FAIRagro UC Submission Form

1. UC general information contact details

1.1. Title & Keywords

Please choose a title, acronym and 3 to 5 keywords.

Title of the project	Increasing FAIRness of FAIRagro data through AI supported metadata enrichment
Short-Title / Acronym	/
Keywords *(min. 3, max. 5.)*	metadata enrichment, text mining, AI, FAIR, Research Data Infrastructures

1.2. Contact details

Please add the contact details of the applicants, co-applicant(s) and associated partner(s) and list all persons involved and their respective roles. Please add further rows to the table, if necessary.

Main UC applicant's name	Juliane Fluck
Main UC applicant's role	FAIRagro Co-applicant
Main UC applicant's Institution[1]	ZB MED – Information Centre for Life Sciences
FAIRragro (co-) applicant or participant	**yes**/~~no~~
Main UC applicant's E-Mail	fluck@zbmed.de
Main UC applicant's ORCID *(optional)*	https://orcid.org/0000-0003-1379-7023
Further persons involved from the main applicant´s institution and their respective roles	
UC-Co-applicant institution[1] first name, last name, ORCID *(optional)*	Leibniz Centre for Agricultural Landscape Research (ZALF) Xenia Specka https://orcid.org/0000-0002-1890-0192
UC-Co-applicant institution[1] First name, Last name, ORCID *(optional)*	Julius Kühn-Institut – Bundesforschungsinstitut für Kulturpflanzen Ulrike Stahl https://orcid.org/0000-0002-5659-910X

1.3. UC format and duration

Please choose the UC type and add the proposed running time.

☒ UC Pilot: ___12_months
☐ UC Project: _____ months

[1] All UC applicants and co-applicants must be part of the FAIRagro Plenary (Co-applicants or Participants) or become FAIRagro Participants in case they want to receive FAIRagro funding

2. UC description and scientific details

2.1. Short UC summary (Abstract)

Please summarise your UC in a short abstract.

[Please keep in mind that Use Cases are not single scientific projects that generate data and have specific needs with respect to data handling. Use Cases must address research data management challenges in agrosystem research and/ or agricultural science and beyond, preferably involving multiple stakeholders, parties and institutions to avoid building isolated solutions. Max. 250 words are accepted[2].]

High-quality metadata is essential for the FAIR principles. As research data management (RDM) becomes crucial, stakeholders recognize the need for meaningful metadata and automated generation processes. These advancements ensure future metadata FAIRness but do not address legacy metadata, which lacks standardized collection practices.

FAIRagro is developing a metadata schema for harmonizing heterogeneous metadata of participating Research Data Infrastructures (RDIs) to facilitate a central search, increasing the Findability of agrosystem resources. To enable an efficient transformation of legacy metadata of the RDIs to make integration into the FAIRagro Central Search Service as efficient as possible, this pilot project tests how far state of the art AI-based text mining techniques, e.g., deep learning models and few-shot learning, are able to automatically extract relevant information from unstructured data (e.g. dataset abstracts, related publications, the data itself, etc.), using Named Entity Recognition (NER). These tasks involve identifying both general and agrosystem domain specific entities and relations. The goal of this pilot is to extract information on two different core entities (Crops, Soil) from two different FAIRagro RDIs (OpenAgrar, BonaRes Repository) and make it available in a structured way. Furthermore, it evaluates if text mining offers a viable method for enriching metadata to the schema developed for powering the Central Search Service.

The outcome will show how far AI methods are ready to make agrosystem resources FAIRer and to assist participating RDIs in extending their provided metadata in a resource efficient way. If the pilot is successful, further developments can be made to support all FAIRagro infrastructures or even other domains in metadata extension, opening up possibilities for e.g. cross-NFDI consortia collaboration in the future.

2.2. UC concept and objectives

What is the overall concept of the Use Case? What are the specific objectives of the UC?

[Recommended: Please add the overall UC concept and overview of the objectives, preferably as graphic/ illustration/ visualisation. This will be helpful in facilitating rapid comprehension of your Use Case by the reviewers.]

[2] Answers by the applicants that exceed the maximum word count will be strictly cut of during the formal check of the application and will not be forwarded to the reviewers. This applies to all sections of the Submission Form and respective application.

