Innovative Datenkreis Infrastruktur (DI)

Konsortium

- Know-Center GmbH (KC) (Prof. Stefanie Lindstaedt)
- Zentralanstalt für Meteorologie und Geodynamik (ZAMG) (Martin Saini)
- Karl-Franzens Universität (KFU), Profilbildender Bereich Smart Regulation (Prof. Matthias Wendland)

Ausgangslage

Big Data, der Umgang mit extrem großen Datenmengen, transformiert viele Bereiche von Forschung, Wirtschaft und Gesellschaft. In der Produktion generieren immer mehr Sensoren immer größere Datenmengen. In Handel und Dienstleistung speisen wir alle mit unseren digitalen Identitäten riesige Datenseen. Unsere dringenden gesellschaftlichen und globalen Herausforderungen, etwa in Klimaschutz und Medizin, lassen sich nur unter Verwendung datengetriebener Ansätze lösen. Der kompetente Umgang mit großen Datenmengen wird also zum zentralen Erfolgsfaktor in beinahe allen Lebensbereichen. Dabei stehen wir vor zahlreichen Herausforderungen, insbesondere **(i) mangelnde Qualität von Datenmanagementprozessen, (ii) fehlende Standards, (iii) Mangel an Lösungen zum sicheren und nachverfolgbaren Teilen von Daten, (iv) Unsicherheit zu rechtlichen Aspekten (Lizenzen, Datenschutzgrundverordnung (DSGVO)) und (v) limitierter Zugang zu Rechen- und Speicherressourcen.** Darüber hinaus bergen die anfallenden Datenmengen enormes Wertschöpfungspotential für die Wirtschaft und versprechen datengetriebene Problemlösungen für viele wissenschaftliche und gesellschaftliche Vorhaben. Unternehmen und Forschungseinrichtungen rüsten sich für diese neuen Aufgaben, bauen Kompetenzen intern auf und bringen sich in Netzwerke im Bereich Datenwissenschaften, Datenanalyse und datengetriebene KI ein. **Um die Daten bestmöglich zu nutzen, müssen die organisationsinternen Datensilos aufgebrochen werden. Organisationen müssen sich auch auf Datenebene vernetzen und über Organisationsgrenzen hinweg mit Daten arbeiten.** Wenn das gelingt, dann können einerseits innovative, digitale Dienste entwickelt werden, und andererseits können in vielen Fällen die Datenbestände selbst monetarisiert werden. Diese Entwicklungen führten zur Etablierung von digitalen Plattformen, die Unternehmen ermöglichen, sowohl Rohdaten als auch verarbeitete Daten unter Einhaltung von rechtlichen Rahmenbedingungen (wie beispielsweise der DSGVO) anzubieten oder von anderen Unternehmen zu beziehen. Dadurch können auch neue innovative, datengetriebene Geschäftsmodelle entwickelt werden. Ein einheitliches Werteverständnis für diese Daten ist essentiell, um Daten über solche Marktplätze handeln zu können. Es gibt derzeit noch keinen allgemein anerkannten Ansatz zur Messung des Datenwertes, dieser wird meist individuell bestimmt. Neben den Chancen, die sich für Unternehmen entwickeln, profitieren auch andere Bereiche von Marktplätzen für Daten. Im Bereich Wissenschaft und Forschung kommt es zu (i) einer Steigerung der sozioökonomischen Wirkung von Forschungsdaten über Domänen und Ländergrenzen hinweg, (ii) Open Innovation wird durch Datenverfügbarkeit gefördert und (iii) auch hier entstehen neue Monetarisierungsmöglichkeiten durch innovative Geschäftsmodelle. Neben der Forschung gibt es auch Vorteile für die Gesellschaft, wie (i) der Zugang zu personalisierten und sektorübergreifenden Business-to-Consumer Diensten und (ii) Vorteile für das Wohlbefinden durch die gemeinsame Nutzung von persönlichen Daten in Schlüsselbranchen. **Datenkreise bieten DatenanbieterInnen und -nutzerInnen in einem klar definierten Anwendungsfeld die Möglichkeit den Austausch und Handel der Daten in einem klar abgegrenzten Raum durchzuführen.** Zentral ist dabei, dass die Datensouveränität beibehalten werden soll. Durch die Entwicklung von Datenkreisen wird der Mehrwert identifiziert, die Hürden durch Open Innovation Entwicklung abgebaut und neue Anwendungsfälle für bestehende Daten entstehen. Hierbei ist es wichtig, alle Betroffenen zu involvieren, um alle Aspekte zu betrachten und in das Konzept miteinzubeziehen.

