

# Streamlining Pest and Disease data

## UC3 Action 2: Towards improved FAIRness



Stefan Kühnel<sup>1</sup>, Christina Wagner<sup>1</sup>, Marissa Lorenzen<sup>1</sup>, Jörg Sellmann<sup>2</sup>, Til Feike<sup>1</sup>

<sup>1</sup>Julius Kühn-Institut – Bundesforschungsinstitut für Kulturpflanzen (JKI), Institut für Strategien und Folgenabschätzung  
<sup>2</sup>Julius Kühn-Institut – Bundesforschungsinstitut für Kulturpflanzen (JKI), Organisationseinheit Digitalisierung und künstliche Intelligenz

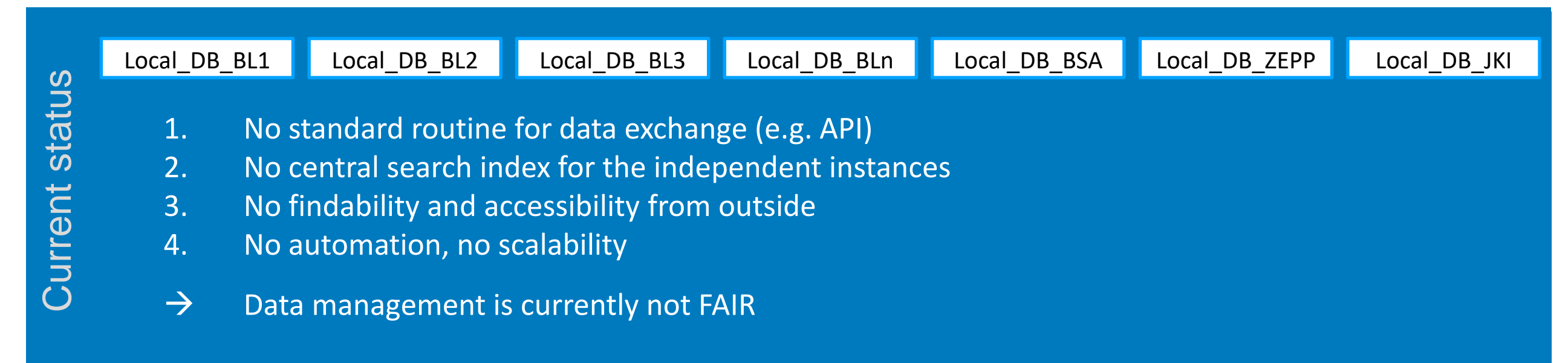
stefan.kuehnel@julius-kuehn.de

### Action 2

#### Establish inventory for and improve accessibility of IPM related data

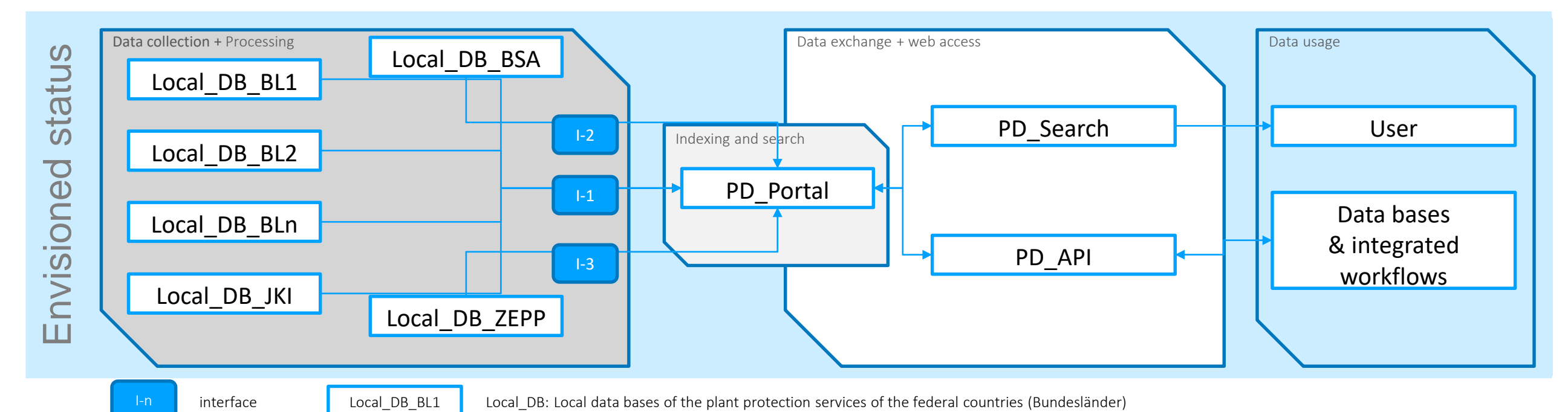
- Central, relational database for indexing (& collection) of existing data
- Develop standard procedures to deal with missing data.
- Manage licensing/ sharing arrangements with data holders.
- Mapping of data to standards vocabularies and ontologies
- Create **standardized output queries** for workflow integration