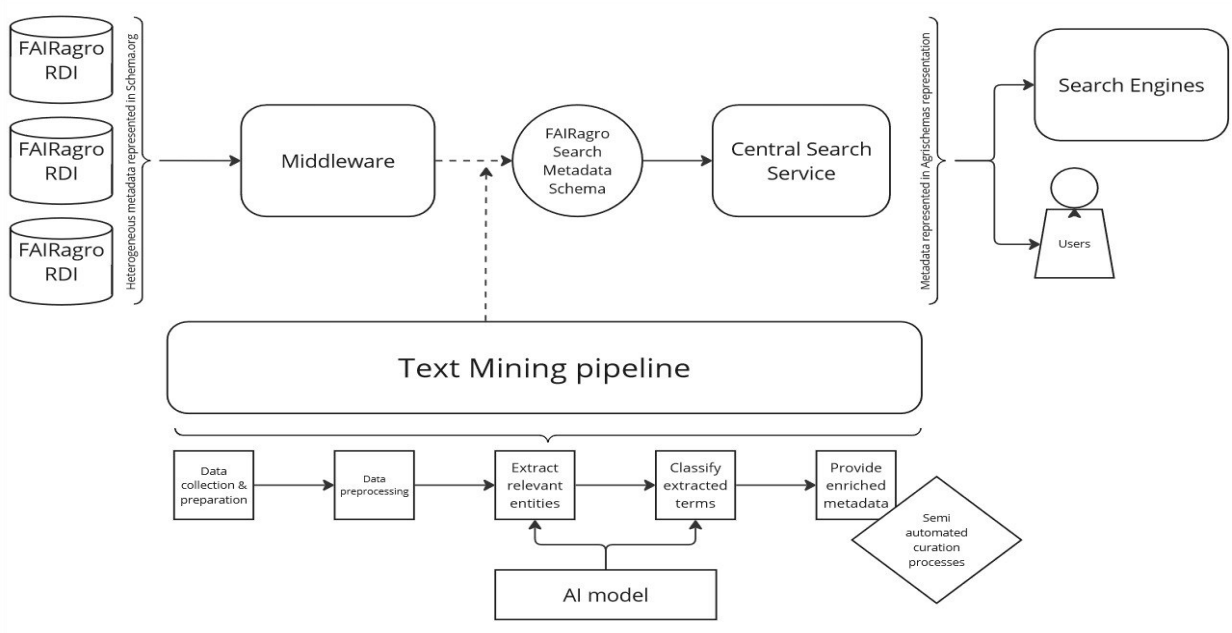

Figure 1 - Integration of the Text Mining pipeline into FAIRagros service infrastructure. The projects Research Data Infrastructures (RDIs) provide metadata as Schema.org markup to the Central Search Service via the Middleware, conforming to the FAIRagro Search Metadata Schema. The Text Mining pipeline collects the unstructured elements provided as metadata, prepares and preprocesses them and extracts and classifies relevant terms based on an AI model. In this way the metadata becomes enriched and will be evaluated with semi automated curation processes, before it will be ingested into the Central Search Service. Users can afterwards execute domain specific, centralized search queries while search engines will consume metadata exposed by the Central Search Service as Schema.org/Agrischemas, increasing Findability for a broader community.

The developments planned in this pilot fit seamlessly into FAIRagro service landscape and helps closing a gap between the current decentralized search capabilities, the envisioned central search service and its benefits towards semantic searches and increased exposure to relevant search engines.

The main components and services for the pilot are either already in a production status and available such as the RDIs and the building blocks of the Text Mining (information extraction) pipeline, as part of ZB MEDs infrastructure, or are progressing according to the project proposal. The combination of the different components (as represented by the dotted arrow) promises an efficient bundling of resources.

The pilot will begin by analysing suitable data for the task in close cooperation with the participating RDIs. During this process the relevant documents and metadata fields will be evaluated for extraction. Annotation guidelines will be developed and used for creating a manually annotated training corpus (**Action 1.1**). The manual annotation will be done using the software INCEpTION which enables the integration of a recommender system, allows curators to correct them, and facilitates the introduction of the main entity classes that will be annotated. This software is already installed in ZB MED's infrastructure and ready to use remotely, enabling the provision of feedback on annotations by domain experts at the RDIs (**Action 1.2**). Simultaneously a literary review will identify the most fitting model for the pilots requirements (**Action 2.1**) and prepare the following Actions by setting up the needed infrastructure for the text mining (**Action 2.2**).

After this preparation is finished, work on using the text mining pipeline for creating automated annotations and extractions will begin by using the annotated text corpus alongside the data imported from the RDI's to develop and test the information extraction pipeline. The main components of this pipeline are::

1. **Data collection**: Gathering annotated data and raw data from RDIs, storing it in an organized manner (e.g., JSON format).
2. **Data pre-processing**: The collected data is pre-processed to ensure that it is in a suitable format for the models. This involves considering the maximum text length each model can

handle and performing necessary pre-processing steps such as tokenization, normalization, and removal of irrelevant data. The output of this step is tabular data containing the different texts and their corresponding annotations. Steps such as tokenization, normalization, and other preprocessing methods produce tabular data in a format fitting for training and testing the models.

3. **Models development**: The models are developed by adapting state of the art deep learning NER methods (supervised learning or few-shot learning using the annotated data as training set).
4. **Models evaluation**: The models are tested using the annotated text corpus as reference data. A comparison between the different methods and an evaluation of which parts of the metadata they perform best on will be the output of this step.
5. **Annotations export**: Export the annotated data with their spans and transfer the different entities into their place in the meta data schema.

As a final step the use-case will evaluate the extracted metadata outputs of the automated annotations and extractions by manually curating the results. Depending on an uncertainty value for the automated annotation, domain experts will assess their correctness, leading to quality results (**Action 3.1**). To benefit from the enriched metadata it has to made available. The requirements for setting the RDIs up for these changes will be explored in close collaboration (**Action 3.2**).

All the insights won during the pilot will support the partners to develop an understanding of needed resources for large scale metadata transformation processes, giving the RDIs insight and support in decision making and planning regarding their approach toward increasing the FAIRness of their metadata collections. To make these findings available and reusable by others (e.g. other RDIs with ties to the agrosystem domain), a strategy will be devised and published (Action 3.3).