Dieses Grobkonzept präsentiert eine Softwarelösung, die für Stakeholder in einem Datenkreis (DatenanbieterIn, -treuhänderIn, -nutzerIn, DatendienstAnbieterIn, RegulatorIn) eine Plattform bietet, um gemeinsam mit Daten zu arbeiten und den Wert zu steigern. Die Kompetenzen im Bereich Datenplattform werden vom KC gestellt, es werden öffentlich Infrastrukturen (ZAMG) integriert und der Prozess wird begleitet von juristischen Experten. Die Softwarelösung ist ein Open Source Produkt und setzt auf offene Standards.

Innovative Softwarelösung

Im Rahmen des von Stefanie Lindstaedt geleiteten Projekts IDE@S (Innovative Data Environment@ Styria[1]) wurde bereits ein Konzept für eine Datenplattform entwickelt (Abb. 1). Es wurde eine Nachfrage nach offenen technischen Lösungen, sowie ein großer Bedarf and technischen Lösungen für IT und Datensicherheit ermittelt.

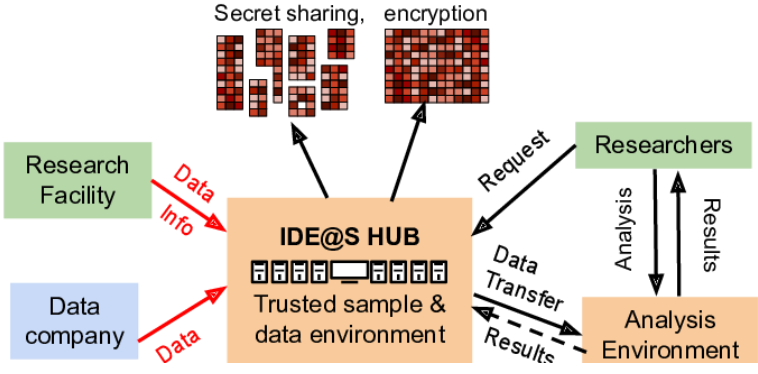

Abbildung 1: Konzept für kollaborative Dateninfrastruktur in IDE@S

[1] https://ideas.tugraz.at

Um diese technischen Anforderungen zu erfüllen bedarf es einer modularen, flexiblen und offenen Softwarelösung. **Die ausgewählte innovative Softwarelösung für Datenkreise Datenkreis Infrastruktur (DI)** basiert auf dem Model CyVerse US (Open Source Lösung, entwickelt von der University of Arizona/US[2]), welches im Team von Stefanie Lindstaedt bereits an der TU Graz, Karl-Franzens Universität Graz und Medizinischen Universität Graz im Rahmen des Hochschulraum Strukturmittel (HRSM) Projektes Integriertes Datenmanagement aufgebaut wurde. Dort haben KC MitarbeiterInnen das System bereits für eine Domäne aufgebaut (Life Sciences) und für eine weitere Domäne (Elektrotechnik) optimiert, was die Flexibilität und Anpassbarkeit des Systems beweisen.

