen eines Experiments sind schlecht nachvollziehbar mit den dazugehörigen Daten verknüpft und sind auch nur unvollständig in den gegebenen Formaten der Daten beschreibbar.

Software: Entscheidende Programme, die alle die gleiche Aufgabe lösen (z.B. Finite-Elemente-Solver), sind weder miteinander kompatibel, noch so einfach zu benutzen, dass sie im wissenschaftlichen Alltag einfach getauscht werden könnten.

Ablage: Daten liegen zwar nach DFG-Kodex gesichert für zehn Jahre, z.B. nach Personen sortiert, auf diversen Platten, aber niemand kann oder will das inhaltlich auswerten, da jeder Ordner eine individuelle Struktur und andere Eigenheiten aufweist.

In all diesen Fällen ist es nicht trivial, datengetrieben Erkenntnisse aus den abgelegten Dateien herauszuziehen. Bereits der Vergleich zwischen zwei ähnlichen Methoden kann scheitern oder ist zumindest unverhältnismäßig zeitaufwändig.

Das Land Baden-Württemberg hatte in seinem Fachkonzept *„E-Science – Wissenschaft unter neuen Rahmenbedingungen"* [1] schon 2014 die Vision für die Wissenschaft skizziert:

„Daten werden nahtlos von Laborgeräten in Datenbanken übernommen und stehen über Kollaborationsplattformen,

1 https://idw-online.de/de/attachmentdata37340.pdf

Henrieke Ribbe, Flowers, 2012, Aquarell auf Papier, 28 x 28 cm

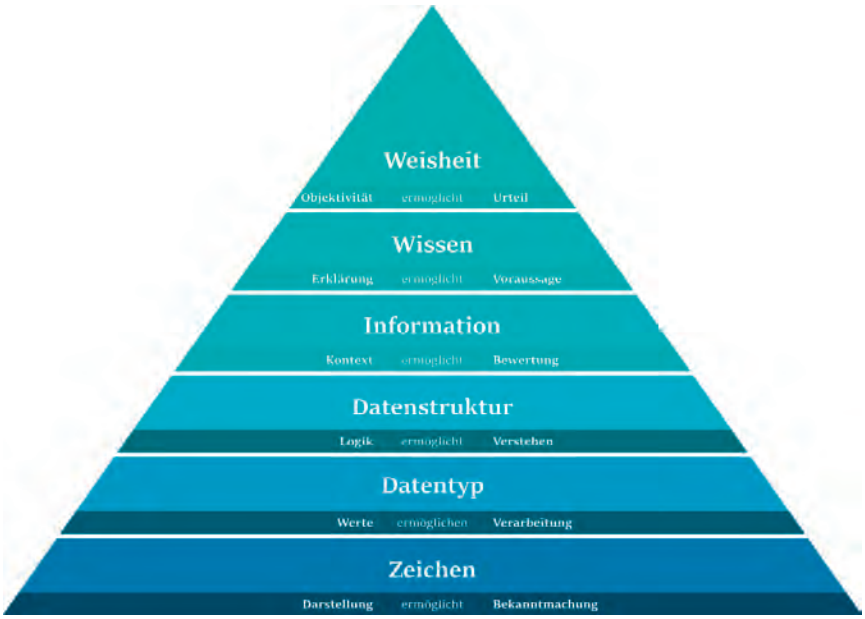

Abbildung 1: Das Konzept der Wissenspyramide fußt auf aufeinander aufbauenden Erkenntnisebenen, die auf jeder Ebene durch Einbeziehung neuer Komponenten neue Möglichkeiten schaffen.

die Schnittstellen zu beliebigen Analysewerkzeugen anbieten, der gesamten – auch über Institutionen und Länder verteilten – Forschungsgruppe zur Verfügung.
Eine solche E-Science-Umgebung verlangt eine neue Infrastruktur. Sie muss aus Sicht der Wissenschaftlerinnen und Wissenschaftler so einfach und zuverlässig verfügbar sein wie die Strom- oder Datensteckdose in der Wand."

Nun sind fast zehn Jahre vergangen und Baden-Württemberg erarbeitet gerade ein neues Fachkonzept. Mit Daten klappt das alles auch so langsam. Aber – wie aus den obigen vier Leitbeispielen ersichtlich – Informationen sind noch immer nicht so leicht verfügbar wie Steckdosen in der Wand. Der Bedarf danach dürfte in den letzten zehn Jahren jedoch noch gestiegen sein. Es fehlt vor allem eine automatisierte Metadatenannotation, um Daten in Datenbanken auch so zu beschreiben, dass sie gefunden und genutzt werden können.

