Schlussbericht zum Verbundprojekt

„Entwicklung eines nachhaltigen Datenökosystems für die Pflanzenzüchtung: BreedFides“

Förderkennzeichen 16DTM111 A-E

Teil II

1. Aufzählung der wichtigsten wissenschaftlich-technischen Ergebnisse und anderer wesentlicher Ereignisse

Arbeitspaket 1: Entwicklung eines züchtungsfokussierten Anforderungsprofils für ein branchenweites Datenökosystem

Meilenstein 1.2. Der Status quo der Datenkuration der einzelnen Akteure ist erfasst und anerkannte Datenqualitätsindizes von phänotypischen Daten entwickelt (Monat 23).
Meilenstein 1.3. Der Status quo der Datenkuration der einzelnen Akteure ist erfasst und anerkannte Datenqualitätsindizes von genomischen Daten entwickelt (Monat 23).
Meilenstein 1.4. Einheitlichen Standards sind für Feldversuche (Bonitur, Messung) erarbeitet bzw. erweitert (Monat 23).
Meilenstein 1.5. Ein Modell des Datenökosystems ist für die Weizenzüchtung entwickelt (Monat 34).

In verschiedenen Workshops wurde das bestehende Datenökosystem in der Pflanzenzüchtung am Beispiel von Weizen aufgezeigt. Darüber hinaus wurde ein Fragebogen konzipiert, in dem weitere Details des Datenmanagements in der Pflanzenzüchtung bei den Mitgliedsunternehmen der GFPi im Bereich Weizenzüchtung abgefragt wurden. Das bestehende Datenökosystem in der Pflanzenzüchtung am Beispiel des Weizens konnte so beschrieben werden.

Die Ergebnisse zeigen, dass das Datenmanagement von einfachen Varianten auf der Basis von Text- oder Excel-Dateien bis hin zu sehr umfassenden, dezidierten Datenmanagementsystemen für genetische, genomische, phänotypische und Umweltinformationen inklusive entsprechender Metadaten reicht. Insbesondere innerhalb größerer Unternehmen wird stärker darauf geachtet, dass alle im Unternehmen erhobenen Daten auch langfristig miteinander interoperabel bleiben.

Leider ist diese Interoperabilität zwischen unterschiedlichen Unternehmen aktuell wenig bis gar nicht gegeben. Beispielhaft sind hier Daten zu nennen, die für firmeneigene SNP-Arrays erhoben wurden. Ein SNP-Array ist eine molekularbiologische Methode, mit der Hunderttausende einzelner genetischer Varianten eines Individuums (Single Nucleotide Polymorphisms, SNPs) gleichzeitig analysiert werden können, um genetische Unterschiede zwischen Individuen zu erkennen. Sofern die Identität der analysierten genetischen Varianten zwischen zwei Firmen aber unterschiedlich ist, lassen sich die Daten nicht oder nur mit hohem Aufwand in gemeinsamen Ansätzen nutzen.

Dagegen ist die Interoperabilität von phänotypischen Daten in der Pflanzenzüchtung aufgrund von behördlichen de-facto Standards des Bundessortenamts für viele Merkmale generell

höher. Im Rahmen des BreedFides-Projekts wurden einheitliche Standards insbesondere für die Erfassung von Krankheitsresistenzen im Weizen erarbeitet, da hier die größten Lücken in der Interoperabilität bestanden.

Zusammenfassend ist die Interoperabilität von Daten in der Pflanzenzüchtung insgesamt mäßig und es erfordert substanzielle Anstrengungen, um eine gemeinsame Datennutzung in der Branche zu ermöglichen. Eine solche gemeinsame Nutzung wird von den Projektpartnern aber als grundsätzlich möglich und von potenziell hohem Wert eingeschätzt.

Auch die Intensität der Datenpflege ist von Unternehmen zu Unternehmen unterschiedlich. Sie variiert von klar definierten Standardverfahren bis hin zu eher einfachen Verfahren, die stark von der einzelnen Person abhängen. Der Trend geht jedoch eindeutig in Richtung einer nachhaltigen Datenkuration in den Unternehmen. Dabei ist die treibende Kraft die Relevanz historischer firmeneigener Daten für die Kalibrierung genomischer Vorhersagemodelle.

Anerkannte Datenqualitätsindizes für phänotypische Daten sind Wiederholbarkeiten und Heritabilitäten. Bei genomischen Daten spielen Allelfrequenzen und Anteil an Fehlwerten die größte Rolle. Die genomische Wiederholbarkeit und Heritabilität sind zentrale Kriterien für die Bewertung der Kombination von genomischen und phänotypischen Daten.

In den gemeinsamen Workshops wurde ein erstes Modell für ein zukünftiges Datenökosystem in der Weizenzüchtung erarbeitet. Dabei werden die Aufgaben für Datengeber, Datentreuhänder bzw. Datenvermittler, Betreiber und Nutzer einer Datenanalyseplattform und Nutzer der aggregierten Ergebnisse getrennt.

Arbeitspaket 2: Entwickeln und Etablieren eines geeigneten Prozederes für die Beantragung, Prüfung und Entscheidung der Datennutzung durch Akteure des Datenökosystems

Im Zuge der Arbeiten an AP2 „Entwickeln und Etablieren eines geeigneten Prozederes für die Beantragung, Prüfung und Entscheidung der Datennutzung durch Akteure des Datenökosystems“ wurden im Laufe des Projektes die Funktionsweisen und Abläufe des Systems konzipiert. Eine Übersicht über das erarbeitete Konzept bietet **Abbildung** 7.

Den Kernbereich des BreedFides-Konzepts stellt eine Katalogumgebung dar. Hier können Datensätze registriert, also durch Metadaten beschrieben werden. Die Dateneigentümer können aus verschiedenen Bereichen stammen, darunter Züchter, wissenschaftliche Einrichtungen oder öffentliche Quellen. Der jeweilige Dateneigentümer zeigt durch die Registrierung eines Datensatzes und die Veröffentlichung des Metadatensatzes im System an, dass ein Datensatz prinzipiell mit anderen Akteuren des Systems gemeinsam genutzt werden kann. Die eigentlichen Daten verbleiben dabei standardmäßig beim Dateneigentümer, es sei denn, dieser möchte die Daten direkt im BreedFides-System speichern. Alle im System registrierten Datensätze sind zur Recherche durch alle registrierten User freigegeben.

Um ein registrierter User im System werden zu können, muss der User im ersten Schritt die Nutzungsbedingungen akzeptieren. In den Nutzungsbedingungen sind allgemeine Rechte und Pflichten des Users und des Systemanbieters dargelegt sowie die allgemeine Funktionalität des Systems beschrieben. Die Nutzungsbedingungen bilden damit den rechtlichen und regulatorischen Rahmen für die Nutzung des Systems für seine Nutzer. Die Zustimmung zu

den Nutzungsbedingungen ist für alle Nutzer des Systems verbindlich. Die Nutzungsbedingungen wurden sowohl mit den Züchtungsunternehmen aus dem Mitgliederkreis der GFPi als auch mit den Konsortialpartnern gespiegelt und das erhaltene Feedback eingearbeitet.

Sofern ein User einen für sich interessanten Datensatz recherchiert hat, kann er über das System beim Bereitsteller eine Nutzungsanfrage stellen. Vor jeder Verfügbarmachung von Daten muss zunächst eine Datennutzungsvereinbarung (Data Use Agreement) zwischen Datennutzer, Dateneigentümer und ggf. Datenbereitsteller geschlossen werden. BreedFides übernimmt als neutraler Datenvermittler die Koordination und Moderation der Verhandlungen dieser Datennutzungsvereinbarung zwischen dem Datennutzer und potenziellen Dateneigentümern auf der Grundlage eines Standardtextes, da ein möglichst großer Teil der Vereinbarungen standardisiert werden soll. Ziel ist es, den Aufwand für den Abschluss einer Datennutzungsvereinbarung für die Beteiligten so gering wie möglich zu halten.

In der Datennutzungsvereinbarung wird Art und Dauer der Datennutzung geregelt, aber beispielsweise auch welche (Nach)Nutzungsrechte an den erarbeiteten Ergebnissen für wen bestehen oder eingeräumt werden.

Nach erfolgreichem Abschluss der Verhandlungen wird eine rechtlich verbindliche Datennutzungsvereinbarung zwischen den Parteien geschlossen. Anschließend kann der Zugang zu den konkret freigegebenen Daten über einen Cloud Data Analysis Room, direkte Datenübertragung oder jegliche andere Form erfolgen, auf die sich die beteiligten Akteure in der Datennutzungsvereinbarung geeinigt haben.

Arbeitspaket 3: Schnittstellen zu weiteren Datenökosystemen

Für die in AP3 notwendige Definition geeigneter Schnittstellen für phänotypische Daten (AP3.1) wurden für die Pflanzenzüchtung vordefinierte Standards wie solche der Breeding Applicable Programmable Interface (BRAPI) (Selby et al., 2019) oder MIAPPE (Papoutsoglou et al., 2020) identifiziert. Um die Schnittstellen anzupassen, wurde geprüft, inwiefern die in AP1 definierten Metadaten mit den BRAPI oder MIAPPE Standards kompatibel sind. Eine JavaScript Object Notation (JSON) Schnittstelle zum nationalen Evaluierungsprogramm EVA2 ist geplant. Diese soll zunächst mittels des in EVA2 hinterlegten Sortiments, der Akzession und des Merkmals erfolgen (AP3.2).

Innerhalb des Projektverlaufes wurden mehrere Varianten der Bereitstellung von Schnittstellen zu weiteren Datenökosystemen evaluiert und Beispiele für Datenharmonisierungsprojekte aufgezeigt. Das ursprüngliche Infrastrukturkonzept wurde an den Bedarfen der Datennutzenden angepasst und im Rahmen des AP5 als Pilot erfolgreich getestet. Das Endergebnis ist eine per API steuerbare Server-Anwendung, welche Daten aus definierten anderen Ökosystemen abrufen und für die Nutzenden bereitstellen kann (siehe GitHub Repositorium).

Im Projektverlauf hat sich Zusammenarbeit mit dem GAIA-X Projekt als schwierig herausgestellt, da bis zum Ende des Projektes gut nutzbare und konkrete Komponenten zur Schaffung eines GAIA-X kompatibler Dienste nur teilweise vorhanden waren. Ein Beispiel für schlecht nachnutzbare Komponenten ist der im Jahr 2022 verwaiste DataSpaceConnector des

AgriGAIA Projektes. Erst zum Ende der Laufzeit (April 2024) konnte eine Demonstrationsplattform von AgriGAIA genutzt werden. Zu diesem Zeitpunkt war der BreedFides Demonstrator bereits im aktiven Test. Die weitergehende engere Verzahnung mit GAIA-X sollte auch über den gemeinsam mit dem ZALF e.V. eingereichten Projektantrag Soil-X erreicht werden, dieser wurde jedoch abgelehnt. Dennoch konnte über die Kooperation mit dem de.NBI Projekt erfolgreich die de.NBI Cloud für das Deployment der ETL Pipeline verwendet werden. Diese ermöglicht ein cloud-agnostisches Deployment, auch durch die zum Projektende erstellte Container basierte Orchestrierung der ETL Pipeline (siehe GitHub).

