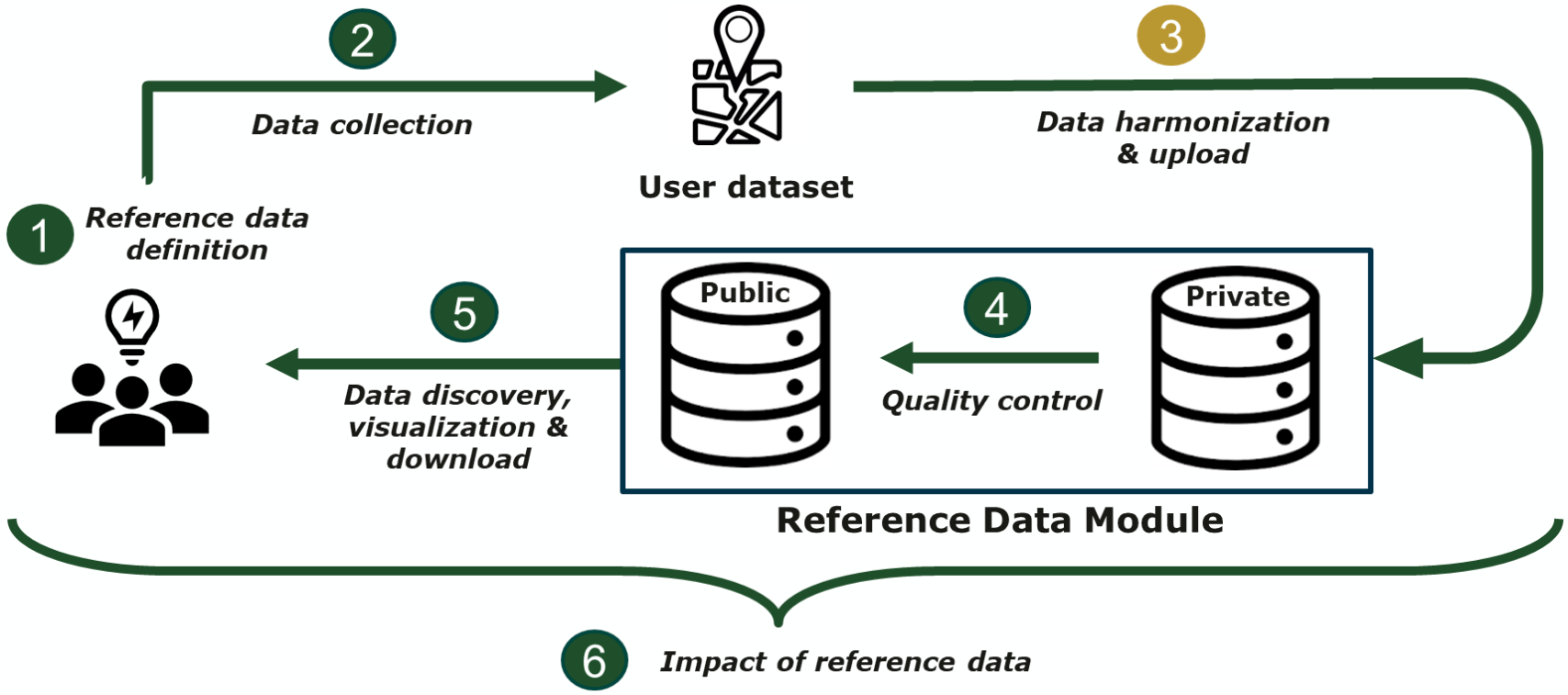
WorldCereal

WorldCereal MOOC I

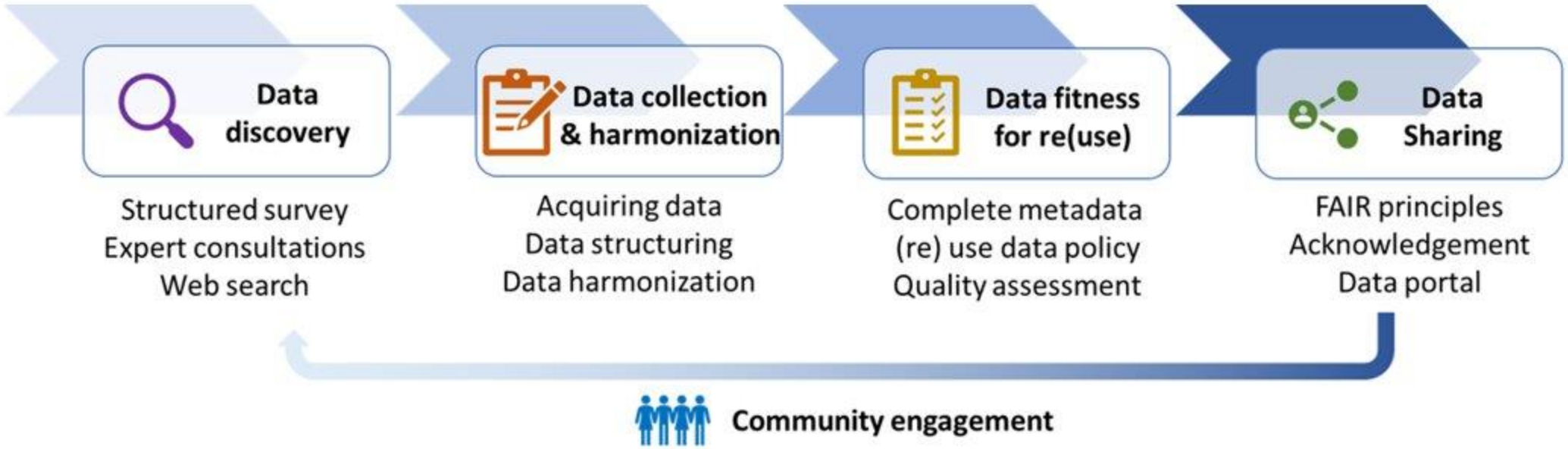
Reference data harmonization and cleaning

