Data management at CIMMYT Wheat Physiology
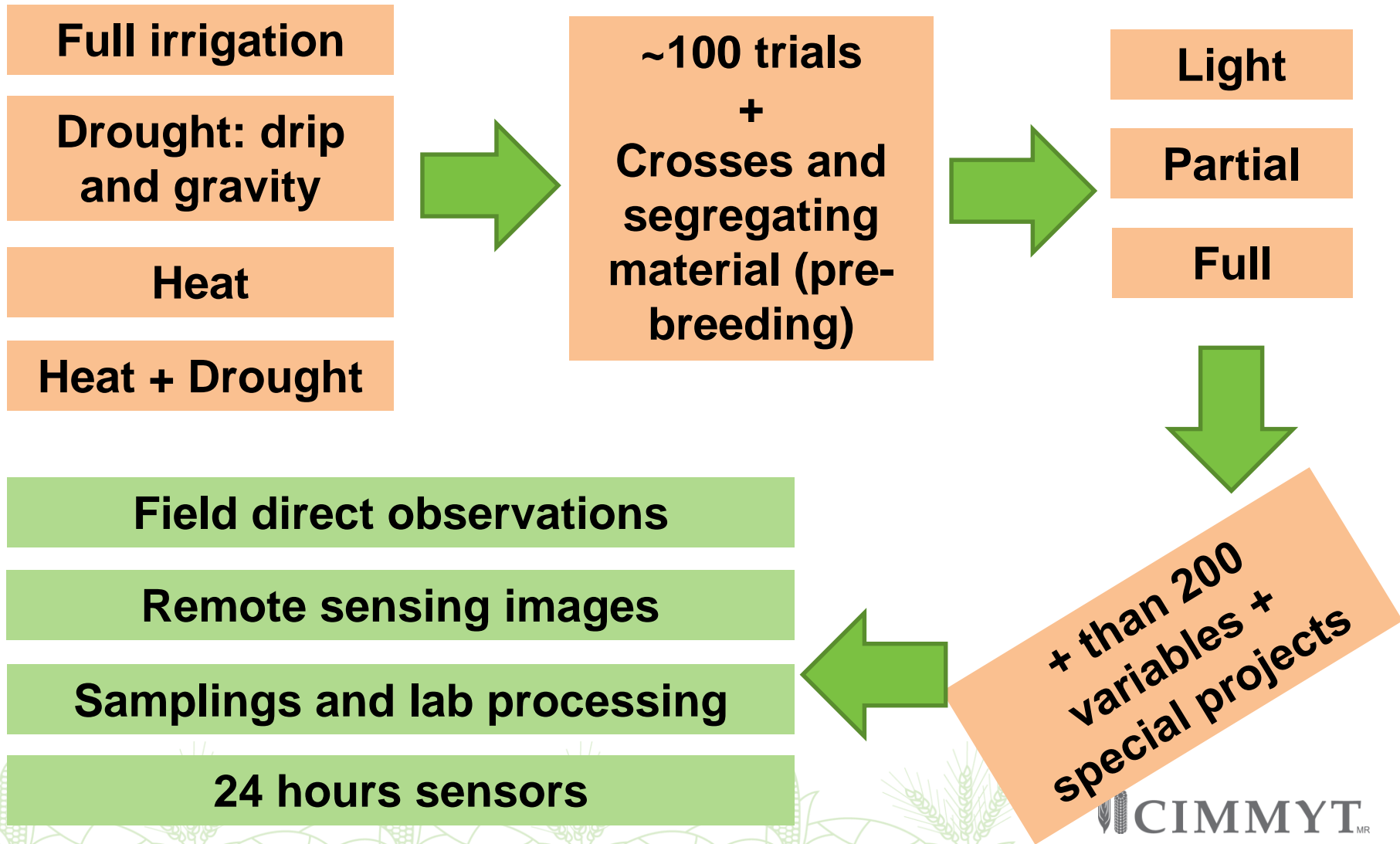
Carolina Rivera