Zur Abstimmung einer Kooperation mit dem Projekt FAIR Data Spaces wurde mit Projektpartner Uni Gießen ein Treffen durchgeführt. Hier wurde der aktuelle Stand des BreedFides-Projektes vorgestellt. Hinsichtlich der Abläufe zur Aushandlung von Datennutzungsverträgen gibt es große Übereinstimmungen zwischen beiden Vorhaben. Dies zeigt, dass die entwickelten Konzepte die generellen Anforderungen sehr gut widerspiegeln. Aus der Domäne Pflanzenzüchtung ergaben sich aber zusätzliche spezifische Anforderungen, welche die Entwicklung und Nutzung einer BreedFides spezifischen Cloud-Infrastruktur notwendig machten.

Als Beispiel für die Datenharmonisierung wurden in projektinternen Veranstaltungen die Themenblöcke Schnittstellen und Harmonisierungs- und Transformationsprozesse (Extract-Transform-Load, ETL) bearbeitet. Im Themenbereich Schnittstellen ist das nationale Evaluierungsprogramm EVA2, das Geoportal des JKI, die BreedingAPI, und das Datenportal des Thünen-Instituts vorgestellt und beschrieben worden. Als Beispiel für ETL Prozesse zur Harmonisierung und Transformation von Daten ist der INSPIRE Datensatz der Bodenzustandserhebung Wald vorgestellt und die technische Herangehensweise dokumentiert worden. Das dort verwendete Verfahren kann als Blaupause für ETL Prozesse zur Datenintegration von verschiedenen Datenökosystemen verwendet werden (Siehe Anhang „Introduction to ETL, Extract-Tranfrom-Load“ vom 06.05.2022).

Die erste Iteration der schnittstellenbereitstellenden Anwendung wurde durch einen Geodatenkatalog realisiert. Hierfür wurde die open-source Software GeoNetwork verwendet und auf der Cloud Sigma gehostet. Die GeoNetwork-Software wurde mit ersten Metadatensätzen aus den Datenökosystemen des Deutschen Wetterdienstes, SoilGrids.eu, dem ISRIC Soil Data Hub, der JKI-Geodateninfrastruktur sowie aus dem Thünen-Atlas befüllt (Siehe Abbildung 1).

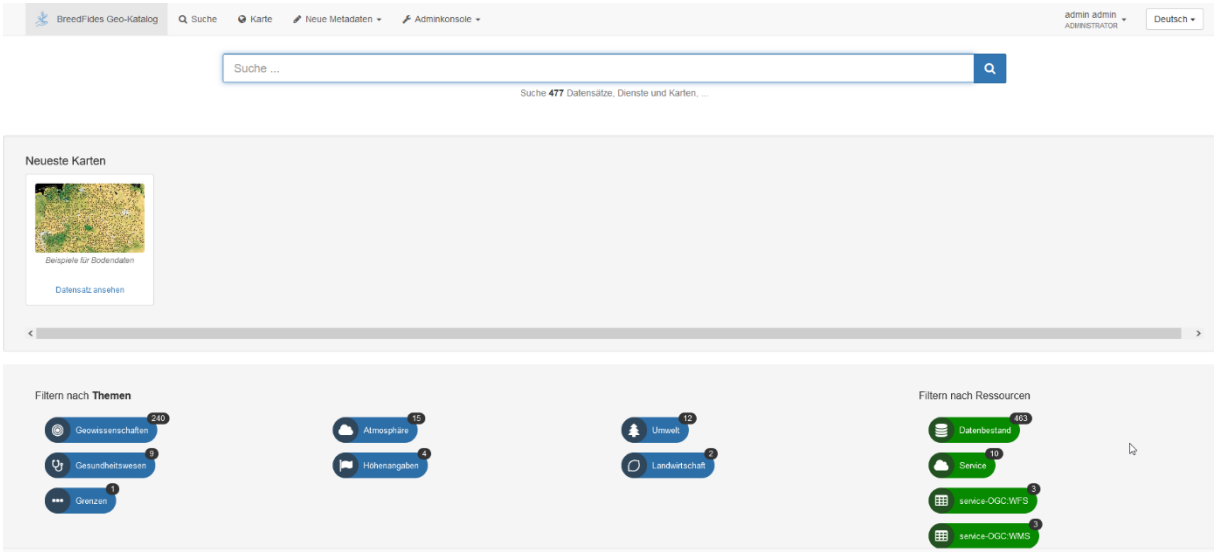

Abbildung 1: Initial evaluierter BreedFides Geo-Katalog Demo Instanz mit derzeit ~450 aus anderen Datenökosystemen abgerufenen Datensätzen. Dieser wurde im Projektverlauf als nicht handhabbar evaluiert und durch andere Komponenten ersetzt.

Es war vorgesehen, über den Geodatenkatalog Schnittstellen in die BreedFides Dateninfrastruktur aufzuzeigen und die enthaltenen Metadaten für Datenabfragen zu Umweltparametern verwendet werden können (Siehe Abbildung 2). Diese waren Teil des Meilenstein 3.2 „Schnittstellenbeschreibungen der weiteren Datendomänen“.

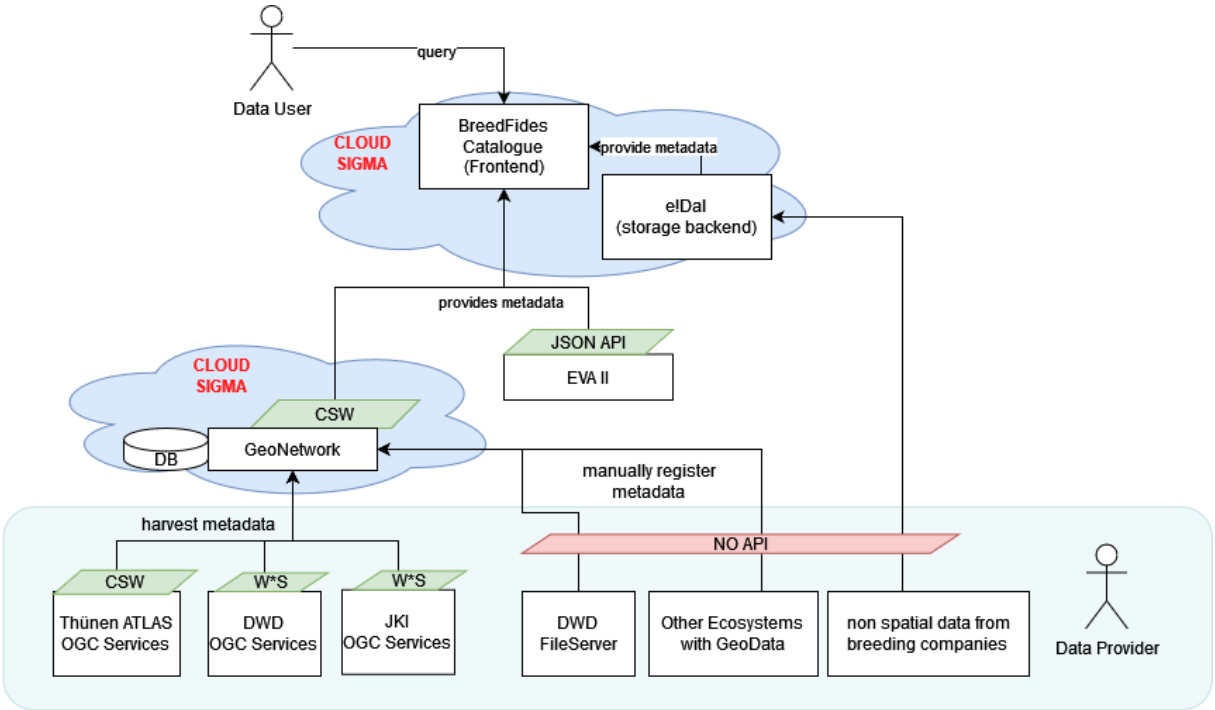

Abbildung 2: Erste Iteration der Verknüpfung mit Daten anderer Datenökosysteme über eine föderierte Infrastruktur. Im Rahmen der Evaluierung und Arbeiten an AP5 wurde diese als nicht gut handhabbar bewertet und überarbeitet.

Jedoch sind in Vorbereitung auf die Arbeiten am Use-Case (AP5) die bereitgestellten Datensätze und Metainformationen des GeoNetwork Knotens kritisch evaluiert worden. Hierbei wurde die Fülle der Informationen als zu groß und nicht handhabbar identifiziert. Das signal-to-noise Verhältnis war zu schlecht und die hohe Zahl an Umweltdatensätzen wurde auf die für den Use-Case relevante Anzahl reduziert. Der GeoNetwork Knoten wurde somit als deprecated und nicht fit-for-purpose klassifiziert (siehe Abbildung 3).

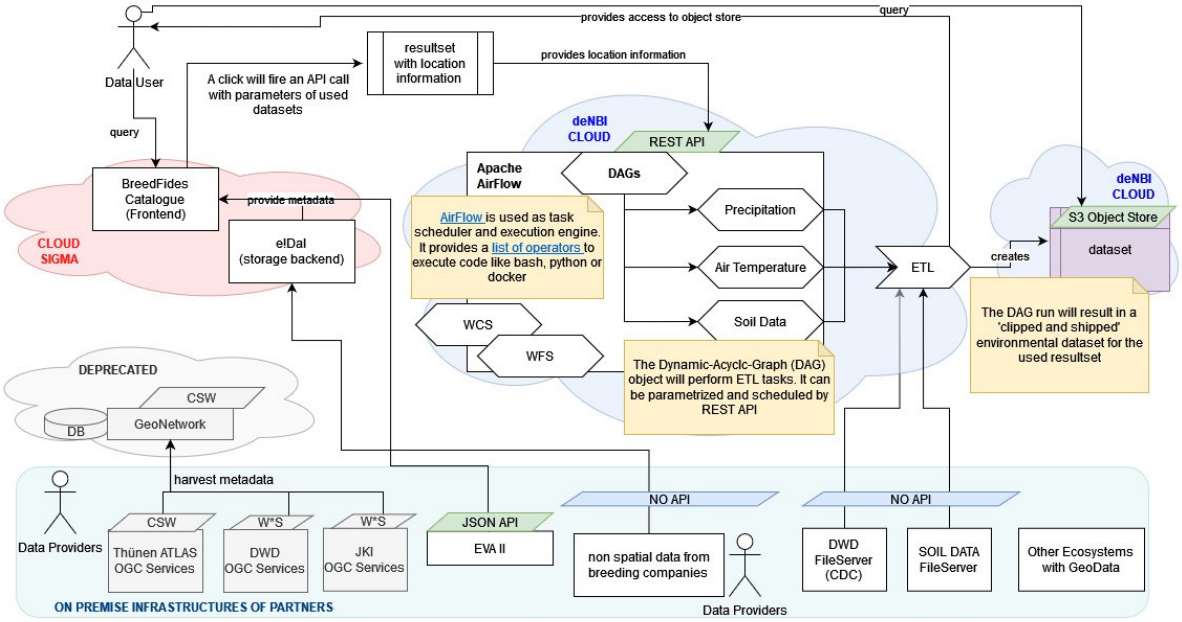

Abbildung 3: *Finale Übersicht der Datenintegrationsinfrastruktur. Neben dem BreedFides Katalog auf Cloud Sigma, ist auch der AirFlow Knoten sowie S3 Object Storage auf der de.NBI Cloud in Betrieb genommen worden. Die Nutzer:innen können über den BreedFides Katalog relevante Daten suchen und sich per Knopfdruck dazu passende, kuratierte, Umweltkovariablen aus anderen Dateninfrastrukturen bereitstellen lassen. Diese werden im S3 ObjectStore ausgeliefert. Der GeoNetwork Katalog wurde außer Betrieb genommen.*

Im Zuge der Aufgabe des GeoNetwork Kataloges sind im BreedFides-Katalog nun nur noch domänenrelevante Datensätze zu finden. Ein Benutzer kann nach relevanten Datensätzen aus der Pflanzenzüchtung suchen und diese auswählen. Da die Auswahl der Daten einen Raumbezug aufweist, entweder direkt über Koordinaten oder indirekt über Ortsnamen, kann einen Bereich von Interesse (AOI) abgeleitet werden. Dieser AOI kann verwendet werden, um die kuratierten Umweltdaten aus den Infrastrukturen der Datenanbieter auszuschneiden und zu versenden.