Hendrik Boogaard (WENR)

MOOC I: Outline



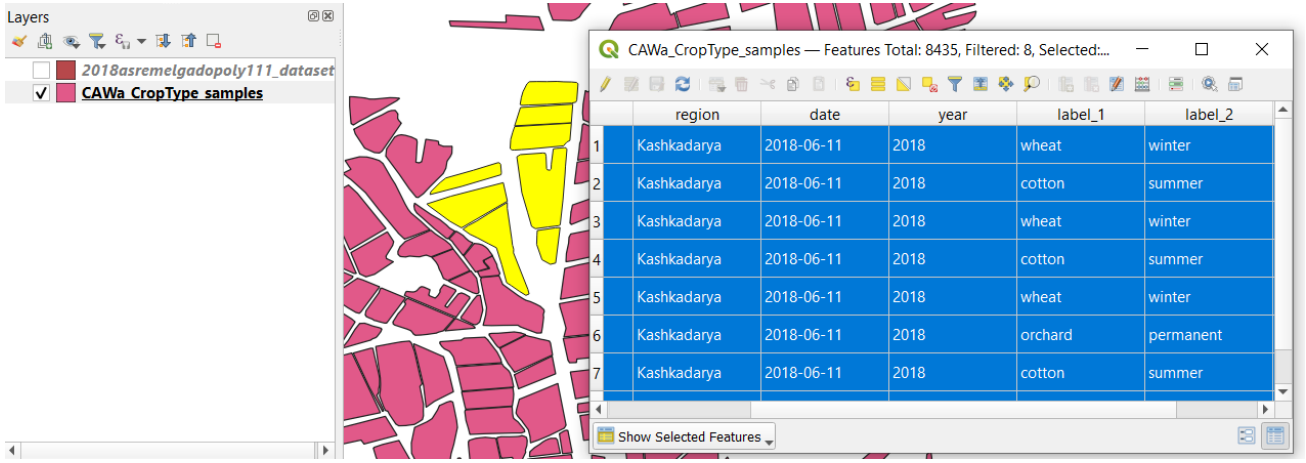
WorldCereal reference data workflow



Generic framework on reference data employed in WorldCereal

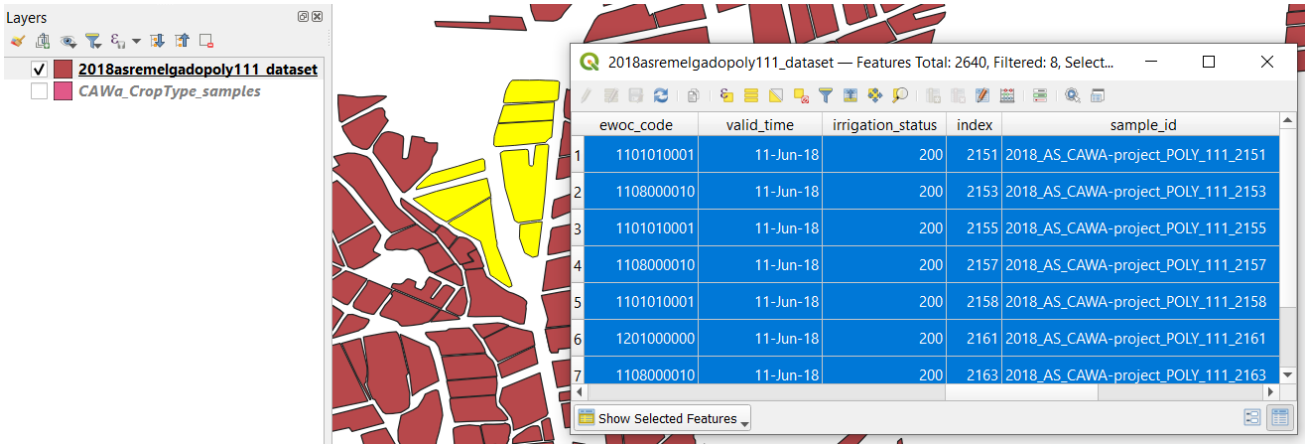
WorldCereal

Harmonizing reference data



- Cleaning data:
  - Remove errors
  - Remove missing labels
  - Remove old years
- Standard filename
- Adding standard attributes
- Filling standard attributes:
  - **WorldCereal crop label**
  - **Validity time**
  - Unique readable ID
- Reprojecting to EPSG:4326

