

# Chapter 2

## From Genotypes to Phenotypes: A Plant Perspective on Current Developments in Data Management and Data Publication



Daniel Arend, Sebastian Beier, Patrick König, Matthias Lange, Junaid A. Memon, Markus Oppermann, Uwe Scholz, and Stephan Weise

**Abstract** Integrative bioinformatics aims to combine information from various sources of different data domains in such a way that a cross-domain analysis becomes feasible. With this approach, insights may be gained, which would not be possible with an analysis restricted to a single domain. For example, relationships between genotypic characteristics (genotypes) and phenotypic characteristics (phenotypes) in their environmental context (environment) could be made visible. The efficient management of such data combined with the supply of corresponding machine-readable access possibilities are essential prerequisites to achieve the outlined goal. This awareness was the nucleus for the development of the concept of data life cycles. In such a cycle, the stages of planning, collecting, processing, analysing, preserving, sharing and reusing are represented. All these steps must be considered, mapped and carried out accordingly in data management.

This chapter will discuss this data life cycle. The description of the individual steps is always based on concrete applications of a modern plant research institution and is therefore allocated to the field of plant bioinformatics. The focus here is primarily on the three data categories “genotype”, “phenotype” and “environment”. The spectrum of activities ranges from local data management to making data available in public archives and thus includes project planning, metadata definition and collection, database storage solutions, data curation processes, data integration technologies, data access interfaces as well as data reusability. The ultimate goal is to make all research results available to the public according to the FAIR principles of Findable, Accessible, Interoperable and Reusable.

**Keywords** Plant genetic resources · Biodiversity · Data management · FAIR · Data life cycle · Plants · Genotype · Phenotype · Environment

---

D. Arend · S. Beier · P. König · M. Lange · J. A. Memon · M. Oppermann · U. Scholz (✉) · S. Weise

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany  
e-mail: [scholz@ipk-gatersleben.de](mailto:scholz@ipk-gatersleben.de)

## 2.1 Introduction

This chapter is based on more than 20 years of data management experiences and activities at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK). The IPK is a leading international research institution in the field of crop plants and their wild relatives. Research focuses on the conservation of biodiversity and the performance of crop plants.

The Institute's distinguishing feature is the German Federal *ex situ* Genebank for Agricultural and Horticultural Crops. This is one of the world's largest genebanks and the largest of its kind in the European Union (EU27). The tasks of the genebank are the conservation of agrobiodiversity and the provision of plant genetic resources (PGR) for research and breeding. The IPK collection comprises about 151,000 samples, so-called accessions, which cover more than 3000 different species. The genebank represents a vault in which the biodiversity of cultivated plants is stored. To maintain this unique collection, regular multiplication trials have to be carried out. This involves recording a wide range of data, in particular phenotypic observations, but also environmental data (e.g. temperature, rainfall or UV radiation). As with all organisms, the phenotype of plants is influenced not only by the genotype but also by the environment. During cultivation, mainly phenotypic traits are recorded. In order to better understand the material, it also becomes useful to use genomic data, e.g. to explain the influence of genotypic variation on the phenotype.

While the data-side focus of the genebank has traditionally been on the passport data of the accessions and on phenotypic observation values, the extension of digital information services makes it possible to integrate data from other domains, e.g. genome or genotyping data, and thus successively develop the genebank into a bio-digital resource centre.

Concretely, in this chapter, we will discuss and include five data domains: plant genetic resources data (1), genomic data (2), genotyping data (3), phenotyping data (4) and environmental data (5). We will briefly explain how we define these terms in the following paragraphs.

1. **Plant genetic resources data:** On the one hand, this includes for each accession so-called *passport data* like country of origin, collection site, the genus/species, the full botanical name and recently also unique identifiers like a DOI (Digital Object Identifier). Furthermore, this includes *characterization data*. These data describe the phenotype and are rather stable, e.g. the properties of organs such as the ear in cereals (e.g. two-row or six-row in barley). The third part is the *evaluation data*. These are phenotypic characteristics that are collected during propagation cultivation. These include, for example, plant height, disease infestation or yield data such as the 1000 kernel weight.
2. **Genomic data:** These are, on the one hand, *sequence data* such as nucleotide sequences of entire chromosomes at pseudomolecule level, and the gene models (genes with their localization on the chromosomes, exons, introns, as well as the coding sequence and translated peptide sequence). Furthermore, it also includes

descriptive *annotations* of the structural regions of the genome like genes and their functions or information about non-coding regions such as repeats.

3. **Genotyping data:** This includes diversity information on how a specific genotype (e.g. one accession) or several genotypes (e.g. several accessions) differ from a reference genotype. The methods used to determine such differences are very diverse, e.g. SNP arrays or genotyping by sequencing (GBS). This also results in very heterogeneous data formats. One example is the so-called variant calling format (VCF). Here, the differences of several genotypes can be mapped to the reference, including qualitative assessments.
4. **Phenotyping data:** This includes all phenotypic traits that are collected outside the classical conservation cultivation in the genebank. This covers experiments both in the field and under controlled conditions, e.g. in the greenhouse.
5. **Environmental data:** These include weather data such as temperature, precipitation, humidity, wind speed or UV radiation. Furthermore, this includes data collected by environmental sensors in isolated environments, e.g. greenhouses. The data from environmental sensors complement the existing weather data and can therefore also be counted as part of it.

These characterized data domains are in the focus of the further described data management processes and systems.

## 2.2 Data Management Concepts in Plant Science

Data management plays an important role in achieving the goal to transform the IPK genebank into a bio-digital resource centre. In the beginning, data was managed analogously on paper or index cards. With the availability of computers, these were rapidly used for this purpose. In particular, database systems were identified as the more effective tool for this task. First databases were created in which different information could be stored and queried in a structured way. Often the results of scientific studies were imported and certain parts could be queried and extracted again. Unfortunately, a description and documentation of how the data acquisition was often missing. However, this is essential in order to be able to reuse the results and feed them as input into new studies.

In recent years, it has become clear that data management is a process that takes place over several stages and can be accompanied and supported by the use of databases. Ultimately, this process is transferable to all scientific fields and is currently a topic in the new scientific discipline of Data Science. Currently, this process is known as the Data Life Cycle (ELIXIR 2021) and is illustrated in Fig. 2.1. Each step of the Data Life Cycle is briefly described in the following paragraphs.

The process step **plan** defines a strategy for managing the data and documentation generated in the research projects. Consideration should be given in advance on how best to avoid problems associated with data management and to create the

**Fig. 2.1** Data Life Cycle  
adapted from RDMkit  
(ELIXIR 2021)



conditions to ensure that all research data continue to have maximum impact in science beyond the end of the research project.

Data **collection** describes the process of gathering information for specific parameters either automatically, i.e., using instruments, as well as manually. During this process, data quality must be ensured regardless of the research field.

Data **processing** is the step in the cycle where data is converted into a format to prepare it for analysis. In addition to format conversion, this stage of the process includes quality checking and pre-processing according to standardized protocols. Furthermore, poor- or low-quality data is discarded in order to create a cleaned dataset that provides reliable results.

In the **analysis** step, the collected data is examined to identify the information contained in a dataset. These investigations can be performed multiple times in the process. Specifically, the data can be analysed directly or indirect analyses can be performed by using models, for example.

Data **preservation** includes all activities necessary to ensure the safety, integrity and accessibility of data for as long as it is required. Data preservation is more than storage and backup. It prevents data from becoming unavailable and unusable over time.

In the **sharing** phase, the data is made available to others. This can be sharing with collaborative partners or publishing the data to the whole research community. It is important to note that data sharing is not the same as making data open access. It is the decision of the data producer how the data will be shared. Thus, restricted access for different user groups is also possible, e.g. only for collaborative partners.

In the **reuse** phase, data is used for a new purpose for which it was not originally intended. This makes it possible to generate and also publish new results based on the same data. Reusability is an essential part of the FAIR principles.

In addition to the steps in the life cycle, the use of data standards as well as data concepts is, of course, essential basics in data management. Furthermore, it is crucial to manage and offer data according to the FAIR data principles. First formulated in 2016 Wilkinson et al. (2016), it is now established in more or less all data domains,

and several funding agencies have also made FAIR their central paradigm (Mons et al. 2017).

Behind this acronym is a guideline for handling research data in a sustainable way.

- **Findable:** Research data needs to be findable by humans via search engines, but also by machines using standardized harvesting formats.
- **Accessible:** A long-term stable access to research data is crucial for sharing research data within the research community and public users and to get credit for the data producers.
- **Interoperable:** Standardized metadata formats are essential for describing research data to integrate them and find possible interconnections.
- **Re-usable:** In order to exploit the full value of research data, it is necessary to provide a full technical description, which guarantees as far as possible a repeatability of the underlying process to create the data and allow users to use them for further investigations.

Therefore, FAIR has also been an important goal during the development of novel standards and updating of existing formats. Some popular examples are the MIAPPE recommendations for describing plant phenotypic experiments (Papoutsoglou et al. 2020) and the MCPD standard for describing plant genetic resources (Alercia et al. 2015). MIAPPE is a descriptive list of recommended minimal attributes, which are helpful to explain and document the experimental setup of phenotypic trials. It was originally described in 2015 and is still under active development. On the other hand, the MCPD (Multi-Crop Passport Descriptors) standard is relatively old and established across genebanks and plant genetic resource providers worldwide. It provides a comprehensive list of stable and well-defined attributes necessary to document genebank accessions.

But of course, meeting the FAIR recommendations requires not only an improvement of data formats and standards but also a re-design and update of existing infrastructures and databases. One obstacle is the homogenization of the vocabularies used in these resources. It is necessary for resources that offer data for exchange to access a standardised vocabulary established by the community. To this end, consortia have been formed with the mission of building such ontologies. For example, there are specialized ontologies that offer a controlled vocabulary for describing plant structures and growth stages (Jaiswal et al. 2005) or ontologies that describe more general concepts, like the Gene Ontology (Gene Ontology Consortium 2004). However, as both language and methodologies continue to evolve, this effort must be supported and sustained.

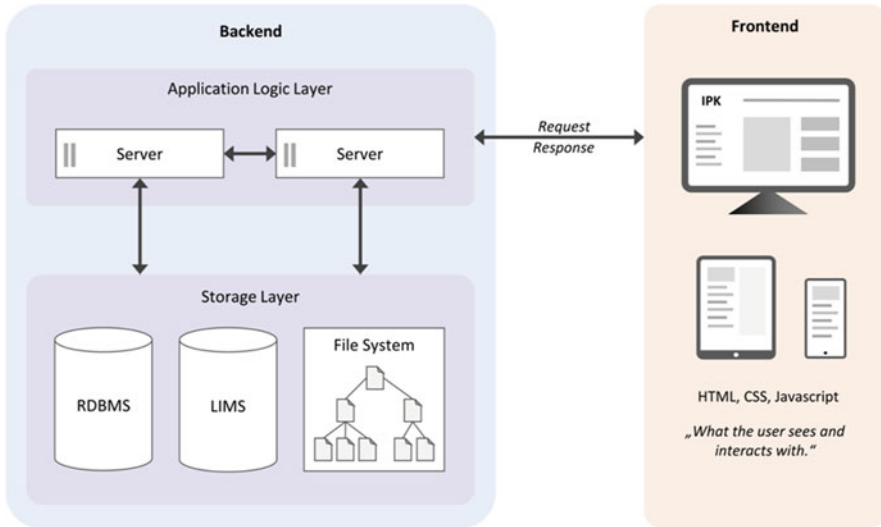
In all research areas, including the life sciences, the tasks of data management and publication are of essential importance. Only in this way can new findings be appropriately substantiated and are traceable. Initially, these tasks were performed exclusively in analogue form. With the broad emergence of computers, it became digital. Along with the triumph of the World Wide Web, these two tasks have received a considerable boost.