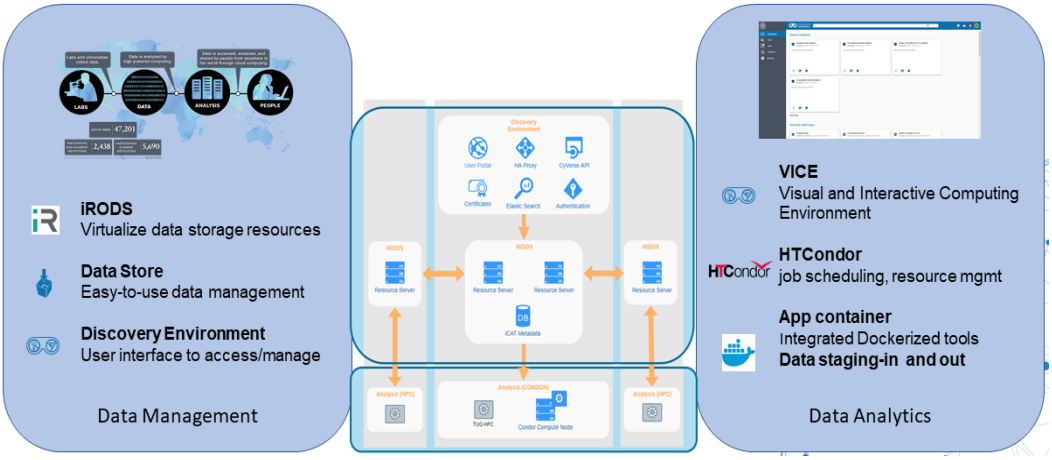

Abbildung 2: Architektur der Softwarelösung

Das Team unter Stefanie Lindstaedt baut im KC eine Datenplattform DI mit denselben zugrundeliegenden Technologien auf. **Die Datenplattform bietet eine einfache Benutzeroberfläche über die unterschiedliche Speicher- und Rechenressourcen an unterschiedlichen Standorten einfach und intuitiv angesprochen werden können.** Es ist keine lokale Installation notwendig, Zugriff erfolgt einfach über den Webbrowser. Die Softwarelösung unterstützt die zwei Hauptkomponenten bei der Arbeit mit Daten: Datenmanagement und Datenanalysen (Abb.2). Zusätzlich ermöglicht die Lösung eine Archivierung von Daten mit Digital Object Identifier (DOI).

- **Datenmanagement**: Hier werden alle Aspekte behandelt von Verwaltung der Daten, Dokumentation durch Metadaten bis hin zu Data Sharing. Das System basiert auf der Integrated Rule-Oriented Data System (iRODS) Technologie und ermöglicht verteilte, dezentrale Speicherressourcen, sowie Föderation von Systemen. Die Metadaten werden in einem zentralen Metadatenkatalog abgelegt, diese können durch Föderation dezentral für jede/n Datenkreis-TeilnehmerIn individuell verwaltet werden. Dadurch können eine Vielzahl von vorhandenen Ressourcen im Datenkreis einfach vernetzt und über eine grafische Oberfläche zugänglich gemacht werden. NutzerInnen haben einen eigenen Ordner in dem Daten verwaltet werden. Diese Daten können mit anderen NutzerInnen geteilt werden. Es gibt die Möglichkeit durch das Usermanagement Teams zu bilden, die Zugriff auf bestimmte Datensätze haben (= spezifizierbare Zugriffsberechtigung innerhalb eines Datenkreises). Das **Datenmanagement auf der Plattform ist domänenagnostisch** – Daten aus jeder Domäne können auf der Plattform verwaltet, dokumentiert und geteilt werden.
- **Data Analytics**: Hier werden alle Prozesse der Datenverarbeitung durchgeführt. Datenanalysen können durch Skripte im Hintergrund durchgeführt werden, aber auch durch eine grafische Oberfläche. Die Analysen werden durch die Dockertechnologie[3], eine der derzeit weltweit führenden Containerstandards, reproduzierbar gemacht. Für die Analysen werden Data Staging-in/out Container, sowie der Container mit der Software verwendet, somit verläuft der Prozess im Hintergrund und der User muss sich nicht um Datentransfer kümmern. Durch das HTCondor[4] Scheduling System können unterschiedliche High Performance Computing (HPC) Ressourcen angesprochen werden, u.a. ZAMG HPC Cluster. An der ZAMG steht zur Nutzung durch externe Anwender ein HPC-Cluster vom Type SGI ICEx-Dakota mit ca. 200 Rechenknoten zur Verfügung, die mit je 2 Intel-Sandybridge Prozessoren bestückt und über ein Infiniband High-Performance-Interconnect vernetzt sind. Das System wurde bis 2017 als operationeller HPC für zeitkritische Berechnungen von Wetterprognosen (Numerical Weather Prediction) eingesetzt und dann als Produktionssystem durch ein neueres HPC-System abgelöst. In Nachnutzung der noch voll funktionsfähigen HPC Hardware kann das Cluster bis zumind. 2022 für externe Nutzer verfügbar gemacht werden, denen ganze Rechenknoten monatsweise zur exklusiven Nutzung angeboten werden können. Die Verteilung der HPC-Ressourcen auf diesen exklusiv bereitgestellten Knoten wird verwaltet durch PBSpro. HTCondor

