DDHT projects 2019: data, models and tools

Lotte de Vos, Jan Top

April 28, 2020

Contents

Introduction

The goal of this document is to obtain a comprehensive but concise overview of datasets, models, tools, infrastructure and ethics-related items in the KB DDHT projects. We focus on *the current status quo*, i.e., things that are actually available in January 2020. To set each project into context, also a brief description of the different domain objectives is given. Moreover, we also list knowledge gaps and needs for additional skills in each project. With this overview we hope that researchers in the different projects will be able to identify issues they may have in common. This will stimulate them to report these issues to the DDHT Knowledge Management team, who will then use this information for organising a number of workshops to reduce knowledge gaps identified across multiple projects or teams. These workshops will also focus on the realisation of the four pillars of the DDHT program.

With respect to the sections *Data* and *Models*, the items mentioned in this project can be seen as possible input for the Metadata Library that is currently under development at WDCC. That web portal will provide access to extensive metadata on shared datasets and models at WUR, including a pointer to the actual item (for example on a WUR shared drive, or in the 4TU repository).

In order to achieve a compact, but also uniform overview, the next section contains a template with guidelines on which information is required. Rather than plans and ideas we emphasize that this document is about currently used data, models, tools, and infrastructure. This will give others the possibility to see what different project use and produce in practice, and to identify shared interests. We do not need full details, but just enough to trigger others. In the subsequent chapters, each project list their current content items, based on this template.

Project content descriptions – Template

[Note: the template provides some examples in the form *'Example' >> 'Preferable example'*]

Domain objectives

Summarize the domain issues addressed by this project, one line per issue.

'Machine learning on remote sensing data' >> 'Understand the occurrence of drought in West-Africa'

Data

List all datasets that are really used and produced in this project in the past year. Indicate for each dataset

- Which things in the real world it is about (in a few words)
- whether this data is in principle also available for others within WUR ('Open for WUR' or `Not open for WUR').

'remote sensing data' >> 'remote sensing data on crops in Europe (not open for WUR)'

'NEVO-table' >> 'NEVO-table with nutritional values of food products (open for WUR)'

Models

A model is a formal representation of knowledge about a domain in the form of an algorithm, mathematical expressions, logical rules, structured text, etc., either derived from data or from explicit expertise.

State in a few words what the models represents.

'BBN model' >> 'Yield prediction for crop x'

Tools

A tool is software that enables data or model processing. For each tool provide:

- Name, if available
- Link to general information about the tool, or if that is not possible a few words on the function of the tool
- Language or library in which it is implemented (e.g. R or Java)

It is no needed to describe how or for what the tool is used in the project.

'Machine learning' >> 'Random forest, https://en.wikipedia.org/wiki/Random_forest, Weka library'

Infrastructure

Which platforms and software do you use to organise your data and share it within your team or with others, or environments to create software.

Examples are ‘Onedrive, Local shares, IRODS, Docker’.

Ethics

List ethical issues that have been explicitly addressed in your project (in a report or notes).

Gaps

Which specific knowledge or skills is missing from your project, for which a workshop (internal WUR or with external experts) would be beneficial? Provide one line per issue.

Project 1 - Data analytics for food chains and consumer-oriented research

Domain objectives

- Generate questionnaires to collect information from consumers
- Assist individual consumers in making healthy nutritional choices
- Monitoring food safety issues
- Monitoring slaughter houses with respect to zoonosis-pathogens
- Support collaboration in agrifood research

Data

- Collected and merged food safety and food fraud data from RASFF, EURSTAT and FAOSTAT (open for WUR)
- Consumer surveys collected over the years (not open for WUR)
- Semantic version of the NEVO-table with nutritional values of food products (open for WUR)
- Zoonosis data, shared with RIVM and NVWA (not open for WUR)

Models

- Bayesian Belief Network (BBN) on food-health relations based on National Health Council
- Workflows in KNIME to automatically extract, process and analyse data on food safety
- Semantic model of nutritional recommendations used by Voedingscentrum