CSIRO-CIMMYT Collaborative workshop

12th June 2017

CIMMYT MR

Traits currently measured at CIMMYT Wheat Physiology

+ calculated/complex phenotypic traits

CIMMYT Wheat Physiology current capacity

Full irrigation

Drought: drip and gravity

Heat

Heat + Drought

→

**~100 trials
+
Crosses and segregating material (pre-breeding)**

→

Light

Partial

Full

↓

+ than 200 variables + special projects

←

Field direct observations

Remote sensing images

Samplings and lab processing

24 hours sensors

CIMMYT MR

Other types of data (metadata or information)

Environmental

Trial management/Exp designs
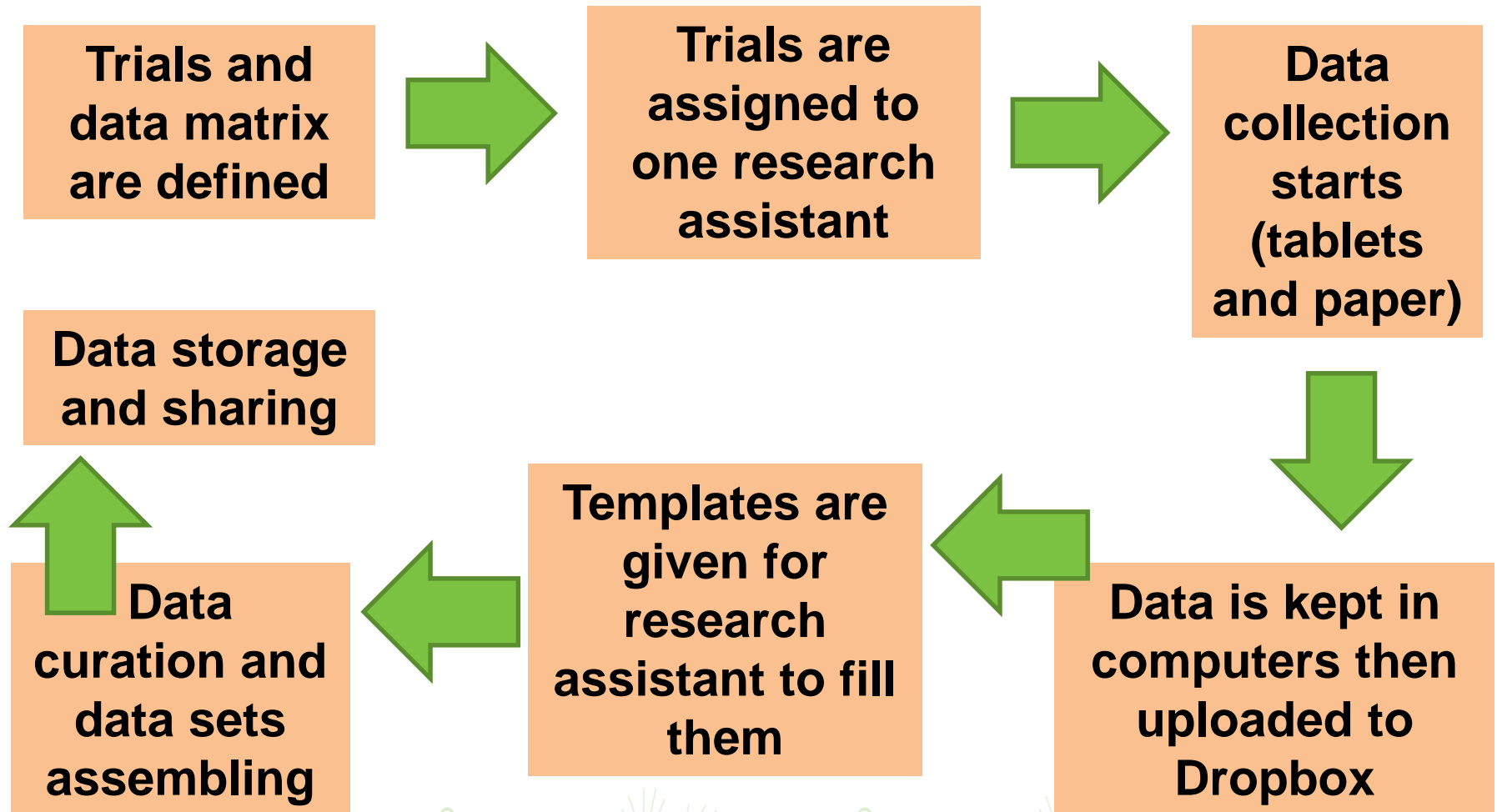

Seed inventories

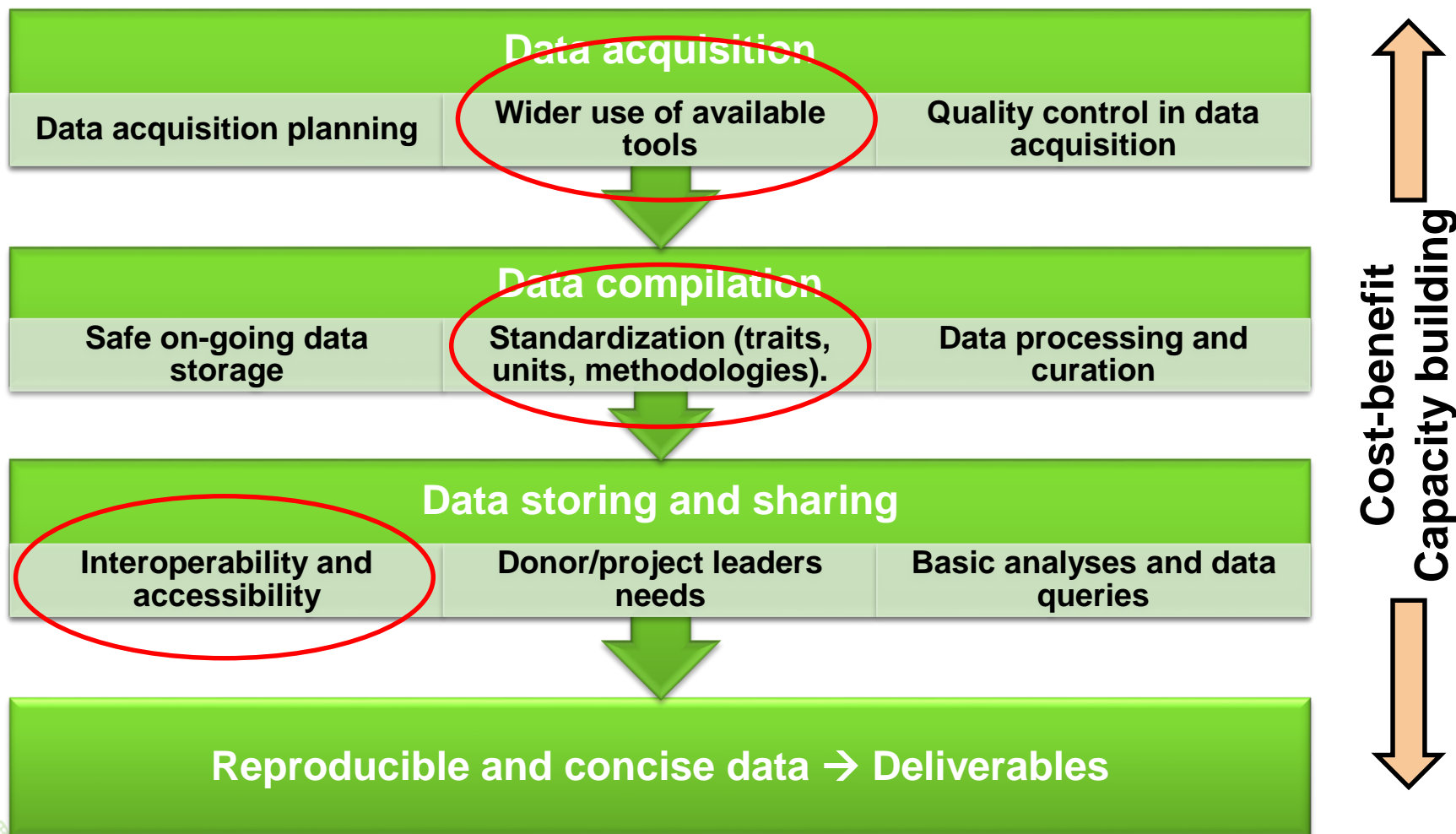**Pedigree/Genealogy**
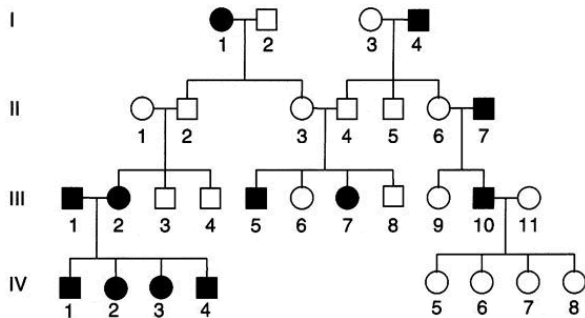
CID
SID
GID
Cross name
Selection history

