Forschungsdaten FAIR verwalten

Die Ergebnisse der ersten Umfrage des Konsortiums NFDI4Phys liefern wertvolle Einsichten in den Umgang mit Forschungsdaten in der Physik.

Holger Israel, Esther Tobschall und Frank Tristram

In die wissenschaftliche Arbeit fließen Daten ein: Messwerte aus dem Labor, Berechnungen von Computern und Forschenden oder Daten anderer Gruppen. Ein Ziel des Datenmanagements muss darin bestehen, wichtige Forschungsdaten dauerhaft zugänglich, nutzbar und nachprüfbar zu machen. Das soll die von Bund und Ländern initiierte Nationale Forschungsdateninfrastruktur (NFDI) vereinfachen.

Diese Forschungsdateninfrastruktur wird in die europäische und weltweite Forschungsdatenlandschaft (z. B. European Open Science Cloud) eingebettet sein. International etabliert ist hierbei die Einhaltung der FAIR-Prinzipien: Daten müssen demnach „findable, accessible, interoperable, reusable" sein, also auffindbar, zugänglich, interoperabel und wiederverwendbar [1]. Beschäftigt man sich dafür mit Dokumentation, Archivierung, Transformation oder der Veröffentlichung von Forschungsdaten, spricht man vom Forschungsdatenmanagement.

In der Informationswissenschaft ist das Konzept der Wissenspyramide gebräuchlich [2, 3]: Ausgehend von bloßen Daten über strukturierte Informationen (z. B. durch geeignete Metadaten annotierte Daten) steigt der Erkenntnisgewinn hin zum Wissen und schließlich zur Weisheit. Daten und Informationen können hierbei Individuen, Gruppen oder einer breiten Öffentlichkeit zugänglich sein [4]. Daraus ergeben sich verschiedene Wege zur Erkenntnis (**Abb. 1**): Forschende können Wissen aus ihren eigenen, individuellen Daten destillieren und veröffentlichen. Andererseits steht die (eigene) Forschung durch Ergebnisse anderer Forschender in einem größeren Kontext und ermöglicht es, auf öffentlicher Ebene Wissen zu gewinnen. Beide Wege haben ihre Berechtigung und sind in der Physik verbreitet.

Das Kernanliegen der Nationalen Forschungsdateninfrastruktur besteht darin, die Wissensgenese aus Communitydaten zu fördern, also die Nachnutzung von Daten. Bei ihrer Ausgestaltung spielen aber auch die Anforderungen der Forschenden an das Management der in der eigenen Gruppe generierten Daten eine wichtige Rolle. Diese Abläufe und Anforderungen unterscheiden sich innerhalb und zwischen den wissenschaftlichen Teildisziplinen deutlich.