Tools

- KNIME (https://www.knime.com/) for data science workflows.
- FAIR Data Edito, part of FAIRification tools as provided by DTL only available for WUR (http://bigdata.wfsr.wur.nl/)
- RDF4J as triple store (https://rdf4j.org/documentation/server-workbench-console/)

Infrastructure

- WUR Metadata Library – initially within WUR
- Docker
- Consumer Data Platform
- WUR-HPC
- Creating WUR subdomains on webpages to provide access to outside parties to services from specific WR-institutes, e.g. https://*.wfsr.wur.nl
- OpenShift
- Elastic Search
- WEcR Datawarehouse
- Git lab

Ethics

- Not considered

Gaps

- Cultural change needed with respect to data sharing

Project 2 - AI in animal and arable systems focused on farm management

Domain objectives

- Recommending texts on innovation in agriculture and forestry (ask-Valerie)
- Automated monitoring nitrogen excretion of cows from feed intake and production (?)
- Monitoring clouds in remote sensing data
- Pest and disease defield crops with imaging
- Combining different layers of information in maps (Akkerweb)
- Disease detection in fresh products, e.g., stem rot disease detection in avocados
- Monitor deforestation due to cocoa production with citizen science
- Metabolite profiling of potato tubers
- Linking satellite images to food safety data (focus grass and maize)

Data

- Large number of references to documents on innovations in agriculture, available through https://www.ask-valerie.eu/ (open for WUR)
- Images of sequence reads (not open for WUR)
- Koeien-en-Kansen or Cows-and-Opportunities (not open for WUR - contact Koos Verloop: farmers have to agree to share the data)
- Ear tag images (not open for WUR)
- Farm data on Akkerweb, >5000 accounts (not open for WUR, farmers have to agree to share the data-contact Thomas Been)
- IBM global weather data on Akkerweb (open for WUR – contact Thomas Been)
- Open data NL on Akkerweb (open for WUR – contact Thomas Been)
- Drone heat images (availability unknown - contact Sander Mucher)
- Hyperspectral and RGB data: Applescap and Tomato datasets (availability unknown - contact Gerrit Polder or Peter Frans de Jong)
- AgroDataCube (processed open data) (open for WUR)
- Satellite images (Sentinel-2, Landsat 8 and BRP) (open for WUR, BRP after one year)

Models

- Agriculture and forestry ontology (10k concepts) as used in ask-Valerie, available on www.foodvoc.org
- BBN models (?)
- Deep Neural Network (DNN) (?)
- Convolutional neural network (CNN) (?)
- Generalized boosted regression model (GBM) (?)
- Cocoasoils ontology
- Tipstar - crop growth model
- WatBal - soil model

Tools

- ROC+ for creating vocabularies https://rocplus.fbresearch.nl/.
- Semantic search as applied in ask-Valerie (domain-specific recommender system) and in Metadata Library (to be found where?)
- NLP (https://stanfordnlp.github.io/CoreNLP/) to support creation of vocabularies and ontologies from selected documents
- Bayesian Belief Networks
- Convolution Neural Network (CNN) for classifying images of sequence reads.
- AI Nimbus cloud masking, developed by Pytorch. It is implemented on the HPC.
- Nanonets https://nanonets.com/
- PCA analysis (https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202)
- Keras package in Python (https://keras.io/)
- Deep learning framework (TensorFlow https://machinelearningmastery.com/introduction-python-deep-learning-library-tensorflow/)
- H2O-gbm function in R https://www.rdocumentation.org/packages/h2o/versions/3.28.0.4/topics/h2o.gbm

- Image alignment software (?)
- YOLOv3 (You only look once) in Pytorch https://medium.com/analytics-vidhya/yolo-v3-theory-explained-33100f6d193 and https://blog.paperspace.com/how-to-implement-a-yolo-object-detector-in-pytorch/
- Metalign https://www.wur.nl/en/show/MetAlign-1.htm
- MetOT (?)
- MSCLUST workflow (?)
- R-routines constructed at WUR (?)
- ask-Valerie (intended for non-scientific users) https://www.ask-valerie.eu/
- GoogLeNet https://research.google/pubs/pub43022/
- Resnet https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

Infrastructure

- Local WUR servers
- WUR-HPC
- Amazon cloud services
- PHIS for pre and post-harvest experimental data
- Akkerweb

Ethics

- Akkerweb complies with *Branche Organisation Akkerbouw* and *Copa Cogea code* on data use.

