DataPLANT: Data in Plant research

Sprecher: Dirk von Suchodoletz; Universität Freiburg, dirk.von.suchodoletz@uni-freiburg.de, ORCID 0000-0002-4382-5104

In der modernen hypothesen-basierten Forschung sind Wissenschaftlerinnen und Wissenschaftler zunehmend auf Dienste und Infrastrukturen des Forschungsdatenmanagements (FDM) angewiesen, die die Erfassung, Verarbeitung, den Austausch und die Archivierung von Forschungsdatensätzen erleichtern. Dabei ermöglicht ein modernes, integriertes FDM erst die Verknüpfung von interdisziplinärer Expertise, sowie Vergleich, Integration verschiedener Analyseergebnisse und Metadatenstudien mit dem darauf beruhenden immensen zusätzlichen Erkenntnisgewinn. Das Ziel von DataPLANT ist es, diesen Mehrwert für den Bereich Pflanzen-Grundlagenforschung zu generieren. Hierzu soll eine nachhaltige, nutzerfreundliche FDM Infrastruktur für diesen Bereich etabliert und durch “Data Stewards” erheblich unterstützt und an konkrere Bedürfnisse angepasst werden.

In der Pflanzen-Grundlagenforschung werden die (molekularen) Prinzipien des pflanzlichen Lebens erforscht, die Pflanzenwachstum, Ernteertrag und Biomasseproduktion bestimmen. Daher ist die Pflanzen-Grundlagenforschung ein multidisziplinärer Forschungsbereich, der unterschiedliche Forschungsfelder wie die Molekulargenetik, Biochemie, Zellbiologie, Systembiologie, Physiologie, Entwicklungsbiologie einschließt und gewonnene Erkenntnisse in der Pflanzenbiotechnologie und Pflanzenzüchtung zur Anwendung bringt (DFG 202-[01, 04, 05, 06, 07]). Um die grundlegenden Prinzipien und Funktionsweise von Pflanzen zu klären, sind zentrale Ansätze in der Pflanzen-Grundlagenforschung unter anderem: (i) mehrere Parameter unter wechselnden (Umwelt)Bedingungen aufzuzeichnen, (ii) die Wirkung genetischer und biochemischer Manipulationen zur Veränderung der Gen- oder Proteinaktivität zu messen, (iii) und die natürliche genetische Vielfalt und Evolution zu analysieren.

Die hierzu eingesetzten Methoden von Transkriptomik, Proteomik und Metabolomik bis hin zu bildgebenden Verfahren und spezialisierten Phänotypisierungsplattformen erzeugen umfassende, hochdimensionale polymorphe Daten, die verarbeitet und integrativ interpretiert werden müssen. Die erfolgreiche Nutzung von Daten unterschiedlicher Modalitäten – aus vielen Quellen und Experimenten, vorverarbeitet oder analysiert mit einer Vielzahl von Algorithmen – erfordert eine Kontextualisierung der Daten. Die “FAIR Data” and Linked Open Data-Prinzipien bieten entscheidende Richtlinien für FDM im Allgemeinen. Darauf aufbauend haben verschiedene Konsortien daher Vorschläge zur besten Vorgehensweise und Erfüllung dieser Grundsätze gemacht, doch ist es fast immer an der Initiative der einzelnen Forscher in seiner Fachwissenschaft, diese auch umzusetzen und in seiner Fachdisziplin zu verankern. Daher stehen umfassende, nutzbare Informationen über die erforderliche Qualität für die Verwendung durch Dritte oft nur in seltenen Fällen zur Verfügung. Eine immer wiederkehrende Erfahrung ist es, dass Forscher praktische Unterstützung bei der Nutzung der fragmentierten und komplexen Ressourcenlandschaft benötigen. Dies erhöht die Notwendigkeit einer maßgeschneiderten (Infra)struktur für FDM.

Durch den Zusammenschluss von technisch-fachlicher Expertise in den Bereichen Pflanzen-Grundlagenforschung, Informations- und Computerwissenschaften und Infrastrukturspezialisten wird DataPLANT Pflanzenwissenschaftlerinnen und -wissenschaftler