Historical data

Methodologies

General data acquisition and management pipeline at CIMMYT WP

Trials and data matrix are defined → **Trials are assigned to one research assistant** → **Data collection starts (tablets and paper)**

↓

Data storage and sharing ← **Data curation and data sets assembling** ← **Templates are given for research assistant to fill them** ← **Data is kept in computers then uploaded to Dropbox**

CIMMYT MR

Data acquisition and management overview of opportunities for improvement

Data acquisition

| Data acquisition planning | Wider use of available tools | Quality control in data acquisition |

Data compilation

| Safe on-going data storage | Standardization (traits, units, methodologies). | Data processing and curation |

Data storing and sharing

| Interoperability and accessibility | Donor/project leaders needs | Basic analyses and data queries |

Reproducible and concise data → Deliverables

Cost-benefit Capacity building

Specific challenges/opportunities in wheat data acquisition

- Centralized planning of measurements

- Wider use of available tools

- On going data quality control

- Capacity building (primary sources)

Increases accuracy and optimizes use of resources

CIMMYT MR

Case: planning measurements and on-going data quality control

Opportunities:

- Follow this scheme for other measurements based on phenology
- Create more automated spreadsheets/templates

Case: Wider use of available tools for data collection

Opportunities:

- Merge with available ontology for physiological traits.
- Use bar codes and specific software for scales

In field scanning

Challenges/opportunities in wheat data processing/curation

Accelerate and improve quality in data processing and curation of several physiological traits.

- Quality check and basic curation done mainly in Excel for general physiological traits → KDXplore? R? Interfaces?

- Access to data visualization

CIMMYT MR

Case: NDVI (GB) data fast track processing

Set parameters to discard outliers

Case: IRGAS data processing

Image and screen prints: Rut Sanchez

Codes and/or programing equipment

Case: historical wheat research data

There is a need to migrate historical data from older software (e.g. UQ123) to more accessible and compatible databases

In the process, we could arrange, process and analyze the data to make it accessible to collaborators and scientist for publications.

- ESWYT historical yield and TGW data, 10 years, ~60 locations, 50 genotypes.
- Data arrangement for complex statistical analyses.
- Manual → R