2.3. Scientific background, added expertise and preliminary work

What is the scientific background and added expertise of the involved parties, stakeholders and institutions? What is relevant preliminary work, which has been conducted by the UC applicants, co-applicants and associated partners?

[Please highlight the relevant scientific background within the state-of-the-art, pointing out the added expertise for the FAIRagro consortium (max. 100 words). Below please list up to 10 relevant publications, products, services, ect. in bullet points.]

The Knowledge Management department at ZB MED has particular expertise in text mining, semantic search, named entity recognition, and information extraction in various life science applications. Most recently this technology has been developed for publications related to the global pandemic caused by the novel coronavirus, SARS-CoV-2.

Relevant publications by the group:

- Langnickel, L., Darms, J., Heldt, K., Ducks, D., Fluck, J. (2022). Continuous development of the semantic search engine preVIEW: from COVID-19 to long COVID. In: Database, Volume 2022. DOI: 10.1093/database/baac048
- Sasse, J., Darms, J., Fluck, J. (2022). Semantic Metadata Annotation Services in the Biomedical Domain — A Literature Review. In: Applied Sciences, 12 (2), p. 796. DOI:10.3390/app12020796
- Julia Sasse, Juliane Fluck: An Annotation Workbench for Semantic Annotation of Data Collection Instruments. In: Studies in Health Technology and Informatics, Volume 302. DOI: 10.3233/SHTI230074
- Kühnel, L., Schulz, A., Hammer, B., Fluck, J. (2022). BERT WEAVER: Using WEight AVERaging to enable lifelong learning for transformer-based models in biomedical semantic search engines. DOI: 10.48550/ARXIV.2202.10101
- Lentzen, M., Madan, S., Kühnel, L., Fluck, J., et al. (2022). Critical assessment of transformer-based AI models for German clinical notes. In: JAMIA Open, Volume 5, Issue 4. DOI: 10.1093/jamiaopen/ooac087

The Information Centre and Library team and the Team FDM at JKI combine expertise in the agricultural domain with handling of domain and publication-specific metadata, subject indexing, the use of subject-specific thesauri and publishing research data at OpenAgrar in a FAIR manner.

Relevant publications by the group:

- Oeltjen, W., Neumann, K., Stahl, U., & Stephan, R. (2019). MyCoRe macht Forschungsdaten FAIR. Bibliothek : Forschung und Praxis, 43(1), pp. 82–90. 10.1515/bfp-2019-2013

The *Research Data Management* and *Data Infrastructures* working groups at ZALF possess extensive expertise in data publication and curation. They operate the BonaRes Repository, a Research Data Infrastructure (RDI) dedicated to the publication of soil and agricultural research data.

Revelant publications by the group:

- Svoboda, N., Schmidt, M., Hoffmann, C., Specka, X. (2022). Citing soil and agricultural research data. BonaRes Series. DOI: 10.20387/bonares-fm2j-c233
- Grosse, M., Hoffmann, C., Specka, X., Svoboda, N. (2020) Managing long-term experiment data: a repository for soil and agricultural research. In: Bhullar, G. S., Riar, A. (eds), Long-term farming systems research. Academic Press, an imprint of Elsevier, London, pp. 167-182. DOI: 10.1016/B978-0-12-818186-7.00010-2
- Specka, X., Gärtner, P., Hoffmann, C., Svoboda, N., Stecker, M., Einspanier, U., Senkler, K., Zoarder, M. A. M., Heinrich, U. (2019) The BonaRes metadata schema for geospatial soil-agricultural research data - merging INSPIRE and DataCite metadata schemes. Computers & Geosciences 132, 33-41. DOI: 10.1016/j.cageo.2019.07.005

2.4. UC specification and maturity

What are the specifications for essential RDM categories?

[Recommended: If applicable, please describe the state-of-the-art of major RDM UC characteristics by filling the table.]

Category	**Description: UC specification and examples**
Disciplines	**Computer Science, Text Mining in agriculture**
Scales (gene, plant, field, farm, landscape, region)	/
Data domains (e.g. weather, soil, crop management)	Metadata of published research data in agriculture
Data types (e.g. tabular, images, gene sequences)	Text, metadata
File formats (open/ proprietary; e.g. xlsx, csv)	Open, JSON, XML, txt, csv, ann
Source data / database / repositories (proprietary/ public RDI/ FAIRagro RDI/ other; e.g. DWD, OpenAgrar)	FAIRagro RDIs (BonaRes Repository, OpenAgrar)
Processing workflows (concept/ prototype/ application/ …)	prototype
Tools/software (open/ proprietary; e.g. R, Phyton, QGIS)	Python, Huggingface

Other	/

How can the maturity of the UC be described in respect to the technology readiness levels (TRL)?

Where a topic description refers to a TRL, the following definitions apply, unless otherwise specified:

- TRL 1 – basic principles observed
- TRL 2 – technology concept formulated
- TRL 3 – experimental proof of concept
- TRL 4 – technology validated in lab
- TRL 5 – technology validated in relevant environment (industrially relevant environment in the case of key enabling technologies)
- TRL 6 – technology demonstrated in relevant environment (industrially relevant environment in the case of key enabling technologies)
- TRL 7 – system prototype demonstration in operational environment
- TRL 8 – system complete and qualified
- TRL 9 – actual system proven in operational environment (competitive manufacturing in the case of key enabling technologies; or in space)

Technology Readiness Level, Source HORIZON 2020 – WORK PROGRAMME 2014-2015:
https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-ga_en.pdf

[Please provide estimates for the TRL today and after completion of the UC. Bullet points are preferred.