## 2.3 Overview of Information Systems

The general architecture of information systems can be divided into two distinct entities: (1) the backend consisting of database management systems (DBMS) comprising application logic, and (2) the frontend, which usually serves as the primary interface for user interaction (graphical user interface). Other solutions have been proposed in so-called tier approaches, where the number of different entities is either reduced for simple applications (all-in-one approaches) or drastically increased for complex applications (n-tier approaches) (Petersen 2001). For the purpose of this chapter, we will focus and discuss the two-tiered approach, which is often also referenced as the client–server architecture.

In information systems, the backend is often synonymous with the database, which the user accesses only indirectly (note that the discussed information systems of the IPK have a more direct access solution integrated). Primary data and metadata are stored and managed here. The DBMS is the software layer of the backend, and one of its tasks is to handle authorization and authentication and thus controls the granularity of data retrieval for specific user groups. For user updates or changes to records in the database, the DBMS is able to enforce constraints that ensure consistency rules are followed. Databases implement different data model and feature paradigms, and have evolved to support application scenario, with relational databases being the dominant class overall (Harrington 2016). The data is accessed indirectly either via application programming interfaces (API) or via special application logic through stored procedures, the specifics of indirect accesses are varied and going into detail here would go far beyond an overview of information systems.

In addition to information systems per se, so-called web-based information systems are playing an increasingly important role. In such systems, the front end is based on web technology. This means that the user interface is a web browser or is accessible via the WWW. Usually, the business logic of such a web application is implemented in a programming language suitable for the use case and deployed by assigning URLs to specific functions or methods. The programming language itself may implement or provide the required HTTP server, or a separate HTTP server such as Apache HTTP Server, Apache Tomcat, or Oracle WebLogic may be used upstream. Classically, business logic often communicates with the storage layer over an internal private network using protocols based on TCP/IP, sockets, or the file system. In complex web information systems, different persistence technologies are often used simultaneously in the storage layer, e.g., in web-based information systems that combine multiple databases in a single web application (Fig. 2.2). The data ingestion and management into the backend of scientific institutions is often realised by a laboratory information management system (LIMS). Its main purpose is to act as a sample management system, but recently data analysis functions and the ability to record digital laboratory documentation (also known as an electronic laboratory notebook or ELN) have also been integrated into some successful LIMS



**Fig. 2.2** Abstract architecture of web-based information systems

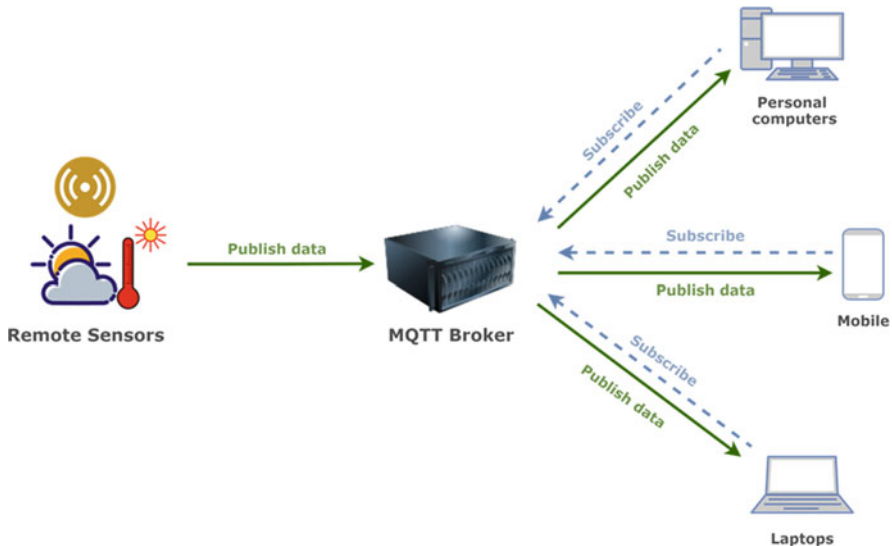
solutions. The actual LIMS implementation of the IPK Gatersleben is described in (Ghaffar et al. 2019).

A further component of backend technology is sensorics data which is mainly used to collect environmental data. The basic idea of sensor networks is based on the idea of the Internet of Things (IoT) (Madakam et al. 2015). Individual sensors are to be networked in a computational interconnected infrastructure. This concept is based on ubiquitous computing. This in turn describes a concept that moves away from the use of one end device to the use of many. This concept thus contrasts with dedicated, application-specific platforms that are designed and installed to combine data collection, storage, exchange and evaluation in one overall system. IoT goes far beyond the original concept of the internet. It is no longer just a network of different computers, but a network of all kinds of devices. Rather, the IoT is a network consisting of different objects that communicate over the Internet to collect and exchange data. This includes both actuators, as a component of a machine that is responsible for moving and controlling a mechanism, and sensors, which detect events or changes in its environment. Some examples of sensors are cameras, weather stations, ground sensors or airborne remote sensing, such as drones or satellites. Active elements are irrigation pumps, fans, lighting or even cooling or heating elements. Usually, both types are combined, like in agricultural machinery and greenhouse controls. These capabilities to build IoT networks are increasingly influencing the nature of experimentation. For example, the detection of phenotypes via sensors is being combined with targeted manipulation of the environment in the field of high-throughput plant phenotyping and breeding research (Fiorani and Schurr 2013; Watt et al. 2020). The concrete interaction of sensors and actors is a practical and technical challenge in terms of system integration that is not to be

underestimated. This is a practical hurdle, because infrastructures that span locations and organizations sometimes use highly heterogeneous interfaces and incompatible systems infrastructures. The homogenization of data formats is done by applying standards as described in Sect. 2.2. The homogenization of transmission protocols plays another central role here, as sensor data are continuous data streams. This affects, among other things, the protocols as well as data exchange formats and units. Application-specific network protocols are the backbone of IoT networks and are responsible for the communication of remote sensors. One of the commonly used network protocols is MQTT (Message Queuing Telemetry Transport). It is a lightweight protocol used to transport data between devices mainly on TCP/IP networks. It was jointly authored by Andy Stanford-Clark (IBM) and Arlen Nipper (Cirrus Link, then Eurotech) in 1999 (MQTT.org 2015). MQTT is an M2M (Machine to Machine) protocol best suited for the remote connections which require a “small code footprint” or in cases where the network bandwidth is limited, such as IoT devices. The publish-subscribe architecture of MQTT described in Obermaier (2018)) and illustrated in Fig. 2.3 is extremely lightweight compared to HTTP’s request/response paradigm.

Where MQTT broker is the central component of the paradigm that acts as a server responsible for passing the messages between the publisher and subscribers. In case of an event, the publisher first transmits the data to a broker with a topic, and if a client requests data of a certain topic, the broker performs matching and then delivers messages accordingly.

Another important layer in information systems is the frontend. It is considered to be everything the user sees and interacts with directly. Especially in web-based



**Fig. 2.3** Principle operation of the MQTT publish-subscribe architecture



information systems, the website rendered in the browser acts as the graphical user interface. The spectrum here ranges from the pure display of data, stored in the storage layer, to highly interactive “Rich Internet Applications” (RIA) (Fraternali et al. 2010), which very often also contain a large proportion of business logic in the form of Javascript. For example, the display of interactive diagrams always requires the use of business logic in the frontend code. In contrast, for the display of text, tables and static images, only the use of HTML and CSS is mandatory. In the age of mobile devices such as smartphones and tablets, the flexibility of the website layout plays an increasingly important role. The necessary flexibility results from the many different display sizes and page formats of mobile devices compared to traditional PC monitors. Therefore, the development is increasingly moving away from static, fixed layouts to so-called adaptive and responsive layouts that adjust as optimally as possible to the different display sizes and page formats.

## 2.4 Selected Data Management Information Systems

The following provides a general overview of some well-known information systems and data warehouses with a focus on plants developed and hosted at our research institute IPK Gatersleben. The description of each system includes the features and architecture, scope and general use cases. The web address where the system can be accessed is stated, as well as the supported data domains. Also explained is how data can be imported and exported and where the system fits into the Data Life Cycle.

### 2.4.1 *The Genebank Information System (GBIS)*

Globally, genebanks play an important role in the long-term conservation of plant genetic resources (Hoisington et al. 1999). They complement the conservation of biodiversity in farmers’ fields and in nature. Besides the preservation of physical samples, data management is one of the most important tasks of a genebank and at the same time one of its greatest challenges (FAO 1997, 2010; Fowler and Hodgkin 2004; Weise et al. 2020). Well-structured documentation of all data and information available on a genebank accession is the basic prerequisite for genebanks to be used. A wide range of data must be taken into account.

The IPK genebank has been in existence for almost 80 years, but is partly based on even older collections, so that material from a period of almost 100 years is preserved. The focus of the documentation has continuously developed over this period, as have the technologies used for this purpose. Furthermore, a number of changes in organizational structures have taken place, and several generations of curators and scientists have maintained the material and constantly added further

parts to the collection. Continuous documentation is indispensable for both the preservation and the exploitation of the material.

The Genebank Information System (GBIS)<sup>1</sup> (Oppermann et al. 2015) is one of the central instruments for documentation and management in the IPK genebank. It was first introduced in 2006 and has been continuously developed ever since. The core of the GBIS is formed by an OnLine Transactional Processing (OLTP) system, which records the data produced in various genebank workflows. This data is compiled into an overall dataset that includes the following areas:

- Pure management data for conservation of collections.
  - Storage quantity and locations.
  - Growth and harvest management.
  - Germination rate, age of the samples, health tests.
  - Reporting and labelling.
- Data of legal significance.
  - Collection permits.
  - Correspondence, documentation of receipt.
- Data to assess the value of the resource.
  - Basic (passport) data.
  - Phenotypic observations.
  - Images of specimens, plants, fruits and seeds.
  - Comprehensive genetic data.

GBIS consists of two areas: (1) a public information and ordering system and (2) an internal system that serves data management and process support. An Oracle DBMS is used for data management; the various application components are implemented both as application server-based web applications and as standalone solutions. Figure 2.4 shows the architecture of the overall system.

From the original idea of documenting, cataloguing and describing plant genetic resources, genebank information systems are increasingly developing into instruments for scientific work and thus reflect the transformation of genebanks into bio-digital resource centres.

GBIS supports all steps of the data life cycle.