[2] https://cyverse.org
[3] https://docker.com
[4] https://htcondor.readthedocs.io

Transfernodes können Analysen von der Datenplattform an den ZAMG HPC Cluster weitergegeben werden. Der **Data Analytics Teil der Plattform ist domänenspezifisch**. Hier können disziplinspezifische Tools hinterlegt werden, sowie spezielle Workflows aus mehreren Analysetools aufgebaut werden.

Die Softwarelösung bietet eine einfache grafische Oberfläche und ist somit intuitiv und einfach nutzbar für Personen ohne IT-Hintergrund. Die Anbindung vorhandener Ressourcen im Datenkreis erfolgt lediglich über die Bereitstellung eines dedizierten iRODS Resource Servers. Dadurch werden Speicherressourcen einfach in das System eingebunden. Über HTCondor Transfernodes können Jobs einfach an Scheduling Systeme von Hochleistungsrechnern weitergegeben werden.

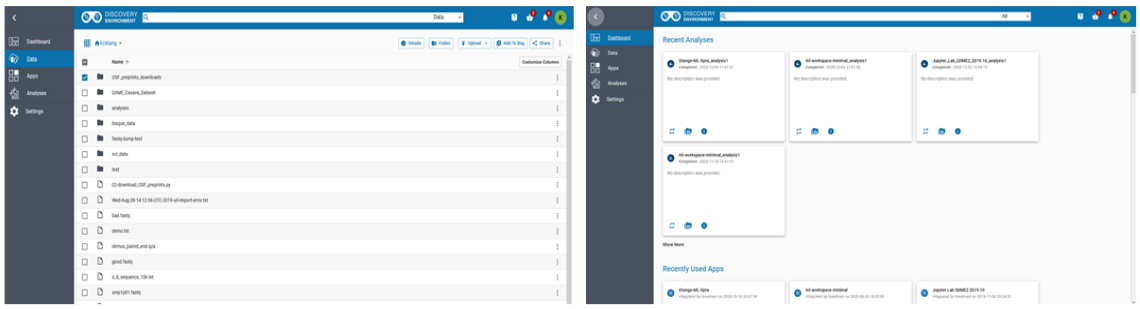

Abbildung 3: Graphische Oberfläche der Softwarelösung (links Datenmanagement, rechts Data Analytics)