Example: Increase findability and accessibility of cultivar trial data conducted by the authorities of the federal states and stored in local data bases:



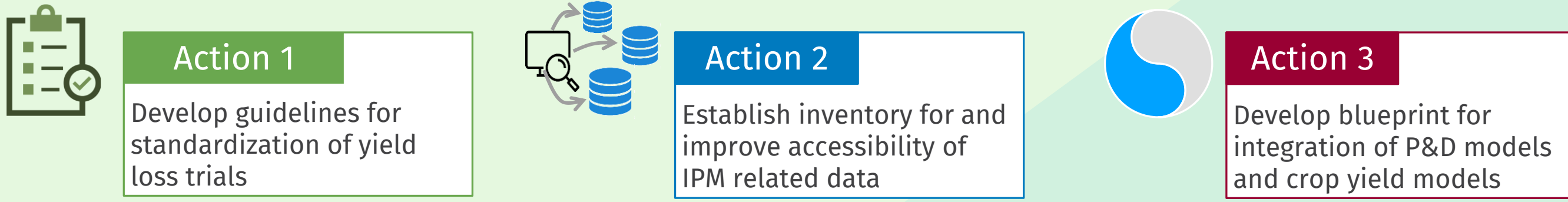
#### Metadata collection and data exchange via a new platform (Codename PD\_Portal)

Indexation or central storage of data from all independent instances of local data bases



### INTRODUCTION

Use Case 3 aims at improving pest and disease related research data management in order to enable a more effective integrated pest management. 3 Actions have been specified, out of which Action 2 is addressed here.



Local database systems are widely used by federal and regional research organisations for the collection and processing of pest and disease related data. These data are generally of high quality, given its standardized experimental design and uniform data structure. To date, no central search index exists, which would summarize the data of the 15 to 20 local, independent database instances. Thus, research data is not easily findable and accessible.

Within Action2, standard output queries are developed to export metadata sets or entire data sets into a newly developed, central database.

### Legacy Data Integration

Often patchy, fragmented, dislocated/ unaccessible and poorly described  
 Sorting the wheat from the chaff: which data is it worth to be recovered?