Für die Ausführung von solchen Extraktion-Transformation-Laden (extract-transform-load, ETL) Prozessen wurden verschiedene technische Lösungen evaluiert. Potenzielle Lösungen waren ARGO CD, Nextflow und Apache AirFlow. ARGO CD ist hierbei aufgrund der hohen Anforderungen des technischen Unterbaus (Kubernetes) nicht weiter verfolgt worden. Nextflow bietet zwar viele relevante Funktionen, aber die REST API Fähigkeit ist nicht in der open-source community edition (kostenlos) verfügbar. Jedoch sollte im Rahmen einer guten user-experience der ETL-Prozess automatisch aus der Katalogumgebung aufgerufen werden können, was eine solche REST API notwendig macht. Entsprechend wurde Apache AirFlow (im Folgenden AirFlow) als geeignetste Lösung identifiziert. AirFlow ist ein open-source Projekt der Apache Foundation und wird in vielen Unternehmen weltweit eingesetzt.

Airflow ist eine Software-Plattform, mit der *Workflows* erstellt und ausführt werden können. Ein Workflow wird in Airflow als DAG (Directed Acyclic Graph, gerichteter azyklischer Graph) dargestellt und enthält einzelne Arbeitsschritte, Tasks genannt, die unter Berücksichtigung von Abhängigkeiten und Datenflüssen angeordnet sind. Ein DAG legt die Abhängigkeiten zwischen den Aufgaben und die Reihenfolge ihrer Ausführung und Wiederholungen fest. Die Aufgaben selbst beschreiben, was zu tun ist, sei es das Abrufen von Daten, die Durchführung von Analysen, das Auslösen anderer Systeme. Ein Beispiel hierfür ist in Abbildung 4 dargestellt.

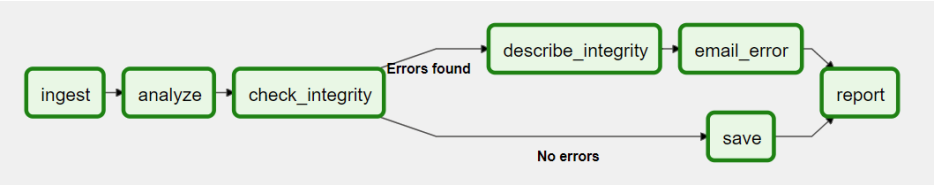

***Abbildung** 4: Beispiel für einen AirFlow DAG Workflow.*

In der Regel werden die Ausführungen von Airflow-Workflows im Voraus zeitgesteuert geplant, können aber auch über eine REST-API explizit ausgelöst werden. Zudem können Airflow-Workflows parametrisiert werden. Diese beiden Funktionen ermöglichen es der BreedFides Infrastruktur über den BreedFides-Katalog gemachte Datenauswahlen und deren AOI als Parameter an die REST API zu senden.

Durch Hindernisse in den Einstellungsprozessen von Software-Entwicklungspersonal musste das Arbeitspaket Verzug aufholen. In Absprache mit dem Projektmittelgeber wurde hierbei eine Mittelumwidmung von Personal zu Sachmittel für Dienstleistungen durchgeführt. Die Sachmittel wurden für die Vergabe eines agilen Software-Entwicklungsprojektes mit dem Rahmenvertragspartner Pixida GmbH verwendet. Dieses Projekt umfasste insgesamt sieben Sprints. Die Dokumentation der Arbeiten sind auf GitHub zu finden (Link).

Der Verzug im Arbeitspaket konnte durch die rasche Arbeit der Firma aufgeholt werden und die in Abbildung 3 gezeigte Pilot-Infrastruktur samt AirFlow ETL Knoten wurde erfolgreich aufgesetzt. Hierin inbegriffen war die Erstellung entsprechender DAGs für die ETL-Workflows zur Datenextraktion. Die Projektergebnisse sind auf Github veröffentlicht: https://github.com/breedfides/airflow-etl

Um zusätzlich Verzug aufzuholen, wurde die AirFlow Infrastruktur nicht zunächst auf der Cloud Sigma Infrastruktur als developer Instanz aufgesetzt, sondern gleich in der besser skalierbaren de.NBI Cloud und als pre-production Instanz. Die Instanz wurde über die Projektlaufzeit live betrieben und konnte nach Authentifizierung über diesen Link aufgerufen werden: https://breedfides-airflow.bi.denbi.de/login/

Die AirFlow Infrastruktur ermöglicht es uns die laufenden DAGs zu Überwachen und Fehler zu erkennen. Ein Beispiel der Administrationsseite ist in Abbildung 5 gezeigt.

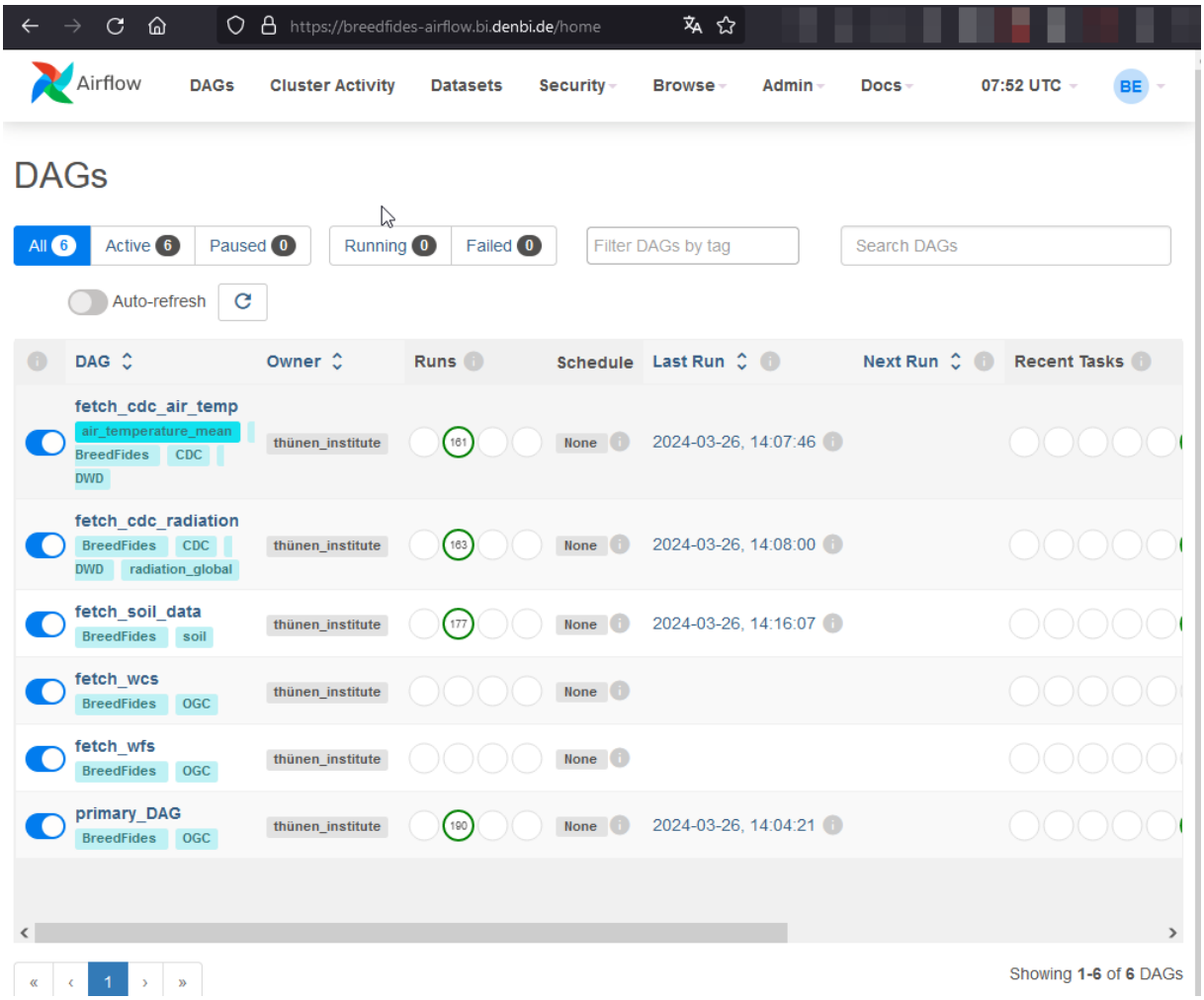

***Abbildung** 5: Administrationsseite der AirFlow Infrastruktur. Die erstellten DAGs sind aktiv und zeigen die erfolgreichen (oder gescheiterten) Läufe an.*

Im Projektverlauf hat sich gezeigt, dass in der Regel alle Umweltkovariablen für jede Auswahl an Daten des BreedFides-Kataloges verwendet werden sollen. Daher wurde ein 'Einstiegs'-DAG erstellt, welcher die anderen Prozesse intern anstößt und überwacht. Dieser Primary DAG ist in Abbildung 6 gezeigt. Eine zusätzliche Erkenntnis war, dass dieser Primary DAG, als durch mehrfache gleichzeitige Aufrufe die Infrastruktur in out-of-memory Fehler gelaufen ist, dafür verwendet werden kann die Ausführung der nachgeordneten DAGs sequenziell anzusteuern.

