KDDart Platform for Data Integration in Pre-Breeding, Breeding and Ecology

www.kddart.org

Grzegorz Uszynski, Puthick Hok, Brian Pierce, David Hunter, Aneal Chandra, Stanley Wijoyo, Manil Chaudhary, Brad Lanham, Frank Detering, Krzysztof Kurczynski and Andrzej Kilian

Diversity Arrays Technology Pty. Ltd. Canberra, Australia

www.diversityarrays.com

KDDart
Knowledge Discovery in Database art

Diversity Arrays presents KDDart, a new IT infrastructure we developed in conjunction with Australian and international partner and collaborator organisations. The KDDart platform, which is in early stages of deployment and testing within several projects, is a modular platform designed to integrate high data volumes that often originate from various sources.

The core module, which collects field data for plant performance, can easily be ported to collect any kind of phenotypic data from both breeding and natural plant and animal populations. The environmental module can store from high to lower resolution spatio-temporal data, obtained through sensor devices. These devices can be configured to communicate with KDDart remotely or from publicly available GIS/remote sensing data sources. Both phenotypic and environmental modules are connected to molecular data module enabled to store marker data from high volume, high resolution genome profiling technologies (e.g. DArTseq). In addition, the system is capable of storing inventory and pedigree/relatedness information.

KDDart is designed as a modern IT infrastructure with 3 major components (layers):

1. Database layer
2. Application layer RESTful API (Data Access Layer or DAL)
3. Client Applications/Software layer

The Application layer API is a heart of the system and is designed as a RESTfull web service which provides and manages access to the KDDart database. DAL is primarily a set of generic methods to perform all possible database operations and is exposed as a web service RESTful API.

Client Applications layer is a collection of scripts, software applets etc. accessing the database through the DAL. These applications can be written as web, desktop or mobile applications using any modern programming language.

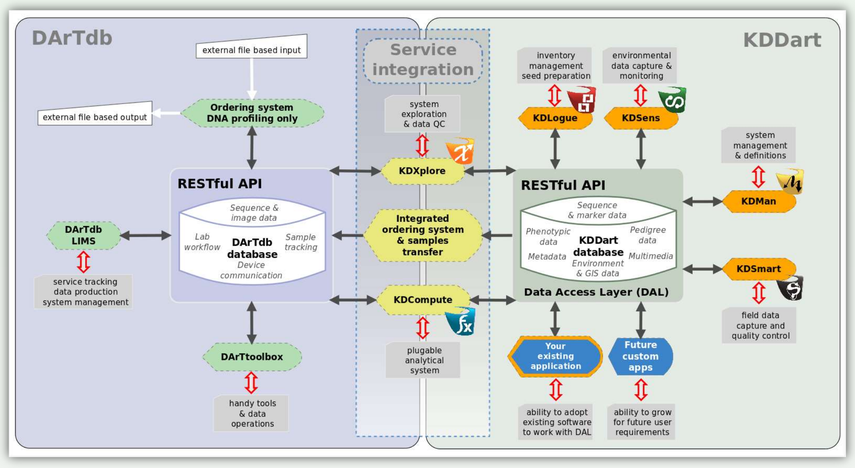
Data collected and stored in KDDart is available for statistical analysis or data-mining using KDCompute, which represents KDDart's generic plug-in platform for batch processing. Processed data/information can be viewed, further analysed and interpreted by project staff and other system users through KDDart's user secure and controlled environment.

Data Integration Platform

Established in 2001 in the Australian capital, Canberra, Diversity Arrays Technology primarily provides genome profiling and IT support services along with consultancy to international private and public organisations. Fostering an atmosphere of openness and collaboration DArT has attracted a very strong interest amongst international agriculture research by providing training, to over 50 visitors, aside technology co-development projects.

The KDDart system provides our clients with a platform to integrate our marker services with customers' phenotypic and environmental data, enabling more efficient data management and analysis. The scenario illustrated below describes how the genotyping data can be connected to the KDDart platform and analysed together with all other data types on clients' platform. This integrated system configuration, facilitated by the web service (RESTful API) architecture of both systems, provides clients a more compact system resulting in a better user experience.

