DGM | Erfahrung · Kompetenz · Wissen
Deutsche Gesellschaft für Materialkunde e.V.

Strategiepapier

Digitale Transformation in der Materialwissenschaft und Werkstofftechnik

Autoren

Stefan Sandfeld, Tim Dahmen,
Frank O.R. Fischer, Christoph Eberl,
Stefan Klein, Michael Selzer, Britta Nestler,
Johannes Möller, Frank Mücklich, Michael Engstler,
Stefan Diebels, Ralf Tschuncky, Aruna Prakash,
Dominik Steinberger, Christian Kübel,
Hans-Georg Herrmann, René Schubotz

www.dgm.de

Digitale Transformation in der Materialwissenschaft und Werkstofftechnik

Strategiepapier

Stefan Sandfeld[1], Tim Dahmen[2], Frank O.R. Fischer[3], Christoph Eberl[4], Stefan Klein[3], Michael Selzer[5], Britta Nestler[5], Johannes Möller[4], Frank Mücklich[6], Michael Engstler[6], Stefan Diebels[7], Ralf Tschuncky[8], Aruna Prakash[1,9], Dominik Steinberger[1], Christian Kübel[10], Hans-Georg Herrmann[11], René Schubotz[2]

[1]Lehrstuhl für Mikromechanische Materialmodellierung, Technische Universität Bergakademie Freiberg
[2]Agenten und Simulierte Realität, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
[3]Deutsche Gesellschaft für Materialkunde e.V.
[4]Fraunhofer-Institut für Werkstoffmechanik IWM
[5]Institut für Angewandte Materialien, Computational Materials Science, Karlsruher Institut für Technologie
[6]Lehrstuhl für Funktionswerkstoffe, Universität des Saarlandes
[7]Lehrstuhl für Technische Mechanik, Universität des Saarlandes
[8]Fraunhofer-Institut für Zerstörungsfreie Prüfverfahren IZFP
[9]Department Werkstoffwissenschaften, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
[10]Institut für Nanotechnologie, Karlsruher Institut für Technologie
[11]Lehrstuhl für Leichtbausysteme, Universität des Saarlandes

Inhalt

1. Management Abstract

Dieses Strategiepapier beschreibt die Vision einer digitalen Transformation in der Materialwissenschaft und Werkstofftechnik (MatWerk) mit dem Ziel, die Transparenz, die Nachhaltigkeit und langfristig auch die Effizienz der Forschung zu verbessern. Wesentliche Bestandteile dieser Transformation sind offene Plattformen, Standards und Technologien für die Datenverarbeitung, den Datenaustausch und die Datenanalyse: Damit können vorhandene Forschungsergebnisse langfristig effizienter genutzt werden. Die Verknüpfung von Daten aus verschiedenen Quellen oder Projekten führt zu zusätzlichen Erkenntnissen. Und der Einsatz neuer Techniken zur Datenanalyse wie Metastudien, Data Mining und maschinelles Lernen erleichtert Studien, die heute nur mit unverhältnismäßig viel Aufwand möglich wären.

Dieses Dokument soll dazu dienen, die derzeitigen Herausforderungen aufzeigen, die die Digitalisierung mit sich bringt, um dann im Anschluss einige individuelle Lösungsvorschläge zu machen.

Für den Erfolg der digitalen Transformation ist es ohnehin besonders wichtig, möglichst die gesamte wissenschaftliche MatWerk-Community in die Gestaltung miteinzubeziehen. In diesem Sinne erhebt dieses Strategiepapier nicht den Anspruch, endgültige Lösungen und eindeutige Wege aufzuzeigen. Vielmehr will das Autorenteam, das unterschiedliche Bereiche der Forschung und Forschungsförderung abdeckt – und neben der Seite der Materialwissenschaft und Werkstofftechnik auch die Seite der Informatik repräsentiert –, eine Reihe von Herausforderungen sowie möglicher Lösungsansätze auf ganz unterschiedlichen Ebenen (von der Lehre über den Forschungsalltag bis hin zu den Förderinstitutionen) aufzeigen.

Einführung

Durch die Digitalisierung haben bereits alle Anwendungsfelder der Ingenieurswissenschaften bahnbrechende Impulse erfahren. Das betrifft namentlich die großen Zukunftsbereiche der Mobilität, Kommunikation, Sicherheit, Gesundheit und Energie – und nicht zuletzt das große, weite Feld der Materialwissenschaft und Werkstofftechnik (MatWerk), die in all diesen Bereichen eine zentrale Rolle spielt. Gerade dort, wo Werkstoffe leicht sein und auch unter extremer Belastung zuverlässig funktionieren müssen, wo also eine genaue Kenntnis der lokalen Eigenschaften von Materialien und Komponenten eine grundlegende Bedeutung hat, zeichnet sich Digitalisierung als Königsweg ab. Studien zufolge basieren nahezu drei Viertel aller neuen Erzeugnisse auf neuen Werkstoffen. Gleichzeitig liegt der Anteil von Materialkosten in der verarbeitenden Industrie bei bis zu 55 Prozent. Wer hier wie dort die Möglichkeiten der Digitalisierung nicht nutzt, oder wer bei seiner strategischen Ausrichtung etwaige Risiken unterschätzt, wird wissenschaftlich – und vor allem wirtschaftlich – abgehängt.

Aktueller Stand

Für die Materialwissenschaft und Werkstofftechnik stellt die Digitalisierung eine besondere Herausforderung dar. Das liegt in dem Umstand begründet, dass sich im MatWerk-Feld traditionellerweise eine Vielzahl von Fachrichtungen – und damit auch eine Vielzahl von unterschiedlichen Standards, Analysemethoden und Modellen – überschneiden. Insbesondere haben Physik, Mathematik, Metallurgie, Chemie, verschiedene Ingenieurswissenschaften sowie die Kristallografie hier eine exponierte Stellung. Gegen Ende des 20. Jahrhunderts wurden diese Disziplinen zusätzlich durch starke Einflüsse aus den entsprechenden rechnergestützten Disziplinen wie „Computational Physics/Chemistry“, „Computational Mechanics“ oder „Computational Materials Science“, aber auch der numerischen Mathematik und der Informatik komplettiert. Dies ermöglicht neben immer realistischeren Simulationen auch moderne Ansätze zur Datenanalyse, von denen beispielsweise experimentbasierte Simulationen heute gerade erst zu profitieren beginnen.

Weitere aktuelle Trendthemen in Forschung und Anwendung reichen von adaptiven und mit Sensoren versehenen Werkstoffen bis hin zu selbstlernenden Fertigungsanlagen oder der Additiven Fertigung. Auch die 3D-Gefügeanalyse mittels Maschinellem Lernen bietet hier ein großes Potential. Dabei überspannt der Prozess die ganze Wertschöpfungskette, und zwar von der atomaren bis zur makroskopischen Ebene: vom Designentwurf über die Erprobung bisher unbekannter Materialkombinationen und die Fertigung mithilfe von datenbankgestützten Charakterisierungsmethoden bis hin zur Qualitätssicherung, der Analyse von Produkt- und Materiallebenszyklen, der Eignungsprüfung und dem Extended Relationship Management (XRM) bei und nach der Auslieferung. Industrie 4.0 und „digitaler Zwilling" (ein virtuelles Abbild eines Werkstoffs, Bauteils oder Produkts, an dem bestimmte Design-Vorstellungen oder die Auswirkungen von Eigenschaften verschiedener Werkstoffe am Rechner noch vor der realen Produktion in jedem Bearbeitungsschritt überprüft und variiert werden können), sowie „Big Data" sind hier Schlüsselwörter.

Die Vorteile der Digitalisierung liegen auf der Hand: Dank Datenbanken zu Eigenschaften und Charakteristiken können unter Berücksichtigung verschiedener Faktoren wie zum Beispiel Temperatur, Alterung oder Beanspruchung Zeit und Energie bei der Entwicklung eingespart werden. Mit Systemsimulationen oder digitalen Werkzeugen können des weiteren Kosten im Produktionsablauf gesenkt und Ressourcen geschont werden. In Zukunft wird es durch die Digitalisierung immer leichter möglich sein, Werkstoffe und Materialien individuell und kombinatorisch maßzuschneidern. Eine gewisse Vorreiterrolle haben dabei bereits größer angelegte Initiativen wie die Materials Genome Initiative (MGI) und der Fraunhofer Materials Data Space (MDS), die beide weiter unten kurz vorgestellt werden.

Herausforderungen

Neben all diesen durchaus verlockenden und neuartigen Möglichkeiten wird aber auch gleichzeitig immer stärker sichtbar, wo möglicherweise fatale Probleme liegen werden. Während die zu bearbeitenden Datenmengen immer stärker anwachsen und die Komplexität der Daten aus möglicherweise multidisziplinären Experimenten und Simulationen immer größer wird, können gleichzeitig die Strategien zum Umgang mit diesen Daten mit den rasanten Entwicklungen nicht mithalten. Dieses Problem ist im Bereich der akademischen Forschung sogar deutlich stärker ausgeprägt als im industriellen Kontext. So ist heutzutage ein Großteil der erzeugten Daten unmittelbar nach der primären Verwendung (zum Beispiel für eine wissenschaftliche Publikation) immer noch im Wesentlichen nicht weiter verwertbar. Neben der fehlenden Möglichkeit, die Daten Dritten zur Verfügung zu stellen, fehlen konsistente Konzepte zur Katalogisierung sowie zur Beschreibung der Daten und jener Prozesse, mit denen diese verarbeitet wurden (sogenannte Meta-Daten).