Die verwendeten Technologien sind offene Lösungen (z.B. iRODS, Docker) und entsprechen den Standards in internationalen Initiativen (e.g. EOSC, Gaia-X). Bei der Plattform wird explizit auf proprietäre Lösungen verzichtet und auf Open Source gesetzt. Diese Open Source Lösung bietet ein hochmodulares System mit vielen Schnittstellen, um die Integration mit anderen Systemen problemlos zu ermöglichen. Durch die Microservice Architektur auf Kubernetes Cluster ermöglicht das System Skalierbarkeit, Flexibilität und einfache Instandhaltung. Offene Technologien haben sich zudem auch schon als sichere Alternative bewiesen.[5] Durch die Kollaboration mit ZAMG können weitere Ressourcen einfach in das System integriert und dadurch angesprochen werden. Durch die vorhandenen Kompetenzen am KC werden alle Aspekte über die Wertschöpfungskette von Daten abgedeckt (Abb. 4).

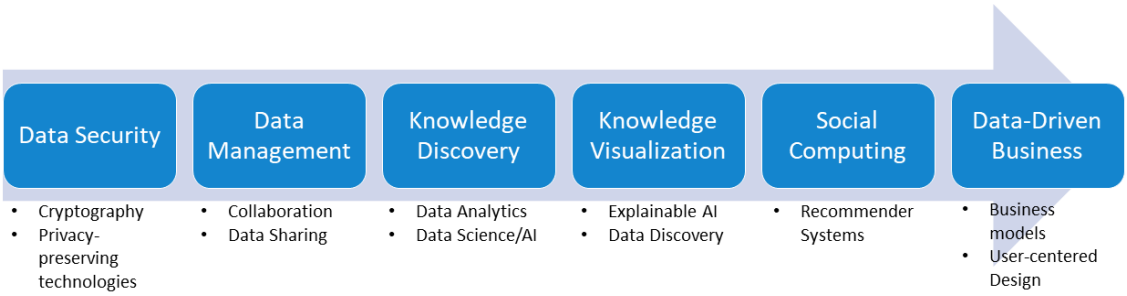

Abbildung 4: Wertschöpfungskette von Daten (abgedeckt durch KC Areas)

Die relevanten Aspekte für die für den Rollout der Softwarelösung sind (i) Datenschutz, (ii) Datensicherheit, (iii) Datensouveränität, (iv) Transparenz und (v) rechtliche Aspekte (Abb. 5) Viele dieser Aspekte können durch technische Lösungen vom KC abgedeckt werden, wie Privacy-Preserving Analytics für Datensicherheit, Kryptographie für Datenschutz, Aktivitätentracking für Transparenz und Rechtemanagement für Datensouveränität. Daneben werden auch noch weitere Kompetenzen benötigt, welche auch die rechtlichen Aspekte abdecken. Daher wurde in das Konsortium mit Prof. Wendland von der KFU auch ein Experte für IT-Recht involviert, um nicht nur eine technische Lösung einzuführen, sondern auch alle rechtlichen Rahmenbedingungen abzuklären. Es werden somit alle DSGVO-relevanten Aspekte abgedeckt für Daten auf der Plattform, sowie NutzerInneninformationen. Zusätzlich werden die entsprechenden Policies der teilnehmenden Organisationen berücksichtigt und Workflows/Anwendungsbeispiele dementsprechend aufgesetzt. Der Austausch von Daten unterschiedlicher Herkunft, Struktur und Schutzbedürftigkeit zwischen privaten wie öffentlichen NutzerInnen wirft neue und komplexe rechtliche Fragestellungen auf, die nach aktuellem Forschungsstand bislang nur ansatzweise geklärt sind. Haftungs- und Compliance Risiken können für weite NutzerInnenkreise daher ein erhebliches Hindernis für die Beteiligung an einer DI darstellen. Umso wichtiger ist die Einbettung des DI in einen geprüften Rechtsrahmen zur Gewährleistung eines hohen Maßes an Rechtskonformität und Rechtssicherheit. Dabei werden alle relevanten rechtlichen Fragestellungen proaktiv adressiert: Datenschutzrecht (DSGVO), Urheberrecht (inkl. Fragen der Inhaberschaft maschinengenerierter Daten), Lizenzrecht, Know-How Schutz, Rechtsrahmen für Plattformbetreiber, Vertragsrecht, Haftungsrecht.