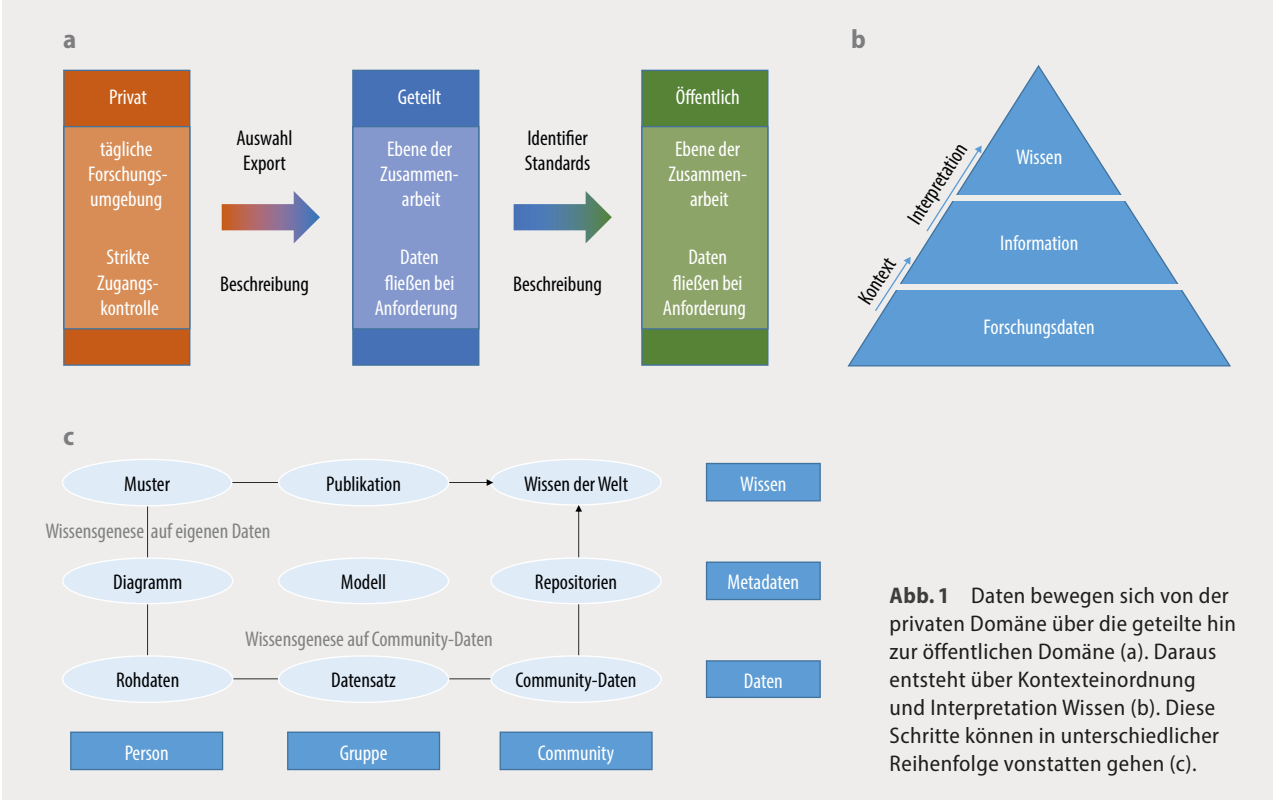

Abb. 1 Daten bewegen sich von der privaten Domäne über die geteilte hin zur öffentlichen Domäne (a). Daraus entsteht über Kontexteinordnung und Interpretation Wissen (b). Diese Schritte können in unterschiedlicher Reihenfolge vonstatten gehen (c).