im Umgang mit Forschungsdaten individuell angepasst unterstützen. Durch seine Infrastrukturspezialisten ist DataPLANT bezüglich der IT und Infrastruktur bereits jetzt sehr gut in die Europäische Forschungslandschaft eingebettet, beispielsweise durch die Teilnahme am Deutschen ELIXIR Knoten und in den Bereichen Cloud-Computing, Tools, Arbeitsabläufen sowie Pflanzen-Bioinformatik und Datenanalyse. Freiburg, Tübingen und Jülich gehören zur Open Science Cloud (EOSC) und engagieren sich unter anderem im Rahmen von EOSC-LIFE sowie sind wichtige Mitglieder des BioDATEN Konsortiums. Ebenso engagiert sich DataPLANT in der pflanzlichen Metadatenstandardisierung im Rahmen von MIAPPE (Minimal Information about a Plant Phenotyping Experiment). Dabei bündelt das DataPLANT-Konsortium das Know-how seiner Mitglieder aus dem Forschungs- und Infrastrukturnetzwerk und bildet die Forschungslandschaft in unterschiedlichen Rollen ab. Die domänenspezifischen Mitglieder sind thematisch kohärent und bringen wichtige Akteure aus der Grundlagenforschung zusammen, darunter z.B. TRR 175-The Green Hub - Central Coordinator of Acclimation in Plants, das Science Data Center BioDATEN und der Cluster of Excellence on Plant Sciences (CEPLAS).

Aufbauend auf einer Basisinfrastruktur (BinAC, bwSFS, de.NBI Cloud, bwCloud) und maßgeschneiderte Ressourcen (ELIXIR, Galaxy, NF-Core) wird mit DataPLANT eine Serviceumgebung geschaffen, um Forschungsdaten nach den FAIR-Prinzipien mit minimalem Zusatzaufwand zu kontextualisieren und den gesamten Forschungszyklus in der modernen Pflanzenbiologie zu unterstützen. Die maßgeschneiderte Servicelandschaft in DataPLANT wird sich aus technisch-digitaler Assistenz sowie personelle Vor-Ort-Assistenz durch sogenannte “Data Stewards” zusammensetzten. Die Data Stewards bilden die Brücke zwischen den Pflanzenforschen hin zu den Informationswissenschaftlern und Infrastrukturspezialisten bzw. der Infrastruktur an sich. Dabei gewährleistet der ständige Austausch eine Plattformentwicklung entlang der Nutzerinteressen. DataPLANT schafft so einen zentralen Einstiegspunkt und eine wertvolle fachspezifische Daten- und Wissensressource (DataPLANT HUB) [Abb.1].

Die kollaborative Forschungsdatenplattform wird Datenprovenienz und Forschungsaustausch gewährleisten und mit Hilfe eines Motivations- und Kreditsystems den Anreiz zur Demokratisierung von Forschungsdaten fördern. Darüber hinaus wird so ein spezifischer Community-Standard für grundlegende Pflanzenforschung (Metadaten) und Workflow-Annotationen, basierend auf generischen, bestehenden und neu entstehenden Standards und Ontologien entwickelt und nutzergetrieben wachsen. Mechanismen zur kollaborativen Forschung unterstützen dabei durch automatisierte Vernetzung von pflanzenforschungsspezifischen (Meta)daten die Erstellung und Verwaltung wohlannotierter Forschungsdatenobjekte. In Verbindung mit Lehre und Trainingskonzepten wird das Sachverständnis im Umgang mit Daten gestärkt und eine Langzeitmotivation zur Schaffung wohlannotierte Datenobjekte erzeugt.

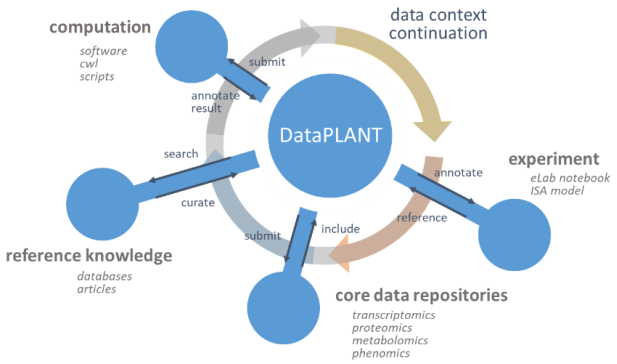

Abbildung 1 Der DataPLANT Hub begleitet und unterstützt den gesamten Forschungszyklus.