---

<sup>1</sup> <https://gbis.ipk-gatersleben.de/>

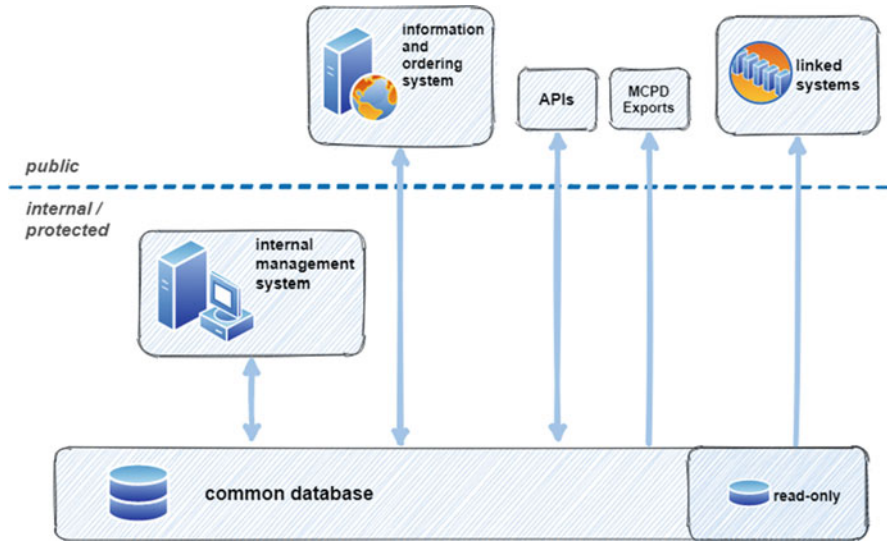


Fig. 2.4 Architecture of the genebank information system

### 2.4.2 *The European Search Catalogue for Plant Genetic Resources (EURISCO)*

Estimates put the number of genebanks worldwide at around 1800, with more than 600 in Europe (Engels and Maggioni 2012). Many genebanks have been in existence, in some cases for decades. Despite the introduction of IT support, especially in the late 1960s and 1970s, most genebanks remained largely isolated from each other. This did not change until the 1980s, when the first attempts were made to make information available across genebanks. It was then that the idea of Central Crop Databases (CCDB, Gass et al. 1997) was born. This idea consisted of strengthening cooperation between genebanks by networking the collections and also making genebank material more accessible to users as well as identifying possible duplicates between the individual collections. However, due to the low quality or lack of data, these goals could only be achieved to a limited extent (van Hintum 1997). One of the biggest difficulties in this context was that for a long time there were no uniform standards for the description and exchange of passport data. A standard that addressed this challenge is the Multi-Crop Passport Descriptors (MCPD). After the presentation of the first draft in 1997 (Hazekamp et al. 1997), the MCPD successively developed into a globally accepted and used standard (Alercia et al., Alercia et al. 2001, 2015). The emergence of MCPD as well as Darwin Core (Endresen and Knüpffer 2012; Wiczorek et al. 2012) represented milestones for

the development of international aggregator systems such as WIEWS,<sup>2</sup> EURISCO<sup>3</sup> or Genesys.<sup>4</sup> They enable the exchange of passport data between genebanks and these systems and thus allow a cross-genebank search for accessions of plant genetic resources.

One of the aggregator platforms mentioned is the European Search Catalogue for Plant Genetic Resources (EURISCO, Weise et al. 2017). This platform is operated within the framework of the European Cooperative Programme for Plant Genetic Resources (ECPGR)<sup>5</sup> and has been available online since 2003. The aim of EURISCO is to provide a central entry point for searching accession-specific passport data and phenotypic data on plant genetic resources accessions maintained in Europe. In addition, EURISCO assists its member countries in fulfilling national obligations, e.g. to the FAO. The majority of European *ex situ* collections are represented in EURISCO. A total of 43 countries are currently part of the EURISCO network. Each country compiles the data of its genebanks in a National Inventory and submits it to EURISCO on a regular basis. The MCPD standard is used for the passport data. Currently, more than two million genebank accessions from about 400 collections are documented in this way in EURISCO, covering more than 6700 genera and 45,000 species. In recent years, work has begun on depositing phenotypic observations collected on accessions in EURISCO in addition to the passport data. Unfortunately, there are no really widely accepted standards for the exchange of phenotypic data so far (Krajewski et al. 2015). This is complicated by the fact that observation values of genebank accessions were partly collected over long periods of time. Various initiatives to harmonize such data have existed since the 1970s, e.g. the IPGRI/Biodiversity descriptor lists (IBPGR 1990; International Board for Plant Genetic Resources (IBPGR) and Commission of the European Communities (CEC) 1984; IPGRI et al. 2001), but they have never achieved general acceptance. More recent approaches aim at mapping different traits and methods onto each other using ontology terms, e.g. CropOntology (Shrestha et al. 2010, 2012), or to put a stronger focus on the description of the material used and the experiments conducted, e.g. MIAPPE (Ćwiek-Kupczyńska et al. 2016; Krajewski et al. 2015; Papoutsoglou et al. 2020). Altogether, this represents a particular challenge that has not yet been conclusively solved. EURISCO uses a minimum consensus approach for exchanging phenotypic data, which is limited to the absolutely necessary data fields.

The provision of data in EURISCO is done using a multi-tier system (Fig. 2.5). The data compiled in the National Inventories is imported into a central staging area through an upload tool. A series of data integrity checks are then performed, most of them at syntactic level, some also at semantic level. Automatically generated error reports help the data providers to successively correct data errors. After release by

---

<sup>2</sup> <http://www.fao.org/wiews/>

<sup>3</sup> <http://eurisco.ecpgr.org/>

<sup>4</sup> <https://www.genesys-pgr.org/>

<sup>5</sup> <https://www.ecpgr.cgiar.org/>

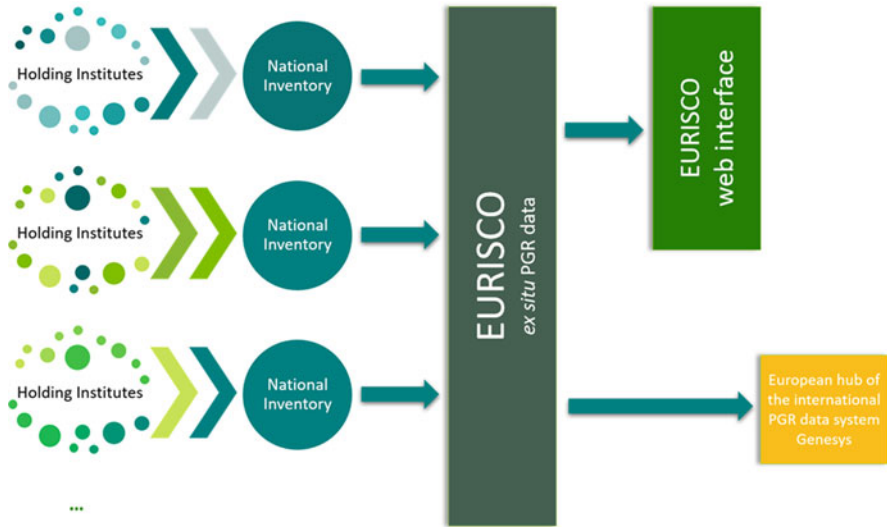


Fig. 2.5 Overview of the EURISCO architecture

the data providers, the new data is integrated into the overall EURISCO database. A web interface is available to the users of the system, which offers a variety of search, visualization and download options. Fuzzy searches are also supported here, for example in the case of scientific plant names and their synonyms (Kreide et al. 2019).

In addition, EURISCO forms the European hub of the international PGR information system Genesys. Passport data is regularly exchanged with Genesys, so that genebank accessions documented in EURISCO can also be found via the Genesys portal.

In terms of the Data Life Cycle, EURISCO can be assigned to the categories Preserve, Share and Reuse.

### 2.4.3 BARLEX

Sequencing and subsequent steps to reassemble the underlying genome sequence for complex plant species have been a lengthy and costly endeavour. Sequencing efforts in the species barley (*Hordeum vulgare* L.) were initiated more than a decade ago (Schulte et al. 2009). At that time, the state-of-the-art approach was to create a comprehensive physical map of overlapping BAC clones carrying small fragments of genome information (Ariyadasa et al. 2014; Schulte et al. 2011), sequence them using NGS technology (Stuernagel et al. 2009; Taudien et al. 2011), and then join the assemblies of the individual BAC clones with mate-pair reads (Beier et al. 2016).

The Barley Explorer (or BARLEX<sup>6</sup> for short) web application was developed to facilitate the process of joining these BAC assemblies (Colmsee et al. 2015). It showed interested users evidence of overlap between adjacent BAC assemblies and all available genomic data associated with each sequence contig. This information was presented in both tabular form and in an interactive graphical edge-node display.

Since its inception in 2015, BARLEX has evolved into the de facto hub for barley genomic sequence information (Beier et al. 2017). With the advent and adaptation of advanced sequencing and assembly techniques such as conformation capture sequencing (Lieberman-Aiden et al. 2009), incorporation of optical mapping (Staňková et al. 2016) or 10X Genomics linked reads (Mostovoy et al. 2016), the speed and accuracy of new complete pseudomolecule sequence assemblies have increased dramatically (Jiao and Schneeberger 2017). To date, new and updated reference barley genome assemblies have been released in 2012 (Mayer et al. 2012), 2017 (Mascher et al. 2017), 2019 (Monat et al. 2019), and 2021 (Mascher et al. 2021), with more than a dozen genotypes being sequenced at the moment to complement pan-genome sequencing efforts (Jayakodi et al. 2020, 2021). The pseudomolecule sequence, genomic scaffold structure, and molecular marker, repeat, and gene annotation (complemented by expression data) for these four different versions of the reference sequence are all available in BARLEX.

BARLEX is built on an Oracle relational database backend and consists of 57 tables, 17 materialized views, 37 stored procedures and more than 95 million rows of data. The web application is implemented with Oracle Application Express (APEX, formerly known as Oracle HTML DB) with custom Javascript procedures. Some of these Javascript procedures use the Cytoscape.js framework (Franz et al. 2016) which enables a graph-based interactive visualization. Additional functionality is supported by cytoscape-qtip, cytoscape-automove, cytoscape-cose-bilkent and cytoscape-context-menus, which help to make the user interface more intuitive and accessible. Tabular data within BARLEX can be exported in various predefined formats such as CSV and HTML. Please note that the download of gene or repeat annotations has been disabled in BARLEX and is distributed via links to long-term stable DOIs deposited at eDAL-PGP (Arend et al. 2016). The import of new data into BARLEX is done via semi-automatic import scripts by the BARLEX team. After manual curation of the data and transformation into the appropriate format, the data is fed into the database using an upload tool (Rutkowski 2005). This manual curation step includes both syntactic and semantic verification. Although ordinary users cannot modify the data via the web application, there is an option to leave feedback on all features and records so that administrators can be notified on feature requests and data inconsistency.

Many types of genomic data are represented in BARLEX, such as sequencing contigs (various technologies and methods), exome capture data, molecular marker data (array-based SNPs), expression data (from Iso-Seq and RNA-Seq), BLAST results, structural information about sequence composition, and sequence order

---

<sup>6</sup> <https://barlex.barleysequence.org>

and orientation in the finished pseudomolecules. With these data domains and the functions BARLEX supports, BARLEX covers the Analyse, Preserve, Share and Reuse fields in the Data Life Cycle.