Sivakumar Sukumaran and Carolina Rivera

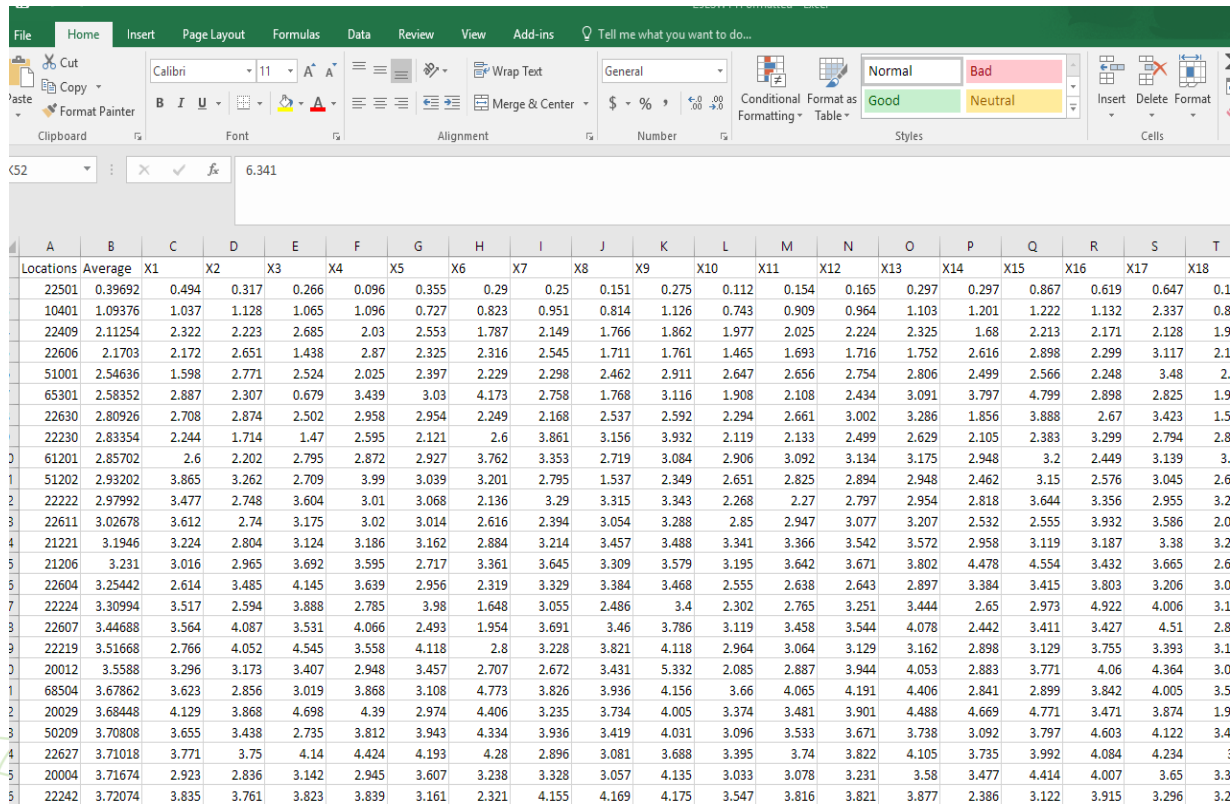| Locations	Average	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18
22501	0.39692	0.494	0.317	0.266	0.096	0.355	0.29	0.25	0.151	0.275	0.112	0.154	0.165	0.297	0.297	0.867	0.619	0.647	0.1
10401	1.09376	1.037	1.128	1.065	1.096	0.727	0.823	0.951	0.814	1.126	0.743	0.909	0.964	1.103	1.201	1.222	1.132	2.337	0.8
22409	2.11254	2.322	2.223	2.685	2.03	2.553	1.787	2.149	1.766	1.862	1.977	2.025	2.224	2.325	1.68	2.213	2.171	2.128	1.9
22606	2.1703	2.172	2.651	1.438	2.87	2.325	2.316	2.545	1.711	1.761	1.465	1.693	1.716	1.752	2.616	2.898	2.299	3.117	2.1
51001	2.54636	1.598	2.771	2.524	2.025	2.397	2.229	2.298	2.462	2.911	2.647	2.656	2.754	2.806	2.499	2.566	2.248	3.48	2
65301	2.58352	2.887	2.307	0.679	3.439	3.03	4.173	2.758	1.768	3.116	1.908	2.108	2.434	3.091	3.797	4.799	2.898	2.825	2
22630	2.80926	2.708	2.874	2.502	2.958	2.954	2.249	2.168	2.537	2.592	2.294	2.661	3.002	3.286	1.856	3.888	2.67	3.423	1.5
22230	2.83354	2.244	1.714	1.47	2.595	2.121	2.6	3.861	3.156	3.932	2.119	2.133	2.499	2.629	2.105	2.383	3.299	2.794	2.8
61201	2.85702	2.6	2.202	2.795	2.872	2.927	3.762	3.353	2.719	3.084	2.906	3.092	3.134	3.175	2.948	3.2	2.449	3.139	3
51202	2.93202	3.865	3.262	2.709	3.99	3.039	3.201	2.795	1.537	2.349	2.651	2.825	2.894	2.948	2.462	3.15	2.576	3.045	3
22222	2.97992	3.477	2.748	3.604	3.01	3.068	2.136	3.29	3.315	3.343	2.268	2.27	2.797	2.954	2.818	3.644	3.356	2.955	3.2
22611	3.02678	3.612	2.74	3.175	3.02	3.014	2.616	2.394	3.054	3.288	2.85	2.947	3.077	3.207	2.532	2.555	3.932	3.586	2.0
21221	3.1946	3.224	2.804	3.124	3.186	3.162	2.884	3.214	3.457	3.488	3.341	3.366	3.542	3.572	2.958	3.119	3.187	3.38	3.2
21206	3.231	3.016	2.965	3.692	3.595	2.717	3.361	3.645	3.309	3.579	3.195	3.642	3.671	3.802	4.478	4.554	3.432	3.665	2.6
22604	3.25442	2.614	3.485	4.145	3.639	2.956	2.319	3.329	3.384	3.468	2.555	2.638	2.643	2.897	3.384	3.415	3.803	3.206	3.0
22224	3.30994	3.517	2.594	3.888	2.785	3.98	1.648	3.055	2.486	3.4	2.302	2.765	3.251	3.444	2.65	2.973	4.922	4.006	3.1
22607	3.44688	3.564	4.087	3.531	4.066	2.493	1.954	3.691	3.46	3.786	3.119	3.458	3.544	4.078	2.442	3.411	3.427	4.51	2.8