- **Assessing confidence/quality**:
  - Spatial accuracy
  - Temporal accuracy
  - Thematic accuracy
- **Complete metadata**:
  - License
  - Citation
  - Provider/holder
  - Background/provenance
- **Adding sampling labels** for large data sets



Standard crop legend

- Standard crop legend to ensure we all speak the same language

- Hierarchical legend based on HCAT

Land cover: Temporary crops
↳ Land cover subgroup: Cereals
↳ Crop group: Wheat
↳ Crop type: Durum wheat
↳ Variety/management: Winter durum wheat

- Dynamic legend

- Check legend here

WorldCereal

AI-assisted crop type mapping

- AI-model, based on OpenAI, suggests WorldCereal legend labels for each original label during user data upload

	Manual Input	Google Translate + RegExp	OpenAI API
Fast	✗	✓	✓
Easy to setup	✗	✓	✓
Good quality	✓	✗	✓
User Friendly	✗	✓	✓
Cheap	✓	✓	✓*

* ~0.02 EUR per request

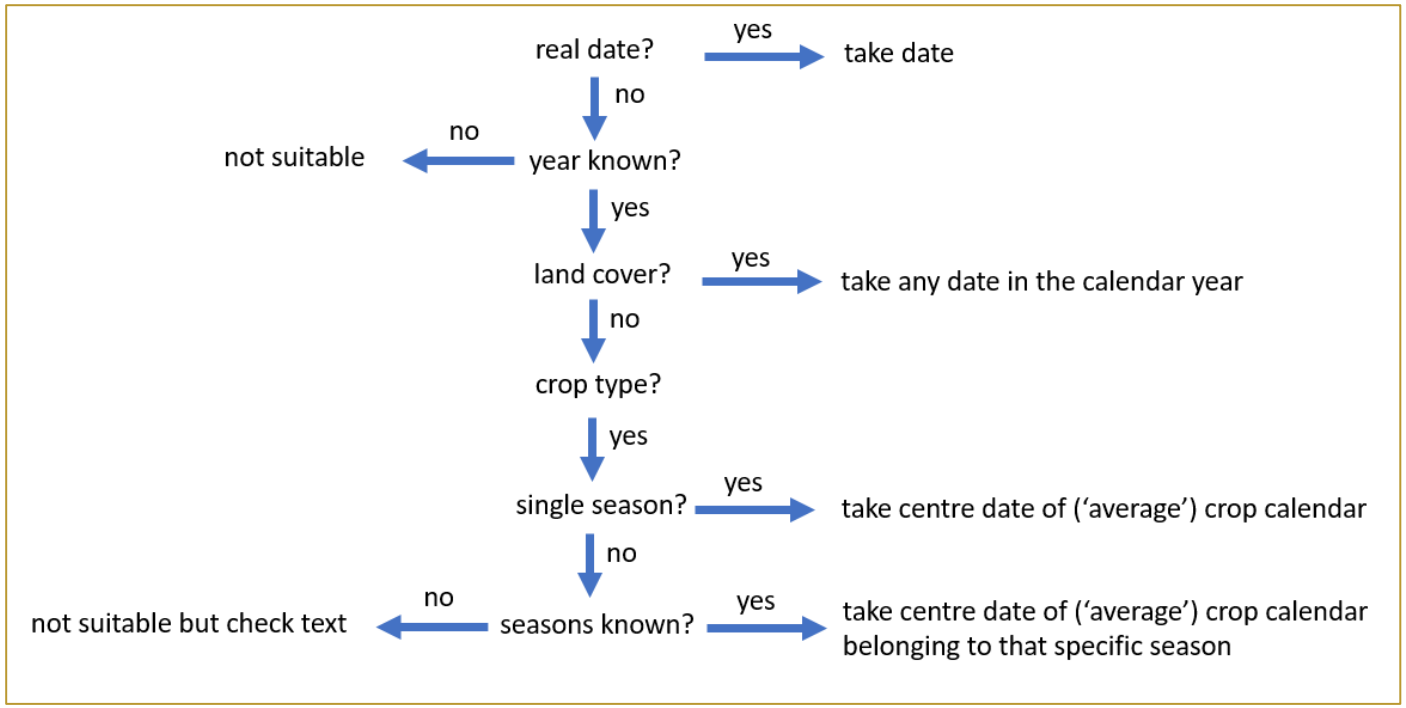
User Input	OpenAI Response
wheat	unspecified_wheat
Durum wheat	durum_hard_wheat
taARWe	unspecified_wheat
wheet	unspecified_wheat
пшениця	unspecified_wheat
WHT	unspecified_wheat
🌾	unspecified_wheat
pszenica 🌽	unspecified_wheat
wheat/maize	cereal_mixed_with
corn (preceded by wheat)	grain_maize_corn_popcorn
smth to feed animals, not wheat	fodder_wheat
wheat + alfalfa	cereal_mixed_with