Darüber hinaus scheitert in vielen Fällen eine Wiederverwertung von Daten durch dritte Gruppen an nicht vorhandenen Verhaltenskodizes, beispielsweise in Bezug auf die Urheberrechte. So ist völlig unklar, wer die wissenschaftliche Anerkennung bekommt oder bekommen sollte: derjenige der die Daten produziert, oder derjenige der ihre Analyse veröffentlicht? Und wem gehören die Daten: dem Steuerzahler, der Heimatinstitution, der wissenschaftlichen Gemeinschaft? Durch diese offenen Fragen ist die Bereitschaft, erhobene Daten anderen Forscherinnen und Forschern zur Verfügung zu stellen, gering. Universitäten sind oftmals strukturell zu schlecht aufgestellt, um die entsprechende Infrastruktur für die Datenspeicherung und -weitergabe zur Verfügung zu stellen. Ähnliches gilt auch für die Förderlandschaft, die zwar die Bedeutung der Digitalisierung im MatWerk-Sektor deutlich wahrnimmt, jedoch derzeit noch keine entsprechenden Fördermittel oder -konzepte zur Verfügung stellt.

Eine verwandte Problematik ist, dass die Digitalisierung bei MatWerk offensichtlich drastisch vom Wissen, den Methoden und Softwarestrategien der Informationstechnologie bzw. der Informatik profitieren könnte. Die für die materialwissenschaftliche Forschung relevanten Algorithmen sind

jedoch in der Informatik meist längst etabliert und daher nicht von Interesse für die Forschung in der Informatik. Ein solches Gefälle zwischen Materialwissenschaften und Informatik ist durchaus problematisch, wenn es darum geht, gemeinsame Fördergelder für die Finanzierung von Projekten einzuwerben, Promovierenden attraktive Themen anzubieten oder Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler für die Rolle zu gewinnen.

All dies führt dazu, dass letztlich in der Forschung ein großer Teil der Ressourcen verschwendet wird, da Daten nur unzureichend ausgewertet werden, Versuche wiederholt werden müssen oder wegen fehlender Dokumentationen Forschungsergebnisse nur unzureichend validiert werden können. Während sich die sozialen Netzwerke, privaten Clouds oder Wissensplattformen im täglichen Leben rasant entwickeln und uns im Privaten einen völlig neuen Umgang mit unseren Daten erlauben, erscheint der Umgang im wissenschaftlichen Umfeld antiquiert, intransparent und kaum nachhaltig.

Eine Vision der digitalen Transformation in der Materialwissenschaft und Werkstofftechnik

Am Ende der digitalen Transformation in der Materialwissenschaft und Werkstofftechnik steht eine Zukunft, in der Computer die Frage „Was wissen wir alles über Material oder Werkstoff X?“ automatisch beantworten können. Die Antwort auf diese Frage wird eine Auflistung aller zu dem Material oder Werkstoff erhobenen Daten sein – sowie eine Beschreibung liefern, wie diese erhoben wurden und wie auf sie direkt zugegriffen werden kann. Das Rechnersystem wird eine Liste aller publizierten Interpretationen der Daten liefern; und es wird den direkten Zugriff auf die Datenverarbeitungskette ermöglichen, die zu diesen Interpretationen geführt hat. In dieser Zukunft wird die Rolle der Wissenschaftlerin oder des Wissenschaftlers noch stärker in der Datenerhebung sowie in der Entwicklung neuer Analysemethoden, Interpretationen und neuer Theorien liegen, denn die Aufgabe der Katalogisierung, Verwaltung und Bereitstellung von vorhandenem Wissen ist weitestgehend automatisiert.

Offensichtlich ist dieser Weg noch weit. Zu dem Ziel führen folgende Bausteine:

- Experimente und Simulationen sowie die jeweils erzeugten bzw. benötigten Daten müssen maschinenlesbar werden. Dazu wird eine physikalisch basierte und standardisierte Beschreibung der Repräsentation eines experimentellen, simulierten oder theoretischen Materials und seines Verhaltens in allen seinen relevanten physikalischen Facetten benötigt. Die Datenerzeugung muss transparent, nachvollziehbar und wiederholbar sein, die automatische Verarbeitung anhand von offen verfügbaren Softwarewerkzeugen nachvollziehbar und die einfache Nutzung durch Dritte mithilfe einer nachhaltigen Speicherung ermöglicht werden, die gleichzeitig die Urheber zitiert. Das sichert die Interoperabilität, Nachhaltigkeit und Validierbarkeit von Experimenten sowohl numerisch als auch physikalisch.

- Zur dauerhaften Archivierung, zur Nutzung oder zum Austausch von Daten müssen offene Plattformen geschaffen werden. Diese ermöglichen es beispielsweise, bereits existierende Daten zur Validierung der eigenen Forschungsergebnisse verwenden zu können.

- Es müssen Konzepte zur Qualitätssicherung der in Datenplattformen abgelegten Daten erarbeitet werden, die Unsicherheiten oder Fehler, die mit der Charakterisierungsmethode oder dem Materialmodell zusammenhängen, quantifizieren und zusammen mit den Daten erfassen.

- Die Fragen, wem Daten gehören oder wem die wissenschaftliche Anerkennung gebührt, bedürfen einer Klärung. Hierzu ist ein neuer Verhaltenskodex und ein entsprechendes Umdenken vonnöten.

- Um zu vermeiden, dass bereits Existierendes dupliziert werden muss, sollten die Methoden zur Datenanalyse für alle zugänglich und verwendbar sein. Dazu sind Aktivitäten auf den verschiedenen akademischen Ebenen erforderlich. Studierende und Promovierende müssen in Bezug auf technische Fähigkeiten zur Datenverarbeitung, Datenanalyse und dem Datenhandling ebenso wie in der Softwareentwicklung angemessen ausgebildet werden.

All diese Schritte erfordern grundlegende Änderungen im Verhalten und im Denken von allen beteiligten Wissenschaftlerinnen und Wissenschaftlern.

2. Herausforderungen

In der Verbindung mit den Fortschritten im Bereich der experimentellen Methoden und der Charakterisierung von Materialien wird die voranschreitende Digitalisierung eine Vielzahl von Entwicklungen in der Materialwissenschaft und Werkstofftechnik erheblich beschleunigen. Hierzu müssen allerdings sowohl experimentelle Ergebnisse als auch die Weiterverarbeitung der Daten digital erfolgen und mit Simulationsmethoden kombinierbar sein.

Wie bereits angedeutet, wird der Weg zur digitalen MatWerk-Welt kein leichter sein. Hier die größten Herausforderungen aus dem Blickwinkel der Autoren:

Heterogenität der Daten

Die Datenverarbeitung in der Materialwissenschaft und Werkstofftechnik ist momentan ausgesprochen heterogen. Dies ist teilweise in der Vielzahl der eingesetzten Geräte bzw. Methoden und den sich daraus ergebenden Anforderungen an die Datenverarbeitung und Datenanalyse begründet. Selbst bei einer Beschränkung auf bildgebende Verfahren kommen zahlreiche Techniken zum Einsatz: neben der optischen Mikroskopie, der Röntgenmikroskopie oder der Atomsonde unter anderem auch die Rasterelektronenmikroskopie (SEM), die Rastertransmissionselektronen-mikroskopie (STEM) oder die Hochauflösende Phasenkontrast Transmissionselektronenmikroskope (HR-TEM). Hinzu kommen Techniken wie Energiedispersive Röntgenspektroskopie (EDX), Elektronenenergieverlustspektroskopie (EELS) und Elektronenrückstreubeugung (EBSD), die sich mit rasterelektronenmikroskopischen Verfahren kombinieren lassen. Die Liste kann beliebig fortgesetzt werden, auch in Richtung nicht-bildgebender Techniken.

Die enorme Vielzahl der eingesetzten Verfahren führt dazu, dass Materialien durch eine große Anzahl mitunter sehr unterschiedlich strukturierter Datensätze beschrieben werden, was insbesondere den Aufbau zentraler Datenbanken und der dafür benötigten Datenschemata bzw. Ontologien erschwert. Dies bezieht sich sowohl auf experimentell gewonnene Daten als auch auf Simulationsergebnisse, was das MatWerk-Feld prinzipiell von anderen Disziplinen unterscheidet.