***Note:** Low TRL does not necessarily imply a lower rating for the UC application; it may indicate a higher potential for improvement and further development.]*

	Description: UC specification
TLR of UC today	TRL 3 - experimental proof of concept
TRL of UC after completion of UC Project/ Pilot	TRL 4 – technology validated in lab

2.5. Self-assessment of FAIRness status quo, data quality and legal aspects

Which FAIR principles are addressed by the UC and which FAIR or RDM elements are relevant (e.g. standards, metadata, ontologies, PIDs, data quality checks, licences, legal aspects, data quality)?

[Recommended: Reflecting on the FAIRness status quo, data quality and/ or legal aspects is helpful to integrate the UC into FAIRagro´s work program and indicates potential room for relevant improvement and further development. List with bullet points is preferred and max. 200 words are accepted.]

- **F2: Data are described with rich metadata:** The UC will support RDIs in providing additional rich, relevant metadata, making additional search queries for finding the data possible
- **I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation:** Developments in FAIRagros TA3 aim at using Linked Data technologies for making metadata machine readable. The UC will develop mechanisms to increase implementation uptake of these technologies.
- **I2: (Meta)data use vocabularies that follow the FAIR principles:** By transforming RDI metadata to agreed upon metadata standards and increasing its structure, making interlinking the data to machine readable ontologies easier, increases its Interoperability

- **R1: (Meta)data are richly described with a plurality of accurate and relevant attributes:** The pipeline developed in the UC will be used to generate metadata, that was identified as highly relevant for finding datasets by current FAIRagro use-cases, ensuring the necessity of results and orientation towards real UC needs

2.6. Expected outcomes

What are the expected outcomes of the Use Case?

[Which output will be generated that can be used by our community (e.g. data, tool, workflow, publications, community guidelines, recommendations, workshop, training, standards)? Please list up to ten outcomes as bullet points.]

- Software source code of the adapted text mining pipeline
- The annotated and quality checked agricultural text corpus
- The trained machine learning models
- The generated enriched metadata resources
- A strategy giving guidance and insight for Research Data Infrastructues on how to enrich their metadata towards the FAIRagro Search Portal Metadata Schema

2.7. Involvement of UC institutions

If additional co-applicants and associated partners are involved in the UC application and implementation, what roles and responsibilities do they fulfil? What are the specific contributions from the UC co-applicant(s) and associated partner(s)?

[A Table or list with bullet points is preferred and max. 200 words are accepted.]

Leibniz Centre for Agricultural Landscape Research, specifically the working groups Research Data Management (FDM) and Data Infrastructures (DIS) will contribute to the UC by:

- Providing metadata from published datasets in the BonaRes Repository via existing harvesting interfaces to be used in text mining approaches in the Use Case Pilot.
- Offering data curation expertise for soil and agricultural research data to aid in developing annotation guidelines.
- Giving feedback on the derived annotations for selected datasets from the BonaRes Repository by providing domain expertise on soil- and agricultural research data

The Information Centre and Library team and the Team FDM at JKI will contribute to the UC by:

- Providing metadata from published datasets and text publication in the OpenAgrar Repository for text mining approaches via existing harvesting interfaces
- Providing annotation of data sets and evaluation of semi-automatic curation workflows regarding the training data from OpenAgrar with respect to publications from the crop and soil domain
- Defining text corpus selection criteria and annotation guidelines
- Giving feedback on the derived annotations for selected datasets from the OpenAgrar Repository by providing domain expertise on soil- and crop research data

3. UC - FAIRagro connection

3.1 UC contributions and benefits

Use Cases are integrated into the consortium's work plan, entailing collaborative efforts with all Task Areas (TA) to accomplish the specific Use Case measures and deliverables over a predetermined duration. The FAIRagro objectives and workplan are published as FAIRagro - A FAIR Data Infrastructure for Agrosystems (proposal).

[We are particularly interested to know how the applied Use Case will interact with the FAIRagro consortium in respect to the different Task Areas, explicit Measures and/ or major activities (in the following described as focal points). Below you can find a summary of a few focal points of FAIRagro´s work programme. Please add a checkmark where you see a connection and provide a short explanation on how you expect to contribute or benefit from FAIRagro (bullet points are preferred).]

Prioritise 1 to 5 FAIRagro focal points where the UC will benefit and/ or contribute. Checkmark as essential, if collaboration is needed for the successful implementation of the UC (The information will help to integrate the UC in the FAIRagro´s work program).