Gaps

- How to parallelize indexing and annotation mechanism in search application for improved performance?
- How to (partly) automate annotation of data (including images)?
- How to explain the output of a deep learning model?
- Experience with python deep learning script for multiclass classification by either Tensor flow Unet multi-class image classification, or improve NIMBUS single class to multi-class and eCognition Deep Learning.
- Training on machine learning (?)
- How to publish data?

Project 3 - Advances in data-driven phenotyping through shared infrastructures and analytics

Domain objectives

- Assist researchers in organising data that are being collected on cocoa-plots in West-Africa
- Detect plant diseases using drones
- Food safety assessment at the level of leaves and fruits
- Observation of phenotype properties of plant

Data

- Disease detection images - from drones and tractors on Unifarm (availability unknown)
- Datasets on apple and grape leaves (10 Mp RGB, 6 band MSI and VNIR HyperSpectral Imaging (HSI)) (availability unknown)
- High-resolution datasets for Akkerweb (phenotype data, but also physical, chemical, spectral and disease data) (availability unknown)
- Hyperspectral images at six timepoint of 80 avocados (availability unknown)
- Multivariate transcriptomics and metabolomics data of potatoes (availability unknown)
- Drone images with RGB, multispectral, and LiDAR measurements of the NPEC crops (availability unknown)
- UNIFARM multispectral drone date throughout the growing season (from May – July 2019) (availability unknown)
- NPEC field vehicle phenotyping measurements (availability unknown)

Models

Tools

- Akkerweb, field/crop zoning application, https://akkerweb.eu/
- Machine learning (?)
- Image alignment software for the multispectral imaging (MSI) camera was made in the computer vision tool Halcon (?)
- PHIS information system for plant phenomics http://www.phis.inra.fr/
- ODK Aggregate server and designed forms and mobile apps for field data collection https://docs.opendatakit.org/aggregate-intro/

Infrastructure

- Local storage
- iDAL, an IPK system (?)

Ethics

- Not considered yet.

Gaps

- Expertise on deep learning

Project 4 - Smart and privacy conserving infrastructures for farm generated data and farm and food safety sensors towards smart and circular agriculture

Domain objectives

- Safely share and reuse (farm-related & research) data, respecting privacy, confidentiality and data ownership, thus facilitating collaborative multidisciplinary research and data scinece on emerging topics like circular agriculture
- Connecting data services into infrastructure that breaks WUR thematical silos and barriers for data sharing with external parties
- Monitoring continuous data collection processes in real time

Data

- Basis Registratie Percelen (BRP) - Parcel geometries (Dutch - RVO via AgroDataCube, European Member states data services) (open for WUR)
- Weather data (KNMI via AgroDataCube, open for WUR, or IBM via Akkerweb, open for WUR)
- Satellite data (Sentinel-2 based NDVI via AgroDataCube, Sentinel-1 based radar) (open for WUR)
- Satellite data (Sentinel-1 based radar) (open for WUR)
- Soil data (Dutch soil map 1:50:000 and BOFEK) (open for WUR)
- Methane concentration at the level of individual cows which is aggregated to a farm level
- Commercially available datasets (e.g. IBM weather data) (not open for WUR)
- Soil quality data (Open Bodem Index) (partly open for WUR)
- Food safety data; RASFF (open for WUR)
- Economic and agricultural data; EUROSTAT (open for WUR), FAOSTAT (open for WUR)

Models

- Crop growth models Tipstar, WOFOST and Lintul
- Crop phenology estimation algorithms
- Food safety/ food fraud prediction models