Fehlende Transparenz und Informationsverlust zu Methoden

Auf die Erhebung von Rohdaten durch Experiment oder Simulation folgt in der Regel eine Verarbeitungskette, die aus dem gewonnenen Material Informationen extrahiert. Während es inzwischen weitestgehend etablierter Standard ist, die Rohdaten auf einem zentralen Server oder Akquisitionsrechner für einige Jahre zu archivieren, erfolgt die Weiterverarbeitung meist dezentral. Auf den Workstations der Wissenschaftler ist deshalb ein mehr oder weniger breit angelegter Werkzeugkasten an verfügbarer Software installiert, der kommerzielle Software, etablierte Open-Source-Werkzeuge, hochspezialisierte und mitunter nicht ganz ausgereifte Forschungssoftware sowie selbstentwickelte Skripte mischt. Einige Institutionen versuchen zwar, diesen Mix durch den Einsatz von virtuellen Maschinen, standardisierten Images oder einen Pool identischer Workstations zu vereinheitlichen. In der Regel aber werden die Workflows zur Datenverarbeitung und -analyse, mit deren Hilfe Rohdaten für abgesicherte, publizierbare Erkenntnisse in Zwischendarstellungen überführt, Kenngrößen oder Features extrahiert und diese statistisch ausgewertet werden können, nur auf der jeweiligen lokalen Workstation implementiert. Zudem werden sie nicht nach einheitlichen Standards dokumentiert – und in der Veröffentlichung zwar nach bestem Wissen und Gewissen, zur Wiederverwendung und Weiterentwicklung der Workflows aber meist unvollständig beschrieben.

Aus dem Vorgehen ergibt sich eine Reihe von Problemen. Das erste und offensichtlichste betrifft die Intransparenz im Rahmen eines Peer-Review-Prozesses. Aus Sicht eines Gutachters ist die Datenverarbeitung meist nur durch die Beschreibung in der Publikation beurteilbar. Diese reicht in aller Regel nicht aus, um die gesamte Auswertung vollständig nachzuvollziehen, sodass nur auf

Plausibilität geprüft werden kann. Selbst in dem seltenen Fall, dass alle Zwischendarstellungen, Projektdateien, Skripte und so weiter mit dem Manuskript zur Verfügung gestellt werden, scheitert eine optimale Bewertung in der Praxis an den Softwarelizenzen oder dem Installationsaufwand zur genauen Rekonstruktion der Arbeitsumgebung, welche die Datenverarbeitung für den Gutachter nachvollziehbar macht.

Da ein wesentlicher Bestandteil der materialwissenschaftlichen Forschung in der Entwicklung von Methoden zum Erkenntnisgewinn über die Eigenschaften eines Materials oder Werkstoffs besteht, wiegt deren Verlust an Methoden, der mit der lokalen Arbeitsweise einhergeht, vermutlich noch schwerer. Diese Methode beinhaltet in zunehmendem Maße und mit wachsender Bedeutung eine Kette aus Schritten zur Datenverarbeitung: einschließlich der benötigten Software, der Ausführungsumgebung und dem Wissen, wie und in welchen Schritten die Software ausgeführt werden muss. Auch diese Implementierung der Methode ist, zumindest insoweit die Datenverarbeitung betroffen ist, in aller Regel an die Workstation des Forschers gebunden: Solange der Forscher noch an der Institution beschäftigt ist, lässt sie sich oftmals noch mit vertretbarem Aufwand durch Anlernen bzw. Einrichten auf eine andere Person bzw. Workstation übertragen. Spätestens mit dem Weggang eines Forschers ist die von ihm verwendete Methode der Datenverarbeitung allerdings zumeist nicht mehr (exakt) rekonstruierbar. Versuche, diesen Verlust durch Dokumentation oder Automatisierung zu verhindern, scheitern in aller Regel an der Komplexität des Sachverhalts – oder schlichtweg an mangelnder Motivation.

Uneinheitliche Qualitätsstandards und Problematiken bei der Qualitätssicherung

Zur besseren Nutzbarkeit und Interpretation von Daten ist es von besonderem Interesse, die erhobenen Daten aus Experimenten und Simulationen auf einen einheitlichen Standard zu heben. Dazu gehört neben einer sorgfältigen Einhaltung der guten wissenschaftlichen Praxis beim Umgang mit Daten vor allem die Schaffung einheitliche Standards bei der Datenerhebung, Verarbeitung und Visualisierung. Nur so kann eine hohe Datenqualität und letztendlich eine Nachhaltigkeit und Wiederverwertbarkeit von Daten ermöglicht werden.

Dies ergibt auch die Frage nach den Verantwortlichen für fehlende oder fehlerhafte Daten. Waren das bisher die bei der Versuchsdurchführung beteiligten Wissenschaftler, kommt nun auch die Frage der Verantwortung bei einer Weiternutzung der Daten für weiterführende Untersuchungen durch andere Wissenschaftler auf. Es ist daher zwingend notwendig, eine Qualitätssicherung zu etablieren, welche die Zuverlässigkeit der Daten gewährleistet. Dazu gehört eine größtmögliche Transparenz der Datenquellen, zu welchem Materialsystem diese zuzuordnen sind. Zu klären ist bei der Qualitätssicherung auch, ob die Materialien wie dokumentiert behandelt wurden – und wie sich die ermittelten Daten nach aktuellen Maßstäben richtig interpretieren lassen. Wie muss mit unvollständigen Datensätzen und fehlenden Metadaten umgegangen werden? Notwendig sind dazu umfangreiche Beschreibungen der Daten und Verarbeitungsmethoden, zum Beispiel der Produktionsbedingungen der Materialien, der Probenherstellung, der Charakterisierungsparameter und der Visualisierungsmethoden. Diese Beschreibungen sind derzeit entweder gar nicht vorhanden, oder nicht ausreichend standardisiert und etabliert, stellen aber letztlich die Voraussetzung für eine branchenübergreifende und multidisziplinäre Verarbeitung von Daten sicher.

Fehlende Kultur des Daten-Sharings

Das heutige System wissenschaftlichen Wettbewerbs basiert in erster Linie auf Veröffentlichungen. Für das Ansehen und damit die Karrierechancen eines Wissenschaftlers oder einer Wissenschaftlerin zählen vor allem Anzahl und Qualität der Publikationen. Nach den heutigen Bewertungskriterien für Veröffentlichungen ist hierbei die bestmögliche Analyse und Interpretation erhobener Daten durch neue Erklärungen, Erkenntnisse oder Theorien wesentlicher Maßstab. Zwar existieren Ausnahmen wie die Zeitschrift *Scientific Data*, die die Veröffentlichung

reiner Daten explizit fördert, diese sind aber noch rar. Ein Anreiz, erhobene Daten anderen Wissenschaftlern oder der Öffentlichkeit zur Verfügung zu stellen, existiert somit kaum. Da die Datenerhebung aber sehr aufwendig ist, ergibt sich daraus als einzig erfolgversprechende Strategie, einmal erhobene Daten selbst zu interpretieren und die erlangten Erkenntnisse danach so oft und prominent wie möglich zu publizieren.

Derzeit sind Anreize zum Daten-Sharing sehr selten, und die Kultur der gemeinsamen Nutzung in der wissenschaftlichen Community ist kaum ausgeprägt. Selbst bei Bereitschaft scheitert ein Datenaustausch oft an fehlenden Richtlinien, die klären, wie die entsprechenden Daten verwendet werden sollen bzw. dürfen – und wie die Urheber solcher Daten wissenschaftlich gewürdigt werden. So ist zum Beispiel bei Zweitveröffentlichungen unklar, wann eine bloße Zitierung des ursprünglichen Artikels der Datenveröffentlichung ausreicht (wie es momentan zum Best-Practice-Modell einer erkenntnisgeleiteten Materialforschung gehört) und wann dieser Urheber als Co-Autor genannt werden müsste. Erschwert wird diese Problematik paradoxerweise durch die öffentlichen Materialdatenbanken: Werden hier akkumulierte Daten oder größere Mengen von Datensätzen verwendet, so werden in der Regel die Datenbanken zitiert, nicht aber alle Forscher, die Daten beigesteuert haben. Dies stellt wiederum einen Nachteil für die Autoren dar, welche die Daten zur Verfügung stellen. Bessere oder andere Methoden zur Bewertung der wissenschaftlichen Ergebnisse sind dringend erforderlich. Die Problematik kann erweitert werden auf die Aggregation von Daten, auf die Verknüpfung von Daten aus verschiedenen Quellen, und die Nutzung von Metadaten.

Mangel an Bewusstsein, Akzeptanz und Beteiligung

Im Allgemeinen ist die Erkenntnis für die Nützlichkeit der Digitalisierung und der gemeinsamen Nutzung von Daten bei MatWerk weniger stark ausgeprägt als in vergleichbaren Disziplinen wie den Lebenswissenschaften oder der Physik. Dies führt zu einer geringeren Akzeptanz und Beteiligung von Wissenschaftlern und Gruppen an gemeinschaftlichen Digitalisierungsprojekten. Dieser Mangel an Bewusstsein wird teilweise auch durch die derzeit mangelnde Ausbildung in Bezug auf Aspekte der Digitalisierung bereits bei der akademischen Ausbildung verursacht. Im Gegensatz hierzu stehen einzelne Wissenschaftler und Gruppen, meist aus den Bereichen Simulation oder aus dem *Integrated Computational Material Engineering* (ICME), die sich in größerer Tiefe für Aspekte der Programmierung, Simulationstechnik, des *High Performance Computing* (HPC) oder auch des Maschinellen Lernens interessieren. Gerade, weil diese Gruppen jedoch an Forschungsthemen arbeiten, die einen natürlichen, sehr nahen Bezug zu Digitalisierungsproblem haben, gelingt es ihnen bislang kaum, die gesamte Community „mitzunehmen“ und Community-weite Änderungsprozesse anzustoßen.