***Note:** The quality and added value of the connection/ collaboration is more important than the quantity.*

FAIRagro Connection through respective TAs and focal points	**UC will benefit**	**UC will contribute**	**Essential**	**Short explanation** (in bullet points)
TA 2: Community Involvement & Networking				
Communication and dissemination (e.g. content production and dissemination, organization of networking events such as workshops and conferences)				
Needs collection and evaluation (e.g. community surveys, usability tests)				
Training and education (e.g. events, workshops, trainings, open educational resources)				
Other…				
TA 3: Standardization, Interoperability & Quality				
Standards for digital resources	X		4	This pilot can reuse elements of ontologies and metadata standards collected in the Standard Inventory developed by M3.1 for developing the annotation guidelines in Action 1.1.
Standards for data management		X	2	M3.2 develops an extension of Schema.org as part of Bioschemas. This pilot will help increasing adoption, by offering RDIs a resource efficient way of adopting the extension.

Measures for data quality and fitness for use				
Data quality annotation and curation				
FAIR workflows and FAIR Digital Objects				
Legal framework and machine-actionable policies				
Other…				
TA 4: Services				
Central Service (´FAIRagro Portal`)				
Networked research data infrastructure		X	1	The Middleware developed in M4.2 harmonizes the metadata of the FAIRagro RDIs and provides it for other FAIRagro services such as the Central Search Service. This pilot supports the Middleware by enriching the metadata of the RDIs, making harmonization easier.
Searchable inventory of services and data		X	3	As outlined in the short explanation of the Middleware Measure (M4.2), the Central Search Service will benefit from developments of the UC.
Scientific Workflow Infrastructure (SciWin)				
Other…				
FAIRagro Task Area 5: Overarching activities				
Internationalisation				
Cross-NFDI networking				
Other…				

3.2 UC requirements

If specific focal points are essential for the implementation of the UC, please elaborate in more detail: What are the UC expectations regarding the collaboration with FAIRagro? What is needed for the successful implementation of the Use Case from the respective TAs?

[Recommended: Please formulate and/ or list precise requirements (max. 200 words). The information will help to evaluate whether FAIRagro can support with the required resources and integrate the UC in the FAIRagro´s work program.]

- Access to metadata of participating research data infrastructures in a format suitable for text mining methods (e.g., via the middleware)
- A list of core metadata entities that should be extracted from unstructured metadata (in cooperation with TA3)
- Support from RDIs/domain experts in creating a manually annotated corpora
- Support from RDIs/domain experts by evaluating a semi-automated curation process

4. UC dissemination and community engagement

4.1 Dissemination & community engagement

One FAIRagro goal is to foster a cultural change towards a FAIR and collaborative RDM in agricultural research. How can the UC contribute to this goal?

[Recommended: A Table or list with bullet points is preferred and max. 200 words are accepted.]

The FAIR principles and the Open Science movement are key pillars of advancing research in a way that enables broader participation and sustainable growth. These developments first and foremost require a cultural change in research practices such as the openness to publish data to make it available for the public, enabling reuse by others, leading to new results and increasing the quality of research in general.

The outcomes of this use case will support such a cultural change by providing tools that decrease the work needed from researchers in achieving FAIRness for their data. Publishing research data in a reusable manner, initially takes more effort for researchers. This can lead to a decision against data publication due to limited resources even if willing to do so.

Another aspect that will foster a cultural change is that the results of this use case will demonstrate the added values of an implementation of the FAIR principles. By enriching metadata and using it to support developments such as easier adoption of the Schema.org extension of M3.2, the use case will support in making datasets more Findable and therefore more Reusable resulting in more willingness of researchers to become part of an open data community.

5. Workplan

Action	Months											
	1	2	3	4	5	6	7	8	9	10	11	12
Action 1			M1 / D1									
Action 1.1												
Action 1.2												
Action 2									M2 / D2			
Action 2.1												
Action 2.2												
Action 2.3												
Action 2.4												
Action 3												M3 / D3
Action 3.1												
Action 3.2												
Action 3.3												

5.1. Actions (Work packages)

If necessary, you can break down the planned UC into actions (work packages) and describe the contribution of each institution and partner involved.

Action 1

Title: Analysis of data and preparation of annotated text corpora (Months 1-3)

Institution 1: ZB MED

Institution/ partner 2: ZALF, JKI

Summary: This action prepares an annotated text corpus for using it in the following actions as a resource in the text mining pipeline. To achieve this, results of this action include selection criteria for building the corpus, annotation guidelines and applying the guidelines to create the corpus.

Action 1.1: Define text corpus selection criteria and create annotation guidelines (Months 1-2)

In close collaboration with domain experts of the BonaRes Repository and OpenAgrar, a set of criteria is defined to determine which datasets and metadata descriptions to include as resources in the text mining corpus. The selection will depend on what information (e.g. metadata fields such as descriptions/abstracts) is available for each dataset.

Additionally, once the resources for the corpus are known, annotation guidelines for **Action 1.2** are proposed by ZB MED and evaluated by domain experts of the RDIs. These include what entities to annotate (e.g. crops, soil types) and include annotation examples from the selected datasets to enable a common understanding and guarantee consistent quality in the following manual annotation of the corpus.

Action 1.2: Creation of manually annotated text corpus (Month 3)

The results of **Action 1.1** will be used to manually annotate the collected text corpus, following the annotation guidelines. ZB MED will provide centralized access to INCEpTION and create the corpus. Domain experts of the RDIs will evaluate these annotations to ensure their quality for the use of the corpus in **Action 2.4**.