22219	3.51668	2.766	4.052	4.545	3.558	4.118	2.8	3.228	3.821	4.118	2.964	3.064	3.129	3.162	2.898	3.129	3.755	3.393	3.4
20012	3.5588	3.296	3.173	3.407	2.948	3.457	2.707	2.672	3.431	5.332	2.085	2.887	3.944	4.053	2.883	3.771	4.06	4.364	3.0
68504	3.67862	3.623	2.856	3.019	3.868	3.108	4.773	3.826	3.936	4.156	3.66	4.065	4.191	4.406	2.841	2.899	3.842	4.005	3.5
20029	3.68448	4.129	3.868	4.698	4.39	2.974	4.406	3.235	3.734	4.005	3.374	3.481	3.901	4.488	4.669	4.771	3.471	3.874	1.9
50209	3.70808	3.655	3.438	2.735	3.812	3.943	4.334	3.936	3.419	4.031	3.096	3.533	3.671	3.738	3.092	3.797	4.603	4.122	3.4
22627	3.71018	3.771	3.75	4.14	4.424	4.193	4.28	2.896	3.081	3.688	3.395	3.74	3.822	4.105	3.735	3.992	4.084	4.234	
20004	3.71674	2.923	2.836	3.142	2.945	3.607	3.238	3.328	3.057	4.135	3.033	3.078	3.231	3.58	3.477	4.414	4.007	3.65	3.0
22242	3.72074	3.835	3.761	3.823	3.839	3.161	2.321	4.155	4.169	4.175	3.547	3.816	3.877	2.386					

Challenges and opportunities in data compilation-after data collection

- Use of templates for data compilation

- Defined protocols for data compilation

- Short term storage, allowing to store different versions of data sets: Dropbox, other clouds, KDXplore

- Define quality control check points when compiling data

- Capacity building (to speed up the data sets compilation process and improve data quality)

CIMMYT

Challenges and opportunities in data storing and sharing

- More extensive use of crop ontology
- Get as close as possible to donors needs
- Achieve efficient long-term storage databases
  - Centralize data within the institution in an accessible way
  - Find data easily inside and outside the institution
  - Store different types of data/information
  - Make data queries and visualize results
  - Compatibility with other software/databases
- Assure computational capacity

CIMMYT MR

Examples of institutional platforms for sharing, storing and visualizing data

Storing and publishing files

Can we merge open access platforms for projects partners and CIMMYT staff?

Data warehouse allowing querying and visualizing of several types of data)