Tools

- Arduino for (near) real-time data collection into MS-Azure https://docs.microsoft.com/en-us/azure/?product=featured or https://portal.azure.com/?configHash=zXr-tHjTB5jq&iepolyfills=true&l=en.en-us&appPageVersion=5.0.303.33218460068.200313-2139#home) Tools supported by WUR FB-IT such as various cloud services offered through MS-Azure Azure (https://powerbi.microsoft.com/nl-nl/ or https://docs.microsoft.com/en-us/power-bi/service-real-time-streaming)
- Microsoft Power BI (laborious and slow in our case where dataset continuously grows) https://powerbi.microsoft.com/
- Farm data authentication/authorisation services: e-Herkenning, Join Data
- GeoTrellis (big geospatial data processing for raster data)
- GeoMesa (big geospatial data processing for vector data)
- Cassandra (distributed, scalable NoSQL database)
- Jupyter Notebooks (open source web application to create and share documents that contain live code, visualizations and narrative text. For data cleaning and transformation) https://jupyter.org/

Infrastructure

- MS-Azure (Microsoft cloud services platform)
- Spark (analytics engine for big data processing)
- AgroDataCube (open data service for agronomic data)
- Akkerweb
- D4Science / Aginfra+ (virtual research environment)
- WUR-HPC

- Gitlab WUR https://git.wageningenur.nl/

Ethics

- Confidentiality and privacy issues and solutions
- Protection of data ownership
- Legislation

Gaps

- Selecting suitable cloud computing services and software tools, including cost estimation
- How to scale up software? External experts are expensive
- Knowledge and experience on Spark, Hadoop and Cassandra
- WUR FB-IT as expertise center for cloud solutions
- Internal experts on cloud solutions are hard to find (not there, or not knowing where to find them)

Project 5 - Autonomous Robots for agri-food processes

Domain objectives

- Autonomous robot systems can reduce the amount of labor required to produce food in
    - open-field agri-food processes (small mobile robots),
    - horticulture (applying explicit knowledge),
    - post-harvest (teaching robots sorting tasks),
    - livestock (how do robots fit into the system),
    - aquatic systems, can reduce the amount of labor required to produce food.

Data

- Robot movement paths (through field or the action of a robot arm) in the form of markers or via computer vision (open for WUR)
- Human actions (positions of elbow, hand, shoulder etc.) in front of a robot, recorded by a camera (open for WUR in 2020, but depending on the quality of data might be released under CC-BY in 2021)
- Parcel data (in-house) (open for WUR)
- Simulated data robot movements and farm maps (open for WUR)
- Interviews on ethical issues related to presence of autonomous robots in livestock and arable farming (open for WUR)

Models

Tools

- Deep learning using neural networks (which tools?). https://github.com/dbolya/yolact , https://github.com/dbolya/yolact, https://github.com/HRNet, https://github.com/andyzeng/visual-pushing-grasping
- ROS (Robot Operating System), an open source software layer for research purposes. https://www.ros.org/
- PyTorch (for deep learning) https://pytorch.org/
- V-REP (for robotic simulation) https://github.com/alexandremstf/ros-vrep, https://www.coppeliarobotics.com/ (free academic version)

Infrastructure

- Data is internally shared on OneDrive and WUR drives
- HPC

Ethics

- This is a substantial element of the project, on which stakeholders are interviewed.
- Privacy is not an issue

Gaps

- Use of HPC is limited by practical, administrative and financial conditions

Project 8 – Sensing Potential

Domain objectives

- Sensing animal (dairy and poultry) welfare, detecting (stages of) lameness
- Measure greenhouse climate at microscale and crop traits – winter light treatment
- Post-harvest product quality of avocado and egg plant
- Measure food intake, i.e. volume and type of food
- Detect food alteration in fresh skimmed powder milk

Data

- Collection of publications on potentially interesting sensors. All released with CC-BY-SA license (open to WUR)
- Wageningen Shared Facilities (https://www.wur.nl/en/Value-Creation-Cooperation/Facilities/Wageningen-Shared-Research-Facilities.htm) (open to WUR)
- Large variety of data on sensors (open for WUR) https://research.wur.nl/en/projects/sensing-potential-kb-38-001-008
- Data from Terahertz sensor collected by OnePlanet (not open to WUR)