Durch die Integration der Pflanzenwissenschaft in das Gesamtnetzwerk NFDI, treibt DataPLANT den digitalen Wandel und die Demokratisierung der Forschungsdaten im Feld voran. Dabei wird sich DataPLANT aktiv an der angedachten Vereinsstruktur der NFDI beteiligen, um einen Rahmen für übergreifender Austausch der Konsortien und eine enge Abstimmung gemeinsamer Themen zu ermöglichen. DataPLANT strebt einen zielgerichteten Austausch in fachgruppen-orientierten Rahmen an, wie er beispielsweise durch NFDI4Life Umbrella geboten wird. Für zukünftige Forschungsvorhaben soll bereits früh eine Einbindung in die jeweilige Community-NFDI erfolgen, um viele Fragen der Standardisierung, Metadaten sowie rechtliche und ethische Aspekte frühzeitig zu klären. An dieser Stelle will sich DataPLANT dafür einsetzen Schnittstellen zu den Fördergebern zu entwickeln, so dass ein gemeinsames Vorgehen in Hinblick auf Open Data und Open Science abgestimmt wird. DataPLANT wird sich einsetzten, den angestrebten Kulturwandel, der ein Reputationsgewinn durch Datenveröffentlichung befördert, durch die NFDI gemeinsam mit den Förderinstitutionen voranzutreiben. Deshalb sollen zu einem frühen Zeitpunkt gemeinsame Richtlinien zu geeigneten Lizenzen für Forschungsdaten abgestimmt werden. Ebenso will sich DataPLANT in den Fragen der Personalgewinnung und -ausbildung einbringen, die auf einem gemeinsamen Fundament für alle Konsortien ruhen sollen. Weitere Herausforderungen liegen in der Motivation für eine umfassende Annotation von Forschungsdaten (mit Metadaten) und die Sicherstellung einer hohen Datenqualität.

DataPLANT ist Mitunterzeichnerin der "Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung" (http://doi.org/10.5281/zenodo.3895208) und sieht dieses Dokument als Grundlage für die notwendigen, wechselseitigen Abstimmungen auf Basis tragfähiger Prozesse und Strukturen. So sollen Themen, die für mehrere Fachkonsortien relevant sind, im Sinne einer nachhaltigen Interoperabilität kooperativ und über einzelne Konsortien hinweg bearbeitet werden. Weitere Fragen aus den Themenbereichen Infrastruktur, Einbindung der Service-Provider und nachhaltige Finanzierung schließen sich an. Zudem sind rechtliche Themen rund um das Aufbewahren, Teilen und die Nachnutzung digitaler Objekte für alle Konsortien von Belang. Daher sollten zentrale Fragen zu Lizenzen, dem Umgang mit sensiblen Daten, zu IT-Recht, zum Datenschutz und zur Vertraulichkeit von Daten für die gesamte NFDI adressiert werden.

Eine für alle Konsortien zentrale Herausforderung besteht im nachhaltigen Langzeitzugang zu Forschungsdaten. Dies schließt Themenfelder, wie Herkunft (data provenance) und Souveränität (data sovereignty) von Daten, Nachhaltigkeit und Sicherheit von Daten, Tools und Services, sowie persistente Identifikatoren (persistent identifiers, z.B. DOIs oder ePICs) sowie die Zertifizierung von Datenrepositorien und Archiven ein. Hierzu gehört insbesondere die Sicherung des technischen Zugangs und produktive Nachnutzungsoptionen für den ursprünglichen Forschungs- und Erstellungskontext (Re-Use). Ebenfalls sind geeignete Maßnahmen für das Management von Forschungssoftware abzustimmen. Dies umfasst u.a. den Betrieb von Entwicklungs- und Archivierungsrepositorien, das Management von Metadaten zur Referenzieren und Zitation, sowie die standardisierte Beschreibung von Software-Code.

DataPLANT wird somit nachhaltige, nutzerorientierte Infrastrukturen bereitstellen, die abgestimmt in übergreifende NFDI Aktivitäten eingebettet sind. Dabei wird DataPLANT die fachspezifischen Anforderungen der pflanzlichen Community vertreten und insbesondere durch die Rückkopplung der Data Stewards auch auf neueste Entwicklungen in der Community zu reagieren.

Co-Sprecher/in	Zugehörige Institution
Prof. Dr. Björn Usadel Director Institute for Bioinformatics IBG-4, Leiter Institute for Biological Data Science (HHU) b.usadel@fz-juelich.de ORCID 0000-0003-0921-8041	Forschungszentrum Jülich, Wilhelm Johnen Strasse 52428 Jülich
Jun.-Prof. Dr. Timo Mühlhaus Computergestützte Systembiologie muehlhaus@bio.uni-kl.de ORCID 0000-0003-3925-6778	Technische Universität Kaiserslautern Postfach 3049 67653 Kaiserslautern
Dr. Jens Krüger Group Leader High Performance and Cloud Computing ORCID 0000-0002-2636-3163	Eberhard Karls Universität Tübingen Geschwister-Scholl-Platz 72074 Tübingen