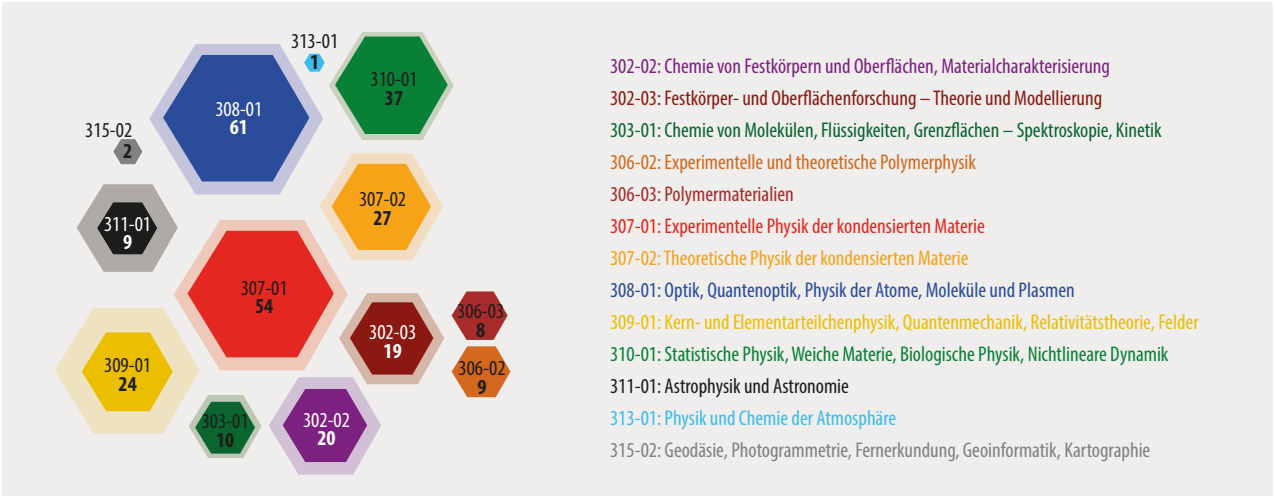

Abb. 2 Viele Teilnehmende an der Umfrage fühlen sich von den bisherigen NFDI-Konsortien nicht vertreten. Die Zahlencodes und Farben zeigen die unterschiedlichen DFG-Fachgruppen. Die Gesamtfläche der Waben ist proportional zur Anzahl der Antworten aus dem jeweiligen Fach. In derselben Skala stellt die farbgesättigte innere Fläche die Anzahl (vgl. Zahlen in Fettdruck) derjenigen Teilnehmenden dar, die ihre Belange in der NFDI im Frühjahr 2020 noch nicht durch eines der Konsortien vertreten sahen.

Daher zielen die Konsortien der NFDI darauf ab, die Forschungsdateninfrastruktur fachspezifisch auszugestalten. Forschende sollen definieren, welche Anforderungen und Wünsche sie daran haben. In drei Förderrunden von 2019 bis 2021 empfehlen Gutachterinnen und Gutachter, die von der Deutschen Forschungsgemeinschaft (DFG) bestellt wurden, die geeignetesten Konsortien zur Förderung im Rahmen der NFDI.[1)]

In mehreren Teilbereichen der Physik gründeten sich früh Konsortien, die ihre Anträge zu den Förderrunden 2019 und 2020 eingereicht haben. In der zweiten Antragsrunde sind dies:

- PUNCH4NFDI für die Teilchen-, Astro-, Astroteilchen-, Hadronen- und Kernphysik,
- FAIRmat in der Physik der kondensierten Materie und der chemischen Physik der Festkörper,
- DaphneNFDI für Daten aus Photonen- und Neutronenexperimenten sowie
- NFDI-MatWerk für die ingenieurwissenschaftlichen Materialwissenschaften.

Doch hat keines der Konsortien den Anspruch, die gesamte Breite der Physik zu vertreten. Daher initiierten die Physikalisch-Technische Bundesanstalt (PTB) und die Technische Informationsbibliothek (TIB) NFDI4Phys als ein Konsortium für die noch nicht vertretenen Bereiche der Physik [5]. Die Leitung liegt seit August 2020 bei der Universität Bremen, vertreten durch den Sprecher Hans-Günther Döbereiner.

Um die gemeinsamen und fachspezifischen Bedarfe an das Forschungsdatenmanagement in der Physik zu ermitteln, gab es eine breit angelegte Online-Umfrage. Diese knüpft an zwei „Datenumfragen der DPG“ aus den Jahren 1973 und 1993 an [6, 7].

1) www.dfg.de/foerderung/programme/nfdi

2) oar.ptb.de/resources/show/10.7795/730.20210511

Ergebnisse der Umfrage

Die vollständige Auswertung mit komplettem Fragenkatalog, Antworten und zugehörigen Datensätzen findet sich im Open Access Repository der PTB.[2)] In diesem Artikel beschränken wir uns auf zentrale Ergebnisse und Leitfragen.