[5] https://doi.org/10.1016/S1361-3723(13)70021-6

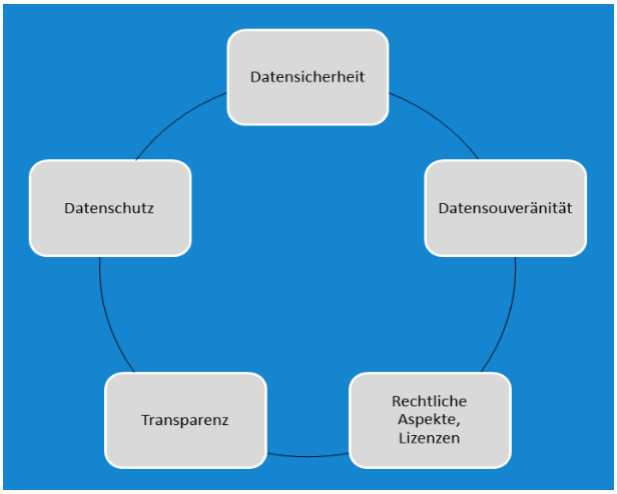

Abbildung 5: Relevante Aspekte für eine Softwarelösung in Datenkreisen

Grober Projekt- und Zeitplan

- **AP1: Projektmanagement (ab Jul 2021).** Koordination der Aktivitäten. Austausch mit und Reporting an BMK.
- **AP2: Stakeholder- und Anforderungsanalyse (Jul – Dez 2021)**. Identifikation und Kontaktaufnahme mit Stakeholder für den Datenkreis in Abstimmung mit BMK. Abhaltung von Arbeitsgruppentreffen zur Anforderungsanalyse.
- **AP3: Identifizierung von ersten disziplinspezifischem Datenkreisen & technische Implementierung (Okt 2021 – Jun 2022).** Nutzung der vorhandenen DI am KC. Integration von ZAMG HPC Ressourcen. Integration weiterer Ressourcen von Organisationen basierend auf Arbeitsgruppentreffen in AP2.
- **AP4: Durchführung von disziplinspezifischen Anwendungsfällen & Evaluierung (Mai 2022 – Jun 2023).** Durchführung von Anwendungsfällen in Begleitung vom KC Team. Anpassung der DI Oberfläche für disziplinspezifische Anforderungen.
- **AP5: Erweiterung der DI für andere disziplinspezifische Datenkreise (ab Apr 2023).**

Operationalisierungsaufwand

Für die Erweiterung und technische Implementierung der DI vom KC auf Organisationen in einem ausgewählten Datenkreis werden **3 VZÄ im Bereich DevOps/Softwaredevelopment** benötigt (300.000 EUR/Jahr). Zusätzlich wird **ein Projektmanager/Data Steward** die Aktivitäten koordinieren, sowie die Entwicklung von domänenspezifischen Anwendungsfällen begleiten. (100.000 EUR/Jahr). Erste Resultate, die aus der Entwicklung der Softwarelösung für Datenkreise hervorgehen, sollen die Grundlage für einen Projektantrag zur Finanzierung des Vorhabens durch öffentliche Gelder bilden. Das KC wird spezialisierte Beratungsdienstleistungen für die Nutzung der DI, sowie für die Entwicklung von Use Cases anbieten. Zusätzlich werden Bewusstseinsbildung, Öffentlichkeitsarbeit und Marketing in diesem Bereich abgewickelt. Um diese Services nachhaltig aufzubauen werden in einem ersten Schritt **zumindest 1 VZÄ im Bereich Business Development und Consulting (KC)** sowie Mittel für Marketingaufwand, und 1 VZÄ im Bereich Legal (KFU) (gesamt 250.000 EUR/Jahr) benötigt.
Für den Aufbau der DI wird hauptsächlich vorhandene Hardware im Datenkreis integriert. Durch die verteilte Infrastruktur ist keine Neuanschaffung von zentralen Ressourcen nötig. Die Abrechnung der genutzten Rechenleistung von ZAMG erfolgt auf Monatsbasis für ganze Rechenknoten, die zur Exklusivnutzung zur Verfügung stehen. Eine feinere Granulierung des HPC-Accounting auf Basis kleinerer Zeiteinheiten und/oder einzelner CPU-cores ist z.Zt. nicht vorgesehen. Der benötigte Speicherplatz wird separat verrechnet.