To learn more about DArT's genotyping service, please see the poster: "**DArT™ and DArTseq™ Genome Profiling for Breeding, Pre-Breeding and Population Genetics**" with number **P0052.**

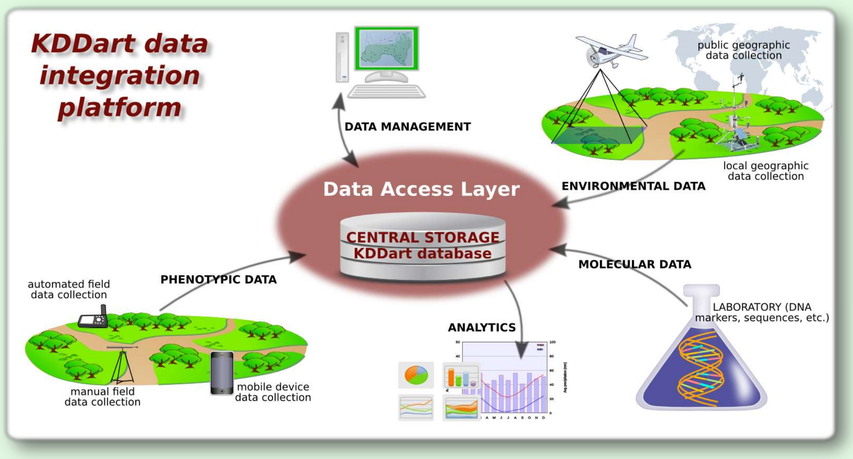


System Availability

KDDart core components (Database Layer with Data Access Layer) is published as open source software and under a GPL3 license. Source is available from: https://github.com/kddart/DAL

DArT has also published number of wrapper libraries for specific programming languages like Java, JavaScript and Perl. Sources of those libraries are available in repositories at: https://github.com/kddart

More online resources are available at: http://www.kddart.org



Conceptual Design

KDDart is our major development project to create a comprehensive and flexible data storage and integration platform.

The KDDart infrastructure, graphically represented above shows the components usually needed in breeding, crop management or ecological applications.

The Data Access Layer with Central Storage form the center of every system implementation, providing efficient data integration to support all the platform application tools and components.

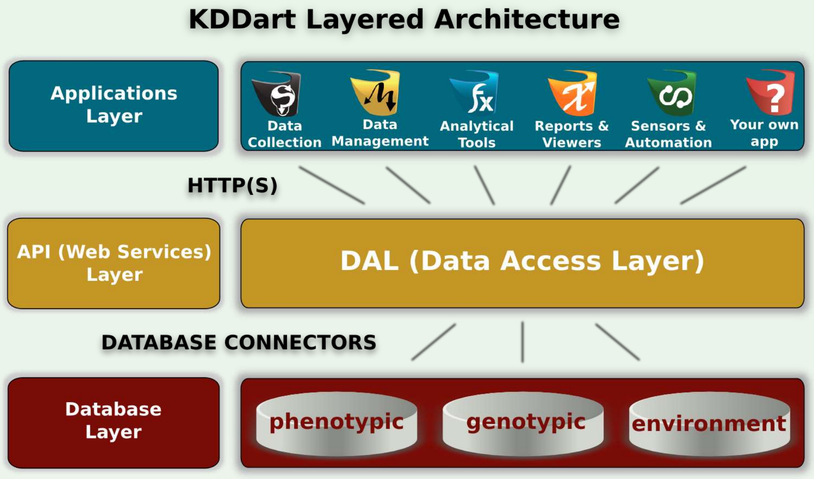
This design allows easy access to all relevant data in a uniform, integrated manner for current and user developed applications.

Technical Design

The following system design requirements were taken into account:

- All components must be freely available and open source
- System must adhere to widely adopted 'open' standards
- The system must be scalable, from small "laptop" to large multi-server installation
- Data must be securely stored with controlled user access to ensure data privacy
- An extensible base design must allow further extensions, enabling users to develop scripts or whole applications as their future needs dictate.

From the above considerations DArT created a layered design with three basic components (layers): Database, API RESTful web service and Applications Client Software.



Analysis and Data Mining

KDCompute is the analytical platform of KDDart, which is a generic tool for the integration of new algorithms. It facilitates the development of "user interfaces" for new and existing algorithms and is designed for use by both experienced and inexperienced IT users.

KDCompute uses a pluggable framework, where each interface with algorithm is a plugin, allowing easy extension and adoption for a program or organisation's needs.

The application and interface is via web browser, using a background queuing server to schedule and process multiuser tasks on powerful computing node keeping the user's workstation free for other tasks.