Schwierige Randbedingungen zur Infrastruktur

Um die zuverlässige und nachhaltige Datenspeicherung zu ermöglichen, werden Speicherpools benötigt, die zentral und für mehrere Forschungsgruppen zur Verfügung stehen. Zudem fehlen Leitfäden, um den richtigen Umgang mit den Daten zu gewährleisten. Richtlinien sind notwendig, um Klarheit über die Datenspeicherung zu schaffen: also ein Gleichgewicht zwischen Datenspeicherung und Wiederherstellung zu finden, da es beizeiten besser und einfacher sein kann, die Daten selbst neu zu generieren.

Die derzeitigen Finanzstrukturen an einer universitären Einrichtung machen strukturelle Änderungen aber schwierig. Die Beschaffung von Repositories und Datenbanken ist eine erhebliche finanzielle Investition in Massenspeicher und deren langfristige Vorhaltung und Wartung. Insbesondere fallen neben den direkten infrastrukturbezogenen Ausgaben wie Hardware, Strom und Bandbreite aufgrund des Personalbedarfs für die dauerhafte Weiterentwicklung von Plattformen, für Datenpflege und Wartung auch mitunter erhebliche und

vor allem langfristige Kosten an. In der Summe liegen diese Ausgaben weit außerhalb der Finanzierungsmöglichkeiten einzelner Forschungsgruppen.

Die Finanzierung über Drittmittel ist jedoch problematisch, da die dauerhafte Vorhaltung der Daten zum Zweck der Nachnutzung nach Projektende mit dem projektbezogenen Förderkonzept der meisten Fördergeber schlecht kompatibel ist. Im Endresultat verweisen Förderinstitutionen meist darauf, dass die langfristige Vorhaltung von Daten aus Mitteln der Grundausstattung erfolgen soll, wohingegen Universitäten auf den vorhabenspezifischen Charakter der Daten verweisen und sich ebenfalls nicht zuständig fühlen. Der Schwarze Peter wird somit hin- und hergeschoben.

3. Lösungen

Zur Bewältigung der im vorangegangenen Kapitel beschriebenen Herausforderungen schlagen wir die folgenden Lösungen vor:

Öffentliche Materialdatenbank

Eine zentrale Aufgabe, der sich die Materialwissenschaft und Werkstofftechnik im Zuge der Digitalisierung stellen muss, ist die Implementierung und Etablierung einer öffentlichen Materialdatenbank, die insbesondere das Problem der fehlenden Transparenz und Nachhaltigkeit von erzeugten Materialdaten beseitigt. Darüber hinaus ermöglicht eine solche zentrale Speicherung von Materialdaten eine eindeutige Verknüpfung der Daten mit ihren Urhebern, was einen ersten wichtigen Anreiz für das Data Sharing darstellt. Im Folgenden werden zwei Ansätze dargestellt, die in manchen Aspekten als Vorbild dienen können:

- Die U.S.-Amerikanische ***Materials Genome Initiative*** (MGI)[1] ist eine 2011 ins Leben gerufene regierungsgestützte Initiative mehrerer Institutionen mit dem Ziel, den Zyklus von der Entwicklung neuer Materialien bis hin zur Herstellung und zur Anwendung drastisch zu beschleunigen. In der MGI nimmt die Bestrebung, Materialdaten zu digitalisieren und mittels Datenbanken verfügbar zu machen, einen zentralen Platz ein. Dabei liegt der Schwerpunkt auf der Kombination von experimentellen und durch Simulationen gewonnenen Daten sowie darauf, den Datenpool durchsuchbar zu machen. Eine Besonderheit der Materials Genome Initiative ist seine Organisationsstruktur, wobei die Liste der aktiv engagierten Partnerorganisationen von Förderinstitutionen wie der National Science Foundation (NSF) über Regierungsorganisationen wie den Departments of State, Interior, Defense und Energy bis hin zu Forschungsinstituten wie dem National Institute for Standards and Technology (NIST) oder der National Aeronautics and Space Administration (NASA) reicht. Das erhebliche Budget ermöglicht unter anderem auch die langfristige Organisation, Finanzierung und Pflege der entsprechenden Materialdatenbanken.
- Die 2017 ins Leben gerufene Fraunhofer-Initiative ***Materials Data Space*** (MDS) ist von der Idee motiviert, dass es für die technologische Souveränität und für geschlossene Wertschöpfungsketten in Deutschland von zentraler strategischer Bedeutung ist, in der aktuellen Werkstoffforschung die Anforderungen einer durchgängigen Digitalisierung zu berücksichtigen. Der MDS ist eine digitale Plattform, die einen digitalen „Lieferschein" für Werkstoffe und Bauteile ermöglicht. Dies umfasst genaue Informationen zur Zusammensetzung und Herstellung von Materialien ebenso wie die gesamte Fertigungshistorie eines Bauteils zur Abschätzung der Werkstoffeigenschaften und des Werkstoffverhaltens im Bauteil – bis hin zu Recycling- und Re-Use-Informationen. Der MDS speist sich aus vier zentralen Schnittstellen: Prozesskettensimulation, mit Sensoren versehene Werkstoffe, lernende Fertigungsanlagen und Monitoring mit Echtzeittechnik.

Von diesen zwei großen, singulären Beispielen abgesehen existieren aktuell bereits einige kleinere Community-getragene Materialdatenbanken. Diese enthalten entweder rein charakterisierende Informationen etwa für kristallografische Daten mit unterschiedlichen Schwerpunkten (zum Beispiel Bilbao[2], oder ICSD[3]), Simulationsdaten für vergleichsweise einfach zu ermittelnde Eigenschaften (zum Beispiel MaterialsProject, OQMD[4]), experimentelle Standardkennwerte (zum Beispiel MatWeb[5]) oder sehr spezielle Daten zu einem einzelnen Forschungsfeld (zum Beispiel DABEF). Ein alternativer Ansatz wird in der NOMAD[6]-Datenbank verfolgt, in der es sowohl ein

[1] https://www.mgi.gov
[2] http://www.cryst.ehu.es
[3] http://www2.fiz-karlsruhe.de/icsd_home.html
[4] S. Kirklin, et al., The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, Npj Comput. Mater. 1 (2015) 15010.
[5] http://www.matweb.com/
[6] https://www.nomad-coe.eu/

Repository zur Ablage beliebiger atomistischer Konfigurationen als auch ein Archiv mit aus diesem Repository generierten normalisierten Informationen gibt – verbunden mit der Möglichkeit, diese Daten online zu analysieren. Eine Variante, die es erlaubt, sowohl experimentell als auch numerisch ermittelte Daten gleichermaßen in einer Datenbank zu hinterlegen und einen programmatischen Zugriff (API) auf diese zu ermöglichen, existiert zurzeit jedoch nicht.

Natürlich braucht eine solche vereinheitlichende öffentliche Materialdatenbank hohe und einheitliche Qualitätsstandards, was Aufwendungen für permanente Plausibilitäts- und Kompatibilitätsprüfungen nötig macht. Dies kann entweder bereits automatisiert beim Upload, vor der Veröffentlichung durch einen Peer-Review Prozess oder danach – etwa durch eine Art „community-review" ähnlich wie in der „Wikipedia" – erfolgen. Offensichtlich haben all diese Begutachtungsprozesse Vor- und Nachteile, und es gilt daher zu untersuchen, welche für die MatWerk-Gemeinschaft am geeignetsten sind. Außer Frage steht jedoch, dass diese enorme Aufgabe nicht von einer Institution alleine gestemmt werden kann (wie es jedoch bei fast allen oben beschriebenen Lösungen der Fall ist), sondern nur unter Beteiligung und Partizipation aller Forscher und Forscherinnen in den Materialwissenschaften und der Werkstofftechnik.

Neben der Limitierung auf entweder experimentelle Messwerte oder Simulationsergebnisse in bestehenden Datenbanklösungen werden dort häufig reinen Materialdaten abgelegt, d.h. die Zusammenfassung von verschiedenen Messungen an einer Probe oder mehrerer Proben. So zielführend ein solches Vorgehen für den Anwendungsfall der reinen Werkstoffauswahl auch ist, so wenig nachhaltig ist es. Ohne jene Werte, die zur Bestimmung eines Kennwertes benutzt wurden, ist schließlich eine nachträgliche Neuauswertung – etwa durch methodische oder theoretische Weiterentwicklungen oder eine umfassende Analytik des gesamten Zustandsraumes – nicht möglich. Insbesondere ermöglichen Daten über einzelne Werkstoffproben die genaue Analyse zum Beispiel von Ausfallursachen oder machen Mechanismen verständlich. Gleichzeitig kann das Abspeichern einzelner Messungen von enormem Wert für spätere Methodenentwicklungen (bzw. Referenztests) oder erneute statistische Auswertungen und für Kreuzkorrelationen unabhängiger Messungen – zum Beispiel von lokaler Härte, Eigenspannungen, Konzentrationen, Versetzungsdichten – sein.

In der zu entwickelnden Materialdatenbank müssen daher sowohl an verschiedenen Stellen ein und derselben Probe gemessene bzw. sich auf das gesamte Probenvolumen beziehende Daten als auch die daraus ermittelten Kennwerte ablegbar und eindeutig unterscheidbar sein. Gleichzeitig muss die Datenbank aber auch so flexibel sein, dass nicht alle diese Informationen erforderlich sind und beispielsweise auch das reine Hinterlegen von Kennwerten ohne die Dokumentation jeder hierzu erfolgten Messung möglich bleibt.