2 Die Hürden auf dem Weg der Daten zur Information

Der „American Standard Code for Information Interchange" (ASCII) beschreibt einen 8-Bit-Kodierungsstandard für einzelne Zeichen, der heute in UTF-8 und anderen Kodierungen verallgemeinert standardisiert ist. Anders als der Name vermuten lässt, ist das Informationsaustauschproblem damit nicht gelöst, denn das Verbinden dieser Zeichen zu inhaltlichen Sinnzusammenhängen, die man wirklich Information nennen kann, benötigt einen ordnenden Geist. Abbildung 1 zeigt die erkenntnistheoretischen Stufen, die dabei von einem niedrigen Ausgangszustand aus zu nehmen sind.

Die Standardisierung auf der untersten Ebene (z.B. anhand von ASCII-Zeichen) ermöglicht es, die Inhalte von Dateien anzuschauen. Der nächsten Stufe sind Standardtypen wie Kommazahl, Ganzzahl, Text etc. zuzuordnen. Auch hier gibt es verschiedene Standards. So könnte beispielsweise ein Datum oder eine Uhrzeit als Typ identifiziert werden, um eine Verarbeitung in Variablen zu erleichtern. Die nächst höhere Einheit sind Datenstrukturen wie verschiedendimensionale Arrays, Graphen oder hierarchische Listen. Es geht also um die richtige Zuordnung bzw. logische Beziehung der Datentypen zu einem großen Ganzen. Diese drei fundamentalen Ebenen sind in praktisch jeder wissenschaftlichen Datei vorhanden, werden aber selten explizit dargestellt (außer beispielsweise im HDF5-Format) und werden in kondensierten Wissenspyramiden meist als „Datenebene" zusammengefasst. Auf der Zeichenebene ist es aber kaum möglich, umfassende Informationen automatisch aus einer beliebigen Datei zu ziehen, deren Datenstruktur nicht bekannt ist. Oft werden dedizierte Programme benutzt, die ein bestimmtes Format lesen können und diese Anreicherung implizit leisten. Doch dadurch ist Interoperabilität – der Austausch von Daten und Programmen zwischen verschiedenen Workflows – in der Praxis oftmals umständlich.
Zwar wird weltweit verhältnismäßig viel Aufwand betrieben, um mittels KI innerhalb der vierten Ebene Informationen aus vielfältigen Daten zu erzeugen, doch oft müssen die Daten dafür erst mühsam gesammelt und vereinheitlicht werden. Im Verhältnis dazu ist die Vorarbeit an Formaten und Daten wenig im Fokus der Aufmerksamkeit, obwohl sie Voraussetzung dafür ist, um den vorhandenen Berg an *Dark Data* effektiv abbauen zu können. Es ist klar, dass die Lösung dieses Problems nicht durch eine Einzelinitiative erfolgen kann. Die NFDI und einzelne Akteure, insbesondere Bibliotheken, können aber wichtige Bausteine liefern, Daten strukturiert zugänglich zu machen, soweit dies im jeweiligen Handlungsfeld des Akteurs liegt.

3 Die höheren Ebenen der Wissenspyramide

Die Abgrenzung von Daten zu Information ist dort erreicht, wo genug Kontext für eine erste Bewertung gegeben ist. Solch ein Kontext kann in derselben Datei wie die Daten stehen oder davon getrennt vorliegen. Das können beispielsweise Parameter einer Messung sein oder Nebenbedingungen, wie das Institut, an dem die Messung stattfand. Solche Informationen dienen zur Bewertung, zum Beispiel der Vertrauenswürdigkeit der Daten. In einem anderen Zusammenhang können solche Parameter, zum Beispiel im Rahmen einer übergreifenden Metaanalyse aller Experimente, selbst wieder Teil der Daten sein.

Das heißt, dass die Grenze zwischen Daten und beschreibender Information (sogenannten Metadaten) abhängig von Bezugspunkten wie der Forschungsfrage gezogen werden muss. Informationen dienen der Bewertung, können aber noch nichts erklären. Zur Wissensebene gehören Einsichten, die Daten erklären können und damit auch Voraussagen ermöglichen. Wesentliche Urteile auf Basis von Vorhersagen sollte man allerdings ausschließlich treffen, wenn auch die Grenzen der Modelle umfassend beachtet und objektiv betrachtet sind (Weisheit).