Die Umfrage erhob einerseits den Status quo des Forschungsdatenmanagements in der Physik und stellte andererseits vorrangige Bedarfe von noch nicht vertretenen Gruppen fest. Die Umfrage bestand aus 27 Fragen und erfolgte anonym über die Open-Source-Plattform LimeSurvey. Insgesamt wurden bei 488 Aufrufen Fragen beantwortet. Davon haben 237 Personen alle Fragen beantwortet. Bis auf eine Ausnahme liegen von allen Universitäten, die Promotionsstudiengänge in Physik anbieten, Antworten vor. Aus Hochschulen und außeruniversitären Forschungseinrichtungen waren die Rückmeldungen weit weniger vollständig. Eine knappe Mehrheit von 53 Prozent der Antwortenden leitet selbst eine Forschungsgruppe.

Viele Teilnehmende sehen sich bislang noch nicht von einem der NFDI-Konsortien vertreten (**Abb. 2**). Zum Zeitpunkt der Erhebung im Frühjahr 2020 war 46 Prozent von ihnen die NFDI noch nicht bekannt. Die bestehenden Konsortien dürften in der Zwischenzeit die Reichweite in ihren jeweiligen Fächern vergrößert haben. Unsere Ergebnisse legen dennoch die Schlussfolgerung nahe, dass ein komplementäres Konsortium im Spektrum der physikrelevanten NFDI-Initiativen notwendig ist.

Welche Forschungsdaten fallen an?

Typischerweise kombinieren physikalische Forschungsgruppen mehrere Methoden (**Abb. 3a**): So verbinden 52 Prozent Experimente und (numerische) Simulation. Jeweils gut 60 Prozent der Forschenden nutzen kommerzielle Software, Open-Source-Software, selbstgeschriebene Software und Programmierskripte zur Datenauswertung. Die geläufigsten Datentypen sind Textformate, Bilddaten (Raster oder Vektor) sowie Software-Code und werden von jeweils

mehr als 70 Prozent der Antwortenden erzeugt. Ein gutes Forschungsdatenmanagement muss diesen Bedingungen Rechnung tragen. Nur wenige Forschende nutzen Daten anderer nach. Die Ursachen dafür sind unzureichende Auffindbarkeit und Zugänglichkeit, proprietäre Datenformate sowie lückenhafte Dokumentation.

Wie ist die Datendokumentation und wie sollte sie sein?

In der Mehrzahl der Gruppen erfolgt die Dokumentation teilweise analog durch persönliche Papierlaborbücher (bei 50 %) bzw. durch Laborbücher, die an Geräte gebunden sind (31 %). Bemerkenswert ist, dass 21 Prozent der Teilnehmenden ein eigenes elektronisches Laborbuchsystem entwickelt und somit vermutlich signifikante Ressourcen investiert haben (**Abb. 3b**). Wo dies noch nicht der Fall ist, besteht der Wunsch nach einer möglichst automatisierten, standardisierten und in die Datenerfassung nutzerfreundlich integrierten Datendokumentation.

Die strukturierte Beschreibung von Daten durch Metadaten ist eine notwendige Voraussetzung für die Nachnutzbarkeit im Sinne der FAIR-Prinzipien. Die Umfrageteilnehmenden identifizieren mehrere Kategorien relevanter Metadaten (**Abb. 4a**). Die Annotation sollte idealerweise automatisch geschehen. Metadaten werden von 73 Prozent der Forschenden auch manuell erfasst. Die Mehrzahl der Befragten nutzt hierzu selbstentwickelte Schemata, standardisierte Schemata kommen bei 38 Prozent zum Einsatz (**Abb. 4b**).

Probleme beim Umgang mit Forschungsdaten

Die Schwierigkeiten der Datennachnutzung sind äußerst vielfältig. Die „Vielfalt der Datenformate und Datenstrukturen“ benannten 69 Prozent der Befragten, 61 Prozent gaben die „Vielfalt der Werkzeuge und Datengenerierungsmethoden“ an und 59 Prozent die „Unübersichtlichkeit der Dokumentation“. Weitere Probleme sind die Vielfalt der Metadaten bzw. unausgereifte Standards dafür, hohe Datenmengen oder die fehlende Rückwärtskompatibilität alter Datenformate.