Standardisierte Datenformate

Eine unabdingbare Voraussetzung für eine erfolgreiche Digitalisierung der Materialwissenschaft und Werkstofftechnik sind verbindliche Standards für die Ablage und den Austausch von Werkstoffdaten. Einheitliche Datenformate tragen einerseits dazu bei, dass die Datenheterogenität verringert wird; andererseits sorgt die durchgängige Konformität von den eigenen Daten mit dem festgelegten Standard für eine einheitlichere Datenqualität. Nicht zuletzt können aus klar definierten Datenformat-Standards auch Richtlinien für den Ausbau der Rechen-, Speicher- und Netzwerkkapazität abgeleitet werden, was dem Problem der heute oft unpassenden Infrastruktur entgegenwirkt.

Bisher entwickelte Formate für Materialdaten sind meist nicht für experimentelle und simulierte Ergebnisse gleichermaßen nutzbar, oft sehr anwendungs- und/oder skalenspezifisch und enthalten häufig aus mehreren Einzelversuchen extrahierte Kennwerte sowie kaum Meta-Informationen darüber, wie die Daten erzeugt wurden. Einige bereits existierende Formate weisen eine starre, wohl-definierte Struktur mit sehr begrenzter Variabilität auf, wie das *Chemical*

Information Format (CIF) zur Speicherung kristallographischer Daten oder das POSCAR-Format zum Ablegen von Atomsorten und -koordinaten. Solche Formate sind zwar vergleichsweise unflexibel und häufig nur begrenzt nutzbar, haben aber den Vorteil, sich gut für eine anschließende Datenanalytik zu eignen. Andere Datenformate, wie das *Physical Information Format* (PIF) oder das *Materials Information Format* (MIF), weisen eine hohe Flexibilität auf, wenn es darum geht, eine Vielzahl von selbst zu definierenden Eigenschaften, aber auch Rohdaten und Meta-Daten zu hinterlegen. Die dort abgespeicherten Dateien sind aber nicht per se direkt miteinander vergleichbar, da möglicherweise gleiche Eigenschaften unterschiedlich benannt oder als Funktion unterschiedlicher Parameter abgelegt wurden.

Um eine breite Akzeptanz in der Community zu schaffen, müssen die zu entwickelnden Formate flexibel sein und über ein begrenztes und eindeutiges Vokabular für die Charakteristika und Eigenschaften von Werkstoffen verfügen. Ebenfalls entscheidend ist, dass das Format für auf verschiedenen Skalen simulierte sowie experimentell gemessene Daten gleichermaßen nutzbar ist. Dem hierarchischen Aufbau von Materialien muss dabei auch in ihrem Datenformat Rechnung getragen werden.

Angesichts der enormen Vielschichtigkeit und Diversität von Werkstoffdaten ist der erfolgversprechendste Weg zu einheitlichen Standards, dass einzelne Disziplinen und Fachgebiete diese zunächst separat definieren (*bottom up*). Im anschließenden Schritt müssen die einzelnen Formate dann zusammengefasst und vereinheitlicht werden. Ein in der Informatik, vor allem im Bereich der Internettechnologie etablierte, dezentral organisierte Prozess zur Formatstandardisierung ist der *Request for Comments* (RFC). Er entstammt dem akademischen Umfeld und hat sich unter technologisch heterogenen, sich schnell ändernden und durch teils gegenläufige Einzelinteressen beteiligter Gruppen geprägte Rahmenbedingungen, die dem MatWerk-Umfeld vergleichbar sind, als pragmatisch erfolgreiches Vorgehen erwiesen. Wir empfehlen daher, den RFC-Prozess zur Standardisierung von Datenformaten in den MatWerk bekannt zu machen und zu etablieren.

Linked Open Science

Praktisch gesehen bringt die fortschreitende Digitalisierung der Wissenschaften mit sich, dass Forscher der unterschiedlichen Disziplinen angehalten werden, ihre Forschungsdaten, Quellcodes und andere Produkte des Forschungsprozesses über das Internet offen zugänglich und nachnutzbar zu machen. Ein technisches Konzept zur Umsetzung dieser Nachhaltigkeitsziele stellt *Linked Open Science* (vernetzte offene Wissenschaft) dar. Darunter versteht man die Kombination von *Linked Data, Open Source Software* und *Web-basierte Arbeitsumgebungen, Cloud Computing, und Open Data.*

Unter Nutzung der Möglichkeiten für die Kommunikation und Zusammenarbeit verändert *Linked Open Science* die traditionellen Forschungsprozesse hin zu einer offeneren, kollaborativeren digitalen Arbeitsweise und bietet Strategien, Werkzeuge und Technologien, mit denen die wissenschaftlichen Disziplinen den Veränderungen der Digitalisierung begegnen können.

Linked Data bezeichnet eine Ansammlung von technischen Vorgehensweisen für die Veröffentlichung sowie die Vernetzung von Daten und Informationen im *World Wide Web*. Dabei ist jeder Datensatz über einen Uniform *Resource Identifier* (URI) eindeutig adressierbar und kann durch einen Hyperlink auf andere Daten verweisen. Die so miteinander verknüpften Daten ergeben ein weltweites *Web of Data*, welches eine maschinelle Datennutzung und -kombination gemäß internationaler *W3C Standards* erlaubt. Mit dem *Web of Data* verbindet sich also die Erwartung, Daten aus ganz verschiedenen Quellen frei nutzen und miteinander verknüpfen zu können. Wurden zuvor spezialisierte Datenaustauschformate entwickelt, beispielsweise mittels eines RFC Prozesses, so erlauben offene Standards und Programmierschnittstellen die Automatisierung von komplexen Recherchen, Überwachungen und Berichterstattungen mit

geringem Aufwand. Zusammenfassend lässt sich sagen, dass mit *Linked Data* eine durchgängige Verknüpfung und Verwaltung von digitalen Forschungsdatensätzen entlang des gesamten Datenlebenszyklus von der Erhebung, Erschließung und Speicherung auf der einen und der Archivierung und des Zugriffs auf der anderen Seite möglich ist.

Open Data definiert rechtliche Rahmenbedingungen und Standard-Lizenzverträge, mit denen ein Autor der Öffentlichkeit auf einfache Weise Nutzungsrechte an seinen Werken einräumen kann. Die *Open Data*-Standards können als rechtliche Grundlage für die Nutzung und Kombination der im Netz bereitgestellten Daten verstanden werden.

Cloud Computing stellt eine technische Vorgehensweise zur Entkopplung der IT-Infrastruktur von der Anwendung dar. Allgemein versteht man unter *Cloud Computing* die Verbindung von virtualisierten und damit skalierbaren Hardware-Ressourcen und serverseitige Ausführung von fachlichen, zum Beispiel wissenschaftlichen Anwendungen als Services. Der Betrieb von IT-Infrastruktur wird in Regel zentral organisiert, da das Thema sehr technisch ist und wenig direkten Bezug zu MatWerk-Kompetenzen besitzt. Die Entwicklung wissenschaftlicher Software hingegen erfordert in der Regel tiefes fachliches Verständnis und ist daher zumeist dezentral zu organisieren. Durch die Entkopplung von IT-Betrieb und Softwareentwicklung kann *Cloud Computing* somit auch im MatWerk-Kontext einen wichtigen Beitrag zur Umsetzung des *Linked-Open-Science*-Konzeptes liefern.

Prozessverbesserungen für wissenschaftliche und Open Source Softwareentwicklung

Ein weiterer Block von Maßnahmen muss der Verbesserung im Entwicklungsprozess wissenschaftlicher Software, insbesondere von *Open Source Software***,** gelten. Momentan kann beobachtet werden, dass der Reifegrad von Anwendungen im MatWerk Umfeld extrem stark variiert. An einem Extrem finden sich kommerzielle Systeme und voll entwickelte *Open-Source*-Pakete, die von darauf spezialisierten Organisationen spezifiziert, entwickelt, gepflegt und im Fall kommerzieller Software auch vertrieben werden. Am anderen Ende des Spektrums finden sich aufgabespezifische Skripte, die oftmals von Promovierenden ohne spezifische Ausbildung in der Softwareentwicklung für ein konkretes Forschungsvorhaben umgesetzt werden, aber nie einen Prozess der Anforderungsanalyse, Spezifikation oder Qualitätssicherung unterlaufen. Auf die Probleme, die mit der lokalen Speicherung dieser Skripte verbunden sind, wurde oben schon verwiesen.

Festzuhalten bleibt an dieser Stelle, dass der Schritt, eine erstellte Software soweit zu abstrahieren und zu verallgemeinern, dass sie auf verwandte aber nicht identische Aufgabenstellungen angewendet werden kann, zumeist unterbleibt. Hierzu müssten zunächst Mindeststandards für die Qualität des Entwicklungsprozesses definiert und eingehalten werden. Dabei ist jedoch zu beachten, dass Standards im Vergleich zu spezialisierten Organisationen vereinfacht werden müssten, da in vielen Fällen kein professionelles Team zur Softwareentwicklung zur Verfügung steht. Andererseits erlauben es modulare Zusatzausbildungen den Wissenschaftlerinnen und Wissenschaftlern, spezifische Software unter Einhaltung definierter Standards zu entwickeln.