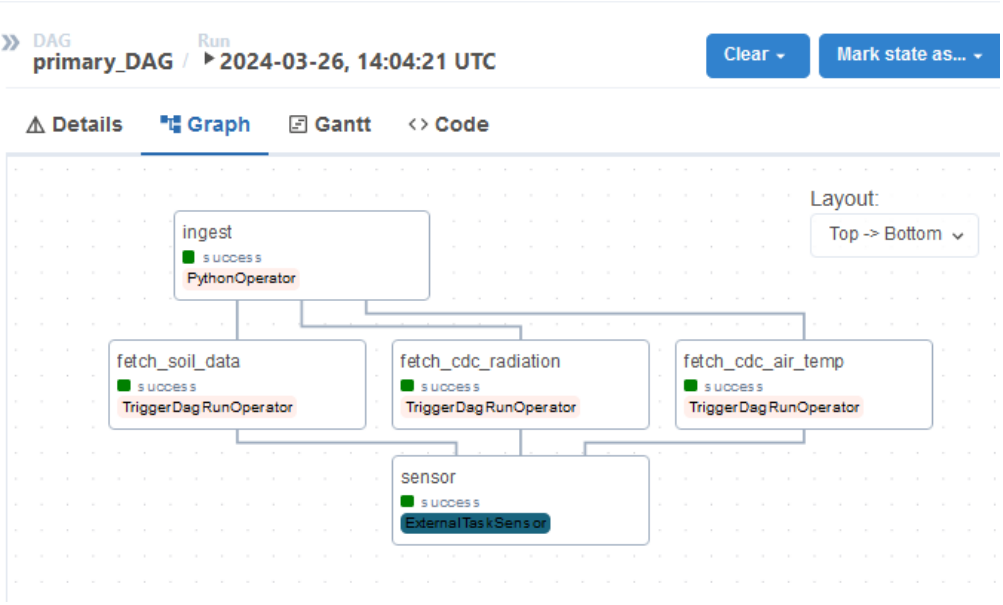

***Abbildung** 6: Flussdiagramm des Primary DAGs. Dieser wird als Einstiegspunkt für die ETL-Pipeline verwendet und triggert intern die anderen DAGs sequenziell. Dies spart Ressourcen auf kleinen Infrastrukturen und sorgt für einen reibungslosen Ablauf ohne out-of-memory Fehler. Für performantere Produktionsinfrastrukturen kann dies auf parallele Ausführung umgestellt werden. Dies verringert die Zeit des Clip-and-Ship Vorgang, zu kosten von höherer Speicher- und CPU-Last.*

Die derzeit bestehende Infrastruktur ermöglichte es uns somit den Meilenstein 3.4 “Funktionalitätsanalysen von Schnittstellen zu angrenzenden Datenökosystemen sind durchgeführt " mit nur leichter Verzögerung erfolgreich abzuschließen. Diese Analyse war Grundlage für die Erstellung der DAGs.

Im weiteren Projektverlauf konnten die Verzögerungen so aufgeholt und der letzte Meilenstein 3.5 “Test der Schnittstellen in der föderierten Dateninfrastruktur wurde durchgeführt” mit dem im Jahr 2024 geplanten Datathon fristgerecht und erfolgreich abgeschlossen werden. Hierfür wurden nochmals die bereitgestellten Datensätze kritisch evaluiert und der zuvor genutzte Bodendatensatz „Bodenübersichtskarte 1:200.000“ (BÜK200) mit Daten aus dem SoilGrids Portal (https://www.isric.org/explore/soilgrids) und dem World Soil Information System (WoSIS, https://doi.org/10.5194/essd-12-299-2020) ergänzt, um eine bessere Aussagekraft der Boden-Umweltparameter zu erhalten.

Im Rahmen des Projektabschlusses und für eine gute Nachnutzbarkeit der Ergebnisse wurde der schnittstellennutzende Anwendungsteil (AirFlow) als Docker Composition (https://docs.docker.com/compose/) veröffentlicht (siehe GitHub). Somit ist eine Nachnutzung durch Folgeprojekte möglich.

Arbeitspaket 4: Föderierte Dateninfrastruktur für eine Datentreuhänderschaft

Gegenstand des Projekts war die Konzepterstellung einer technischen Lösung für eine föderierte Dateninfrastruktur gemäß dem in AP1 und AP2 entwickelten Anforderungsprofil für ein Datenökosystem sensibler Daten aus der Pflanzenzüchtung. Dazu wurden gemeinsam mit dem Projektpartner GFPi im ersten Halbjahr 2022 6 thematische Anforderungsworkshops mit Züchtungsunternehmen durchgeführt. Für das Arbeitspaket 4 konnten so in den Workshops IT-Anforderungen, Anforderungen der Datendomäne sowie Schnittstellen in bestehende

Datenbanken verdichtet werden. Darüber hinaus wurde eine Umfeldanalyse zu relevanten Forschungsprogrammen im Bereich Datentreuhänderschaft und Infrastrukturen für sensible Daten durchgeführt. Insbesondere die Deep Dive Workshops zu den GAIA-X Föderierungsservices (GSFX) wurden genutzt, um potentiell einsetzbare Technologien und Konzepte zur föderierten Authentifizierung und Autorisierung auf Passfähigkeit zu den Anforderungen an ein Datenökosystem für die Pflanzenzüchtung hin zu untersuchen. Dafür waren die Erfahrung des Projektpartners vit zur datentreuhänderischen Dateninfrastruktur für die Rinderzucht besonders wertvoll.

Auf dieser Basis wurde die Architektur und Datenstruktur eines Datenkatalogsystems entworfen, das als migrationsfähiger Dienst in neutralen Cloudumgebungen einsetzbar ist. Ein Hauptaugenmerk war dabei, eine cloud-agnostische Infrastruktur zu entwickeln. Dies wurde in Zusammenarbeit mit AP3 umgesetzt. Es sollte eine möglichst transportable, container-basierte und an keine spezifische Serverlandschaft gebundene Technologie erarbeitet werden. Ein weiterer in den Anforderungsworkshops formulierter Aspekt war, dass alle möglichen Varianten von kollaborativen oder auch bilateralen Datennutzungsvarianten gemäß der in AP2 konzipierten Verfahren (Prozederes für die Beantragung, Prüfung und Entscheidung von Datennutzungen) agil umsetzbar sind.

Im Verlauf des Projekts wurde diese Anforderung zum Konzept eines zweiphasigen Recherche- und Datennutzungs-Systems konkretisiert. Der Kern des Konzeptes ist ein mehrphasiges Recherche- und Datennutzungs-System, das Komponenten für die Registrierung von Datensätzen, der Aushandlung von Datennutzungsverträgen für konkreten Datenkooperationen und zur Bereitstellung Anwendungsfall-spezifisch zusammengestellter Datensätze in Datenanalyseräumen umfasst. Dieses besteht aus a) einem Katalog- und Rechercheportal für Datensätze, das für alle registrierten Nutzer nach Akzeptanz der Nutzungsbedingungen nutzbar ist und b) die Bereitstellung individueller Datensätze gemäß individueller Datennutzungsverträge in einem für die Vertragsparteien exklusiven Datenanalyseraum (siehe auch AP2). Dabei werden durch den Datentreuhänder Datensätze nach Abschluss einer Datennutzungsvereinbarung zwischen den Datenprovider und den Datennutzer in einen separaten, exklusiv den Vertragsparteien zugänglichen Analyseraum geladen und die Zugänge gemäß Vertragsbedingungen bereitgestellt. Die Authentifizierung und Bindung an die Nutzungsbedingungen und Datennutzungsvereinbarung wird zweistufig über Zertifikate geregelt. Diese berechtigen den Nutzer nach Akzeptanz der BreedFides Nutzungsbedingungen zum Zugang zum Rechercheportal oder exklusiv für den Zugang zu einem vertraglichen vereinbarten Datenanalyseraum gelten.

Anschließend wurde die Implementierung eines Demonstrators der entworfenen Zwei-Ebenen-Architektur, bestehend aus der Datenkatalogumgebung und dem Datenzugriffsaushandlungsservice, durchgeführt. Dabei war eine sichere Umsetzung eines digitalen Authentifizierungs- und Autorisierungsprozesses nötig. Zur Registrierung für den Zugriff auf den Datenkatalog wird der Nutzer aufgefordert, Informationen zur eigenen Person über ein Registrierungsformular zu machen. Die Felder des Registrierungsformulars entsprechen der Struktur eines x509 Zertifikats. Nach Akzeptanz der Nutzungsbedingungen wird dieses Formular an das Backend weitergeleitet, welches ein x509 Zertifikat ausstellt und dessen Seriennummer als anonymisierten Identifikator serverseitig ablegt. Beim entworfenen Anmelde-Protokoll wird zuerst überprüft, ob es sich um ein gültiges, vom Backend ausgestelltes Zertifikat handelt. Sofern dies der Fall ist, wird ein zeitlich befristetes JSON Web Token (JWT) erstellt, welches an den Nutzer des Datenraumes übergeben wird. Dieses JWT

enthält nur minimal vertraglich notwendigen personenbezogenen Daten. Eine serverseitige Speicherung datenschutzrelevanter Daten kann so verhindern werden. Im Verlauf der Nutzung des Datenkatalogs wird dieses JWT zwingend benötigt, einerseits um neue Datensätze über einen Eingabedialog hochzuladen/zu registrieren, andererseits um Datensätze über die Recherchefunktionalität zu finden.

Der Registrierungsprozess für neue Datensätze fragt in der ersten Implementierungsstufe des Demonstrators technische Metadaten zu einem Datensatz beim Nutzer ab, wie z.B. Titel, Beschreibung, Schlagwörter oder Autoren. Diese Metadaten werden serverseitig neben den Datensatzdateien persistiert. Die Recherchefunktion des Demonstrators umfasst eine indizierte Freitextsuche, mit welcher gezielt nach Datensätzen gesucht werden kann. Vorherige Suchen können hierbei für die Dauer der Gültigkeit des JWTs im Client für die spätere Nutzung hinterlegt werden. Die Recherchefunktionalität soll in weiteren Implementierungsstufen noch um eine facettierte Suche erweitert werden.