WorldCereal

Validity time

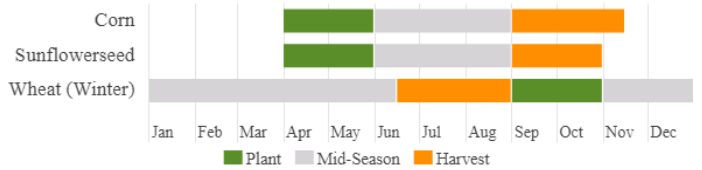
- Each sample has its own validity time
- If missing, the date must be derived from year (and season) of observation
- Especially relevant for crop type

Data type: Field Observation Survey



Data type: Virtual Interpretation

Date of image or photo

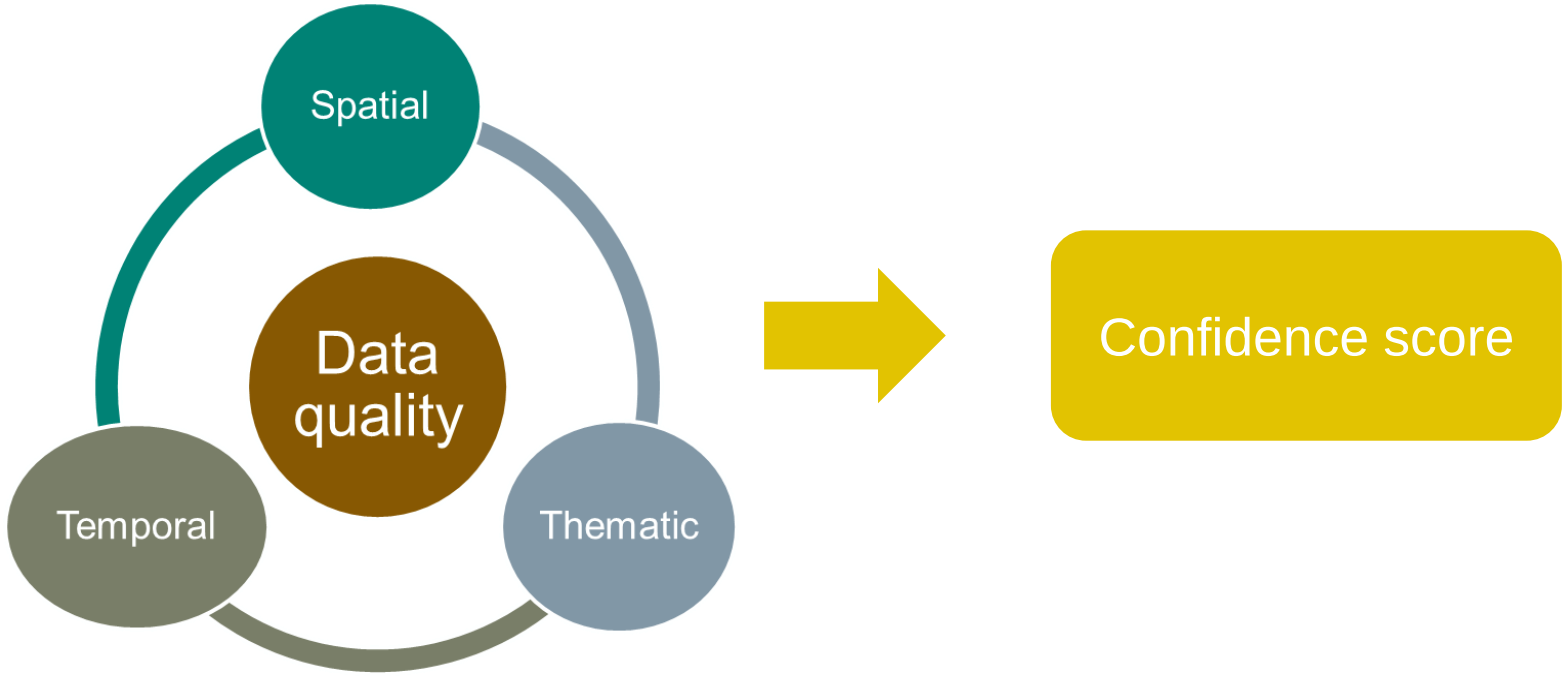


For instance USDA crop calendars, in future WorldCereal calendars can be used

WorldCereal

Quality assessment

- Generic scheme for assessing fitness for use of reference data
- Confidence scores for spatial, temporal, and thematic accuracy



See chapter 4 "Quality assessment of reference data"