Zur Entwicklung eines Standards bietet sich ein Modell in verschiedenen Stufen an:

- *Stufe 1: Versionskontrolle mit kontrolliertem lokalen build.* Der Quelltext der Software wird in einem Versionsverwaltungssystem (git, svn, mercurial etc.) gespeichert. Wenn die Software kompiliert werden muss, erfolgt dies über ein Build-Werkzeug (cmake, scons, autotools etc.). Es existiert eine Dokumentation, welche Entwicklungsumgebungen oder Bibliotheken zur Entwicklung und zum Einsatz der Software erforderlich sind.
- *Stufe 2: Buildserver und Testautomation.* Die Software wird durch automatisierte Komponenten- und Integrationstests gesichert, aber mit uneinheitlichem Testaufwand. Eine

zentrale Entwicklungsumgebung kompiliert und testet jede neue Version im Versionsverwaltungssystem automatisch.
- *Stufe 3: Qualitätsmanagement.* Fachliche und technische Anforderungen an die Software werden schriftlich fixiert. Art und Umfang der Softwaretests sind schriftlich definiert.

Die Realisierung dieses Szenarios erfordert recht überschaubare Maßnahmen. Dabei sind die Maßnahmen der Stufe 1 durchaus schon in vielen software-affinen Forschungsgruppen zumindest aus dem simulativen Bereich etabliert, und sogar kommerzielle Programme, wie MATLAB, unterstützen verschiedene Arten von Versionskontrollen. Komponenten, Unit,- und Integrationstests (Stufe 2), die auf unterschiedlichen Ebenen die korrekte Softwarefunktionalität sicherstellen, sind in der MatWerk-Gemeinschaft deutlich seltener anzutreffen. Mögliche Ursache für die spärliche Nutzung ist neben der Unkenntnis der Methodik auch der (scheinbare) erhöhte Arbeitsaufwand, der (scheinbar) in etwas investiert werden muss, das nicht unmittelbar mit der Lösung des wissenschaftlichen Problems zusammenhängt. Gleichwohl existieren komfortable Verwendungsmöglichkeiten in verschiedenen Integrierten Softwareentwicklungsumgebungen, oder auch gekoppelt an Versionsverwaltungssystems (zum Beispiel von GitHub bereitgestellt).

Integrierte Datenverarbeitungsplattform

Ein Beispiel für die Umsetzung dieses *Linked-Open-Science*-Konzeptes ist eine integrative Plattform, welche die Datenverarbeitung konsolidiert. Diese Infrastruktur besteht typischerweise aus einer Instanz pro Arbeitsgruppe, kann prinzipiell aber auch institutsweit zentralisiert oder bei Projekten mit sehr speziellen Anforderungen pro Vorhaben separat aufgesetzt werden.

Dabei stellt eine Plattform eine Arbeitsumgebung zur Verfügung, die einen Zugriff auf die im Einsatz befindlichen Softwarewerkzeuge bietet. Die einzelnen Funktionen der Software werden hierbei derart gekapselt, dass ein standardisiertes Protokoll zur Automatisierung entsteht. Das bedeutet, dass zwar weiterhin eine Vielzahl unterschiedlicher kommerzieller, selbst entwickelter oder über *Open Source* verfügbarer Softwarewerkzeuge zum Einsatz kommen kann. Jedes Modul wird jedoch derart gekapselt, dass es durch eine einheitliche Schnittstelle automatisiert werden kann. Indem Skripte gegen diese Schnittstelle entwickelt werden, können Arbeitsabläufe softwareübergreifend automatisiert werden. Das Vorgehen wurde beispielsweise von der elektronischen *Labbook Jupyter*[7] erfolgreich umgesetzt.

Der Ansatz einer vollständigen Integrierung und Automatisierung der Datenverarbeitung hat den Vorteil, dass die Implementierung von Methoden selbst in einem Umfeld mit hoher Fluktuation an Mitarbeitern wesentlich besser erhalten bleibt. Es stellt sich jedoch das Problem, dass der Ansatz, für jede Aufgabe ein Skript zu entwickeln, aus Sicht der meisten MatWerker keine eingängige oder effiziente Art ist, mit Software zu arbeiten. Um das Problem zu lösen, müsste eine graphische Programmiersprache entwickelt werden, die die standardisierte Automatisierungsschnittstelle als Backend nutzt. Die Sprache bedient sich einer Graph-Metapher; Algorithmen oder Rechenschritte werden als Knoten repräsentiert, Daten als Kanten.

Das Paradigma ist von Lösungen wie LabView her bekannt und wurde auch in zahlreichen Anwendungen aus der Computer Grafik bereits erfolgreich und mit hoher Akzeptanz bei auch wenig programmieraffinen Nutzerinnen und Nutzern angewendet. Die für die Implementierung einer Methode erforderliche Datenverarbeitung wird als „Datenverarbeitungsgraph“ (DVG) bezeichnet. Die beschriebene graphische Programmiersprache wird durch eine browserbasierte Benutzerschnittstelle bedient. Somit kann der Zugriff auch einzelne Abläufe in der Datenverarbeitung von jedem Terminal erfolgen.

[7] http://jupyter.org

Um den eingangs angesprochenen Verlust von Methoden vorzubeugen, ist es erforderlich, dass die Implementierung der Datenverarbeitung einer Versionsverwaltung unterliegt. Das bedeutet, dass alle erstellten DVG im Server versionskontrolliert gespeichert werden, sodass Änderungen nachverfolgt werden können.

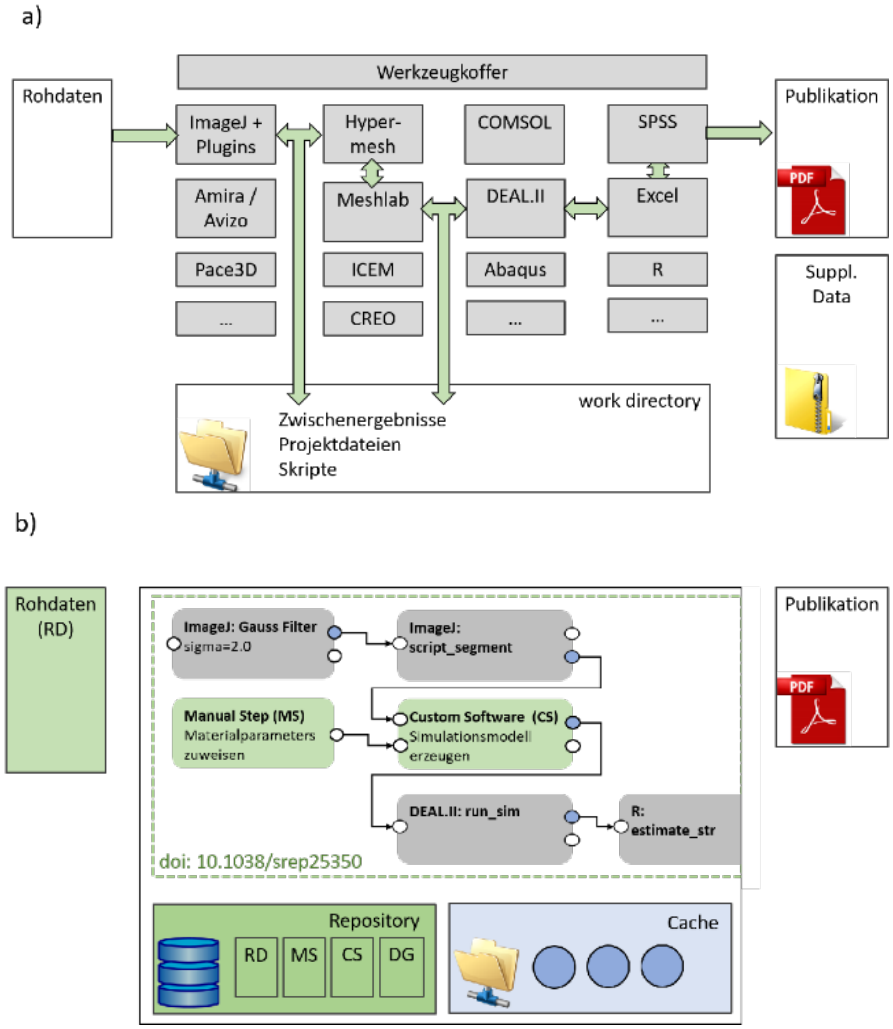

Abbildung 1. Das Konzept einer integrierten Datenverarbeitungsplattform ändert die Sicht auf Datenverarbeitung und -analyse: a) In der heutigen Sicht findet die Datenverarbeitung in erster Linie auf den lokalen Workstations der Forschenden statt. Die Datenverarbeitung wird als Arbeitsablauf betrachtet. b) Zukünftig soll die Datenverarbeitung auf einer integrierten Plattform stattfinden, deren Benutzung mittels einer graphischen Programmiersprache im Browser erfolgt. Die Datenverarbeitung wird als eigenständiges Dokument betrachtet, das archiviert, geteilt, begutachtet und veröffentlicht werden kann.