Acknowledgements

DArT would like to acknowledge its clients, visitors, students and the DArT team. **Thank you!**

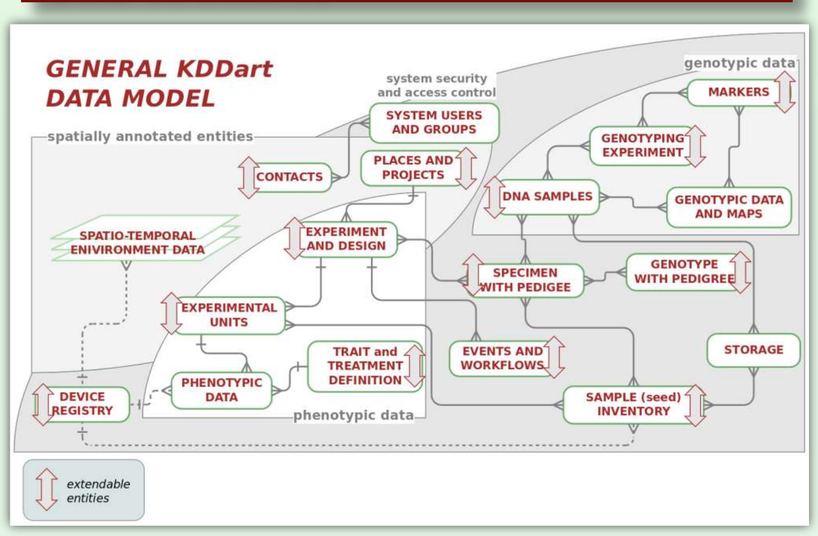
Supported initially by DArT PL – Horticulture Australia VC program: 'Varietal Evaluation and Crop Improvement System'

Further development with DArT PL's resources, GRDC project for trial system database and CIMMYT SeeD project

KDDart Features

- Integrating all data types relevant to breeding, crop management and/or ecology (genetics, field data and environment)
- Remote sensing and high density environmental monitoring devices and other sensors can access the system directly
- Seamless **integration of analytical / data mining tools** (e.g. through our own KDCompute plugin infrastructure, but also as stand alone tools)
- Immediate integration of modelling technologies with KDDart
- Enabling **informed decisions** to breeders, farmers and research community
- Creating an **"environment" for system growth** and improvement through API (RESTfull services)

Data Structures



- **Phenotypic data** is organised into experiments
- **Experimental designs and traits** are definable features of every implementation and may utilise existing ontologies
- Granular **germplasm information with pedigree** is connected to both experimental setup and DNA data and markers
- **Genotypic data and maps** (both genetic and physical) can be saved in the system
- All **experimental data** can be spatially annotated and experiments organised into **workflows**
- All germplasm can be organized into **inventory** and all samples can be organised into **storage locations**
- **Sensors and other devices** can be used to enhance capture of phenotypic and environmental data and help in inventory management
- **Environmental and other spatial data** can be stored in **GIS** layered structures and related to other data through coordinates
- **Spatio-temporal data** can be stored in environmental module
- Most data **entities can be extended** with custom defined additional data types and multimedia can be attached

Applications

KDSmart - designed to operate on a variety of handheld devices for field data collection. Originally developed from a stable version of DataKapture software jointly developed by the Q/NSW DPI, it has been adapted to meet a variety of new user requirements. Online mode allows direct synchronisation of trial data with the KDDart database. When off-line the trial data is available for the user to work in the field to capturing trial results.

KDMan - the general system management utility. A web based application designed to setup system environment to satisfy entity dependencies, perform day-to-day data managerial tasks, manage and curate datasets.

KDXplore - a versatile application, useful for breeders, technicians, curators and developers. Able to assist with your trial selection and manage the distribution of those copies onto multiple KDSmart devices. Data collected in the field with KDSmart can move back to KDXplore for data curation then upload, to ensure only quality trial data is stored in KDDart.

KDSens - an application providing an interface between various generic environmental sensors, such as weather stations, soil probes, etc. and the KDDart database using the Data Access Layer (DAL). Sensor definitions are maintained within KDDart.

Your own application – open architecture and Data Access Layer allow users to create their own scripts and software with custom interfaces and functionalities. They may range from simple ad hoc scripts up to large, complex software tools with entirely custom designed interfaces.

Organically growing business, mainly through partnerships. Our mission is to facilitate the development of a network of people and organisations as a venue to spread the benefits of genotyping and information technologies across the agricultural and environmental sectors in an equitable manner.