FAIRagro runtime

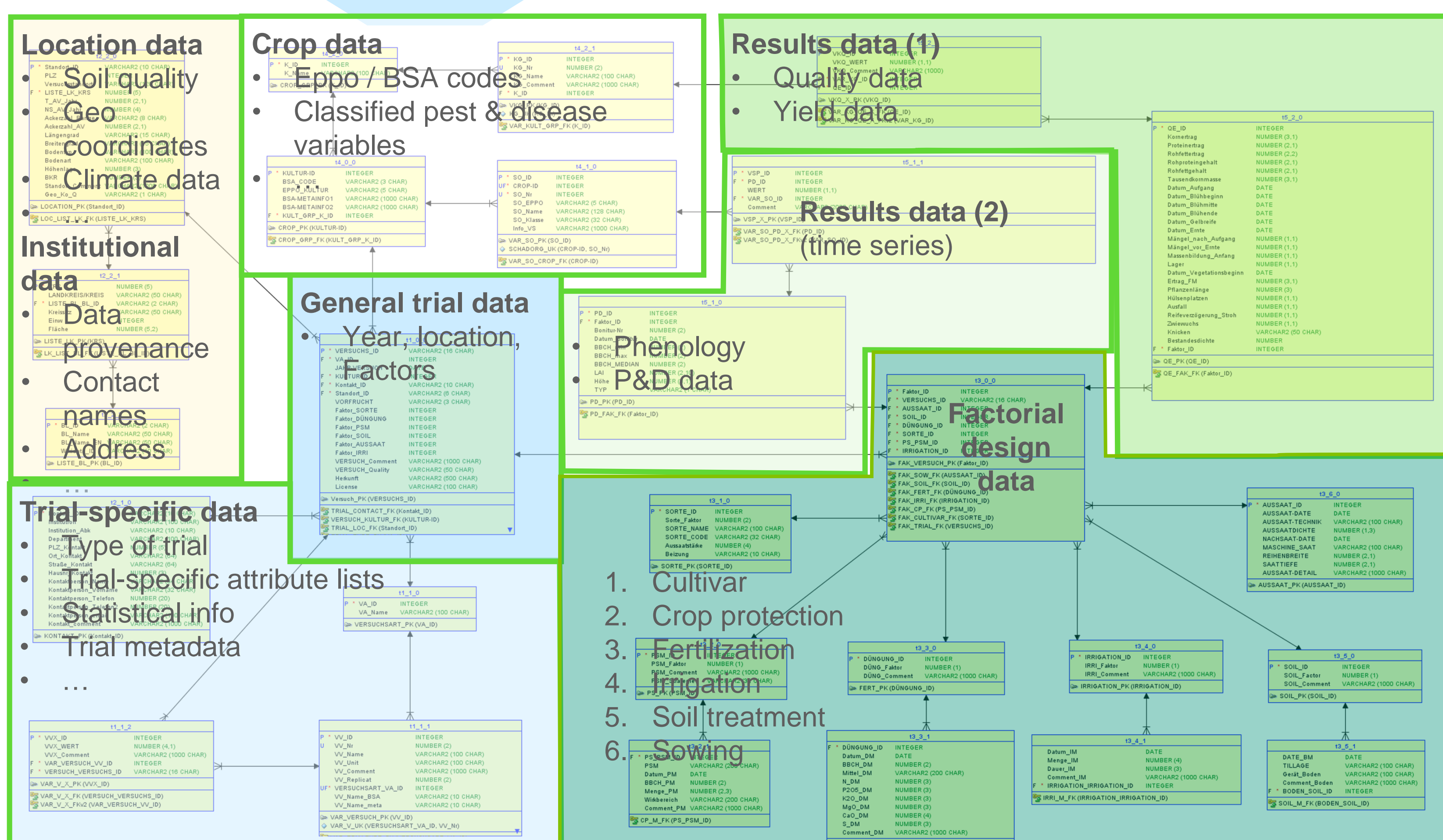
### The big agricultural data stream

#### Towards a FAIR data MGMT by design:

Data lakes, non-relational databases & ARC

- Standard folder structure for research projects using annotated research containers (ARC): (semi-) automatic generation of machine readable Metadata (e.g. xml, RDF or JSON-LD)
- Automated QC-Procedures and data quality assessment
- Authentication procedures, accessibility and licensing
- Crawling of input data and instantaneous workflow integration

### PD\_Portal: a consolidated platform to manage P&D data (Entity relationship diagram, Work in progress)



- #### Dedicated output queries
- Standardized formats and vocabularies
  - Standardized units

- #### Standardized input for workflow integration (e.g., modeling, LCA)
- Comparable
  - Reproducible
  - Fully traceable raw data