4 Die Aufarbeitung der Datenebenen

In der täglichen Datenarbeit in unserem naturwissenschaftlich geprägten Exzellenzcluster **3D Matter Made to Order** steht die Automatisierung innerhalb der unteren drei Datenebenen im Fokus.

Wir haben dafür ein Programmpaket entwickelt, das eine Bitfolge auf die Strukturebene heben soll und damit Anwendungen unterstützt, die Informationen extrahieren. Wir beschreiben hier anhand der vier Leitbeispiele aus Abschnitt 1, wie sich die Fähigkeit, Zeichenfolgen zu strukturieren, auf vielfältige Weise dazu einsetzen lässt, Anwendungen zur Informationsgewinnung zu speisen.

Formatvielfalt: (Hunderte Messgeräte mit eigenen, einzigartigen Formaten)

Lösung: Es können die zu vergleichenden Datenstrukturen aus heterogenen Datenformaten extrahiert werden. Damit ist eine Datenanalyse unabhängig vom Format, beispielsweise mit den Daten einer anderen Forschungsgruppe ohne zusätzlichen Aufwand möglich. Dabei werden Formatunterschiede (auch Versionen) mindestens halbautomatisch identifiziert und sorgen nicht sofort für Inkompatibilitäten.

Metadaten: (Konfigurationen oder Bedingungen eines Experiments sind schlecht nachvollziehbar)

Lösung: Es können beliebige Parameter automatisch gefunden und in dafür vorgesehene Felder eingetragen werden. Dadurch lassen sich Beschreibungen von Experimentordnern standardisieren, Lücken in der Beschreibung identifizieren und in Echtzeit auch Beschreibungsvorschläge ableiten.

Software: (Entscheidende Programme sind weder kompatibel, noch einfach zu benutzen)

Lösung: Es können Programme, die sich in In- und Outputformaten unterscheiden, gleichartig angesprochen werden. Damit lässt sich Workflowcode vom individuellen Code, der nur für ein bestimmtes Programm gebraucht wird, auf hoher Abstraktionsebene trennen. Damit wird auch eine Standardisierung der Aufrufparameter und Ergebnisbeschreibung erleichtert.

Ablage: (individuelle Struktur und andere Eigenheiten)

Lösung: Es können gleichartige Inhalte formatunabhängig identifiziert und gleichartige Informationen daraus extrahiert werden. So können aus einem Ordner mit digitalen Laborbuchaufschrieben und Messdaten eines Experiments alle vorhandenen Informationen herausgezogen werden, ohne zu wissen, wo sie genau stehen.

Im Einzelnen erhoffen wir uns in den nächsten Jahren von dem Zusammenspiel dieser Anwendungen, dass

1. die Benutzung von Fremdformaten keinen Mehraufwand mehr darstellen und barrierefrei interoperabel wird,
2. bewertungsrelevante Parameter durch Datenbankabfragen besser verfügbar und schneller ersichtlich werden,
3. wissenschaftliche Software, auch mit verschiedenen Schnittstellen, modular einsetzbar und dadurch auch besser vergleichbar wird,
4. Metaanalysen von gleichartigen Daten auf Dark Data ohne händisches „Data Wrangling“ möglich werden.

Diese vier idealtypischen Ergebnisse setzen voraus, dass Programme der Analyse (ab der Informationsebene) gewisse generische Strukturen verarbeiten können und nicht auf ganz bestimmte proprietäre Formate, z.B. eines bestimmten Messgerätes beschränkt sind. All diese Analyseanwendungen sollten auf einer von wenigen Schnittstellen arbeiten können (CSV, JSON, YML, XML, SQL, HDF5 etc.), die umfassend über die zugrundeliegenden Daten informiert (siehe Abbildung 2).

Wir haben unseren aktuellen Prototyp einem ersten Wirklichkeitstest unterzogen und 10.000 materialwissenschaftliche Dateien auf Zenodo analysiert. Dabei ging es zunächst darum festzustellen, wie die Zeichenerkennung in der Praxis funktioniert und ob gleiche Datenstrukturen auch als gleich erkannt werden. Das ist Grundvoraussetzung, um eine darauf aufbauende Anwendungslogik zu definieren und anzuwenden. Tatsächliche Anwendungen wie die Metadaten aller Zenodo-Dateien qualitätsgesichert zu verbessern, sind aber momentan außerhalb unseres Projektfokus.