Freitext-Antworten verweisen unter anderem auf fehlende Community-Standards, was bei einem bestimmten Experiment zu annotieren sei, und dass oft a priori unklar ist, welche Daten langfristig wertvoll sein könnten. Um die Qualität und Zuverlässigkeit eigener Forschungsdaten zu belegen, gelten Reproduzierbarkeit, Standards (z. B. der guten wissenschaftlichen Praxis) und statistische Methoden als wichtigste Faktoren. Das Erwähnen fehlender Terminologiestandards zeigt ein deutliches Interesse an einer gemeinsamen Datenkultur der Physik auf. Herausfordernd ist dabei die (dynamische) Heterogenität der Datentypen und Formate, der Methoden und Geräte.

Ein weiteres Manko ist, dass zum Zeitpunkt der Umfrage erst 15 Prozent der Antwortenden ein „strukturiertes“ Forschungsdatenmanagement anwendeten. Für 38 Prozent der Befragten war dies im Arbeitsalltag noch kein Thema. Die Akzeptanz ließe sich vor allem durch „Effiziente Werkzeuge und eindeutige Leitlinien“ sowie die „Anerkennung als wissenschaftliche Leistung“ erhöhen. Die Forschenden wünschen sich ein Forschungsdatenmanagement, das einen „echten Mehrwert“ bietet.

Diskussion und Ausblick

Die Umfrage zeichnet ein heterogenes Bild zum Stand der Digitalisierung in der physikalischen Forschung. Einerseits ist die numerische Simulation weit verbreitet. Andererseits werden Experimente und Forschungsschritte noch manuell und oft ad hoc dokumentiert. Die Daten, Metadaten und Werkzeuge sind so vielfältig wie die Inhalte physikalischer Forschung. Dennoch gibt es Gemeinsamkeiten und physikspezifische Schwerpunkte. Diese zeigen sich auch im Vergleich zu den Ergebnissen von Umfragen der NFDI-Konsortien aus der Chemie [8] und den Ingenieurwissenschaften.[3] So korreliert die „IT-Affinität“ von physikalisch Forschenden mit der Vielzahl der benutzten Werkzeuge und Do-it-yourself-Lösungen. Der Mangel an kontrollierten Vokabularen und darauf aufbauenden Ontologien für die Physik stellt eine zentrale Herausforderung für das Forschungsdatenmanagement in der Physik dar.

3) vgl. doi.org/10.25534/tudatalib-104 und nfdi-matwerk.de/wp-content/uploads/2020/08/2020-08-18_NFDI-MatWerk_Digitalsierungsumfrage_Publikation.pdf