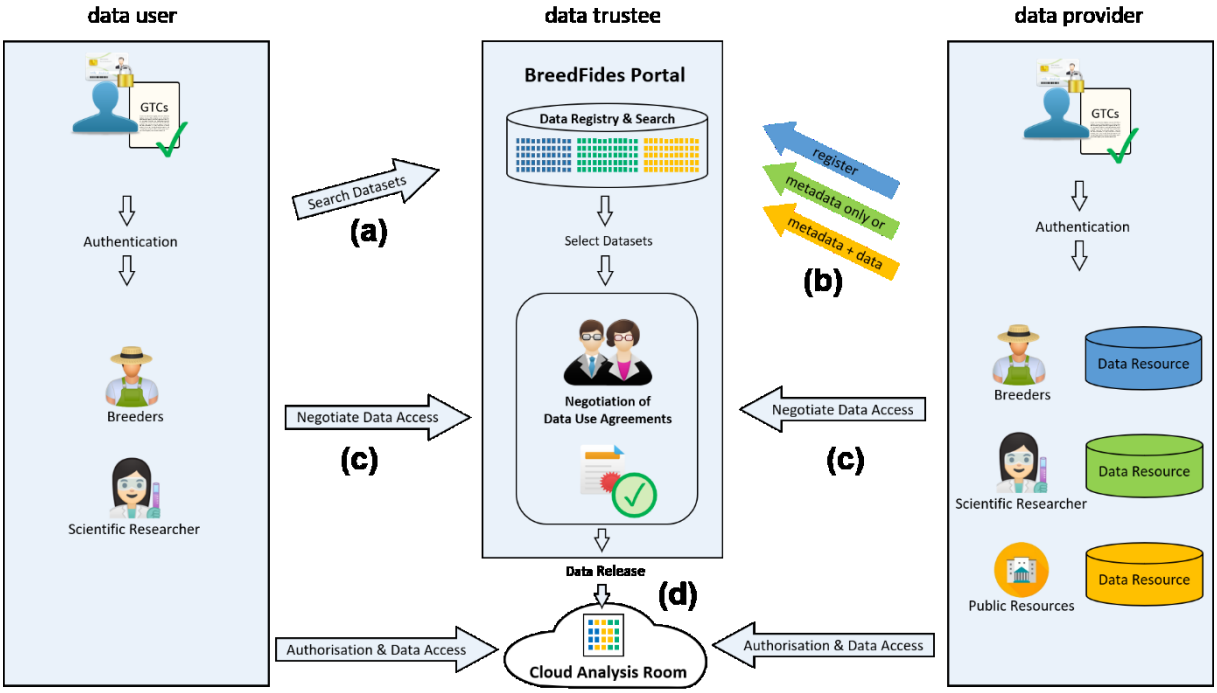

***Abbildung** 7: Gesamt-Architektur der BreedFides Datentreuhandplattform – Das BreedFides Portal bietet Datennutzern und Datenbereitstellern die Möglichkeit einer Recherche (a) von Datenbereitstellern registrierten Datensätzen (b). Für aus Suchanfragen hervorgegangenen, für den individuellen Nutzungszweck des recherchierenden Nutzers relevante Datensätze, werden, unter Vermittlung der Plattform, Datennutzungsvereinbarungen ausgehandelt (c). Nach der Ausfertigung und Abschluss von Datennutzungsverträgen können die ausgewählten Datensätze, gemäß der getroffener Nutzungsvereinbarung, in einem Cloud-basierten Datennutzungsraum (d) für die Datennutzer und Datenbereitsteller exklusiv zugänglich bereitgestellt werden.*

Zur prototypischen Umsetzung der Architektur wurde eine Evaluierung nutzbarer Softwaresystemen vorgenommen. Es wurden verschiedene technische Konzepte für

(1) den geschützten Zugriff auf die bei den Akteuren verteilten Daten,
(2) deren Katalogisierung in einer datentreuhänderisch betriebenen Katalogumgebung mit individuellen und vertraglich geregelten Recherchemöglichkeiten und
(3) den technisch nachprüfbaren und gesicherten Zugang zu ausgewählten Datensätzen in neutralen Datenräumen

untersucht.

Zu den untersuchten Konzepten zählten verschiedene Backendsysteme und Storage-Infrastrukturen, die hinsichtlich der ausgearbeiteten Anforderungskriterien und notwendigen Bedarfen evaluiert wurden. Es wurden u.a. die Bewertung von Umsetzungsmöglichkeiten eines Zertifikats-basierten Nutzerregistrierungs- und Authentifizierungs-Protokolls, die Erfassung von technischen Metadaten und die Verfügbarkeit von geeigneten Suchfunktionalitäten bewertet. Außerdem wurde der Aufwand für eine notwendige Anpassung der jeweiligen Plattformen an die spezifischen Bedarfe der Datentreuhandinfrastruktur abgewogen. Aufgrund von mehreren, verschiedenen konzeptionellen Vorteilen, wie dem flexibel erweiterbaren Sicherheitssystem, der Möglichkeit komplexe Datensätze inklusive technischer Metadaten zu erfassen, einer umfangreichen Volltextsuche und der, beim Projektpartner IPK verfügbaren, technischen Expertise wurde die e!DAL Infrastruktursoftware (Arend et al. GigaScience, 2022) als geeignetste Lösung für das Backend der Dateninfrastruktur selektiert.

Im nächsten Schritt wurden mögliche Konzepte und Technologien für die Vorlagen-basierte Erstellung von Datennutzungsvereinbarungen und deren Abbildung auf digital bearbeitbaren Datenstrukturen untersucht. Dazu wurde die Open Digital Rights Language (ODRL) untersucht. ODRL ist ein Standard bzw. eine formale Sprache, um Richtlinien für den Umgang mit Dateninhalten und Diensten zu formulieren und wurde u.a. für die Anforderungen im DRM (Digital Right Management) für digitale Medieninhalten entworfen und wird sowohl in weiteren Projekten der Datentreuhandmodelle als auch in der Nationalen Forschungsdateninfrastruktur (NFDI) verwendet. Diese Technologie ermöglicht es, im Datentreuhänder-Portal erstellte Datennutzungsvereinbarungen persistent zu speichern, sowie auf Basis vordefinierter Templates konkrete Vertragsdokumente zu generieren.

Im Folgenden wurde mit der Implementierung eines Prototyps der BreedFides-Datentreuhänderplattform begonnen, um Erfahrungen in Umsetzung, Akzeptanz bei den Züchtern und Integration den Zuarbeiten aus AP 1,2,3 und die Anwendung in AP5 zu sammeln und mögliche Synergien mit verwandten Projekten, wie z.B. NFDI oder GAIA-X, zu untersuchen. Konkret wurde u.a. sowohl für die Registrierung neuer Datensätze durch einen Datenbereitsteller als auch für das Aushandeln der spezifischen Datennutzungsvereinbarungen zwischen involvierten Parteien eine Wizard-Komponente vorgesehen. Mit Hilfe der von der Firma Pro-Corn über AP5 bereitgestellten Beispieldatensätze für die Bundessortenversuche sowie des von AP1 und AP5 erarbeiteten Metadatenschemas konnten semantisch und technisch relevante Metadaten übergreifend über alle potenziellen Datenbereitsteller harmonisiert und bei der Umsetzung des ersten Implementierungsentwurfs der Wizard-Komponente berücksichtigt werden.

Die erste Implementierungsphase (Meilenstein 4.4) wurde in Form eines Mock-Ups zur Demonstration des Gesamtkonzepts durchgeführt, wobei die Ergebnisse aus den AP1, AP2, AP3 integriert werden konnten. Der Mock-Up konnte so zum Test des Gesamtkonzepts mithilfe des Use-Cases aus AP5 genutzt werden. Konkret wurde das spezifizierte Metadatenschema und das abgestimmte Datenformat für phänotypische Züchtungsdaten und Umweltdaten als Basis für die Datenregistrierung im Portal verwendet (AP1). Darüber hinaus konnte der rechtliche Rahmen für Datennutzungsverträge und die allgemeinen Nutzungsbedingungen als konzeptionelle Grundlage für das Datenrechercheportal und den Daten-Zugang verwendet werden (AP2). Für das Laden von öffentlichen Datenquellen wurden Cloud-basierte Workflow-

Dienste (AP3) genutzt. In Rückkopplung mit den Partnern GFPi, JKI und Thünen Institut wurden alle beschriebenen Lösungen iterativ getestet und schrittweise verfeinert. Neben dem kontinuierlichen Austausch wurde hierzu am JKI am 10.05.2023 ein Workshop zu Interfaces & Standards und am IPK am 4.10.2023 ein Projekttreffen durchgeführt.

Die agile Methode dieser ersten Phase der Implementierung mündete in der Niederschrift einer Implementierungsspezifikation, die in der zweiten Implementierungsphase als Basis für die Umsetzung einer produktiven Lösung genutzt werden kann. Diese Spezifikation wurde zur Prüfung wirtschaftlicher Verwertbarkeit als Erfindungsmeldung am 22.12.2023 am IPK eingereicht.

Die Implementierungsspezifikation war die Basis für die Implementierung einer Testinstanz in einer Cloud-Umgebung mit Zugriffsmöglichkeit für Kooperationspartner. Hier wurden zusätzlich zum Prototypen aus dem zweiten Projektjahr die im Folgenden beschriebenen technischen Fähigkeiten der föderierten Dateninfrastruktur als Proof-of-Concepts gemäß Meilenstein 4.3 umgesetzt und mit Daten aus den in AP5 entwickelten Anwendungsfällen zum Test durch die Projektpartner und Präsentation in Workshops versehen.

Als erstes wurde dabei der beschriebene Prozess zur Registrierung von Datennutzern realisiert und eine Maske (Abb.7) zur Eingabe der benötigen Attribute für die Erstellung eines X-509 Zertifikates abgefragt. Diese Angaben wurden anschließend im Serverbackend registriert und persistent abgelegt. Um diesen initialen Registrierungsprozess abzuschließen, muss jeder Nutzer außerdem die im Projekt erarbeiteten Nutzungsbedingungen des Systems akzeptieren.

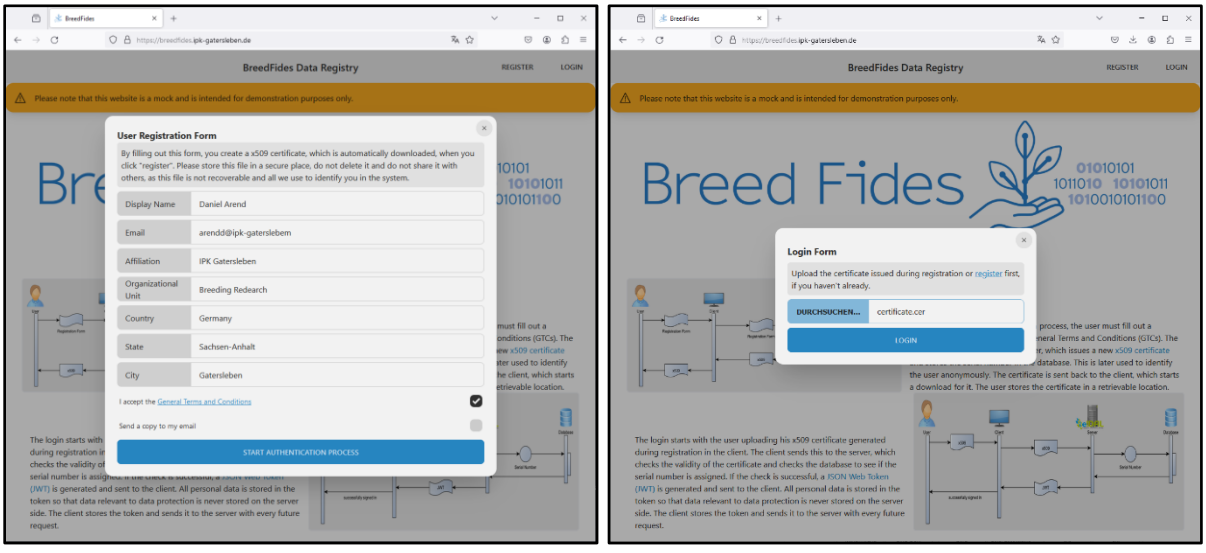

Abbildung 8: Benutzerregistrierung und Zertifikatübergabe in der Testinstanz des BreedFides Portal. Für den Zugang zum BreedFides Portal werden von jedem Nutzer wenige, persönliche Angaben zur Erstellung eines X509-Zertifikates benötigt. Dieses wird anschließend dem Nutzer als Download ausgehändigt und außerdem serverseitig persistiert. Anschließen kann jeder Nutzer das Zertifikat als persönlichen Zugangsschlüssel nutzen, um sich für den Zugang zum Portal zu authentifizieren

Im nächsten Schritt wurde die für die Registrierung neuer Datensätze beschriebene Wizard-Komponente zur Metadaten-Erfassung und Annotierung von Datensätzen implementiert. Diese sollte nicht nur die erarbeiteten Metadatenattribute abfragen, sondern auch eine nutzerfreundliche Eingabe gewährleisten, um die Hürde für die Dateneigentümer zu senken

und eine einfache, aber qualitative Datenbeschreibung sicherzustellen. Dies wurde durch eine schrittweise Erfassung der notwendigen Daten realisiert, die Formular-basierte Eingabeschritte mit einem vorlagenbasierten Uploadprozess kombiniert (Abb 8.)