5 Technische Implementierung

Der beschriebene Workflow befindet sich noch in einem frühen Entwicklungsstadium, vergleichbar mit Technology Readiness Level (TLR) 3-4. Als ersten Großversuch haben wir 10.000 strukturierte und unstrukturierte „ASCII-verdächtige“ materialwissenschaftliche Dateien von Zenodo heruntergeladen (‘.csv‘, ‘.json‘, ‘.dat‘ und ‘.txt‘-Dateien). In 98,6% der Fälle konnten wir den Zeichenstandard (Encoding) automatisch identifizieren. Eine manuelle Stichprobe der übrigen Dateien ergab, dass proprietäre Formate vorzuliegen scheinen, sodass 9.860 Dateien weiterverarbeitet werden konnten.

Eine erste Idee war nun, dass gleiche Dateiformate an-

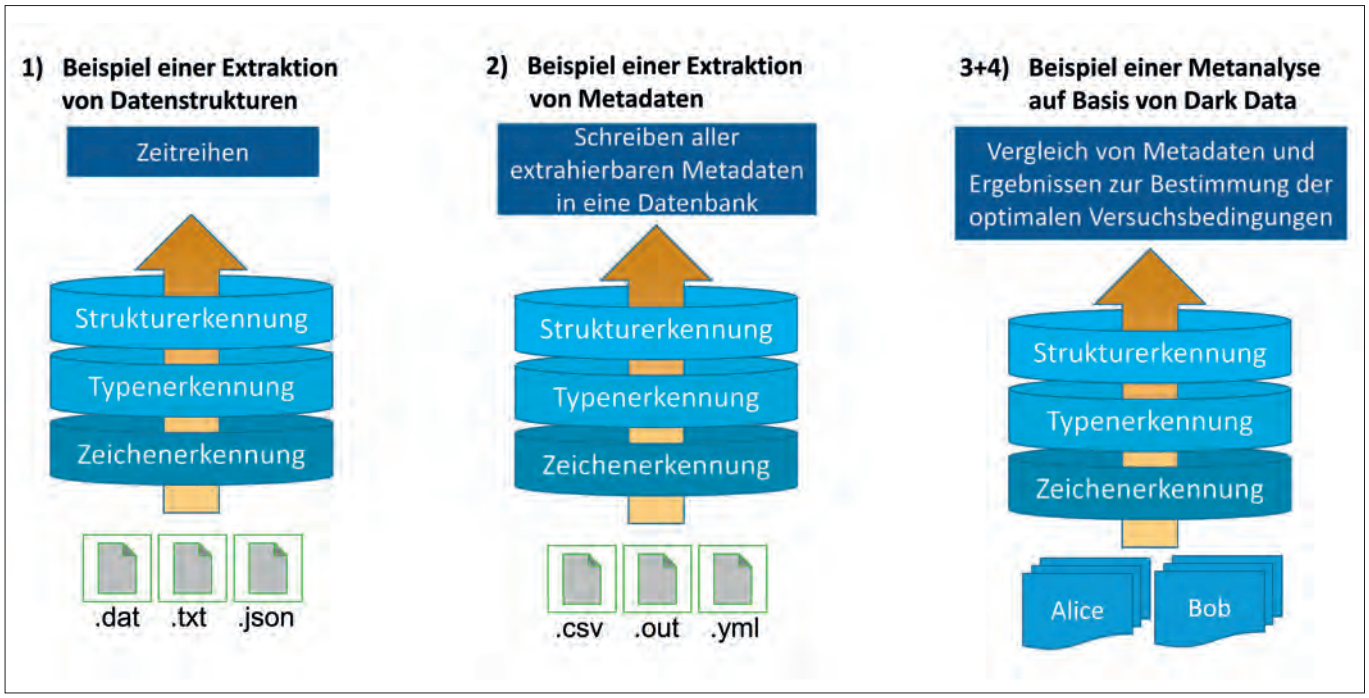

Abbildung 2: Leitbeispiele grafisch veranschaulicht. Die Kernkomponente unter dem Verarbeitungsprozess auf Informationsebene besteht immer aus den drei Schritten, die Rohdatenstrukturen automatisiert zu erkennen.