#### **2.4.4 BRIDGE**

Although a wide diversity of landraces and PGRs are stored in genebanks, there has been little success in utilizing them and incorporating them into breeding programmes. One of the challenges here is the availability of information on molecular and phenotypic profiles of the entire seed stock. Apart from the fact that transferring beneficial alleles from PGRs to modern elite varieties is a challenge in itself (Wang et al. 2017), this availability is a prerequisite for incorporating PGRs into commercial plant breeding. Therefore, genebanks have begun to systematically categorize and catalogue their germplasm collections at both the molecular and phenotypic levels (Mascher et al. 2019; Romay et al. 2013). An example of one of these pioneering projects was carried out on the crop barley, where 22,626 accessions of the genebank hosted at the IPK Gatersleben were surveyed and analysed based on genotyping-by-sequencing (Milner et al. 2019). The resulting molecular profiles could now be combined for the first time with passport data, historical and newly collected phenotypic data to draw conclusions about the global barley diversity and to find interesting genes and loci for plant breeding. This information resource was adapted into the web portal BRIDGE<sup>7</sup> (König et al. 2020).

BRIDGE is both a data warehouse and exploratory data analysis tool for large-scale barley genomics. Through a unified collection manager for user-defined germplasm datasets, various analyses can be performed or visualized. One of the core features is the quick selection of collections either using the lasso selection tool on the provided graphical output or by setting different filters over the complete set based on passport data, phenotypic traits or molecular markers (SNPs). In addition, BRIDGE uses a concept known as “interactive brushing and linking”, where changing parameters in one visualisation results in a direct response in other visualisations that are dynamically linked (Keim 2002). All of this supports the interactive user experience and enables rapid analyses of more than 9000 data points of phenotype data, visualisation of genetic diversity by PCA and t-SNE, or exploration of Manhattan plots to genome-wide association studies. Also integrated is an intuitive variant browser for the study of SNP data based on the GBS sequence data of 22,626 genotypes. Genotypic data can be exported in VCF files (Danecek et al. 2011) for custom collections of genotypes and specific genomic regions of interest, e.g. for a whole gene or single exons. The Java library “isa4j” (Psaroudakis et al. 2020) is used to realise a customised export of phenotypic data in the ISA-Tab format (Sansone et al. 2008, 2012). Based on the user’s custom collections

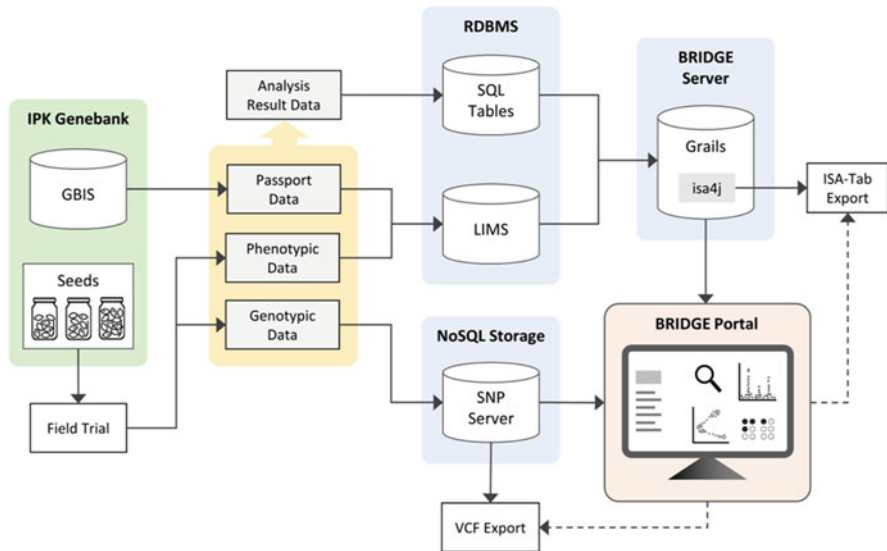
---

<sup>7</sup> <https://bridge.ipk-gatersleben.de>

of genotypes, a ZIP-archive containing ISA-Tab formatted text files and additional phenotypic images is generated on the fly for download via the user’s web browser.

The BRIDGE web portal utilises the client–server model as general architecture with REST-like HTTP-APIs as the communication layer between client and server. HTML5, CSS3 and Javascript are used for client-side development. Groovy, Java and Python programming languages are used to implement the server-side counterparts of the HTTP-API. The web application framework “Grails” is used to implement all aspects except the server-side API of the integrated SNP browser. The server-side part of the SNP browser is implemented in Python using the Flask library. Well-established libraries like Numpy (Harris et al. 2020), Pandas (McKinney 2010), Zarr (Miles et al. 2020) and Scikit-learn (Pedregosa et al. 2011) are used for the performant handling of large SNP data matrices and scientific computing aspects like calculation of minor allele frequencies or principle component analysis.

All passport and phenotypic data are provided via the IPK-LIMS through project-specific logical relations to GBIS. Analysis result data like the outcome of GWAS or PCA is stored in standard tables in the Oracle RDBMS (Fig. 2.6). The import of analysis result data is performed via customised import scripts by using CSV files. Data of genomic diversity is imported by the conversion of VCF files to Zarr archives that are then used by the server-side part of the SNP browser. The VCF files can be optionally annotated by SnpEff (Cingolani et al. 2012). The import of



**Fig. 2.6** Overview about the general architecture and data flow in BRIDGE



gene annotations is performed by directly using GFF3<sup>8</sup> files. Data that gets imported into the system is checked automatically for syntax errors. The responsibility for the general plausibility of the data belongs to the data provider who wants to present his project data in the portal. The process of data import is managed by the administrator of the web portal. As BRIDGE was designed to present the results of specific genebank genomics projects, it is currently not possible, nor is it intended, for end users to import and view their own data in the portal.

Regarding the Data Life Cycle, BRIDGE can be assigned to the categories Collect, Process, Analyse, Preserve, Share and Reuse.

### 2.4.5 *e!DAL-PGP*

The FAIR data principles are widely accepted by the scientific community for supporting long-term stable research data handling. Although established infrastructures such as the ELIXIR Core Data Resources and Deposition Databases provide comprehensive and stable services and platforms, a large quantity of research data is still inaccessible or at risk of getting lost. Currently several high-throughput technologies, like plant genomics and phenomics are producing research data in abundance, the storage of which is not covered by established databases.

The eDAL-PGP<sup>9</sup> (Plant Genomics and Phenomics) research data repository is a comprehensive infrastructure providing diverse datasets of plant-related research data. It has no general data type or data volume limitations, and therefore, it provides genomic sequences, phenotypic images, metabolite profiles and also research software and scripts. It started in productive mode in 2016 (Arend et al. 2016) and based on the previously developed JAVA-based eDAL infrastructure<sup>10</sup> (Arend et al. 2014), which follows an “infrastructure to data” (I2D) approach to provide an on-premise data management and publication system. This approach can in comparison to the common data publication-as-service model also feature a FAIR data publication culture, but it differs in costs and effort for establishment and maintaining (see Fig. 2.7).

The data publication-as-a-service model usually costs a fee, needs data property control and provides storage capacity limits. In contrast, the data publication premises model keeps data in-house and can use internal server and storage hardware by installation of the eDAL software. The fully embedded data submission and review process allows to easily store and publish research data by using persistent DOIs. To make the data FAIRly available, eDAL supports several relevant features (Arend et al. 2020).

---

<sup>8</sup> <https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md>

<sup>9</sup> <https://edal-pgp.ipk-gatersleben.de/>

<sup>10</sup> <https://edal.ipk-gatersleben.de/>

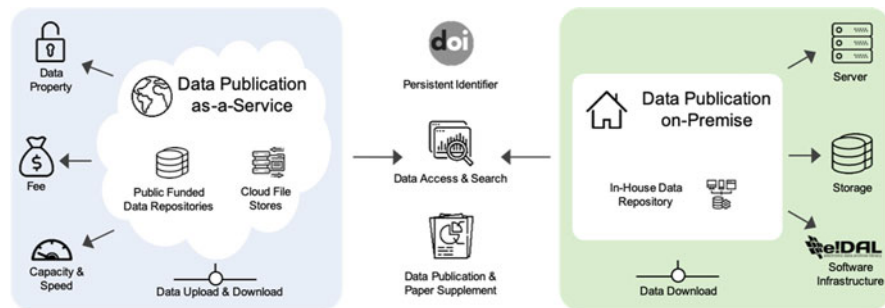


Fig. 2.7 Different data publication approaches

Providing machine-readable metadata, which are based on the DublinCore standard and are automatically embedded into the provided content pages of every published dataset, e!DAL guarantees that the contained research data is easily **FINDABLE** by common search engines. By using the well-established DOIs, the findability is further increased due to the diverse network and interactions of the DataCite services like ORCID or CrossRef. Furthermore, DOIs are persistent and guarantee the long-term stable **ACCESSABILITY** of published datasets. The DOI resolver provides simple access to the referenced datasets, e.g., in a research or data publication, even if the physical location of the underlying data changes over the time. On top of this, the e!DAL's web server takes care that the datasets are accessible via comprehensive content pages, which allow users to navigate through the dataset and download certain files or metadata. The content pages not only provide the metadata directly on the page, but also embed the metadata in the sources to provide the **INTEROPERABILITY** of the datasets. By using standardised schema (Guha et al. 2016) and format (Lanthaler and Gütl 2012) the information about relationship between datasets can be aggregated. The DublinCore is well-established and therefore e!DAL guarantees the long-term stable **REUSABILITY** of the datasets by collecting a minimal set of technical metadata, which are crucial to open and read the data files. The support of different licences makes it easy to clearly define by whom and how the datasets can be used.

The success of I2D Approach is shown by the constantly increasing number of datasets, accesses and downloads of the e!DAL-PGP instance. The comprehensive functionality of e!DAL as well as the simple installation and configuration, e.g. by using powerful and user-friendly infrastructures such as the ELIXIR AAI, are the reasons that in the meanwhile further institutional instances based on e!DAL were planned or already established. Nevertheless, even if scientists are getting more opportunities to exchange their research data within the community, the incentive is still quite low (Cousijn et al. 2019). The procedure of data publication and citation is in contrast to the established peer-review process for research articles not very common (Tenopir et al. 2015), which has of course cultural reason, but also technical limitations (Parsons et al. 2019). Beside the commercial Data Citation Index, also some open, community-driven initiatives like Make Data Count

(Cousijn et al. 2019) were developed to overcome these limitations and improve the incentives for researchers. Additionally, more and more publishers demand authors reference their research data as citations in the common reference list of their articles (SciData Editorial 2019). All these developments will help to increase the acceptance of research data as an important scientific asset and to establish a FAIR research data publication culture.

#### ***2.4.6 The IPK Weather Database: Collection and Provision of Meteorological Data***

We encounter weather data every day and they often seem trivial. However, they are essential for interpreting the results of field trials, as the expression of traits can be weather-dependent (Philipp et al. 2018). The measurement of meteorological data represents a special type of data collection, as the data is continuously recorded over a very long period of time. As a result, the processes of the life cycle from the collect to the reuse of the data take place in parallel. Another special aspect is the change in data collection and processing methods.