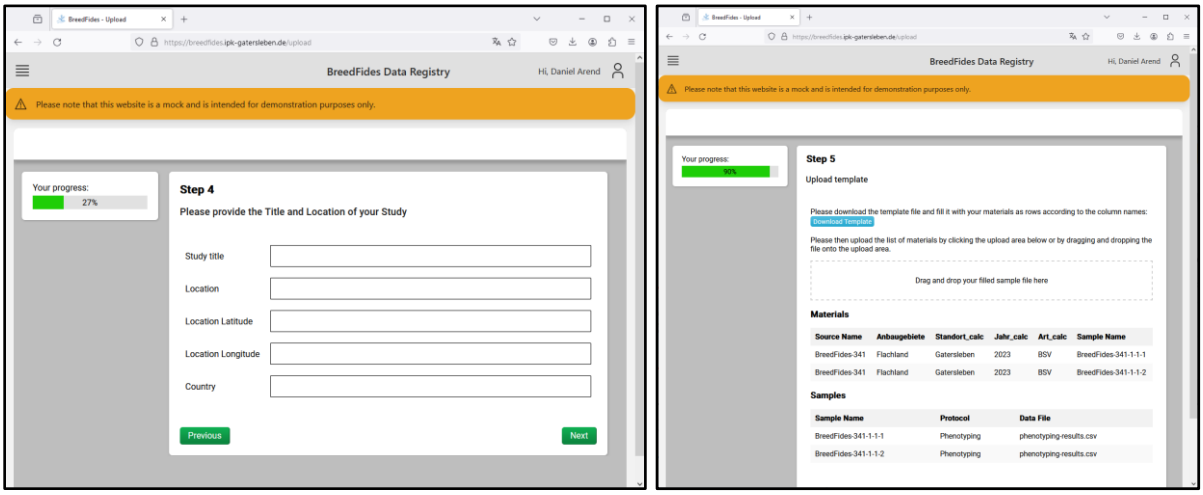

Abbildung 9: Datenregistrierungs-Wizard der Testinstanz des BreedFides Portal. Für die Erfassung von Datensätzen im BreedFides Portal wird dem Datenprovider ein mehrstufiger Prozess in Form eines Wizards bereitgestellt. Die erfassten Metadaten basieren auf den vorab ausgearbeiteten und abgestimmten Attributen sowie die im Rahmen des UseCases bereitgestellten Beispieldaten.

Anschließend können die registrierten Datensätze im BreedFides Portal anhand der hinterlegten Metadaten über eine bereitgestellte Suchfunktion (Abb 9.) gefunden werden. Die Suchfunktion bietet verschiedene grundlegende Filtermöglichkeiten und wurde prototypisch implementiert. Die jeweiligen Filter und Kategorien leiten sind vom Metadatenschema ab. Des Weiteren bietet die Suche die Möglichkeit, gefundene Datensätze zu selektieren und das Aushandeln eines Datennutzungsvertrages (DUA = Data Use Agreement) mit den Datenprovider zu initiieren.

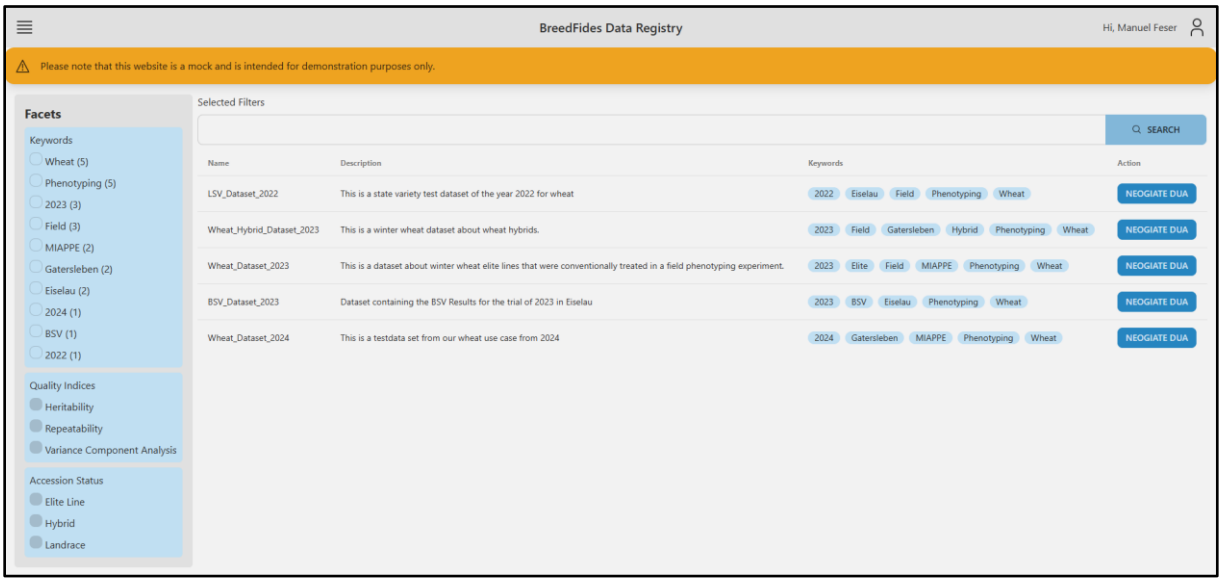

Abbildung 10: Suchoberfläche der Testinstanz des BreedFides Portal. Neben der Freitextsuche (oben), können über eine facettiere Suche (links) alle Datensätze nach bestimmen Kategorien gefiltert werden. Über den „Negotiate DUA“ Button (rechts) können Datennutzer für ausgewählte Datensätze der Prozess zur Aushandlung eines Data Use Agreements gestartet werden, um so Zugang zu den gewählten Daten zu erlangen.

Im nächsten Implementierungsschritt wurde eine weitere Wizard-ähnliche Komponente zur Erstellung eines geeigneten Datennutzungsvertrages auf Basis des Vertragstemplates aus AP2 entwickelt. Diese ermöglichte ein einfaches Definieren der gewünschten Rahmenbedingungen für die Datennutzung über eine schrittweise Abfrage und Auswahl von einfachen Kriterien in verständlicher Sprache durch den Datennutzer. Neben festen Eingabefeldern, um beispielsweise Kontaktinformationen einzugeben, bietet der Dialog eine Reihe von Auswahlfeldern, über die der Nutzer definierte Optionen selektieren kann. Außerdem können zusätzliche Dokumente als Anlage ergänzt werden.

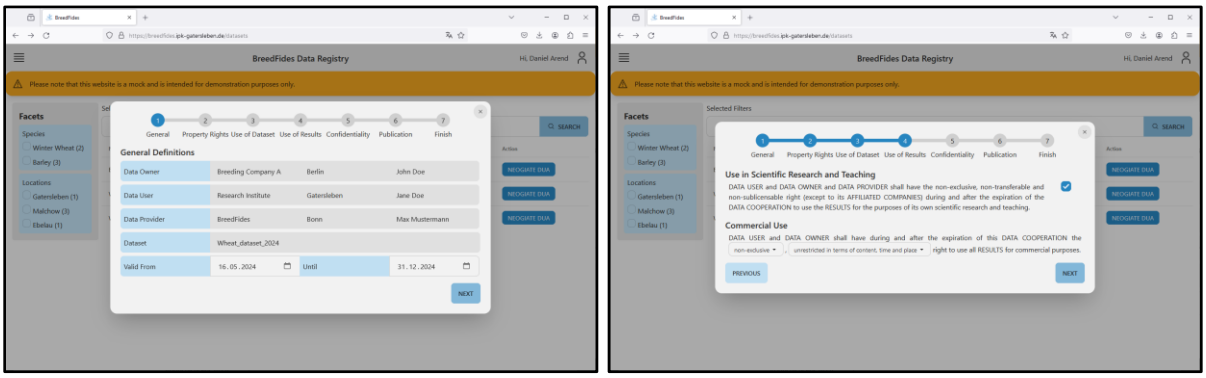

Abbildung 11: Data Use Agreement Wizard der Testinstanz des BreedFides Portals. Jede Anfrage auf einen Datenzugriff initiiert einen mehrstufigen Prozess, indem der Datennutzer seine Nutzungsinteressen definieren kann.

Anschließend werden die gewählten Optionen mit Hilfe der definierten Templates über eine Template-Engine in vollständige Datennutzungsverträge mit den entsprechenden juristischen Formulierungen übersetzt. Die gewählte Option wird als Textbaustein im dazugehörigen Abschnitt der Datennutzungsvereinbarung hinterlegt (Abb. 10). Das erstellte Dokument wird dem Nutzer über eine Vorschauansicht dargestellt und kann anschließend als PDF-Dokument abgespeichert werden (Abb. 11). Außerdem wird die dazugehörige Datennutzungsanfrage im Profil des Datennutzers hinterlegt, solange sie sich in der anschließenden Prüfung durch den Dateneigentümer befindet. Der Dateneigentümer wird über die eingegangene Nutzungsanfrage informiert und erhält eine Kopie des vom anfragenden User vorgeschlagenen Datennutzungsvertrages mit der Bitte um Prüfung.