Action 2

Title: Automatic extraction and annotation of metadata (Months 1-9)

Institution 1: ZB MED

Summary: This action aims to research the most suitable AI methods for the pilot and integrate them into a Named Entity Recognition pipeline. The main steps are to conduct a literature review, decide on the candidate models to use and determine the required infrastructure to run them. Next, the data from the RDI's and other external sources is prepared, followed by training the candidate models using this data. The models are then evaluated using the annotated data. Finally, the models' outputs are made available to be used in **Action 3**.

Action 2.1 Literature Review and Model Selection for Named Entity Recognition in agriculture (Months 1-3)

This action includes a comprehensive literature review on the state of the art in Named Entity Recognition (NER) in the field of agriculture. It involves determining which current state-of-the-art models can be used directly off-the-shelf and identifying which models need further adaptation for the requirements of this specific pilot. Additionally, it defines the technical specifications required to run those models.

Action 2.1 operates in parallel to **Action 1** and its results, aiming to stay up-to-date and continuously update the list of models simultaneously. It is important to find through the literature review the number of samples per entity that is required to train models in a supervised manner (learn from positive examples) or evaluate if few-shots training (learn form few examples) is sufficient. At the same time this action is identifying additional annotated agricultural datasets in the literature which might be included in the following actions.

Action 2.2: Set up the information extraction pipeline infrastructure (Month 2)

To develop and test the models selected in **Action 2.1**, the right computational infrastructure needs to be set up. The main parts of this setup include data storage and computing power:

- Data Storage: Enough storage space is needed to hold all the data, and it should be connected to the computing systems for quick access.
- Computing power: High computational power is required to run, train, and test the advanced models. The key specifications to consider are the size of the GPU (graphics processing unit) and the corresponding CPU (central processing unit) and RAM (memory) that go along with it.

The product of this Action is selecting the suitable infrastructure and setting up all the users and the storage options to use it.

Action 2.3: Data preparation (Months 4-5)

The pilot relies on data from different RDIs and their heterogeneous data formats. This requires to process each dataset based on its source and create a unified dataset for training and testing the models. By applying a unified format and set of attributes to all data, regardless of the source, we ensure the system works smoothly and can potentially incorporate new RDI's. This approach guarantees scalability and reproducibility.

The result of this Action is a unified dataset that enables minimum steps for data preprocessing in the following step of model development.

Action 2.4: Model development and data reprocessing and evaluation of results (Information Extraction Pipeline) (Months 6-9)

Using the infrastructure established in **Action 2.2**, the design of the models using AI frameworks (i.e. Huggingface) is evaluated by executing the pipeline. Its main components are:

- **Data preprocessing:** Based on each model, different data format and different data size should be adapted to the model. The data is split into training and testing data.
- **Model parameters optimization:** Set up the model parameters for the application. These parameters vary from input/output sizes to optimization methods and cost functions.
- **Exporting output:** Make sure that the output format is comparable to the data format implemented in **Action 2.3** so that it can be evaluated and compared to other models. The output of each model should be a list of entities in a format that includes their label (enitity class), text, position in original text and a certainty score of the label. From ZB MED's expertise in biomedical text mining, existing pipelines can be used as reference or integrated in this step.
- **Evaluation:** Defining evaluation criteria based on matching entity labels and their spans in the text, counting correct (True Positive), incorrect (False Positive), and missed (False Negative) predictions. Common metrics used are Precision, Recall, and F1 Score. The models are then tested on the dataset, and their outputs are compared to the manually annotated text corpus from **Action 1.2** as reference data. ZB MED has existing evaluation scripts for biomedical data that can be adapted for this purpose.

The outputs of this action are:

- The model(s) that will be published
- The metadata according to the schema from **Action 1.1** and a simple prototype for a tool for curating uncertain results (for **Action 3.1**)
- The information extraction pipeline.

Consideration for the pipeline are:

- Enable continuous training on any new data that will be available in the future to improve the performance.
- It should be modular meaning all components could be adapted in other applications.
- It should be extensible meaning that new feature (i.e. new RDI's and new entity classes) could be added in the future.

Action 3

Title: Evaluation of semi-automatic curation approaches and designing large scale metadata transformation processes (3 Months)

Institution 1: ZB MED

Institution/ partner 2: ZALF, JKI

Summary: Action 3 uses the results of **Action 2** and makes them available in different ways. By evaluating uncertain results of the automated extraction, **Action 3.1** ensures the quality of the extracted metadata. This opens the possibility of ingesting it into the RDIs in a structured way, which will be explored in **Action 3.2**. To plan large scale metadata enrichment tasks for the future and to make the insights available for a broader community, **Action 3.3**. will develop and communicate a strategy, which other (FAIRagro) RDIs can profit from.

Action 3.1: Definition and evaluation of semi-automatic curation workflows

The extracted metadata from **Action 2.4** marked as uncertain will be evaluated by domain experts BonaRes Repository and OpenAgrar. They will decide if the extraction of the entity was accurate or not and correct it if necessary. By approaching this task in a structured and reproducible manner, it will create insight on the time and resources needed, enabling estimations for possible future use-cases and RDIs (see **Action 3.3**).