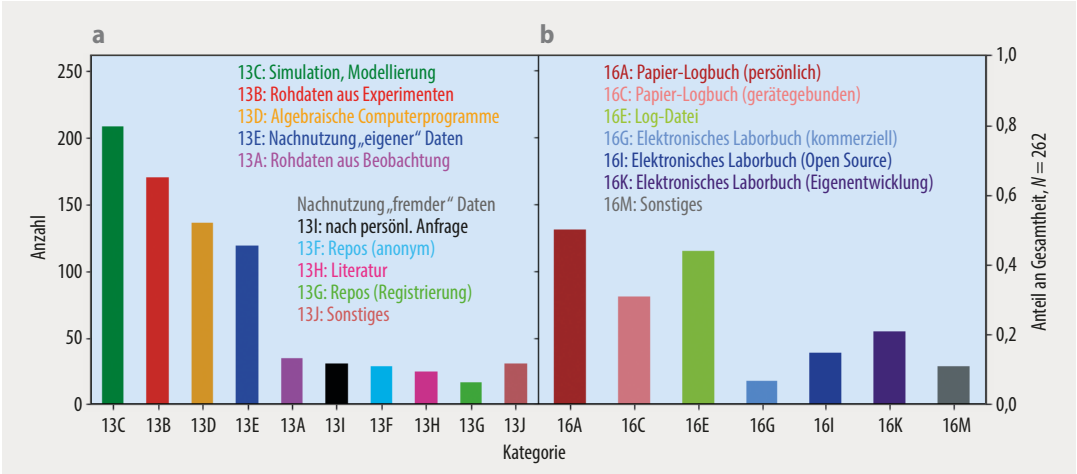

Abb. 3 Die gängigste Art der Datenerzeugung (a) sind die Simulation bzw. Modellierung sowie Rohdaten aus eigenen Experimenten. Die Dokumentation (b) erfolgt nach wie vor am häufigsten mittels Papier-Laborbuch. Bei beiden Fragen waren Mehrfachauswahlen möglich.

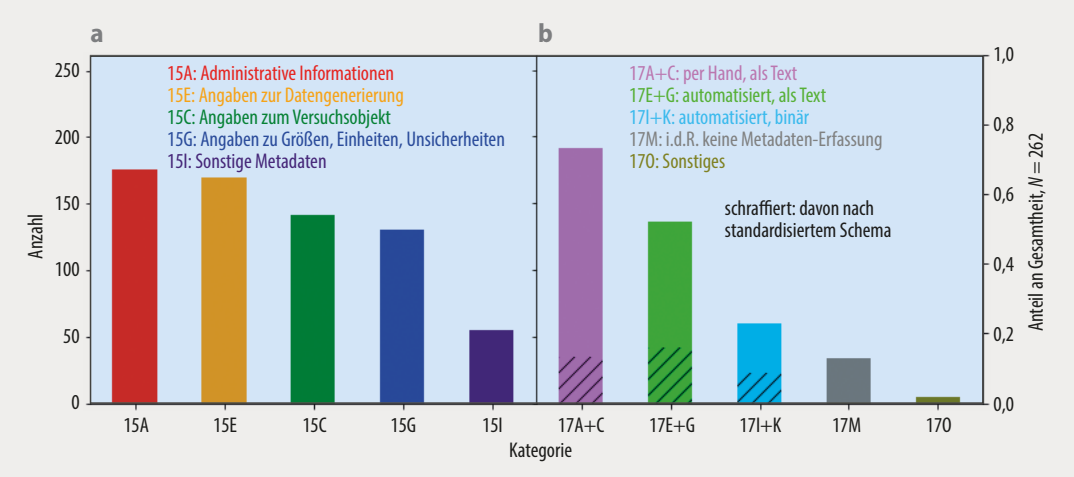

Abb. 4 Die Umfrage erfasste mehrere Metadatenklassen, die genutzt wurden, um Datensätze zu beschreiben (a). In der Mehrzahl werden Forschungsdaten in der eigenen Arbeitsgruppe manuell annotiert, häufig mittels selbstentwickelter Metadatenschemata (b). Mehrfachauswahlen waren möglich.

Die Umfrage bestätigte unsere Vermutung, dass wesentliche Bereiche der Physik, wie die Optik und Quantenoptik, die Physik der Atome, Moleküle und Plasmen, aber auch die biologische und medizinische Physik und nicht zuletzt die statistische Physik und nichtlineare Dynamik sich nicht durch andere Konsortien repräsentiert sehen. In diesen Teilbereichen der Physik basieren die experimentellen oder theoretischen Studien auf Proben und Simulationsserien mit einer großen Anzahl an Parametern. Bei der Konzeption und Auswertung spielen selbstentwickelte Geräte, Modelle und Methoden eine wichtige Rolle. Diese gemeinsamen Charakteristika sind bei der „FAIRifizierung" physikalischer Forschungsdaten zu beachten. Proben und Objekte, die Prozesse der Vorbereitung und Durchführung der Experimente und Studien, aber auch die zur Analyse eingesetzten experimentellen Methoden und Modelle gilt es, durch disziplinspezifische Metadaten zu beschreiben. Ein modularer und flexibler Ansatz zum Verknüpfen bestehender und noch zu entwickelnder Vokabularien, Ontologien und Metadatenschemata berücksichtigt diese Vielfalt am besten. Harmonisierte Metadaten sind die Voraussetzung, um Versuche und Simulationen eindeutig und nachvollziehbar zu beschreiben.