hand eines für sie typischen „Fingerabdrucks" identifiziert werden könnten. Z.B. ist Software, die eine „.dat"-Datei weiterverarbeiten soll, nicht einfach anhand der Dateiendung zu identifizieren. Eine „.dat"-Datei könnte eine Form von CSV-Datei sein, oder etwas „Selbstgemachtes". Kann man diese interne Datenstruktur unterscheiden, so kann man einmal vorhandene Informationen über bestimmte Dateien (z.B. mit welchem Programm sie zu öffnen sind) im Prinzip auf die ganze Klasse übertragen. Es ist dann auch möglich Konvertierungen und einfache Überprüfungen durchzuführen, wenn für einen Dateisubtyp schon ein Prüfskript vorhanden ist. In der Tat konnten wir die unterschiedlichen technischen Metadaten der vorhandenen Datenstrukturen (z.B. Abfolge, Trennzeichen, Spaltenanzahl, Schachteltiefe, enthaltene Datentypen etc.) nutzen, um Dateien zu kategorisieren. Eine automatische Standardisierung der Granularität der Klassifizierung scheint aber schwierig: Nehmen wir das bekannte Beispiel von CSV-Dateien: Man könnte diese beispielsweise nach Trennzeichen, Spaltenzahl, Überschriften oder enthaltenen Datentypen unterscheiden und sagen, Dateien „gleicher" Klasse müssen bei allem exakt übereinstimmen. Das ist sicherlich gut, um für wissenschaftliche Messdaten festzustellen, welches Programm diese Tabelle geschrieben hat oder analysieren kann, doch Excel kann all diese Dateien öffnen. Es hängt also von der Anwendung ab, auf welche Klasse man eine vorhandene Information übertragen kann. Bei genügend Daten sollte sich das statistisch herauskristallisieren, ansonsten muss das vom Anwender vorgegeben werden. Je nach (willkürlichem) Feinheitsanspruch konnten wir für die 10.000 heruntergeladenen Testdaten von Zenodo „mindestens 60 Formate" oder aber fast Tausend finden.

Hier zeigt sich auch das Problem, dass für die Aufgabe Informationsstrukturen in Dateien zu finden, bis auf triviale Fälle wie CSV-Dateien etc., kein allgemeines Verständnis oder auch nur vollständig annotierte Beispiele existieren. Es ist noch nicht einmal standardisiert, welche Strukturen „entdeckungswürdig" sind, also welche Klassen von Datenstrukturen analog zu „Tabelle" oder „Fließtext" existieren.

Aus diesem Grund haben wir uns für ein kaskadisches Abarbeiten entschieden: Ein mehrstufiger Algorithmus, der logischerweise mit einer Autodekodierung des Encodings beginnen muss, sorgt unter Zuhilfenahme verschiedenster statistischer und fest programmierter Verfahren für ein konsistentes Ergebnis. Zu den statistischen Verfahren gehören Ähnlichkeitsanalysen der Datentypen, Mustererkennung und Clusteralgorithmen. Unser Programm ist grundsätzlich so aufgebaut, dass verschiedene Detektoren jeweils versuchen, spezifische Strukturen zu erkennen. Die Detektoren arbeiten höchst unterschiedlich: Während ein Tabellendetektor vor allem nach gleichmäßigen Spalten, möglichen Trennzeichen und gleichmäßigen Typen schaut, ist der Key-Value-Detektor vor allem an bestimmten Signalen (Klammern, Einrückungen, Zuordnungen) interessiert. Zudem gibt es semistrukturierte Bereiche und Fließtexte, die nicht fälschlicherweise als Tabelle oder dergleichen erkannt werden dürfen. Der Vergleich der

Universität Marburg

TU München

Universitätsbibliothek
Salzburg

Berlin-Brandenburg
International School

EINFACH MACHEN. AUS METALL.

Zambelli Bibliotheken

Lernen und Wohlfühlen

Die Zambelli Bibliothekseinrichtungen begleiten wissenschaftliche und öffentliche Bibliotheken, die sich mit neuen Gegebenheiten auseinandersetzen und sich weiterentwickeln wollen. Wir helfen Ihnen Ihre Bibliothek so auszustatten, dass attraktive und funktionale Lernräume entstehen. Dabei können Sie sich auf in Sicherheit und Funktion bewährte Einrichtungs lösungen verlassen.
Gemeinsam schaffen wir gestalterisch-kreative Raumkonzepte.