The long tradition in recording meteorological data is accompanied by some changes in measurement intervals, sensor technology and data archiving. Manual recording of the values of analogue sensors on paper at fixed hours of the day is now replaced by continuous recording of electronic data in databases. This results in special requirements for statistical evaluation and error analysis (World Meteorological Organization (WMO) 2017).

Nationally and globally defined standards exist for the design of the measuring station and the data to be recorded, which in particular ensure the comparability of the measured values (Löffler 2012; World Meteorological Organization (WMO) 2018).

Meteorological observations have been recorded at the IPK since 1953. It is not difficult to conclude that these data are not primarily recorded digitally. The measurement results have been stored in databases only since 1993. For the period before 1993, at least the monthly values were subsequently captured and incorporated into the database. The result is that evaluations since 1993 are possible with a resolution accurate to the day, but analyses of the long-term measurement are only possible with a lower resolution (Fig. 2.8).

Today, data collection is done through an acquisition pipeline that stores, processes and aggregates the data collected by the data loggers to display and provide it to users in an appropriate way.<sup>11</sup> For this purpose, a series of plausibility checks are carried out on the raw data, and the time-based aggregations are calculated and saved from this cleansed dataset (Fig. 2.9).

---

<sup>11</sup> <https://wetter.ipk-gatersleben.de/>

	monthly	daily (7:00,12:00:21:00)	daily (every hour)	hourly
1953–03/1993	✓	✓ on paper	✗	✗
04/1993–12/1999	✓	✓	✗	✗
01/2000–present	✓	✓	✓	✓

Fig. 2.8 Resolution of the measurement for the IPK weather station



Fig. 2.9 Acquisition pipeline

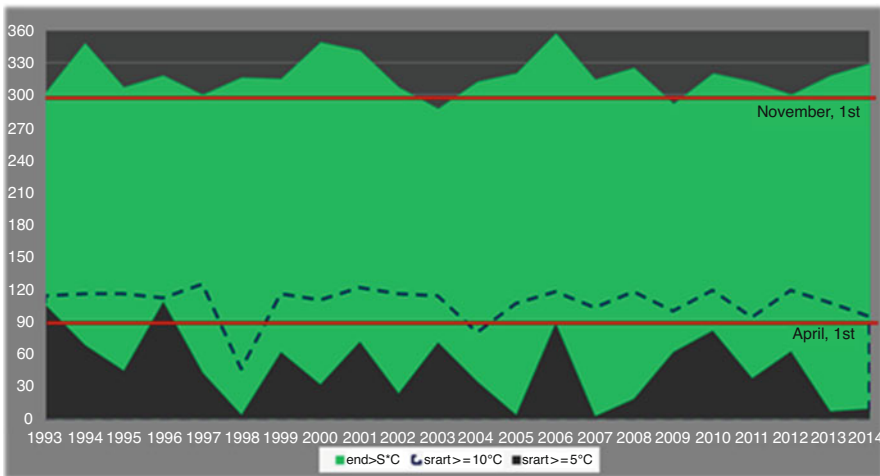


Fig. 2.10 Development of the growing season in Gatersleben 1993–2014

The data provided are not only used in the context of scientific experiments at IPK, but also serve as a basis for decisions on the conservation of biodiversity in the genebank.

Thus, the analysis of the meteorological data itself also offers insights into the climate development at the Gatersleben site, such as the development of the growing season (Fig. 2.10).

The above example shows the fluctuations in the start and duration of the growing season in the period from 1993 to 2014 as calculated. The recognizable variations from the threshold value of 1 April or 1 November influence the time at which traits are expressed. It also becomes clear that the calculation method influences

the result (start  $\geq 5$  °C or start  $\geq 10$  °C as dashed line). But it shows also a specific characteristic of meteorological data: The dataset is not finalized, but represents a daily extended data series on various meteorological parameters. This also means that conclusions drawn at an earlier point in time may have to be supplemented or revised for new studies.

With regard to the data life cycle, the IPK Weather Database covers the categories Collect, Process, Analyse, Preserve and Share.

### 2.4.7 Plant Phenotyping Portal

In addition to the IPK Weather Database, environmental data from high-throughput phenotyping facilities can be collected too. This is done in greenhouses or growth chambers. Two important plant growth facilities are the Plant Cultivation Hall and LemnaTec greenhouses (Altmann 2020). Here, the environment can be controlled to various degrees. For data acquisition, the MQTT protocol is used and plays a crucial role in the communication within the interconnected sensoric infrastructure at IPK. The Plant Cultivation Hall and LemnaTec greenhouses have 498 and 130 soil and climate sensors respectively, which generate data every 5–10 min. Additionally, 13 environmental sensors are transferable from one facility to another. The environmental data is essential for contextual and statistical analysis, aiding in the improvements in the agricultural use cases when shared in standardised formats. The idea is to store the raw sensory data in an interoperable and reusable way (Memon 2020). Therefore, using Node-RED, a flow-based programming tool, the MQTT protocol is implemented to communicate the data between the vendor-specific sources and the database. The MQTT protocol transmits the data as messages. Hence, the data is enveloped in messages (Fig. 2.11) and published to the broker through the assigned MQTT topic. For a permanent recording of the data, an authorised client subscribes to the topics that contain the relevant data and stores them in the database.

The topics are designed to contain the metadata related to the sensor data. For example, in Fig. 2.11, IPK\_G.1300 is determined as the building where the sensor is located, followed by the room number, container, type of sensor (such as temperature, humidity or moisture), sensor's node id, and the sensor port, since a single sensor node may have multiple ports. Whereas the message body includes the captured sensor data.

```
Topic: IPK_G.1300/0.001/CONTAINER/TEMPERATURE/node_name15/4
Message: {
  "Value": "19.2",
  "Unit": "°C",
  "Timestamp": "2020-01-20T17:15:09.0"
}
```

Fig. 2.11 An example of MQTT published message (Memon 2020)

In order to permanently record sensoric data and make them accessible for downstream data analysis, they must be stored in databases (Stöbe 2019). Because sensor data is streaming data, i.e. continuously delivered, its archive can only be conducted by aggregation over windows as a discrete snapshot. Usually, the resolution of such windows differs from seconds to hours. This depends on the expected fluctuation rate of measured values. In respect to weather data, i.e. wind, temperature, humidity and solar radiation, aggregation over 5 min is common. By doing so, subscribed environmental data is averaged over 10 min and stored in a relational database. Its metadata, like sensor placements or locations, is maintained in LIMS. On top of the database backend, the web application “Plant Phenotyping Portal” was developed. It integrates the aggregated sensor data, the metadata, and the experimental setup with the goal of breaking down individual plants, the installed sensor, and related environmental measures over time. Figure 2.12 shows the interface of the application.

The application’s interface allows users to download and view the sensory data between any two given time points for any available sensor(s). Furthermore, the application supports visualising the sensory data. For example, in Fig. 2.13, the chart above displays the temperature of different sensors, showing the sensor’s operating status and the chart below presents the light intensity between specific durations.

The application uses the Oracle Application Express (APEX) framework for these visualizations and covers the Collect, Preserve, Share and Reuse parts of the Data Life Cycle.

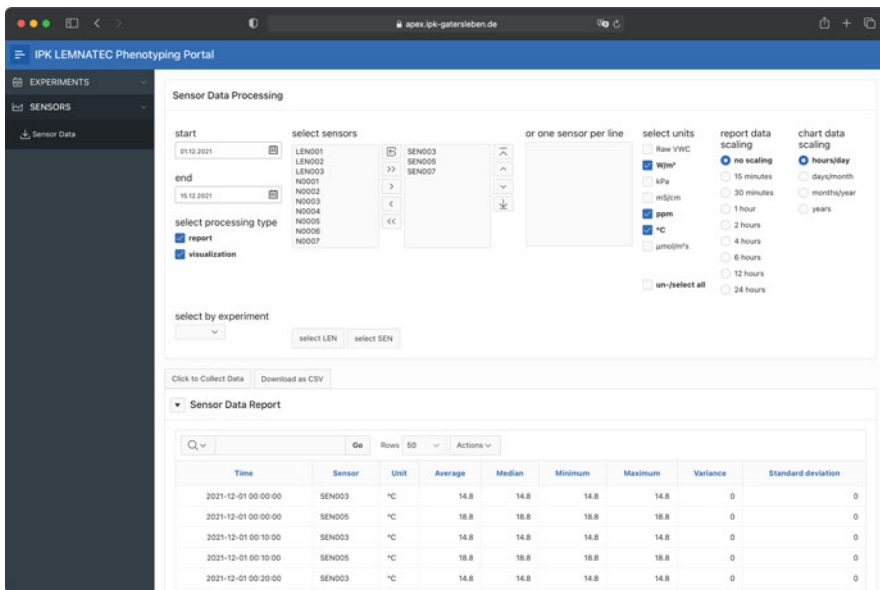


Fig. 2.12 User interface of the Plant Phenotyping Portal

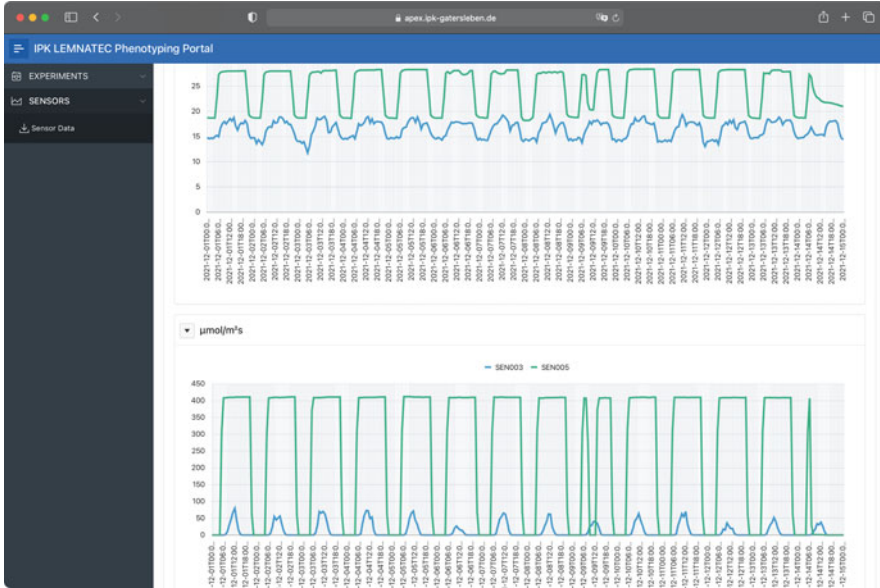


Fig. 2.13 Visualisation of archived sensor data in a Web Information System

Table 2.1 Systems and Data Domains

	Plant genetic resources data	Genomic data	Genotyping data	Phenotyping data	Environmental data
GBIS	✓			✓	
EURISCO	✓			✓	
BARLEX		✓			
BRIDGE	✓		✓	✓	
e!DAL-PGP	✓	✓	✓	✓	
IPK Weather DB					✓
Plant Phenotyping Portal				✓	✓

## 2.5 Summary and Outlook

Data management and the applications described here are diverse and yet serve the purpose of preparing data under consideration of the FAIR principles and offering it to its users. The requirements and the functions of the individual system are of course closely coupled with the data domains covered. Table 2.1 gives an overview of the different combinations of data domains in the information systems that have been worked on at IPK over the last 20 years. Unsurprisingly, most of the systems presented are focused on plant genetic resources and phenotypic data, but more recently genomic, genotypic, and environmental data have increasingly been added.