Models

- Unidentified models
- Protocol for describing sensors

Tools

- Pytorch: https://pytorch.org/
- Tensorflow: https://www.tensorflow.org/
- MIRA https://www.perclass.com/apps/perclass-mira. This is a paid tool for hyperspectral data analysis. WPR, WFBR and WFSR have purchased the license to use these tools
- Perclass (MATLAB) https://www.perclass.com/perclass-toolbox/product
- Unscrambler - for multivariate analysis, standard tool for spectroscopy (WUR can provide access) https://www.camo.com/unscrambler/?utm_expid=.u3mRHJ7aRgaLopdzAYcHYQ.0&utm_referrer=
- Spectral-python for spectral data http://www.spectralpython.net/
- Sensors
    - for greenhouse: Net radiometers, PAR sensors, Slab sensors (water content- temperature), Weighing gutters(Aquabalance), Crop scales, Chlorophyll fluorescence, Crop temperatures sensors, Stem diameter sap flow meter sensors
    - for other use-cases: RGB (normal color cameras), Thermal camera (Livestock temperature measurement), several spectral cameras (Specim FX-10, Specim-FX17, IMEC SWIR) and hand-held spectrometers (ASD-Labspec, SCIO)

Infrastructure

- 4TU repository for survey data

Ethics

- Has not yet been discussed

Gaps

- Tools for automatic and continuous detection of relevant publications and reports

- How to arrange IP, for example when working with OnePlanet?
- How to merge data from multiple sensors?

Project 9 - Community management of natural resources using high tech, mobile technology & connectivity solutions in small-holder farming context

Domain objectives

- Monitoring and communicating deforestation due to the production of cacao
- How to deal with the effect of cloud coverage on the satellite data?

Data

- Data from satellites LandSat (data too coarse), Sentinel 1 (optical) and Sentinel 2 (radar) (open for WUR)
- Satellite data from SIMIT (Open for WUR)
- ISRIK soil map (Open for WUR) https://www.isric.org/explore/isric-soil-data-hub
- Crowdsourced observations via smartphones from local farmers, including photos (Open for WUR)
- Online predictions from www.cbm.wur.nl, a platform to obtain historical data for certain locations (Open for WUR)

Models

Tools

- LSTM, a deep learning memory-based model to analyze time-based satellite data (Long short-term memory (LSTM) is an artificial recurrent neural network architecture used to predict forest change)
- Groenmonitor for automatic cloud detection. http://www.groenmonitor.nl/
- Opendatakit.org (ODK), a customizable data collection application on smartphones.
- Sentinel toolset to translate raw sentinel data (https://earth.esa.int/web/guest/software-tools/content/-/article/sentinel-toolboxes)

Infrastructure

- Facebook is used for communication
    - Data is shared within WUR (Weblink with username and password to exchange the data)
- Ethics
- Focus group discussions on deforestation
- A survey will determine the current socio-economic variables and how they evolve over time
- There are no privacy issues

Gaps

- Image recognition

Project 13 (now in 7) - Sensing the City

Domain objectives

- Discover patterns in rat occurrences in Amsterdam, in relation to other factors such as fast food locations and waste bins, in order to reduce nuisance

Data

- AMS (Amsterdam Metropolitan Solutions) data sets, including rat sightings (location and time) over 3 years
- Data from the municipal waste service in Amsterdam
- Crowd-sourced data on rat sightings or proxies thereof
- Information from Rentokil (pest control)

Models

Tools

- ArcGIS is used to visualize data https://www.arcgis.com/
- QGIS is a free variant, but has fewer facilities https://qgis.org/

Infrastructure

- The AMS data can be obtained via the Open Data Portal

Ethics

- No personal data, but maps may show 'less attractive' neighbourhoods
- Data relate to public health

Gaps

- More sensor information needed (possible with OnePlanet as a partner)
- Interoperability techniques for using external data

Project 14 (now in 7) - ClimateImpactMonitor.eu

Domain objectives

- The aim of the project is to provide insight into the socio-economic impact of weather, specifically drought

Data

- No data available at this point. Relevant data are: land use, agricultural production, insurance data, drought index, weather, groundwater levels

Models

Tools

Infrastructure

Ethics

- Data from social media and insurance company can be privacy sensitive

Gaps