Gleiches gilt für den Einsatz selbst entwickelter Skripte und Software. Jede Änderung an einem genutzten Skript stellt eine Änderung der Methodenimplementierung dar und soll der Versionskontrolle unterliegen. Zusammenfassend kann gesagt werden, dass eine Reihe von Daten unter Versionskontrolle gespeichert und dauerhaft archiviert werden. Diese Daten sind: (1) alle experimentellen Rohdaten, (2) alle Daten, die von einem Menschen manuell bearbeitet wurden, (3) die Komponenten der Ausführungsumgebung, auf der die DVG laufen – einschließlich installierter Software und selbst entwickelter Skripte – sowie (4) der DVG selber.

Ein wesentlicher Vorteil dieser Strategie liegt darin, dass der Datenverarbeitungsteil implementierter Methoden referenziert werden kann. Da die DVG über ein browserbasiertes *User Interface* bearbeitet werden können, können einzelne DVG mit einem Persistent Identifier (PI), z.B. Permalinks, versehen werden. Anhand dieser PIs kann ein DVG mit einem anderen Forscher geteilt und Zugriff gewährt werden. Im Rahmen einer Veröffentlichung kann der PI zu einem DVG als

Supplementary Material eingereicht werden. Prinzipiell ist auch eine Verknüpfung von PI zu DVG mit dem DOI System denkbar, sodass langfristig DVG auch eigenständig veröffentlicht und zitiert werden können. Dieser Absatz scheint vor allem dann sinnvoll, wenn sich zukünftig ein System modularer DVG Fragmente vergleichbar den Modulen einer *Standardbibliothek* herausbildet und sich einzelne Forscher darauf spezialisieren, wiederverwendbare DVG Komponenten zur Verfügung zu stellen.

Deep Learning und *Digitale Realität*

Ein wesentlicher Trend der letzten Jahre, der auch auf die Materialwissenschaft und Werkstofftechnik abstrahlt, stellen die enormen Fortschritte im Bereich der Künstlichen Intelligenz (KI) dar. Hierbei sind vor allem Künstliche Neuronale Netze unter dem Schlagwort *Deep Learning* hervorzuheben. Mittels *Deep Learning* ist es möglich, große Datenmengen automatisiert und mit hoher Komplexität hinsichtlich der Semantik des untersuchten Phänomens zu untersuchen. Die Anwendungen sind zahlreich; am offensichtlichsten sind die Möglichkeiten zur automatischen Klassifizierung, Objekterkennung und pixelgenauen Markierung von Objekten in Mikroskopie-Daten. Insbesondere im Bereich der *in-situ*-Mikroskopie gerade auch mit hoher Zeitauflösung, im Bereich von *High-Throughput* Experimenten, der korrelativen Mikroskopie und überall dort, wo die manuelle Auswertung der anfallenden großen Datenmengen an ihre Grenzen stößt, werden zukünftig verstärkt Künstliche Neuronale Netze zumindest eine erste, grobe Auswertung übernehmen. Doch gerade auch im Bereich von Mustererkennung bei nicht bildgebenden Verfahren, zum Beispiel bei mikromechanischen Versuchen – und generell bei Daten, die von menschlichen Experten nur mit Mühe zu beurteilen sind – kann eine automatisierte Klassifizierung nützlich sein.

Deep Learning hat die Grenze des Machbaren in der automatischen Datenanalyse in den letzten Jahren wesentlich verschoben; aber die Verfahren sind nicht ohne Probleme. Das Training der Netzwerke, die generell dem Bereich des Überwachten Lernens zuzuordnen sind, erfordert große Mengen an Rechenleistung und noch größere Mengen bereits korrekt analysierter Trainingsdaten. Während der Bedarf an Rechenleistung durch die zunehmende Verfügbarkeit von Parallelrechnern in Form von Rechenclustern und auf *Deep-Learning*-Anwendungen spezialisierte GPU-Lösungen adressiert werden kann, stellt die Verfügbarkeit von Trainingsdaten eine ernstzunehmende Herausforderung dar. Zum einen sind die experimentellen Kosten für das Erzeugen ausreichender Mengen an Trainingsdaten zu hoch, zum anderen stellt die manuelle Analyse der Trainingsdaten einen nicht zu rechtfertigenden Aufwand dar. Auch treten die für die Fragestellung relevante Konstellationen experimentell nur mit sehr niedriger Wahrscheinlichkeit auf. Und wurde das fragliche Phänomen theoretisch vorhergesagt, aber experimentell noch nicht beobachtet, stehen erst gar keine Trainingsdaten zur Verfügung.

In diesen Fällen bietet es sich an, Trainingsdaten synthetisch zu erzeugen. Das Vorgehen kommt in zahlreichen Anwendungen in der Informatik bereits zum Einsatz: So wurden gerenderte Tiefenbilder parametrischer Modelle genutzt, um führende Systeme zur Gestenerkennung zu trainieren. Dreidimensionale Objektmodelle wurden angewandt, um Systeme zur Objekterkennung zu trainieren. Synthetische Bilder von Verkehrsszenen wurden genutzt, um *Deep-Learning*-Systeme zur semantischen Segmentierung im Anwendungsfall autonomer Fahrzeuge zu trainieren.

Einen *systematischen* Ansatz zur Erzeugung synthetischer Trainingsdaten stellt das Konzept *Digitale Realität* dar. Hierzu werden parametrische Modelle, die jeweils einen Teilaspekt der Realität abbilden, zu sogenannten Szenarien zusammengesetzt. Durch Festlegung aller Parameter eines Szenarios kann ein solches instanziiert werden, es entsteht eine virtuelle Probe. Diese stellt jeweils einen Punkt im Parameterraum möglicher Mikrostrukturen dar. Durch eine Simulation eines Messvorgangs (*in-silico*-Experiment) werden aus der virtuellen Probe nun synthetische Daten erzeugt, die dann zum Training der Künstlichen Intelligenz verwendet werden können. Der Vorteil

des Ansatzes liegt einerseits darin, dass die manuelle Auswertung der Daten entfallen kann da das erzeugende System das zugrundeliegende Modell ja bereits kennt. Außerdem sind die statistischen Eigenschaften des Parameterraumes vollständig bekannt und können beliebig beeinflusst werden. Das bedeutet: Für jedes Phänomen können bestimmte Instanzen mit der Häufigkeit erzeugt werden, die optimale Trainingsergebnisse liefert, statt auf die Parameter eines Fertigungsprozesses angewiesen zu sein.

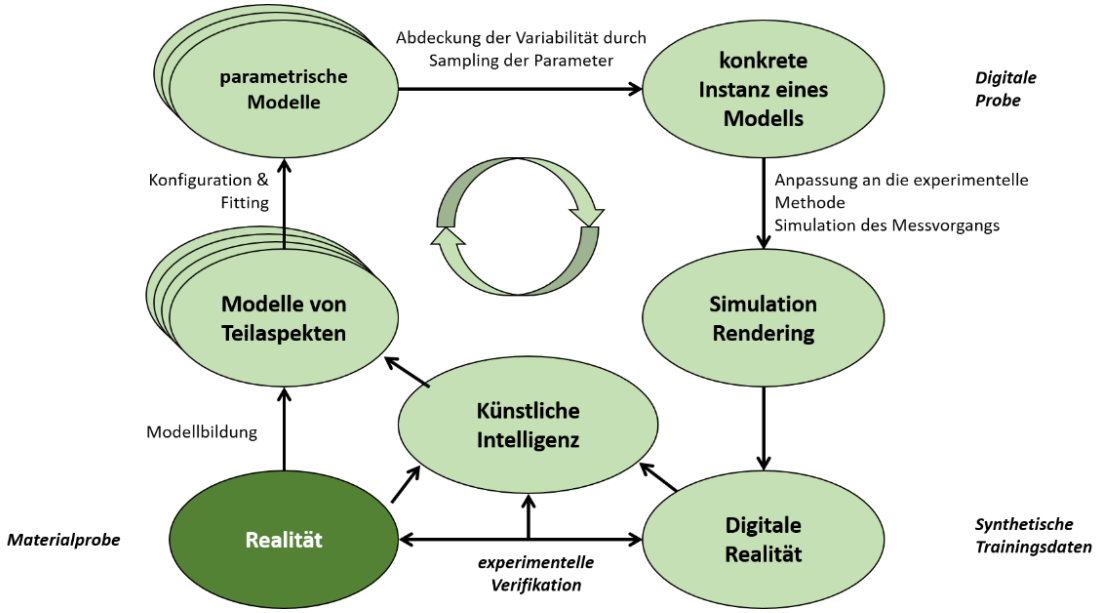

Abbildung 2. Das Digital-Reality*-Konzept zur synthetischen Erzeugung von Trainingsdaten für* Deep Learning*: Partielle Modelle der materialwissenschaftlichen oder werkstofftechnischen Realität werden zu parametrischen Szenarien zusammengesetzt. Durch Festlegung aller Parameter eines solchen Modells entsteht eine* Digitale Probe*. Mittels Vorwärtssimulation eines Messvorgangs werden aus der Probe synthetische Messdaten erzeugt, die dann zum Training einer Künstlichen Intelligenz eingesetzt werden können.*

4. Schritte zur Umsetzung

Die in diesem Dokument vorgeschlagenen Lösungen sind nicht ad-hoc umsetzbar. Dementsprechend ist eine dringende strategische Frage, wie eine schrittweise Umsetzung möglichst effizient und effektiv realisiert werden kann. Aus der Frage ergeben sich Aufgaben für verschiedene Gruppierungen und Institutionen.