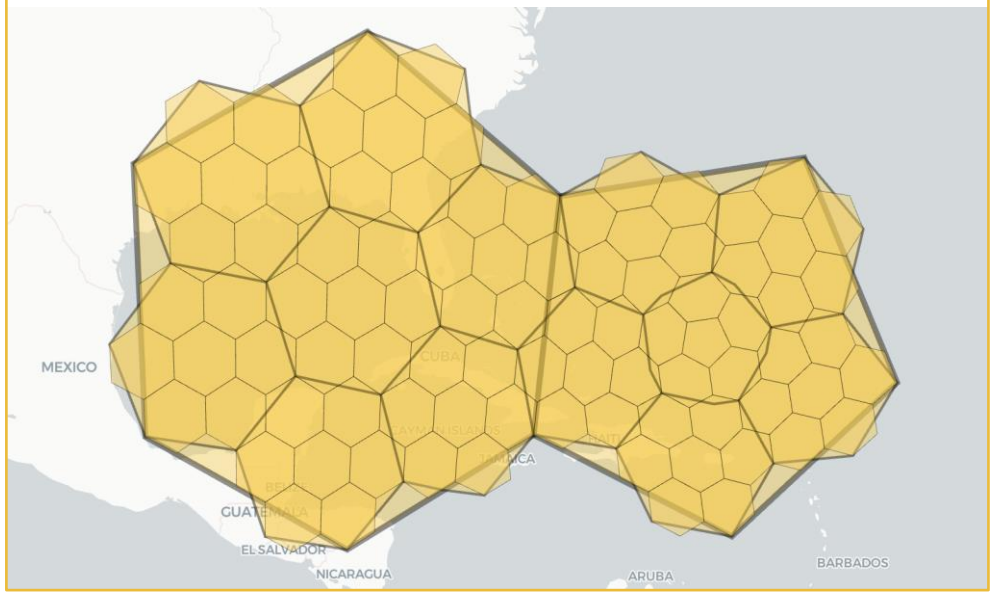
WorldCereal

Dataset sampling

Large datasets are automatically **sampled**, taking into consideration the spatial distribution and land cover/crop type label, ensuring that a relevant subset is automatically available for downstream crop mapping tasks

- Every sample in the dataset gets assigned a corresponding **H3 grid cell** at resolution 3 (H3 L3 cell)
- Occurring crop types in the dataset are grouped into crop categories, e.g. vegetables, dry pulses, oilseed crops
- For each H3 L3 cell and for each crop category, 50 random spatially distributed samples are selected for further processing

Hexagonal Grid Representation

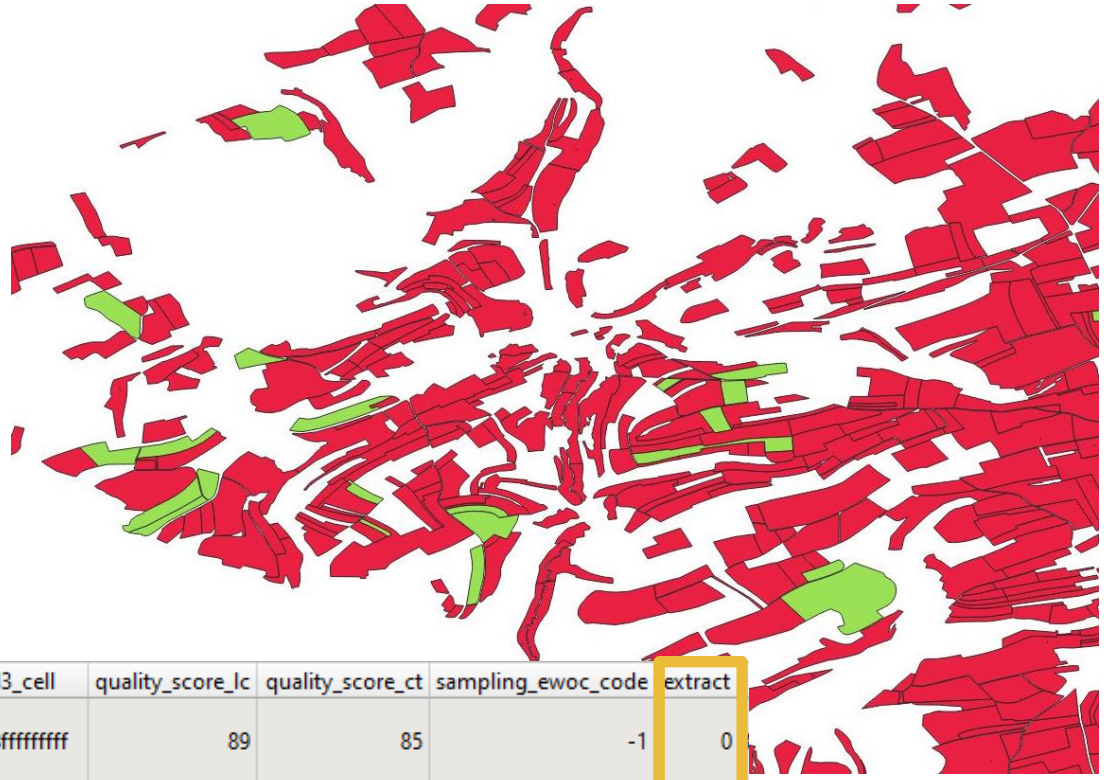


WorldCereal

Dataset sampling

Example of sampling result for spatially dense dataset (LPIS Luxembourg)