Action 3.2: Develop plan for ingesting enriched metadata into the RDIs

Once the extracted metadata has been evaluated as correct, **Action 3.2** will explore ways of ingesting it back into the respective RDI, to make it available for the users of the services. To make this possible, metadata models of the RDIs have to be re-evaluated and extended. Due to the dependencies of other services, that rely on the RDIs and orientation towards developments such as FAIRagro metadata efforts, this Action can only achieve conceptional results instead of finished implementation and will be carried out in intensive cooperation between all pilot partners.

Action 3.3: Conceptualize processes for large scale metadata enrichment

Based on the experiences made and statistics collected during the previous Actions, a strategy for enriching other datasets of the RDIs is developed. The strategy will describe requirements, estimated resources and devise a timeline for implementing the developed metadata enrichment processes on a large scale. The outcome will be published and not only support the co-applicant RDIs in enriching the rest of their collections but can also act as a basis for developing similar approaches for other FAIRagro RDIs.

5.2. Milestones

Please define milestones for the whole UC n and add the time in months needed for completion.

A Milestone is an event or achievement that marks an important stage in a process in respect to the UC Project or UC Pilot.

#	Milestone *(Name and description)*	Achieved after x month
M1	*Text corpus is manually annotated following developed annotation guidelines*	*After 3 months*
M2	*Tests on automated metadata extraction and annotation are completed and enriched metadata is available*	*After 9 months*
M3	*Strategy and semi-automated processes on metadata extraction are designed and communicated*	*After 12 months*

5.3. Deliverables

Please define deliverables for the whole UC and the expected time in months until the delivery.

A Deliverable is an object e.g. publication, report, event, tool, software) produced as a outcome of the UC pilot or UC project.

#	Deliverable *(Name and description)*	Achieved after x month
D1	*Report on manually annotated training corpus is published*	*After 4 months*
D2	*Report on test results of automated metadata extraction and annotation are evaluated and published*	*After 10 months*
D3	*Strategy on semi automated metadata extraction and curation processes is published*	*After 12 months*

6. Cost planning

6.1. FAIRagro funding

Please define the types of funding and name the institution and estimate the respective costs.

The usage guidelines of DFG for NFDI apply here, which states that e.g. basic equipment is not fundable: https://www.dfg.de/formulare/nfdi300/v/dfg_nfdi300_de_v0721.pdf

*[**Note:** In case of the approval of the UC application, the cost planning will be revised according* to the FAIRagro personal-, travelling and material rates (can be requested).*]*

Please provide a short but specific description:

Type of Funding	**Description/ justification**	**Institution**	**Estimated costs [€]**
Personnel	Research assistant (15h/week for 12 months) for actions 1-3. The position is the main responsible staff for setting up the annotation tool, the text mining pipeline with all its components, training the models and making the results available for the RDIs.	ZB MED	12.244,32 €
Personnel	Research assistant (9,75h/week for 6 months) for actions 1 and 3. The position is the main responsible staff for annotation work and evaluation of semi-automatic curation workflows regarding the training data from OpenAgrar	JKI	4.000 €
Personnel	Research assistant (9,75h/week for 6 months) for actions 1 and 3. The position is the main responsible staff for annotation work and evaluation of semi-automatic curation workflows regarding the training data from the BonaRes Repository	ZALF	4.000 €
Total estimated funding [€]			20.244,32 €

6.2. UC in-kind contribution

If applicable, please list and define the in-kind contributions from the UC (co-)applicant institution(s) and associated partners.

Please provide a short but specific description:

Type of in-kind contribution	**Description/ justification**	**Institution**	**Estimated Value [€]** *(optional)*
Personnel	Coordination between project partners and supervision of research assistant staff	ZB MED	
Personnel	Researcher, Defining text corpus selection criteria and annotation guidelines in close collaboration with ZBMed and ZALF, supervision of Research Assistant	JKI	

Personnel	– Data steward: defining annotation guidelines; provide domain expertise to evaluate the derived annotations; – Researcher: supervision of research assistant;	ZALF	

7. Policies and Permission rights

This section contains an overview of the consent, policies, and permissions rights that the submitter is consenting to when applying with the Use Case submission form.

Acceptance of FAIRagro Use Case application procedure

By submitting your Use Case application, you hereby expressly confirm that you have read and understood the FAIRagro UC call published on the FAIRagro website: ´Oboarding new Use Cases`

Only UC applications are accepted, which were submitted on time, latest until July 5 2024 6pm to UC_fairagro@listserv.dfn.de. The use of the FAIRagro UC Submission form and its completeness is a prerequisite for the formal acceptance of the UC application.

Acceptance of FAIRagro Privacy Policy

By submitting your Use Case application, you hereby expressly agree to the FAIRagro Privacy Policy.

We will process the following data from you:

- Surname, first name
- contact details
- Description of the use case and other related information, as shown in the form
- Further comments from you

We use this data to check your submission for suitability and to contact you if necessary for consultation and queries. In the course of the evaluation, the FAIRagro Community Advisory Board (CAB) and all members of the FAIRagro Steering Committee (SC) will have access to your submission in order to evaluate it. If your submission is successful, we will process your data for project implementation. The legal basis for data processing is Art. 6 para. 1 lit. b GDPR. If we accept your Use Case, we will retain your data until the end of the project in 2028. In the event of rejection, we will delete your data immediately.