**Table 2.2** Systems and categories in Data Life Cycle

	Plan	Collect	Process	Analyse	Preserve	Share	Reuse
GBIS	✓	✓	✓	✓	✓	✓	✓
EURISCO					✓	✓	✓
BARLEX				✓	✓	✓	✓
BRIDGE		✓	✓	✓	✓	✓	✓
e!DAL-PGP					✓	✓	✓
IPK Weather DB		✓	✓	✓	✓	✓	
Plant Phenotyping Portal		✓			✓	✓	✓

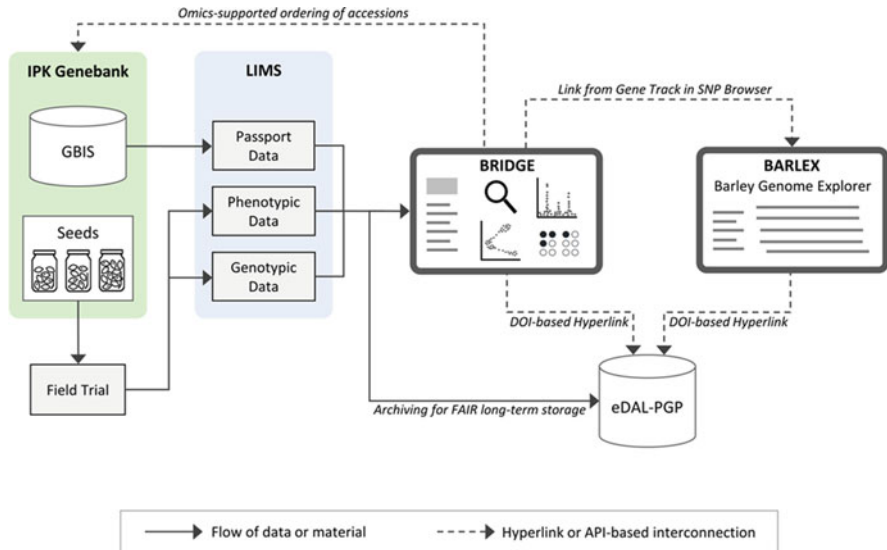
Accordingly, the classification of the systems examined into the individual phases of the Data Life Cycle also varies. While all systems support the later *Preserve and Share* steps, the *Plan* phase is underrepresented (Table 2.2). This can be explained by the fact that most systems were not designed to collect new data and start the data collection process, but rather to document and present data in an appealing form and manner and to exchange it with the community.

In summary, the presentation of the information systems has shown for which data domains data management solutions have been developed at IPK in Gatersleben. These were developed in general independently of each other and have thus grown historically. However, it can be stated that all steps of the data life cycle are served by the systems.

Generally, the applications described do not stand alone, but are designed via various interfaces in such a way that interaction between information systems is possible. One such example is depicted here in Fig. 2.14, where the interconnections between the IPK Genebank, GBIS, BRIDGE, BARLEX and eDAL-PGP are illustrated. The IPK Genebank and its GBIS serve as a primary data and material resource for genebank genomics experiments and field trials. The phenotypic and genotypic data derived from experimental field trials is then fed into visualization and analysis web tools like BRIDGE and BARLEX, while phenotypic observations of regular genebank multiplication trials are directly integrated into the GBIS. The genotypic data in the form of SNP-matrices (VCF files) is also deposited in eDAL-PGP for FAIR-compliant long-term storage. DOI-based hyperlinks from the SNP-browser in BRIDGE to the corresponding datasets in eDAL-PGP allow the users to download the original VCF files to their personal computers or HPC-servers for their own analysis. Hyperlinks from the visualised gene features in the BRIDGE SNP-browser to BARLEX allow the users to retrieve further information about the barley genome and genes.

Important challenges for the future are, on the one hand, the consistent semantic interlinking of the various information systems specialised in their use cases via unique identifiers and, on the other hand, the creation of central entry points for data research and data analysis. Currently we are using the IPK LIMS system as a central repository to implement a unique management of identifiers of biological objects. To increase efficiency, it is also important to develop reusable generic





**Fig. 2.14** Visualization of interconnections between IPK Genebank and its GBIS, BRIDGE, BARLEX and eDAL-PGP in the frame of genebank genomics experiments

software components for recurring tasks of interactive research data presentation and visualisation.

We have presented the approach of a research institution. It is obvious that an institute like the IPK Gatersleben does not exist autonomously. There are connections to cooperation partners all over the world. In order for the entire scientific community to be able to use the data, this data must be offered accordingly and thus be reusable. The foundation for knowledge discovery and innovation is good data management, because it allows data to be reused and new connections to other data to be formed by the community. One challenge is to make datasets not only understandable to humans but also readable and actionable by machines (Mons 2019). Open (non-binary) formats and richly annotated metadata are a prerequisite for this. However, in many areas of the life sciences, one or both of these requirements are not met, hindering both knowledge discovery and progress in general. The FAIR data principles (Wilkinson et al. 2016) are a start to making such a vision of the future a reality. To properly understand the FAIR principles, it is important to distinguish between two cases: First, FAIRification of existing data and FAIR-by-design, data created with FAIR principles in mind (Jacobsen et al. 2020). FAIRification of existing data is arguably the more challenging task to accomplish, as it requires updating data and metadata.

An example of a project focused on FAIRification of data is the ELIXIR (Crosswell and Thornton 2012) implementation study FONDUE. The task is to link available plant genotyping and phenotyping data using stable identifiers and to document those links in the repository metadata thus enabling search, retrieval

and reuse of such linked data. In this study, the main focus is on the so-called ELIXIR Core Data Resources (Drysdale et al. 2020), which are widely used in the life sciences and include such well-known repositories like the European Nucleotide Archive (Leinonen et al. 2011). The idea is to trigger a shift in thinking among data producers through this top-down approach by changing policies at key (genomic) data entry points. One obstacle to be overcome is that such further developments do not remain isolated cases and are adapted by other data providers and repositories. Only in this way can profound progress be made.

Many promising results have already been achieved with the approaches and data management systems described above. However, the integration of a wide variety of data is only at the beginning of the development. The long-term research goal of IPK Gatersleben is to develop into a bio-digital resource centre. For this purpose, a central entry point for accessing the IPK data needs to be established. Furthermore, the stored information about the biological objects should be provided with identifiers in such a way that traceability and integrability beyond the IPK institute boundaries are possible. These challenges will be the focus of activities for the bioinformaticians, data stewards and data scientists in the future.

**Acknowledgements** This work was supported by grants from the German Federal Ministry of Education and Research to, Uwe Scholz (SHAPE 2: FKZ 031B0884A, de.NBI: FKZ 031A536A) and Matthias Lange (AVATARS: FKZ 031B0770A) and from European Union's Horizon 2020 Research and Innovation Program to Matthias Lange (AGENT project: grant agreement no. 862613). The initial development of GBIS was jointly funded by the German Federal Ministry of Education and Research and the German Federal Ministry for Food and Agriculture. The development and maintenance of EURISCO are being funded by the European Cooperative Programme for Plant Genetic Resources.

## References

- Alercia A, Diulgheroff S, Metz T (2001) FAO/IPGRI multi-crop passport descriptors [MCPD]. Food and Agriculture Organization of the United Nations (FAO); International Plant Genetic Resources Institute, (IPGRI), Rome
- Alercia A, Diulgheroff S, Mackay M (2015) FAO/bioversity multi-crop passport descriptors V. 2.1 [MCPD V. 2.1]. Food and Agriculture Organization of the United Nations (FAO); Bioversity International, Rome
- Altmann T (2020) Forschungsbericht—Research Report—2018 - 2019. Leibniz-Institute of Plant Genetics and Crop Plant Research, Gatersleben, pp 78–81
- Arend D, Lange M, Chen J, Colmsee C, Flemming S, Hecht D, Scholz U (2014) e!DAL—a framework to store, share and publish research data. *BMC Bioinform* 15:214. <https://doi.org/10.1186/1471-2105-15-214>
- Arend D, Junker A, Scholz U, Schüler D, Wylie J, Lange M (2016) PGP repository: a plant phenomics and genomics data publication infrastructure. Database 2016:baw033. <https://doi.org/10.1093/database/baw033>
- Arend D, König P, Junker A, Scholz U, Lange M (2020) The on-premise data sharing infrastructure e!DAL: Foster FAIR data for faster data acquisition. *GigaScience* 9:giaa107. <https://doi.org/10.1093/gigascience/giaa107>

- Ariyadasa R, Mascher M, Nussbaumer T, Schulte D, Frenkel Z, Poursarebani N, Zhou R, Steuermagel B, Gundlach H, Taudien S, Felder M, Platzer M, Himmelbach A, Schmutzer T, Hedley PE, Muehlbauer GJ, Scholz U, Korol A, Mayer KFX, Waugh R, Langridge P, Graner A, Stein N (2014) A sequence-ready physical map of barley anchored genetically by two million single-nucleotide polymorphisms. *Plant Physiol* 164:412. <https://doi.org/10.1104/pp.113.228213>
- Beier S, Himmelbach A, Schmutzer T, Felder M, Taudien S, Mayer KFX, Platzer M, Stein N, Scholz U, Mascher M (2016) Multiplex sequencing of bacterial artificial chromosomes for assembling complex plant genomes. *Plant Biotechnol J* 14:1511–1522. <https://doi.org/10.1111/pbi.12511>
- Beier S, Himmelbach A, Colmsee C, Zhang X-Q, Barrero RA, Zhang Q, Li L, Bayer M, Bolser D, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Sampath D, Heavens D, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Houben A, Doležel J, Ayling S, Lonardi S, Langridge P, Muehlbauer GJ, Kersey P, Clark MD, Caccamo M, Schulman AH, Platzer M, Close TJ, Hansson M, Zhang G, Braumann I, Li C, Waugh R, Scholz U, Stein N, Mascher M (2017) Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Sci Data* 4:170044. <https://doi.org/10.1038/sdata.2017.44>
- Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* 6:80–92. <https://doi.org/10.4161/fly.19695>
- Colmsee C, Beier S, Himmelbach A, Schmutzer T, Stein N, Scholz U, Mascher M (2015) BARLEX—the barley draft genome explorer. *Mol Plant* 8:964–966. <https://doi.org/10.1016/j.molp.2015.03.009>
- Cousijn H, Feeney P, Lowenberg D, Presani E, Simons N (2019) Bringing citations and usage metrics together to make data count. *Data Sci J* 18:9. <https://doi.org/10.5334/dsj-2019-009>
- Crosswell LC, Thornton JM (2012) ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol* 30:241–242. <https://doi.org/10.1016/j.tibtech.2012.02.002>
- Ćwiek-Kupczyńska H, Altmann T, Arend D, Arnaud E, Chen D, Cornut G, Fiorani F, Frohberg W, Junker A, Klukas C, Lange M, Mazurek C, Nafissi A, Neveu P, van Oeveren J, Pommier C, Poorter H, Rocca-Serra P, Sansone S-A, Scholz U, van Schriek M, Seren Ü, Usadel B, Weise S, Kersey P, Krajewski P (2016) Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods* 12:44. <https://doi.org/10.1186/s13007-016-0144-4>
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST (2011) The variant call format and VCF tools. *Bioinformatics* 27:2156–2158
- Drysdale R, Cook CE, Petryszak R, Baillie-Gerritsen V, Barlow M, Gasteiger E, Gruhl F, Haas J, Lanfear J, Lopez R, Redaschi N, Stockinger H, Teixeira D, Venkatesan A, Elixir Core Data Resource Forum, Blomberg N, Durinx C, McEntyre J (2020) The ELIXIR Core data resources: fundamental infrastructure for the life sciences. *Bioinformatics* 36:2636–2642. <https://doi.org/10.1093/bioinformatics/btz959>
- ELIXIR (2021) Research Data Management Kit. A deliverable from the EU-funded ELIXIR-CONVERGE project (grant agreement 871075) [WWW Document]. <https://rdmkit.elixir-europe.org>. Accessed 28 May 21
- Endresen DTF, Knüpffer H (2012) The Darwin Core extension for genebanks opens up new opportunities for sharing genebank datasets. *Biodivers Inform* 8:12–29. <https://doi.org/10.17161/bi.v8i1.4095>
- Engels JMM, Maggioni L (2012) AEGIS: a regionally based approach to PGR conservation. In: Maxted N, Dulloo ME, Ford-Lloyd BV, Frese L, Iriando JM, Pinheiro de Carvalho MAA (eds) *Agrobiodiversity conservation: securing the diversity of crop wild relatives and landraces*. CABI, Wallingford, pp 321–326