Both the full and sampled dataset are available for download in RDM



valtime	sampleID	ewoc_code	valid_time	irrigation_status	sample_id	h3_l3_cell	quality_score_lc	quality_score_ct	sampling_ewoc_code	extract
2017-08-01	2017_AF_OAF_POINT_1103	1000000000	01/08/2017	0	2017_AF_One-Acre-Fund...	837a48fffffffff	89	85	-1	0
2017-08-01	2017_AF_OAF_POINT_1104	1101060000	01/08/2017	0	2017_AF_One-Acre-Fund...	837a48fffffffff	89	85	110106000	2
2017-08-01	2017_AF_OAF_POINT_1105	1000000000	01/08/2017	0	2017_AF_One-Acre-Fund...	837a48fffffffff	89	85	-1	0
2017-08-01	2017_AF_OAF_POINT_1106	1101060000	01/08/2017	0	2017_AF_One-Acre-Fund...	837a48fffffffff	89	85	110106000	2

WorldCereal

Metadata

Completing metada (following FAIR principles)

- Title
- Unique ID
- Creator (name, URL, e-mail)
- License
- Citation
- Description (DataSet Name)
- Related publication (ReferenceDataSet)
- Coverage (spatial, temporal)

Title
A crop type dataset on Central Asia, 2018 (Remelgado et al, 2020)

Collection ID
2018asremelgadopoly111

Feature...	Dataset...	Sample...	Metada...
2639	1	0	2

Region
AS

Geometry type
Polygon

Observation Time
Real Date

Date Range of Observations
1/3/2018 to 1/9/2018

Worldcereal Reference Documents

- Crop type legend
- Irrigation Status legend
- About observation date
- Dataset confidence score calculation

Downloads

Metadata Excel	Download
Harmonized Dataset	Download
Harmonization Steps	Download
Sample Extracts	Download

Citation

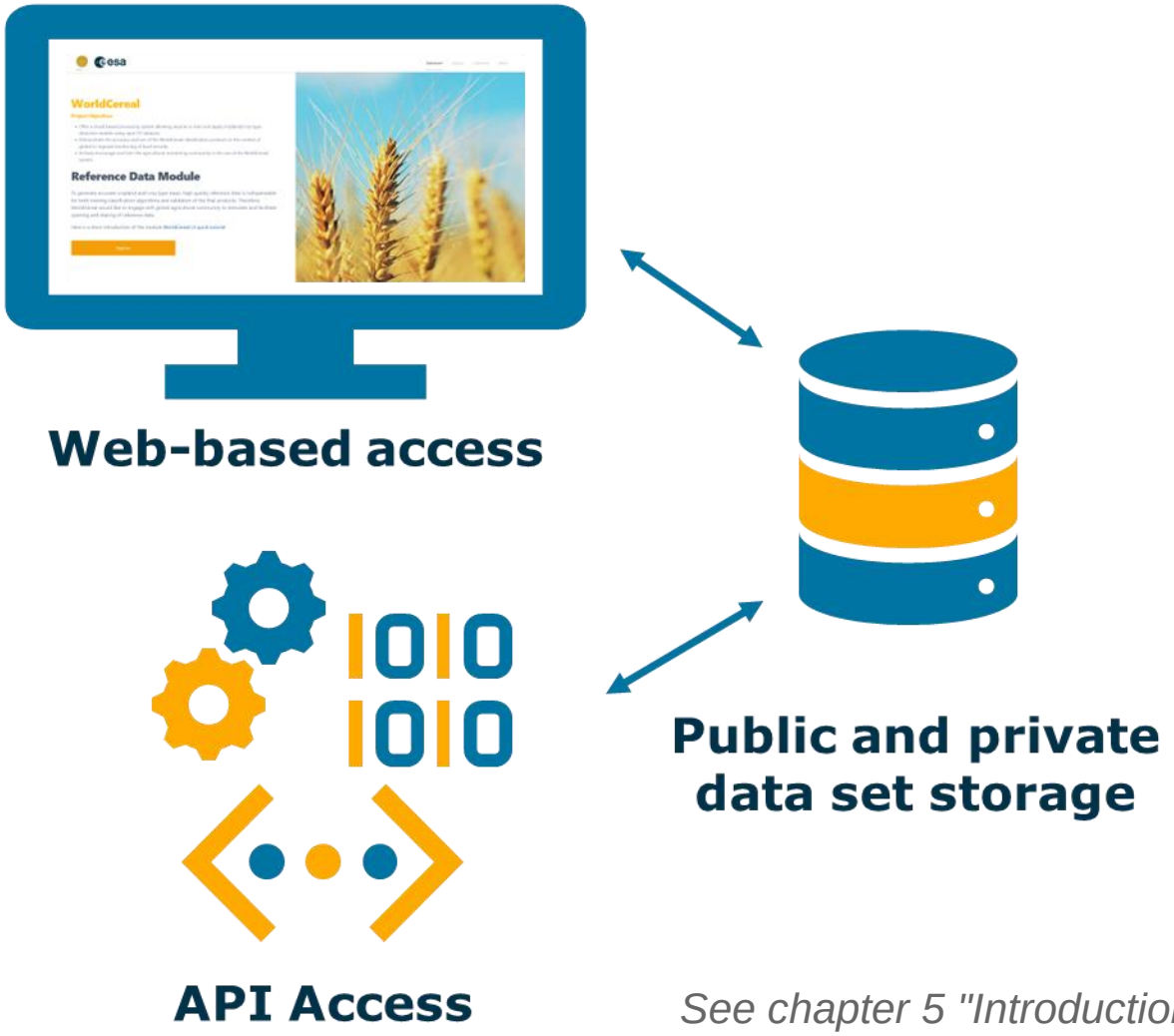
Remelgado, R., Zaitov, S., Kenjabaev, S., Stulina, G., Sultanov, M., Ibrakhimov, M., Akhmedov, M., Dukhovny, V. and Conrad, C., 2020. A crop type dataset for consistent land cover classification in Central Asia. Scientific Data, 7(1), pp.1-6.