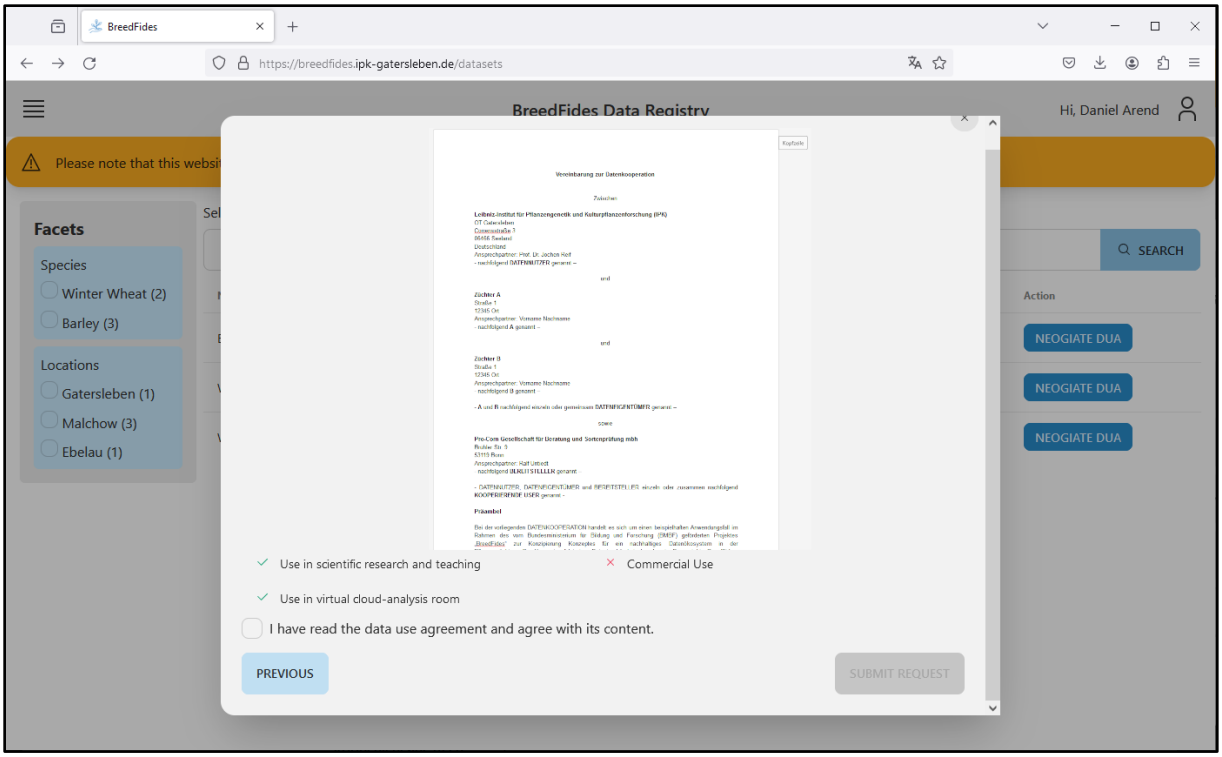

Abbildung 12: Vorschaudialog am Ende des Data Use Agreement Wizard der Testinstanz des BreedFides Portals. Jeder Datennutzer erhält eine Zusammenfassung der von ihm gewählten Datennutzungskriterien, sowie eine Vorschau auf den fertig erstellten Datennutzungsvertrag. Anschließend kann der Datennutzungsvertrag absenden werden, wodurch der Datenprovider informiert wird und seine Prüfung beginnt.

Für den Fall, dass die Prüfung des Datennutzungsvertrages erfolgreich war, steht der Datensatz dem anfragenden Datennutzer entsprechen den ausgehandelten Bedingungen zur Verfügung. Dabei stehen ja nach Vereinbarung verschiedene Möglichkeiten des Datenzugangs bereit. Neben dem einfachen Download der Daten, können diese auch in einen geschützten Analyseraum übertragen oder mittels des in AP3 entwickelten ETL-Prozesses z.B. mit Klimadaten aus öffentlichen Datenräume verschnitten werden.

Diese Komponenten wurden durch Workshops mit den GFPi-Mitgliedern, Projektpartnern und Rückkopplung aus den Workshops mit der Begleitforschung in 2024 verfeinert. Dazu gehörte u.a. die GFPi Sommertagung am 16.05.2024, der Workshop mit GFPi-Mitgliedern am 03.12.2024, der Begleitforschung am 13.5.2024 und der DTM Vernetzungsveranstaltung am 11.9.2024.

Arbeitspaket 5: Test der föderierten Dateninfrastruktur anhand von Use Cases

Im Rahmen der Use-Case Entwicklung und des Datathons am 15.07.2024 bei Partner GFPi in Bonn konnten die in AP3 identifizierten Datenschnittstellen kritisch evaluiert werden. Die erste Iteration der schnittstellennutzenden Applikation konnte beim Workshop „Standards und Schnittstellen“ am 05.10.2023 bei Partner JKI in Quedlinburg vorgestellt werden. Hierbei hat sich gezeigt, dass der Fokus eher auf wenige und dafür qualitativ hochwertige Datenschnittstellen gelegt werden sollte. Im Zuge dessen wurde die erste Implementierung überarbeitet und mit einer kleinen Auswahl von wichtigen Datenschnittstellen zu Umweltkovariablen realisiert. Im BreedFides Datathon konnte der automatisierte Datenabruf über die Schnittstellen vorgeführt werden. Abermals konnten die Anforderungen der

Nutzenden mit der Implementierung abgeglichen und die Implementierung nachgeschärft werden. Hierbei wurde der Bodendatensatz als nicht ausreichend aussagekräftig identifiziert. Daher wurde ein neuer Bodendatensatz (WoSIS, soilgrids) über eine Schnittstelle in die ETL-Pipeline integriert und die bereitgestellten Umwelt-Datensätze so ergänzt. Die so optimierte Infrastruktur und der Workflow für Datennutzende wurde in der BreedFides Abschlussveranstaltung vorgestellt und erhielt positive Rückmeldungen.

In Online-Workshops, die von der GFPi moderiert wurden, wurde ein Fallbeispiel entwickelt, das dazu diente, die Eckpfeiler des föderierten Datenökosystems in enger Interaktion mit verschiedenen Züchtungsunternehmen zu testen. Als Grundlage dafür wurden die vollumfänglichen phänotypischen Daten des „Bundessortenversuchen Winterweizen“ (BSV) gewählt, die um zusätzliche externe Daten, welche die Versuchsstandorte charakterisieren, ergänzt wurden. Das Fallbeispiel reflektiert in ausgezeichneter Weise die Größe und Heterogenität der durch verschiedene Pflanzenzüchtungsunternehmen erhobenen Daten.

Auch für die Erprobung der im Projekt erarbeiteten „Datennutzungsvereinbarung“ war der BSV-Datensatz geeignet, da die Daten von der Gesamtheit der in Deutschland ansässigen Weizenzüchtungsunternehmen als Dateneigentümer erhoben werden. Durch den Abschluss einer einzigen Datennutzungsvereinbarung konnte so ein Maximum an Unternehmen eingebunden werden. Nach Abschluss der Vereinbarung erfolgte der Datentransfer. Die Daten wurden unter Berücksichtigung der entwickelten Datenqualitätsindizes kuratiert, wobei an allen Standorten sehr hohe Wiederholbarkeiten verzeichnet wurden. Im Anschluss daran wurde die entwickelte ETL-Pipeline aus AP3 genutzt, um die Informationsdichte für die einzelnen Umwelten (Wetter- und Bodendaten) zu erhöhen. Vor der Analyse über Umwelten hinweg wurde die Konnektivität der Prüfglieder zwischen den Einzelumwelten überprüft. Diese war ausreichend hoch (Abbildung 13) und ermöglichte integrierte Analysen.

Abbildung 13: Anzahl überlappender Genotypen über die Jahre hinweg.

Die Analyse der Varianzkomponenten unterstrich die herausragende Qualität der phänotypischen Daten mit einer Heritabilität im weiteren Sinne von 97 % für den Kornertrag. Die Varianz der Interaktion zwischen Genotypen und Orten wies mit 7,3 % eine größere Größe auf als die Varianz der Interaktion zwischen Genotypen und Jahren (vgl. Abbildung 14).

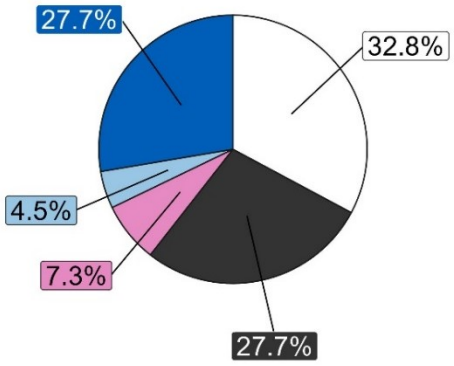

Abbildung 14: Verhältnis der Varianz der Genotypen (dunkelblau), deren Interaktion mit Jahren (hellblau), Orten (rosa) sowie Jahren und Orten (schwarz), und des Restfehlers (weiß) für den Kornertrag.

Für die integrierte Analyse wurde ein linear gemischtes Modell verwendet, bei dem die Effekte für die Faktoren Jahr, Ort, Genotyp und Wiederholung sowie deren Interaktionseffekte geschätzt wurden. Die Effekte der Interaktionen zwischen Genotyp und Jahr wurden anschließend genutzt, um euklidische Distanzen zwischen den Orten zu schätzen. Diese euklidischen Distanzen dienten wiederum als Grundlage für eine Clusteranalyse. Die Daten

weisen auf zwei divergierende Gruppen im Hinblick auf den Kornertrag hin (Abbildung 15). Die Resultate unserer Analysen wurden den beteiligten Pflanzenzüchtungsunternehmen präsentiert.

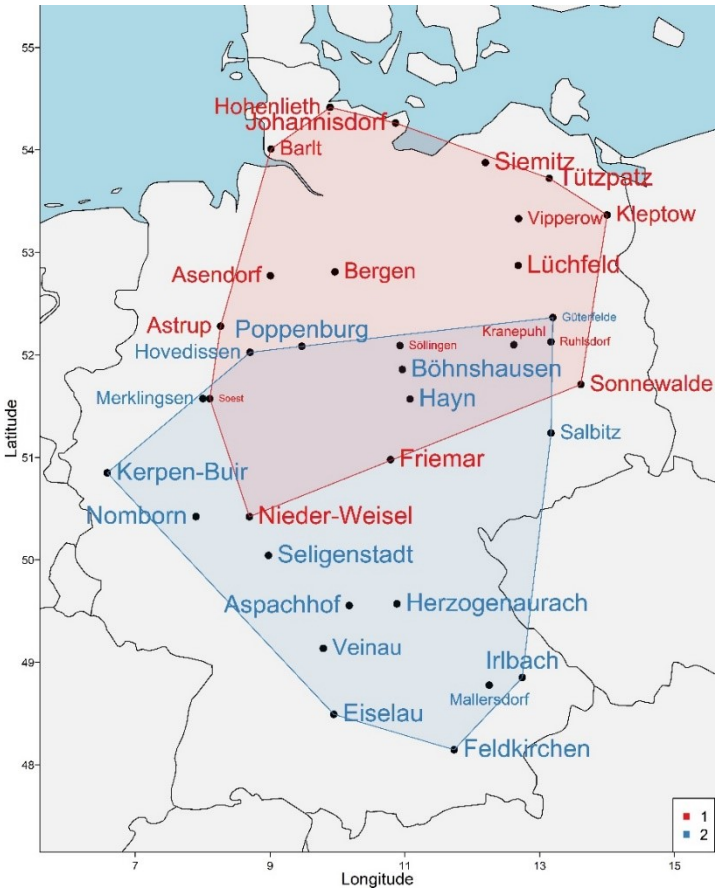

Abbildung 15: Clusteranalyse aufbauend auf den Interaktionseffekten zwischen Genotypen und Umwelten.