Business Model

Um die innovative Softwarelösung nachhaltig einzuführen und zu gewährleisten, dass sich das System erhält, muss es ein klares Betriebs-/Kostenmodell geben. Die Area Data-Driven Business am KC arbeitet gemeinsam mit der Universität St. Gallen an diesen Themen und verfügt über Experten für die Entwicklung von datengesteuerten Geschäftsmodellen. Ein Schwerpunkt liegt hierbei bei der Gestaltung von Datenplattformen und -märkten. Um die nachhaltige Etablierung einer Datenplattform zu gewährleisten, müssen regelmäßige Einnahmen zur Deckung laufender Kosten (z.B. Updates/Maintenance der Software, laufende Kosten für den Betrieb der Hardware, Kernteam) lukriert werden. Um qualitativ hochwertige Geschäftsprozesse zu ermöglichen, können Smart Contracts für Austauschprozesse eingeführt werden. Der Wert von Daten kann mithilfe von KC Kompetenzen für individuelle Fragestellungen ermittelt werden (Data Value Check).

Kooperation zwischen BMK und Konsortium

Im Rahmen von Arbeitsgruppentreffen sollen Beteiligte aus unterschiedlichen Domänen, die zugehörigen Intermediäre, sowie das BMK gemeinsam aktiv in die Entwicklung von Anwendungsfällen auf der Softwarelösung involviert werden. Dabei sollen Methoden aus dem Bereich des Design Thinking zum Einsatz kommen. Die so erarbeiteten Anwendungsfälle werden anschließend im Open-Source Framework implementiert und ermöglichen eine weitere Optimierung der Softwarelösung für unterschiedliche Domänen.

Kompetenzen im Konsortium

Know-Center GmbH ist mit 20 Jahren Erfahrung eine Innovationsdrehscheibe zwischen Wissenschaft und Industrie und bietet als non-Profit Unternehmen anwendungsorientierte Forschung in Kooperation mit akademischen Institutionen und Partnern aus der Wirtschaft. KC verfügt mit über 130 MitarbeiterInnen über umfangreiche Erfahrung in nationalen aber auch internationalen, kooperativen F&E Projekten im Bereich Big Data, ML und AI. Weiters gründete KC das Europäische Netzwerk der nationalen Big Data Center of Excellence und wurde von der BDVA/DAIRO mit dem iSpace-Label in Gold als eines der führenden Big Data-Forschungszentren in Europa ausgezeichnet. Die Expertise des Kompetenzzentrums beruht auch auf langjähriger Erfahrung in nationalen und internationalen Projekten in den Bereichen Datenaufbereitung bzw. Analyse, Data Sharing, Aufbau und Betrieb von Datenplattformen, Datenwertermittlung, Technologien zum DSGVO konformen Datenaustausch. Aktuell wird auch in mehreren Forschungsprojekten zum Aufbau von Datenplattformen (IDE@S, HRSM, Data Market Austria) mitgearbeitet.