Grob entlang des „Lebenszyklus" von Forschungsdaten geordnet, ergeben sich also die folgenden Handlungsfelder für das Forschungsdatenmanagement in der Physik:

- Aufbau eines definierten Vokabulars als Grundlage einer harmonisierten Metadatenauszeichnung, hin zu einer (vernetzten) Ontologie der Physik (Terminologiearbeit)
- Interoperable Messgeräte mit offenen Schnittstellen und Dateiformaten lassen sich zu modularen, flexiblen Workflows verbinden (Smart Lab).
- Datenbank- und Forschungsdatenmanagement-Systeme sowie Analysetools, die mittels fundierter Terminologie und Ontologie synoptische Multi-Messenger-Forschung (nach dem Vorbild der Astroteilchenphysik [9]) und den Einsatz von maschinellem Lernen und künstlicher Intelligenz erlauben
- Communitystandards für die Physik: Wie wollen wir Datenqualität definieren, messen und belohnen?
- Aufbau eines Systems föderierter Repositorien für die Physik und ihre Integration in NFDI und EOSC
- Förderung der Data Literacy in der Physik: Forschende in der Physik sollen ihre eigenen Data Scientists sein. Entsprechend hoch muss der Stellenwert von „Datenthemen" in Lehre und Weiterbildung sein. Darauf aufbauend gilt es, ein Modell für „Data Stewards" in der Physik zu entwerfen, das auf der Anerkennung von Expertise beruht.
- Ein weiteres Ziel ist die Konzeption eines „Quanten-Forschungsdatenmanagements", das die Einführung von Quantencomputern in die Forschungspraxis begleitet.

Die Umfrage hat wertvolle Impulse für die Priorisierung der Aufgaben im Konsortium NFDI4Phys geliefert und den Grundstock für die Weiterentwicklung gelegt.

*

Für die Mitarbeit beim Erstellen und Unterstützen der Umfrage bedanken wir uns bei Elke Brehm, Lutz Doering, Georg Düchs, Benjamin Gloger, Uwe Kahlert, Christian Krause, Giacomo Lanza, Joachim Meier und Jan Straßburg.

Literatur

[1] *M. D. Wilkinson* et al., Scientic Data **3**,160018 (2016)
[2] *R. L. Acko,* J. Appl. Syst. Anal. **16**, 3 (1989)
[3] *S. Baskarada* und *A. Koronios*, Australas. J. Inf. Syst. **18**, 5 (2013)
[4] *A. Sesartić*, *A. Fischlin* und *M. Töwe,* ISPRS Int. J. Geo-Inf. **5**, 91 (2016)
[5] *H. Frahm,* Physik Journal, März 2019, S. 3
[6] *P. Staehelin*, Physikalische Blätter, Juli 1973, S. 328
[7] *H. Behrens* et al., Physikalische Blätter, Januar 1993, S. 56
[8] *S. Herres-Pawlis, J. C. Liermann* und *O. Koepler*, Z. Anorg. Allg. Chem. **646**, 1748 (2020)
[9] *M. Spurio*, Particles and Astrophysics, Springer (2015)

Die Autoren

Dr. Holger Israel, Physikalisch-Technische Bundesanstalt, Wissenschaftliche Bibliotheken, Bundesallee 100, 38116 Braunschweig; **Dr. Esther Tobschall**, Technische Informationsbibliothek, Welfengarten 1B, 30167 Hannover und **Frank Tristram**, Karlsruher Institut für Technologie, Schlossplatz 19, 76131 Karlsruhe