Maßnahmen durch die Community

Ein wesentlicher Aspekt bei der Umsetzung der digitalen Transformation ist es, die MatWerk-Gemeinschaft für das Thema zu sensibilisieren. Dies soll durch bewusstseinsbildende Maßnahmen initiiert werden – beispielsweise auch in Form dieses Strategiepapiers. Dabei nehmen Demonstratoren – also Erfolgsgeschichten von durchgeführten Projekten, die besonders von neuartigen Digitalisierungsansätzen profitieren – einen wichtigen Stellenwert ein. Solche Demonstratoren haben die wichtige Funktion, in der MatWerk-Gemeinschaft eine entsprechende intrinsische Motivation zu erzielen. Ohne überzeugende Erfolgsgeschichten wird ein Großteil der Forschenden nur die zusätzlichen kurzfristigen technischen und finanziellen Investitionen sehen.

Eine unmittelbare Maßnahme zur Standardisierung von Datenformaten besteht darin, das Thema in den bereits existierenden Gremien, beispielsweise in den entsprechenden Arbeitskreisen der Deutschen Gesellschaft für Materialkunde e.V. (DGM), auf die Tagesordnung zu setzen. In dem Rahmen könnte auch der enge Kreis der direkt an der Standardisierung beteiligten Wissenschaftler Schulungen über das Vorgehen im RFC-Prozess erhalten, um danach schrittweise auf eine Standardisierung von Datenformaten hinzuarbeiten.

Eine erfolgreiche Implementierung der unterschiedlichsten Digitalisierungsansätze erfordert Kenntnis der technischen Details und Vorgehensweisen, etwa in Bezug auf die Versionsverwaltung von Auswerteskripten oder der Implementierung von Datenbanken, aber auch von Konzepten, die über die im vorigen Abschnitt aufgezeigten noch Hinausgehen. Hier ist die akademische Lehre in der Pflicht, eben diese Konzepte und Strategien in den Lehrplan aufzunehmen.

Fokussierte Workshops und thematische Sommer-/Winterschulen sind eine sehr geeignete (und mit begrenzten Mitteln realisierbare) Möglichkeit, um nicht nur die Theorie, sondern gleichzeitig auch die Anwendung in *Hands-On*-Sitzungen zu vermitteln. Dabei bedarf es qualifizierter Dozenten, sodass hier die dringende Empfehlung ist, einen engen Schulterschluss mit der angewandten Informatik zu vollziehen.

Der erforderliche Struktur- und Mentalitätswandel vor allem bezüglich der Anerkennung wissenschaftlicher Datenerzeugung und der Bereitschaft zur Zusammenarbeit und gemeinsamen Nachnutzung von Daten und Software erfordert einen langfristigen Diskussionsprozess. Die Diskussionen sollten offen und mit hoher Sichtbarkeit in der MatWerk-Gemeinschaft geführt werden – insbesondere dort, wo Kontroversen bestehen. Als geeignete und einfach umsetzbare Formate bieten sich Podiumsdiskussionen auf Konferenzen an.

Maßnahmen durch Fördergeber

Um die Digitalisierungsstrategie nachhaltig realisierbar zu machen, müssen auch auf Seiten der Fördergeber eine Reihe von Voraussetzungen geschaffen werden. Ein erster Schritt ist die Öffnung vorhandener, projektbezogener Fördermaßnahmen für Aufwände und Ausgaben, die primär projektbezogen sind, ihren vollen Nutzen aber erst in einer digitalen Nachnutzung der Ergebnisse nach Projektende entfalten. Dies bezieht sich auf die Verallgemeinerung, Verbesserung und Veröffentlichung von projektbezogener Software aber auch auf die Aufbereitung und Bereitstellung von Datensätzen.

Diese Dinge sind zwar prinzipiell bereits heute förderfähig; praktisch scheitert eine Bewilligung in der Regel aber daran, dass der zeitliche Horizont von Gutachtern auf den Projektzeitraum begrenzt ist. Somit ergibt sich für solche langfristigeren Maßnahmen zur Steigerung der Nachhaltigkeit zumeist ein ungünstiges Kosten-Nutzen-Verhältnis, das einer Förderung in kompetitiven Verfahren im Wege steht. Eine Lösung könnte darin bestehen, mittels geänderter Begutachtungskriterien Bewusstsein für die Problematik zu schaffen; eine andere Lösung wäre, Maßnahmen zur digitalen Nachhaltigkeit aus anderen Töpfen zu finanzieren.

Während der *Bottom-Up*-Ansatz, projektbezogene Lösungen nachträglich zu verallgemeinern, für wissenschaftlich komplexe, softwaretechnisch aber einfache Anwendungsfällen erfolgversprechend ist, greift das Vorgehen nicht bei Vorhaben, die mit erhöhten technischen Schwierigkeiten einhergehen. So sind Maßnahmen zur Förderung der Entwicklung wissenschaftlicher Software im MatWerk-Kontext ein zweiter wichtiger Punkt. Hier wird ein Projektformat benötigt, das sich stärker an den Rahmenbedingungen komplexer Softwareentwicklung orientiert und technische Entscheidungen (etwa zur Softwarearchitektur) früher und zentraler betrachtet. Insbesondere müssen derartige Vorhaben als Werkzeugentwicklungen verstanden werden und dürfen als solche nicht nach dem angestrebten Erkenntnisgewinn innerhalb des Projektes gebunden sein: Sie müssen sich vielmehr daran orientieren, welchen Nutzen die erstellten Werkzeuge für andere Projekte erzeugen.

Ein weiterer Punkt, in dem Mittelgeber einen wesentlichen Beitrag leisten können, ist die Förderung einer Bereitschaft zur Freigabe und Nachnutzung erstellter Daten und Software bei den Antragstellern. Diese kann ganz wesentlich durch Maßnahmen zur digitalen Nachhaltigkeit wie die Veröffentlichung erzeugter Daten in öffentlichen Datenbanken gesteigert werden, aber auch dadurch, dass die Zusage, die innerhalb von beantragten Projekten erzeugten Daten oder Software als *offen zugänglich, oder als Open Source* zur Verfügung zu stellen, bei der Bewertung und Begutachtung eine Rolle spielt, bzw. gefordert wird.

Maßnahmen auf politischer Ebene

Ein Punkt, der auf politischer Ebene geklärt werden muss, betrifft die Zuständigkeit für die langfristige Finanzierung des Betriebs wissenschaftlicher Datenbanken. Hiervon sind direkte Infrastrukturausgaben für Hardware, Strom und Bandbreite betroffen, aber auch die Personalkosten für die kontinuierliche Weiterentwicklung der Datenbankanwendungen, Schnittstellen, Oberflächen und Ontologien, die dauerhafte Pflege der Datensätze und die langfristige Optimierung von Datenstandards. Für diese Aufgabe ist eine auf kurze Zeiträume ausgerichtete projektbezogene Förderung naturgemäß prinzipiell ungeeignet.

Neben den klassischen Fördergebern wie dem Bundesministerium für Bildung und Forschung (BMBF), der Deutsche Forschungsgemeinschaft (DFG) und Förderstiftungen kämen hierfür die Hochschulen und Fakultäten als Träger in Frage, aber auch wissenschaftliche Gedächtniseinrichtungen wie Bibliotheken, Rechenzentren und Archive sowie existierende beziehungsweise neu zu gründende Institutionen. Aus der Vielzahl der Möglichkeiten leitet sich für die Politik der klare Auftrag ab, hier eine Klärung herbeizuführen und eine klare Verantwortlichkeit mit entsprechenden Finanzierungsmöglichkeiten zu schaffen.

Alle diese Maßnahmen komplett umzusetzen wird sicherlich noch einiges an Zeit benötigen. Und natürlich wird auch nicht jedes Institut und jede Forschungsgruppe alle Maßnahmen umsetzen können oder müssen. Unsere Hoffnung ist jedoch, dass nach und nach ein Umdenken erfolgt, das sich mittelfristig positiv auf alle Forschungsabläufe im Bereich der Materialwissenschaft und Werkstofftechnik auswirken wird – vielleicht auch mit Hilfe dieses Strategiepapiers.

DGM | Erfahrung · Kompetenz · Wissen
Deutsche Gesellschaft für Materialkunde e.V.

Deutsche Gesellschaft für Materialkunde e.V.

Die Deutsche Gesellschaft für Materialkunde e.V. vertritt die Interessen ihrer Mitglieder – als Garant für eine kontinuierliche inhaltliche, strukturelle und personelle Weiterentwicklung des Fachgebiets der Materialwissenschaft und Werkstofftechnik.

Besucheranschrift

Deutsche Gesellschaft für Materialkunde e.V.
Wallstraße 58/59
10179 Berlin
069 / 75306 750
dgm@dgm.de

Postanschrift

Deutsche Gesellschaft für Materialkunde e.V.
c/o INVENTUM
Marie-Curie-Straße 11 - 17
53757 Sankt Augustin
069 / 75306 750