- FAO (1997) The State of the World's Plant Genetic Resources for Food and Agriculture. Food and Agriculture Organization of the United Nations, Rome
- FAO (2010) The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Commission on Genetic Resources for Food and Agriculture, Food and Agriculture Organization of the United Nations, Rome
- Fiorani F, Schurr U (2013) Future scenarios for plant phenotyping. *Annu Rev Plant Biol* 64:267–291. <https://doi.org/10.1146/annurev-arplant-050312-120137>
- Fowler C, Hodgkin T (2004) Plant genetic resources for food and agriculture: assessing global availability. *Annu Rev Env Resour* 29:143–179. <https://doi.org/10.1146/annurev.energy.29.062403.102203>
- Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD (2016) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32:309–311. <https://doi.org/10.1093/bioinformatics/btv557>
- Fraternali P, Rossi G, Sánchez-Figueroa F (2010) Rich internet applications. *IEEE Internet Comput* 14:9–12. <https://doi.org/10.1109/MIC.2010.76>
- Gass T, Lipman E, Maggioni, L (1997) The role of Central Crop Databases in the European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR). In: Lipman, E et al. (eds) *Central Crop Databases: Tools for Plant Genetic Resources Management*, European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR); International Plant Genetic Resources Institute, Rome (Italy), pp 22–29
- Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32:D258–D261. <https://doi.org/10.1093/nar/gkh036>
- Ghaffar M, Schüler D, König P, Arend D, Junker A, Scholz U, Lange M (2019) Programmatic access to FAIRified digital plant genetic resources. *J Integr Bioinform* 16:20190060. <https://doi.org/10.1515/jib-2019-0060>
- Guha R, Brickley D, Macbeth S (2016) Schema.org: evolution of structured data on the web. *Commun ACM* 59(2):44–51. <https://doi.org/10.1145/2844544>
- Harrington JL (2016) Relational database design and implementation. Morgan Kaufmann, London
- Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. *Nature* 585:357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Hazekamp T, Serwinski J, Alercia A (1997) Multicrop passport descriptors. In: Lipman E et al. (eds) *Central Crop Databases: Tools for Plant Genetic Resources Management*, European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR); International Plant Genetic Resources Institute, Rome (Italy), pp 40–44
- Hoisington D, Khairallah M, Reeves T, Ribaut J-M, Skovmand B, Taba S, Warburton M (1999) Plant genetic resources: what can they contribute toward increased crop productivity? *Proc Natl Acad Sci* 96:5937–5943. <https://doi.org/10.1073/pnas.96.11.5937>
- IBPGR (1990) Descriptors for Brassica and Raphanus. International Board for Plant Genetic Resources, Rome
- International Board for Plant Genetic Resources (IBPGR), Commission of the European Communities (CEC) (1984) Plum descriptors. Rome
- IPGRI, ECP/GR, AVRDC (2001) Descriptors for Allium (*Allium* spp.). International Plant Genetic Resources Institute, Rome; European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR), Asian Vegetable Research and Development Center, Taiwan
- Jacobsen A, Kaliyaperumal R, da Silva Santos LOB, Mons B, Schultes E, Roos M, Thompson M (2020) A generic workflow for the data FAIRification process. *Data Intell* 2:56–65. [https://doi.org/10.1162/dint\\_a\\_00028](https://doi.org/10.1162/dint_a_00028)
- Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, Vincent L, Ware D, Zapata F (2005) Plant ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp Funct Genom* 6:388–397. <https://doi.org/10.1002/cfg.496>

- Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, Ens J, Zhang X-Q, Angessa TT, Zhou G, Tan C, Hill C, Wang P, Schreiber M, Boston LB, Plott C, Jenkins J, Guo Y, Fiebig A, Budak H, Xu D, Zhang J, Wang C, Grimwood J, Schmutz J, Guo G, Zhang G, Mochida K, Hirayama T, Sato K, Chalmers KJ, Langridge P, Waugh R, Pozniak CJ, Scholz U, Mayer KFX, Spannagl M, Li C, Mascher M, Stein N (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* 588:284–289. <https://doi.org/10.1038/s41586-020-2947-8>
- Jayakodi M, Schreiber M, Stein N, Mascher M (2021) Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res* 28. <https://doi.org/10.1093/dnares/dsaa030>
- Jiao W-B, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. *Genome Stud Mol Genet* 36:64–70. <https://doi.org/10.1016/j.jpbi.2017.02.002>
- Keim DA (2002) Information visualization and visual data mining. *IEEE Trans Vis Comput Graph* 8:1–8. <https://doi.org/10.1109/2945.981847>
- König P, Beier S, Basterrechea M, Schüller D, Arend D, Mascher M, Stein N, Scholz U, Lange M (2020) BRIDGE—a visual analytics web tool for barley Genebank genomics. *Front Plant Sci* 11:701. <https://doi.org/10.3389/fpls.2020.00701>
- Krajewski P, Chen D, Ćwiek H, van Dijk ADJ, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP, van Oeveren J, Pommier C, Scholz U, van Schriek M, Usadel B, Weise S (2015) Towards recommendations for metadata and data handling in plant phenotyping. *J Exp Bot* 66:5417–5427. <https://doi.org/10.1093/jxb/erv271>
- Kreide S, Oppermann M, Weise S (2019) Advancement of taxonomic searches in the European search catalogue for plant genetic resources. *Plant Genet Resour Charact Util* 17:559–561. <https://doi.org/10.1017/S1479262119000339>
- Lanthaler M, Gütl C (2012) On using JSON-LD to create evolvable RESTful services. In: *WS-REST '12: proceedings of the third international workshop on RESTful design*, April 2012. pp 25–32. <https://doi.org/10.1145/2307819.2307827>
- Leinonen R, Akhtar R, Birney E, Bower L, Cerdano-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G (2011) The European nucleotide archive. *Nucleic Acids Res* 39:D28–D31. <https://doi.org/10.1093/nar/gkq967>
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289. <https://doi.org/10.1126/science.1181369>
- Löffler H (2012) *Meteorologische Bodenmesstechnik (vormals: Instrumentenkunde), Leitfaden für die Ausbildung im Deutschen Wetterdienst Nr. 6*. Selbstverlag des Deutschen Wetterdienstes, Offenbach am Main (Deutschland)
- Madakam S, Ramaswamy R, Tripathi S (2015) Internet of things (IoT): a literature review. *J Comput Commun* 3:164–173. <https://doi.org/10.4236/jcc.2015.35021>
- Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang X-Q, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutz T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544:427–433. <https://doi.org/10.1038/nature22043>

- Mascher M, Schreiber M, Scholz U, Graner A, Reif JC, Stein N (2019) Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat Genet* 51:1076–1081. <https://doi.org/10.1038/s41588-019-0443-6>
- Mascher M, Wicker T, Jenkins J, Plott C, Lux T, Koh CS, Ens J, Gundlach H, Boston LB, Tulpová Z, Holden S, Hernández-Pinzón I, Scholz U, Mayer KFX, Spannagl M, Pozniak CJ, Sharpe AG, Šimková H, Moscou MJ, Grimwood J, Schmutz J, Stein N (2021) Long-read sequence assembly: a technical evaluation in barley. *Plant Cell* 33:1888–1906. <https://doi.org/10.1093/plcell/koab077>
- Mayer KFX, Waugh R, Langridge P, Close TJ, Wise RP, Graner A, Matsumoto T, Sato K, Schulman A, Muehlbauer GJ, Stein N, Ariyadasa R, Schulte D, Poursarebani N, Zhou R, Steuernagel B, Mascher M, Scholz U, Shi B, Langridge P, Madishetty K, Svensson JT, Bhat P, Moscou M, Resnik J, Close TJ, Muehlbauer GJ, Hedley P, Liu H, Morris J, Waugh R, Frenkel Z, Korol A, Bergès H, Graner A, Stein N, Steuernagel B, Scholz U, Taudien S, Felder M, Groth M, Platzer M, Stein N, Steuernagel B, Scholz U, Himmelbach A, Taudien S, Felder M, Platzer M, Lonardi S, Duma D, Alpert M, Cordero F, Beccuti M, Ciardo G, Ma Y, Wanamaker S, Close TJ, Stein N, Cattonaro F, Vendramin V, Scalabrin S, Radovic S, Wing R, Schulte D, Steuernagel B, Morgante M, Stein N, Waugh R, Nussbaumer T, Gundlach H, Martis M, Ariyadasa R, Poursarebani N, Steuernagel B, Scholz U, Wise RP, Poland J, Stein N, Mayer KFX, Spannagl M, Pfeifer M, Gundlach H, Mayer KFX, Gundlach H, Moisy C, Tanskanen J, Scalabrin S, Zuccolo A, Vendramin V, Morgante M, Mayer KFX, Schulman A, Pfeifer M, Spannagl M, Hedley P, Morris J, Russell J, Druka A, Marshall D, Bayer M, Swarbrick D, Sampath D, Ayling S, Febrer M, Caccamo M, Matsumoto T, Tanaka T, Sato K, Wise RP, Close TJ, Wannamaker S, Muehlbauer GJ, Stein N, Mayer KFX, Waugh R, Steuernagel B, Schmutz T, Mascher M, Scholz U, Taudien S, Platzer M, Sato K, Marshall D, Bayer M, Waugh R, Stein N, Mayer KFX, Waugh R, Brown JWS, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N, The International Barley Genome Sequencing Consortium, Principal investigators, Physical map construction and direct anchoring, Genomic sequencing and assembly, BAC sequencing and assembly, BAC-end sequencing, Integration of physical/genetic map and sequence resources, Gene annotation, Repetitive DNA analysis, Transcriptome sequencing and analysis, Re-sequencing and diversity analysis, Writing and editing of the manuscript (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491:711–716. <https://doi.org/10.1038/nature11543>
- McKinney W (2010) Data structures for statistical computing in python. In: van der Walt S, Millman J (eds) *Proceedings of the 9th Python in Science Conference*, pp 56–61. <https://doi.org/10.25080/Majora-92bf1922-00a>
- Memon JA (2020) Concept and implementation of homogeneous Sensorics infrastructure for the analysis of environmental factors in plant phenotyping (Masterabschlussarbeit). Universität Bielefeld, Bielefeld
- Miles A, Jakirkham, Durant M, Bussonnier M, Bourbeau J, Onalan T, Hamman J, Patel Z, Rocklin M, Shikharsh, Abernathy R, Moore J, Schut V, Raphael D, de Andrade ES, Noyes C, Jelenak A, Banihirwe A, Barnes C, Sakkis G, Funke J, Kelleher J, Jevnik J, Swaney J, Rahul PS, Saalfeld S et al (2020) Zarr-developers/zarr-python: v2.5.0. Zenodo. <https://doi.org/10.5281/zenodo.4069231>
- Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpfner H, Basterrechea M, König P, Schüler D, Sharma R, Pasam RK, Rutten T, Guo G, Xu D, Zhang J, Herren G, Müller T, Krattinger SG, Keller B, Jiang Y, González MY, Zhao Y, Habekuß A, Färber S, Ordon F, Lange M, Börner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N (2019) Genebank genomics highlights the diversity of a global barley collection. *Nat Genet* 51:319–326. <https://doi.org/10.1038/s41588-018-0266-x>
- Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, Ens J, Li C, Muehlbauer GJ, Schulman AH, Waugh R, Braumann I, Pozniak C, Scholz U, Mayer KFX, Spannagl M, Stein N, Mascher M (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. *Genome Biol* 20:284. <https://doi.org/10.1186/s13059-019-1899-5>