Dataset Provider Details

Code	Description	Url	Contact
CAWa project (Remelgado et al, 2020)	Central Asia Waters (CAWa) (Remelgado et al, 2020)	www.cawa-project.net	ruben.remelgado@idiv.de

DataSet Name	ReferenceDataSet	Type Of License	ReferenceToLicense
A crop type dataset for consistent land cover classification in Central Asia	https://doi.org/10.1038/s41597-020-00591-2	CC_BY	

Storage & access



Reference data is managed in the WorldCereal Reference Data Module (RDM)

See chapter 5 "Introduction to the WorldCereal Reference Data Module"

WorldCereal

Storage & access

Reference Data Collections

Public Collections available as input for processing

Collections	Features	Year
97	53346399	All

https://rdm.esa-worldcereal.org/

Title

A crop type dataset on Central Asia, 2018 (Remelgado et al, 2020)

Collection ID

2018asremelgadopoly111

Featu...	Datas...	Samp...	Meta...
2639	1	0	2

Region

AS

Geometry type

Polygon

Observation Time

Real Date

Date Range of Observations

1/3/2018 to 1/9/2018

Worldcereal Reference Documents

- Crop type legend
- Irrigation Status legend
- About observation date
- Dataset confidence score calculation

Downloads

Metadata Excel	Download
Harmonized Dataset	Download
Harmonization Steps	Download
Sample Extracts	Download

Citation

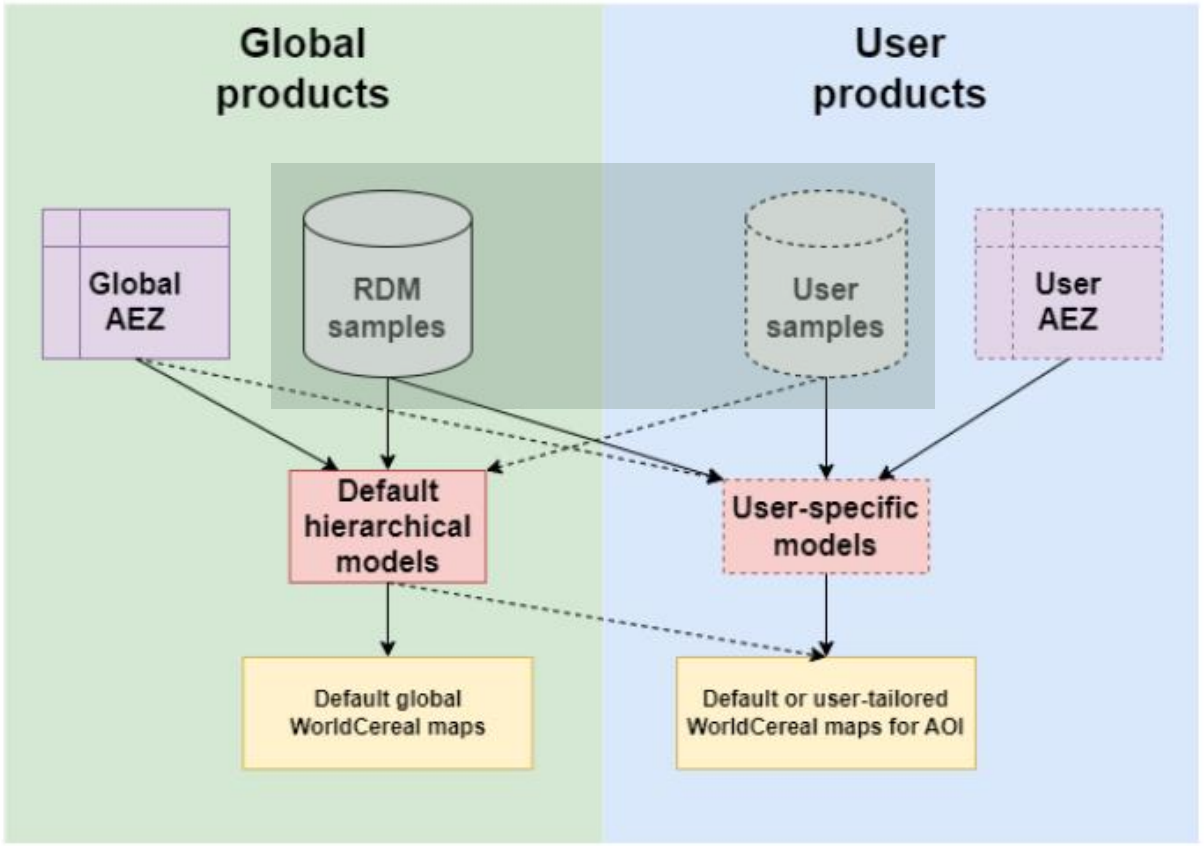
Remelgado, R., Zaitov, S., Kenjabaev, S., Stulina, G., Sultanov, M., Ibrakhimov, M., Akhmedov, M., Dukhovny, V. and Conrad, C., 2020. A crop type dataset for consistent land cover classification in Central Asia. Scientific Data, 7(1), pp.1-6.