Arbeitspaket 6: Entwicklung eines nachhaltigen Betriebskonzepts und Einbindung der Branche

Die Diskussionen und Arbeiten zum Thema des AP6 „Entwicklung eines nachhaltigen Betriebskonzepts und Einbindung der Branche“ wurden über den gesamten Projektzeitraum durchgeführt. Durch eine Vielzahl von Gesprächen mit unterschiedlichen Stakeholdern, Interessierten und Beteiligten des Datenökosystems konnten daraus einige wichtige Rahmenbedingungen für den zukünftigen Betrieb des BreedFides Ökosystems abgeleitet werden:

Eine Weiterführung des Konzeptes wird von allen Projektpartnern als sinnvoll und erfolgversprechend erachtet. Für eine nachhaltige Etablierung des Systems erachten die Beteiligten ein organisches Wachstum der Anzahl der Nutzer und Datenkooperationen als notwendig und zielführend. Damit einhergehen soll eine sukzessive Steigerung des Vertrauens der Nutzenden in das System. Um dies zu erreichen, wird eine größtmögliche Neutralität des Systembetreibers gefordert, die durch möglichst wenig Eigeninteressen gewährleistet werden kann. Darüber hinaus ist eine finanzielle Absicherung des Betriebes erforderlich, die möglichst unabhängig von öffentlichen Zuwendungen und befristeten Verträgen ist.

Diese Voraussetzungen können nach derzeitigem Kenntnisstand am besten durch den Betrieb des Systems im Umfeld des oder durch den Branchenverband der Pflanzenzüchtungsunternehmen in Deutschland erfüllt werden. Der Verband vertritt die Gesamtinteressen seiner Mitglieder, ist diesen aber gleichermaßen verpflichtet. Entsprechend vertritt er keine Individualinteressen einzelner Unternehmen und handelt nach dem Gleichbehandlungsprinzip. Darüber hinaus deckt sich der satzungsgemäße Zweck der GFPi e. V. als „Forschungsverband“ der in Deutschland tätigen Pflanzenzüchtungsunternehmen „die Förderung von Pflanzenwissenschaft und –forschung mit dem Ziel, Pflanzen vor allem züchterisch zu verbessern“ mit dem Zweck des konzipierten Datenökosystems.

Der Betrieb soll sich zunächst auf den Kern des Datenökosystems, d.h. den Betrieb und die Pflege des BreedFides-Portals konzentrieren. Analysedienste und Datenkuration sind nicht als vorrangiges Dienstleistungsangebot des Betreibers vorgesehen und sollen stattdessen von den Systemteilnehmern übernommen werden. So kann z.B. die BreedFides-Partnerin IPK ihr Know-how in Form von Dienstleistungen den Pflanzenzüchtungsunternehmen zur Verfügung stellen.

Das Konsortium geht davon aus, dass die Komponente „niedrige Einstiegsschwellen“ für die Etablierung des BreedFides Systems entscheidend ist. Dazu gehören vor allem eine durchdachte und intuitive Benutzeroberfläche sowie möglichst geringe Kosten für die Nutzerinnen und Nutzer.

Die in AP6 gesammelten Erkenntnisse werden von den Projektpartnern GFPi und IPK in das BMBF-geförderte Projekt “DRIVE” eingebracht. Hier wird GFPi das in BreedFides konzipierte System praktisch etablieren und als Systembetreiber fungieren.

Arbeitspaket 7: Rechte- und Datenmanagement

Die Bereitstellung einer Cloudumgebung ist durch die Beauftragung beim Anbieter SigmaCloud erfolgt. Weiterhin wurden die geforderten Dokumente zur Erfüllung der Auflage „Dokumente zur Nutzung der externen, kommerziellen Cloud-Umgebung“ eingereicht. Die Dokumente im Einzelnen waren ein Verzeichnis der geplanten datenschutzrechtlich relevanten Verarbeitungstätigkeiten (vtt), eine Freigabe durch den Datenschutzbeauftragen des IPK sowie der Nachweis eines IT-Sicherheitskonzeptes. Die Cloud-Instanz wurde für die im Arbeitspaket 4 beschriebenen Aktivitäten eingesetzt. Bei der Entwicklung wurde auf größte Portabilität des Containers geachtet.

Arbeitspaket 8: Projektkoordination

Im Berichtszeitraum wurden monatliche Online-Meetings durchgeführt, um den Fortschritt des Projekts nachzuhalten und die Zusammenarbeit der Partner kontinuierlich zu verbessern. Zusätzlich zu diesen Treffen wurden Vorträge zu spezifischen Themen präsentiert, die für den Fortschritt des Projekts wichtig sind. Außerdem gab es folgende persönliche Treffen bei den Projektpartnern:

- 12. und 13.07.2022, vit, Verden

- 04.10.2023, IPK, Gatersleben
- 05.10.2023, JKI, Quedlinburg
- 15. und 16.07.2024, GFPi, Bonn
- 29. und 30.10.2024, Thünen Institut, Braunschweig

Im Rahmen von Veranstaltungen der DTM-Förderinitiative und in Workshops mit der Begleitforschung wurden Beiträge zu Konsortien übergreifende Ableitung von Verfahren und Lösungsansätzen in Abstimmung mit den Projektpartnern geleistet und Ergebnisse ins Konsortium gespiegelt.

2. Wichtigsten Positionen des zahlenmäßigen Nachweises

3. Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

4. Voraussichtlicher Nutzen, insbesondere die Verwertbarkeit des Ergebnisses auch konkrete Planungen für die nähere Zukunft im Sinne des fortgeschriebenen Verwertungsplans

Es wurde am 22.12.2023 eine Erfindungsmeldung am IPK mit dem Titel „Konzept und Spezifikation von technischen Diensten zum Aufbau geschützter, datentreuhänderisch betriebener Datenräume zum kollaborativen Arbeiten mit sensiblen Daten aus der Pflanzenzucht“ eingereicht. Diese wird zum Zeitpunkt der Berichtserstellung von der IPK-Administration geprüft.

Im Rahmen der Ausschreibung BAnz AT 26.10.2023 B4 „Moderne Züchtungsforschung für klima- und standortangepasste Nutzpflanzen von morgen“ wurde der Projektantrag „DRIVE“ (FKZ: FKZ031B1537A) ausgearbeitet, in den nach Abschluss des BreedFides-Projektes Teile des entwickelten Datentreuhand-Konzeptes einfließen und in einem übergeordneten und produktiven Datenaustauschportal genutzt werden sollen. Dieser wird seit dem 1.11.2024 gefördert.

5. Während der Durchführung des Vorhabens dem Zuwendungsempfänger bekannt gewordenen Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen (auf Begleitforschung eingehen)

In Rahmen von Workshops mit dem von der EU Kommission geförderten Gaia-X Programm und der gemeinsamen Beteiligung der Antragssteller an Konsortien aus dem Förderprogramm „Nationale Forschungsdateninfrastruktur“ (NFDI) sowie der Beteiligung als Servicezentrum des Deutschen Knotens der Europäischen Bioinformatik-Infrastruktur (ELIXIR) konnten Aspekte zu technischen Infrastrukturen und Konzepten in das BreedFides-Projekt einfließen. Die geplante Kooperation mit dem Projekt FAIR Data Spaces wird nicht weiterverfolgt. In einem Treffen am 04.12.2024 mit den FAIR-Data Spaces Projektpartner Uni Gießen wurde der aktuelle Stand des BreedFides-Projektes vorgestellt. Hinsichtlich der Abläufe zur Aushandlung von Datennutzungsverträgen gibt es große Übereinstimmungen. Dies demonstriert, dass die entwickelten Konzepte die generellen Anforderungen sehr gut widerspiegeln. Aus der Domäne Pflanzenzüchtung ergeben sich aber zusätzliche spezifische

Anforderungen, welche die Entwicklung und Nutzung einer BreedFides spezifischen Cloud-Infrastruktur notwendig machen.

6. Erfolgte oder geplante Veröffentlichungen des Ergebnisses nach Nr. 5 der NKBF/NABF

Publikationen

Poster “Development and practical testing of data trustee model for plant breeding” ELIXIR All Hands 2023, Dublin Irland

Veranstaltungen

- 28.02.2022 Anforderungsworkshops mit Züchter und Projektpartnern
- 07.03.2022 Anforderungsworkshops mit Züchter und Projektpartnern
- 10.03.2022 Treffen mit dem Bundessortenamt zu BreedFides
- 11.03.2022 Anforderungsworkshops mit Züchter und Projektpartnern
- 23.03.2022 GAIA-X DeepDive Identity Trust
- 07.04.2022 GAIA-X DeepDive Federated Catalogue
- 16.05.2022 Anforderungsworkshops mit Züchter und Projektpartnern
- 07.06.2022 GAIA-X DeepDive Federation Service
- 09.06.2022 GAIA-X DeepDive Compliance
- 20.06.2022 Anforderungsworkshops mit Züchter und Projektpartnern
- 23.06.2022 DTM-Auftaktveranstaltung
- 7./8.09.2022 GSFX Connect
- 20.10.2022 Workshop mit Züchtern „Use-Cases”
- 30.03.2023 Schnittstellen EVA2 (virtuell)
- 13.04.2023 Schnittstellen EVA2 (virtuell)
- 06.06.2023 ELIXIR All Hands Meeting, Dublin, Irland
- 12./13.07.2023 Projekttreffen bei Projektpartner VIT in Verden
- 06.09.2023 DTM_Vernetzungsveranstaltung (in Person in Berlin)
- 04.10.2023 BreedFides Projekttreffen (in Person am IPK) am
- 05.10.2023 Workshop „Standards und Schnittstellen“ (in Person beim Partner JKI)
- 22.11.2023 Verzahnung von Datennutzungsvereinbarung/Data Use Agreements und IT-Infrastruktur am
- 04.12.2023 BreedFides Meets FAIR Data Spaces (virtuell)
- 06.05.2024 GFPi Sommertagung (in Person beim Partner IPK in Gatersleben)
- 13.05.2024 Fachgruppenworkshop mit der DTM-Begleitforschung (virtuell)
- 15.07.2024 BreedFides Datathon (in Person beim Partner GFPi in Bonn)
- 29.10.2024 Projektabschlussmeeting (in Person beim Partner Thünen-Institut in Braunschweig)
- 11.09.2024 DTM_Vernetzungsveranstaltung (in Person bei DTM-Konsortium in Berlin)
- 03.12.2024 Vorstellung Ergebnisse für GFPi-Mitglieder (Webinar)