- Mons B (2019) FAIR science for social machines: let's share metadata Knowlets in the internet of FAIR data and services. *Data Intell* 1:22–42
- Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD (2017) Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science cloud. *Inf Serv Use* 37:49–56. <https://doi.org/10.3233/ISU-170824>
- Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, Cao H, Schlebusch SA, Giorda K, Schnall-Levin M, Wall JD, Kwok P-Y (2016) A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Methods* 13:587–590. <https://doi.org/10.1038/nmeth.3865>
- MQTT.org (2015) 10th birthday party | MQTT [WWW document]. <https://web.archive.org/web/20150315025826/>. <https://mqtt.org/2009/07/10th-birthday-party>. Accessed 18 May 2021
- Obermaier D (2018) MQTT 5—Die Neuerungen für das IoT-Standardprotokoll. JAXenter. <https://jaxenter.de/mqtt-5-internet-of-things-protocol-74891>. Accessed 18 May 2021
- Oppermann M, Weise S, Dittmann C, Knüpfner H (2015) GBIS: the information system of the German Genebank. *Database* 2015:bav021. <https://doi.org/10.1093/database/bav021>
- Papoutsoglou EA, Faria D, Arend D, Arnaud E, Athanasiadis IN, Chaves I, Coppens F, Cornut G, Costa BV, Ćwiek-Kupczyńska H, Droesbeke B, Finkers R, Gruden K, Junker A, King GJ, Krajewski P, Lange M, Laporte M-A, Michotey C, Oppermann M, Ostler R, Poorter H, Ramirez-Gonzalez R, Ramšak Z, Reif JC, Rocca-Serra P, Sansone SA, Scholz U, Tardieu F, Uauy C, Usadel B, Visser RGF, Weise S, Kersey PJ, Miguel CM, Adam-Blondon A-F, Pommier C (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytol* 227:260–273. <https://doi.org/10.1111/nph.16544>
- Parsons MA, Duerr RE, Jones MB (2019) The history and future of data citation in practice. *Data Sci J* 18:52. <https://doi.org/10.5334/dsj-2019-052>
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
- Petersen J (2001) Benefits of using the n-tiered approach for web applications. *Benefits Using N-Tiered Approach Web Appl.* <https://web.archive.org/web/20060618183914/https://www.adobe.com/devnet/coldfusion/articles/ntier.html>. Accessed 20 May 2021
- Philipp N, Weise S, Oppermann M, Börner A, Graner A, Keilwagen J, Kilian B, Zhao Y, Reif JC, Schulthess AW (2018) Leveraging the use of historical data gathered during seed regeneration of an ex situ Genebank collection of wheat. *Front Plant Sci* 9:609. <https://doi.org/10.3389/fpls.2018.00609>
- Psaroudakis D, Liu F, König P, Scholz U, Junker A, Lange M, Arend D (2020) isa4j: a scalable Java library for creating ISA-Tab metadata [version 1; peer review: 2 approved]. *F1000Res* 9. <https://doi.org/10.12688/f1000research.27188.1>
- Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, McMullen MD, Holland JB, Buckler ES, Gardner CA (2013) Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol* 14:R55. <https://doi.org/10.1186/gb-2013-14-6-r55>
- Rutkowski T (2005) Konzeption und Implementation einer plattformunabhängigen Importsoftware für ORACLE Datenbanken (Projektarbeit). Fachhochschule Harz, Wernigerode
- Sansone S-A, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Jones P, Lister A, Miller M, Morrison N, Rayner T, Sklyar N, Taylor C, Tong W, Warner G, Wiemann S (2008) The first RSBI (ISA-TAB) workshop: “can a simple format work for complex studies?” *OMICS. J Integr Biol* 12:143–149. <https://doi.org/10.1089/omi.2008.0019>
- Sansone S-A, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, Begley K, Booth T, Bougueleret L, Burns G, Chapman B, Clark T, Coleman L-A, Copeland J, Das S, de Daruvar A, de Matos P, Dix I, Edmunds S, Evelo CT, Forster MJ, Gaudet P, Gilbert J, Goble C, Griffin JL, Jacob D, Kleinjans J, Harland L, Haug K, Hermjakob H, Sui SJH, Laederach A, Liang S, Marshall S, McGrath A, Merrill E, Reilly

- D, Roux M, Shamu CE, Shang CA, Steinbeck C, Trefethen A, Williams-Jones B, Wolstencroft K, Xenarios I, Hide W (2012) Toward interoperable bioscience data. *Nat Genet* 44:121–126. <https://doi.org/10.1038/ng.1054>
- Schulte D, Close TJ, Graner A, Langridge P, Matsumoto T, Muehlbauer G, Sato K, Schulman AH, Waugh R, Wise RP, Stein N (2009) The international barley sequencing consortium—at the threshold of efficient access to the barley genome. *Plant Physiol* 149:142. <https://doi.org/10.1104/pp.108.128967>
- Schulte D, Ariyadasa R, Shi B, Fleury D, Saski C, Atkins M, deJong P, Wu C-C, Graner A, Langridge P, Stein N (2011) BAC library resources for map-based cloning and physical map construction in barley (*Hordeum vulgare* L.). *BMC Genomics* 12:247. <https://doi.org/10.1186/1471-2164-12-247>
- SciData Editorial (2019) Data citation needed. *Sci Data* 6:27. <https://doi.org/10.1038/s41597-019-0026-5>
- Shrestha R, Arnaud E, Mauleon R, Senger M, Davenport GF, Hancock D, Morrison N, Bruskiwich R, McLaren G (2010) Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB Plants* 2010:plq008. <https://doi.org/10.1093/aobpla/plq008>
- Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, Arnaud E (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the crop ontology developed by the crop communities of practice. *Front Physiol* 3:326. <https://doi.org/10.3389/fphys.2012.00326>
- Staňková H, Hastie AR, Chan S, Vrána J, Tulpová Z, Kubaláková M, Visendi P, Hayashi S, Luo M, Batley J, Edwards D, Doležel J, Šimková H (2016) BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol J* 14:1523–1531. <https://doi.org/10.1111/pbi.12513>
- Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, Schulte D, Petzold A, Felder M, Graner A, Scholz U, Mayer KF, Platzer M, Stein N (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics* 10:547. <https://doi.org/10.1186/1471-2164-10-547>
- Stöbe E (2019) Konzeption und Implementierung einer nachrichtenorientierten Sensorikinfrastuktur für eine Pflanzenphänotypisierung und einer adaptiven Bewässerungssteuerung (Masterarbeit). Hochschule Midweida, Fakultät: Angewandte Computer- und Biowissenschaften, Leibniz-Institut für Pflanzen-genetik und Kulturpflanzenforschung, Mittweida
- Taudien S, Steuernagel B, Ariyadasa R, Schulte D, Schmutzer T, Groth M, Felder M, Petzold A, Scholz U, Mayer KF, Stein N, Platzer M (2011) Sequencing of BAC pools by different next generation sequencing platforms and strategies. *BMC Res Notes* 4:411. <https://doi.org/10.1186/1756-0500-4-411>
- Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, Pollock D, Dorsett K (2015) Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS One* 10:e0134826. <https://doi.org/10.1371/journal.pone.0134826>
- van Hintum TJJ (1997) Central crop databases—an overview. In: Lipman E et al. (eds) *Central Crop Databases: Tools for Plant Genetic Resources Management*, European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR); International Plant Genetic Resources Institute, Rome (Italy), pp 18–21
- Wang C, Hu S, Gardner C, Lübberstedt T (2017) Emerging avenues for utilization of exotic germplasm. *Trends Plant Sci* 22:624–637. <https://doi.org/10.1016/j.tplants.2017.04.002>
- Watt M, Fiorani F, Usadel B, Rascher U, Muller O, Schurr U (2020) Phenotyping: new windows into the plant for breeders. *Annu Rev Plant Biol* 71:689–712. <https://doi.org/10.1146/annurev-arplant-042916-041124>
- Weise S, Oppermann M, Maggioni L, van Hintum T, Knüpfper H (2017) EURISCO: the European search catalogue for plant genetic resources. *Nucleic Acids Res* 45:D1003–D1008. <https://doi.org/10.1093/nar/gkw755>



- Weise S, Lohwasser U, Oppermann M (2020) Document or lose it—on the importance of information management for genetic resources conservation in genebanks. *Plants* 9:1050. <https://doi.org/10.3390/plants9081050>
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglaiss D (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* 7:e29715. <https://doi.org/10.1371/journal.pone.0029715>
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3:160018
- World Meteorological Organization (WMO) (2017) Challenges in the transition from conventional to automatic meteorological observing networks for long-term climate records. WMO-No. 1202, Geneva
- World Meteorological Organization (WMO) (2018) Guide to instruments and methods of observation, Volume V: quality assurance and management of observing systems. WMO-No. 8, Geneva