Die ZAMG ist der nationale österreichische meteorologische und geophysikalische Dienst und eine nachgeordnete Dienststelle des Bundesministeriums für Bildung, Wissenschaft und Forschung (BMBWF). Die ZAMG hat den Hauptsitz auf der Hohen Warte in Wien und Kundenservicestellen in Graz, Innsbruck, Klagenfurt und Salzburg. Der Tätigkeitsbereich der rund 300 Mitarbeiterinnen und Mitarbeiter erstreckt sich von Wettervorhersagen und Wetterwarnungen über angewandte meteorologische, klimatologische und geophysikalische Forschung bis hin zum Erdbebendienst und zu umweltmeteorologischer Gutachtertätigkeit. Die ZAMG wurde 1851 gegründet und ist der älteste selbstständige Wetterdienst der Welt. Die ZAMG betreibt ein meteorologisches (rund 270 Stationen) und ein seismisches (rund 40 Stationen) Messnetz. Außerdem betreibt sie das Sonnblick Observatorium in Salzburg und das Conrad Observatorium in Niederösterreich. Die Expertinnen und Experten der ZAMG vertreten Österreich in zahlreichen internationalen Organisationen und Vereinigungen wie z.B. WMO (Weltmeteorologische Organisation der Vereinten Nationen), ECMWF (Europäisches Zentrum für Mittelfristige Wettervorhersagen) und EUMETSAT (Europäische Vereinigung zur Entwicklung von Wetter- und Klimasatellitensystemen). Die ZAMG ist die führende meteorologische Institution in Österreich und bietet ihre Leistungen unter anderem öffentlichen und privaten Fernseh- und Rundfunkanstalten, Tageszeitungen und Dienstleistern wie Versicherungen, Energiewirtschaft, Winterdienstfirmen, Bauunternehmen und Gemeinden an. Mit hochverfügbaren IT-Systemen gehört die ZAMG zur kritischen Infrastruktur des Bundes.

Smart Regulation ist ein interdisziplinärer profilbildender Bereich der Karl-Franzens Universität Graz. Die Mitglieder des Profilbildenden Bereiches sind in rund **30 verschiedene Forschungsprojekte** eingebunden, die sowohl auf nationaler als auch internationaler Ebene in Kooperation mit anderen Institutionen und Einrichtungen durchgeführt werden. Fragestellungen, die behandelt werden, umfassen beispielsweise die Untersuchung von rechtlichen Maßnahmen und Instrumente, um das Risiko einer zweckwidrigen Datenverwendung gering zu halten.

Relevante Vorarbeiten/Projekte im Konsortium

Data Market Austria (KC, ZAMG)[6]: Das Projekt Data Market Austria legte den Grundstein für eine breite Palette an relevanten Themen und Aspekten im Kontext von Datenkreisen, indem es die technologischen Grundlagen für sichere Datenmärkte und Cloud-Interoperabilität vorantrieb und ein Umfeld schuf, das datenzentrierte Innovationen ermöglicht.
IDE@S (KC, TUG)[7]: In diesem vom Land Steiermark finanzierten Projekt werden die relevanten Player in Wirtschaft, Wissenschaft und Öffentlichkeit in der Steiermark in Workshops vernetzt und deren Anforderungen im Bereich Dateninfrastrukturen erhoben. Ziel des Projektes ist ein steirisches Modell für die kollaborative Nutzung großer Datenmengen zu erstellen.
HRSM Projekt – Integriertes Datenmanagement (KC, TUG)[8]**:** In diesem Projekt wird eine Datenplattform für Forschende aus den Life Sciences an den Grazer Universitäten etabliert.
Austrian DataLab and Services (KC, TUG)[9]**:** Im Austrian DataLab and Services Projekt wird auf die vorhandene Infrastrukturschicht der österreichischen Universitäten aufgebaut. Es werden Datenservices aufgebaut und etabliert und für alle zugänglich gemacht.

Kontakt

Prof. Dr. Stefanie Lindstaedt, CEO Know-Center GmbH
E-Mail: slind@know-center.at

[6] https://datamarket.at
[7] https://ideas.tugraz.at
[8] https://cyverse.tugraz.at
[9] https://forschungsdaten.at/adls