Also accessible via GEOSS portal

- https://www.geoportal.org
- Thematic area: “harmonized agronomy in-situ data”
- Catalog: WorldCereal

WorldCereal

Adding reference data



Users can upload their private data and combine it with public data to train custom models (customized growing season and crop type)

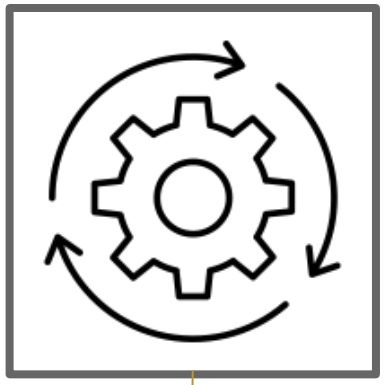
WorldCereal

Semi-automated harmonization tool

User dataset

user-data.gpkg

Semi-automated harmonization



Harmonized dataset with consistent naming, data attributes, legend and metadata

2020_MEX_SAMPLE_POINT_111.geoparquet

sample_id	ewoc_code	valid_time

Questionnaire

Attribute identifying crop type label?

Short identifier?

Attribute identifying date label?

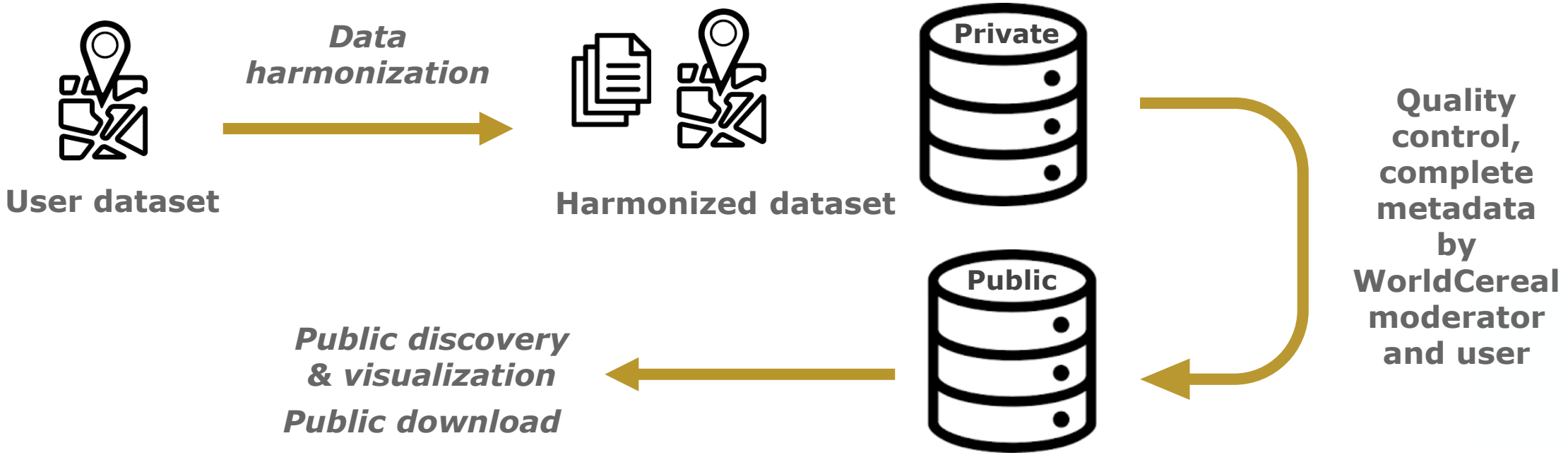
AI model based mapping

original_label	ewoc_code
Frijol	1105010010
Maiz	1101060000

user checks generated mapping

Sharing reference data

- Public: share your reference data with public → **Additional metadata quality control !**
- Restricted: the WorldCereal consortium and you can use the data
- Private: only you can use the data



Take home messages

- Harmonization, curation and quality assessment is crucial for proper re-use of published reference data
- WorldCereal has collected, harmonized and curated many datasets which are available for the mapping community
- However large spatial gaps exists
- WorldCereal offers user-friendly upload procedure to upload and use own reference data and if possible share data with public or consortium (restricted)
- WorldCereal promotes the sharing of reference data to further improve the accuracy of cropland and crop type maps across the globe
- Learn more through our dedicated documentation page

WorldCereal

WorldCereal

THANK YOU

Interesting links:

About ref data → https://esa-worldcereal.org/en/reference-data

RMD UI → https://rdm.esa-worldcereal.org/

Documentation → https://worldcereal.github.io/worldcereal-documentation/rdm/overview.html

Questions? → WorldCereal Forum MOOC I

Subscribe to our mailing list

esa vito gisat IIASA Université de Strasbourg Universitat de València Wageningen University & Research