Publication on the website

If your use case submission is successful, we will publish your name and details (UC title, participating institutions, UC summary and, if available, image/graphic) of your use case on the FAIRagro website. The legal basis for data processing is Art. 6 para. 1 lit. f GDPR. We have a legitimate interest in the transparent presentation of our activities. Once the project has been completed, we will cease publication.

You have the right to object to the processing (Art. 21 GDPR), your particular situation causes you to object to the processing for this reason, as you consider your fundamental rights and freedoms to be impaired. In some cases, we may demonstrate compelling legitimate grounds for the processing which override your rights and freedoms.

Upon successful acceptance, we will reach out to you for specific input for your Use Case subpage and graphic(s) and or pictures.

By submitting your Use Case application now, you agree to further communicate with FAIRagro to provide curated information about project details of this Use Case and you agree to openly publish this on the FAIRagro webpage using a Creative Commons by 4.0 licence thus ensuring that the FAIRagro community can effectively contribute to this Use Case. You will receive assistance from FAIRagro in this regard.

8. Appendix

Please check mark, list and specify the references you would like to add to your UC application.

- ☰ Letter of Commitment (LoC) of all institutions involved as applicant and co-applicant(s)

Letter of Commitment

ZB MED plans to submit an UC for the project “FAIR Data Infrastructure for Agrosystems” (FAIRagro, 2023-2028). The FAIRagro consortium is part of the National Research Data Infrastructure (NFDI) and is organized in five Task Areas (TA) and 25 Measures.

As Main UC applicant of the pilot **“Increasing FAIRness of FAIRagro data through AI supported metadata enrichment” ZB MED** will actively lead the proposed work throughout the entire UC running time and contribute the following appropriate collaborative measures:

- Coordination of the project with involved co-applicants
- Expertise and support in creating annotation guidelines and resources in development of a manually annotated text corpus
- Design, implementation and execution of an AI model based text mining pipeline to provide the results to co-applicants
- Guidance and support in developing curation processes for uncertain extracted entities and coordination for developing a large scale metadata enrichment strategy

Place, Date: Bonn, 04.07.2024

Prof. Dr. Juliane Fluck

Main UC applicant

Deputy Scientific Director, Head of Knowledge Management

Letter of Commitment

ZB MED – Information Centre for Life Sciences plan to submit a Use Case Pilot for the project "FAIR Data Infrastructure for Agrosystems" (FAIRagro, 2023-2028). The FAIRagro consortium is part of the National Research Data Infrastructure (NFDI) and is organized in five Task Areas (TA) and 25 Measures.

ZB MED – Information Centre for Life Sciences has established contact with ***Leibniz Centre for Agricultural Landscape Research (ZALF),*** represented by the executive board Prof. Dr. Frank Ewert and Martin Jank. ZALF expressly supports the Use Case Pilot application ***"Increasing FAIRness of FAIRagro data through AI supported metadata enrichment".***

Leibniz Centre for Agricultural Landscape Research, specifically the working groups *Research Data Management (FDM)* and *Data Infrastructures (DIS)* will actively participate in the FAIRagro Use Case Pilot within the limits of available resources. Throughout the Use Case Pilot's duration, they will support it through the following collaborative measures:

- Provide metadata from published datasets in the BonaRes Repository via existing harvesting interfaces to be used in text mining approaches in the Use Case Pilot.
- Offer data curation expertise for soil and agricultural research data to aid in developing annotation guidelines.
- Give feedback on the derived annotations for selected datasets from the BonaRes Repository by providing domain expertise on soil- and agricultural research data.

Leibniz Centre for Agricultural Landscape Research (ZALF)

Müncheberg, July 5th, 2024:

_____

Prof. Dr. Frank Ewert
Scientific Director

_____

Martin Jank
Administrative Director

Letter of Commitment

ZBMED plan to submit an UC for the project “FAIR Data Infrastructure for Agrosystems” (FAIRagro, 2023-2028). The FAIRagro consortium is part of the National Research Data Infrastructure (NFDI) and is organized in five Task Areas (TA) and 25 Measures.

ZBMED established contact with UC-Co-applicant institution ***Julius Kühn-Institut*** *represented by* UC-Co-applicants name/ *Heike Riegler,* expressly supports the UC application ***Increasing FAIRness of FAIRagro data through AI supported metadata enrichment.***

Julius Kühn-Institut will actively participate in ***Increasing FAIRness of FAIRagro data through AI supported metadata enrichment*** throughout the entire UC running time and support through the following appropriate collaborative measures:

The Information Centre and Library team and the Team FDM at JKI combine expertise in the agricultural domain with handling of domain and publication-specific metadata, subject indexing, the use of subject-specific thesauri and publishing research data at OpenAgrar in a FAIR manner und will therefore contribute as follows:

- Providing metadata from published datasets and text publication in the OpenAgrar Repository for text mining approaches via existing harvesting interfaces
- Providing annotation of data sets and evaluation of semi-automatic curation workflows regarding the training data from OpenAgrar with respect to publications from the crop and soil domain
- Defining text corpus selection criteria and annotation guidelines
- Giving feedback on the derived annotations for selected datasets from the OpenAgrar Repository by providing domain expertise on soil- and crop research data

UC-Co-applicant name */ **Heike Riegler***

Place, Date: Quedlinburg, 08.07.2024

_____

Name, Title