Nehmen Sie mit uns Kontakt auf!
regalsysteme@zambelli.com

www.zambelli.com/regalsysteme

Detektorergebnisse miteinander liefert dann einen Strukturierungsvorschlag. Nutzer können bei Uneindeutigkeit eine eigene Entscheidung treffen, welche Struktur sie bevorzugen.
Als instruktives Beispiel für ein mehrdeutiges Ergebnis, sei dieser Zweizeiler aus einer realen Datei gegeben:[2]
total CPU-time : 0.45 seconds
total wall-time : 1.47 seconds

Das Programm erkennt ein doppeltes Key-Value-Paar, aber auch ein gewisses, sich wiederholendes Muster, das kein reiner Text zu sein scheint, als auch die Strukturvariante einer zweizeiligen und fünfspaltigen Tabelle. Defaultmäßig würde das Programm hier die Key-Value-Paare als Struktur ausgeben. Falls man aber einen Text oder eine Tabelle haben will, so könnte es auch diese Struktur mit den entsprechenden Metadaten ausliefern. Voraussetzung ist natürlich, dass das Programm irgendeine Struktur erkennt. Wenn z.B. der Tabellendetektor nichts fände, dann kann das Programm auch keinen sinnvoll als Tabelle strukturierten Vorschlag machen. Entsprechend verhält es sich mit den anderen Detektoren.

6 Ergebnis und Ausblick

Obwohl wir noch ziemlich am Anfang stehen, sind wir bereits in der Lage, jede beliebige ASCII-artige Datei in eine HDF5-Datei zu konvertieren. Das HDF5-Format ist für unsere Zwecke geeignet, weil es die Beschreibung von Datentypen und Datenstrukturen ermöglicht, sodass diese Elemente einheitlich gespeichert werden können. Diese einheitliche Erkennung der Typen und Strukturen gelingt weitestgehend. Jede Weiterverarbeitung und Analyse braucht allerdings ein eigenes Programm, das, auch wenn es HDF5 lesen kann, gewisse Voraussetzungen an die Datenbenennung und -strukturierung setzt, die noch nicht automatisch einheitlich geliefert werden können. Z.B. könnte man eine Anschrift als Tabelle, hierarchisch, als Text oder uneinheitlich strukturieren. Ein „Anschriftenverarbeitungsprogramm“ bräuchte in der Regel jedoch eine ganz bestimmte Form des Inputs, selbst wenn es HDF5 lesen könnte.
Trotzdem kommen wir dem großen Ziel freiströmender Informationsflüsse langsam näher. Wir denken, dass die Entwicklung in den nächsten Jahren in die Richtung gehen wird, die Grenzen zwischen Datenformaten zu überwinden, um einen Innovations- und Produktivitätsschub in Wissenschaft, Industrie und Verwaltung auszulösen, der seit Einführung des Computers einzigartig ist. Das ist wohl auch unabdingbar, um die vielgepriesenen Versprechungen des Informationszeitalters einzulösen. Wir werden unseren Prototyp weiterentwickeln und in einem Onlinedemonstrator zur Verfügung stellen. Auf s.kit.edu/deraw wird es ab Januar 2024 möglich sein, eigene Dateien hochzuladen und Extraktionen oder Konvertierungen durchzuführen.

Frank Tristram hat von 2014 bis 2019 strategische Forschungsdatenprojekte für das Land Baden-Württemberg geleitet. Die Webseite forschungsdaten.info und die Konferenzreihe „E-Science-Tage“ wurden durch diese Projekte initiiert. Seit 2019 schafft er am Exzellenzcluster 3D Matter Made to Order praxisnahe Strukturen zum FAIRen Umgang mit Forschungsdaten. Eine der größten Herausforderungen liegt im Vermeiden von „Dark Data“. Das sind Daten, die kein zweites Mal „angeschaut“ werden und nicht automatisch interpretiert werden können. Besondere Relevanz haben dabei komplex strukturierte Messdaten, die nicht für KIs zugänglich sind.

Diese Arbeit wurde gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder – 2082/1 – 390761711.

Frank Tristram
ist am Exzellenzcluster 3DMM2O der Universität Heidelberg und des KIT hauptverantwortlich für den Umgang mit Forschungsdaten. Das beinhaltet die Definition und Unterstützung einer exzellenten wissenschaftlichen Praxis mit Daten, sowie die Schaffung von Rahmenbedingungen, um dies allen Clusterbeteiligten zu ermöglichen.
frank.tristram@kit.edu

2 Es sei hier schon angenommen, dass das Programm erkannt hat, dass diese zwei Zeilen irgendwie zusammengehören und sich vom Rest der Datei abgrenzen. In der Realität ist dieser Prozess ein iterativer.