GCP Templates task: Rationale

- A significant quantity of data is already being continuously produced by GCP
  - But with no consistent way of storing the data.
- To avoid information loss, we can't wait for
  - Databases to be fully developed and/or
  - A complete set of import scripts to be produced.
- We need to manage these data immediately
  - So that it can be easily be stored, read, analyzed and then later consolidated into databases.

Templates task: The problem

- Researchers, analytical tools and databases all want data in **different formats**
- Ideally, we only want one way to **capture and store** the definitive version of the data
- Long term solution is store the data directly in a database
- However, often we need a short term solution
  - Since the data is being produced now!
  - Users don't have database experience

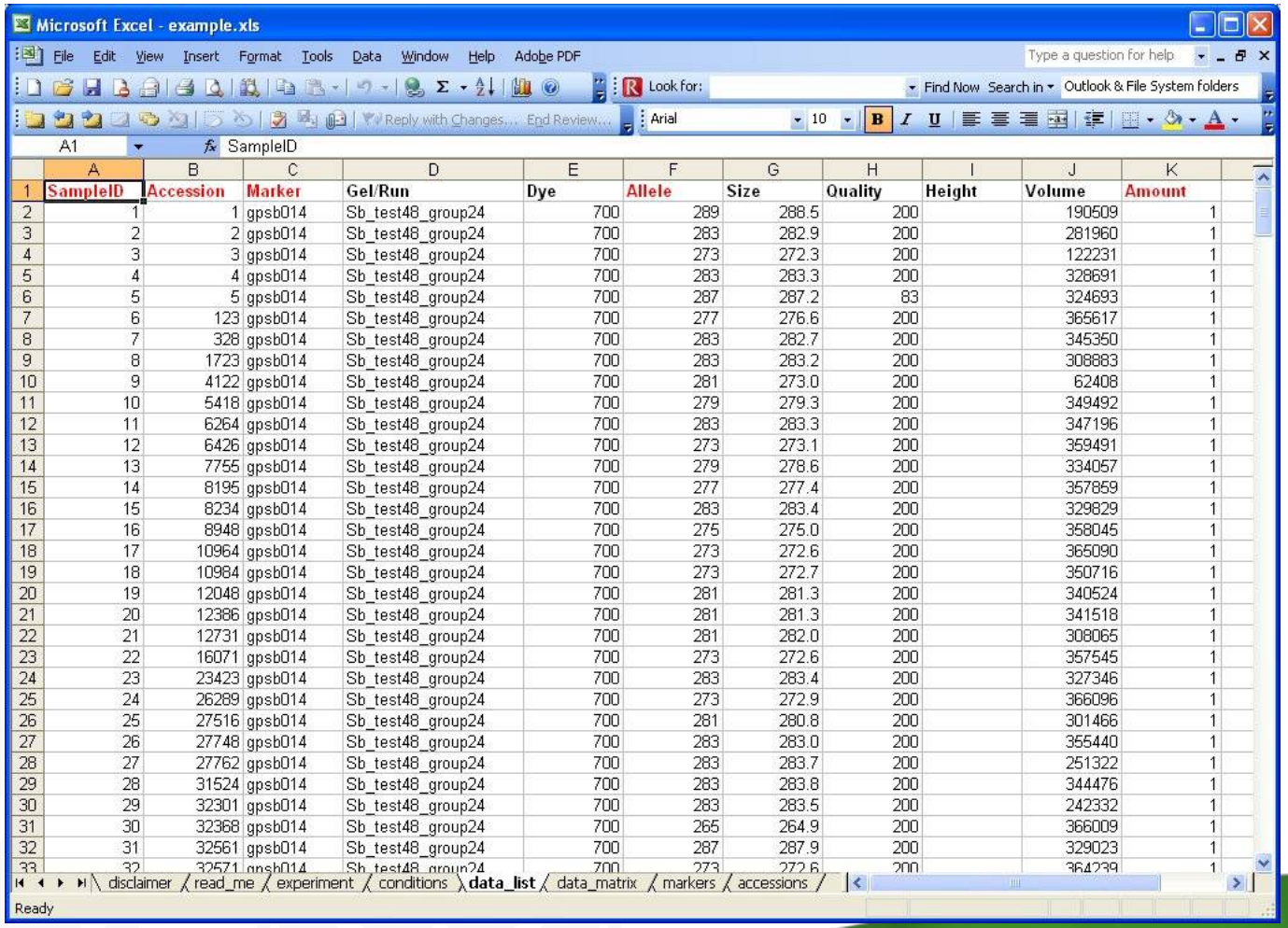
What is a template?

- A template is a type of form or formatted blank document
  - e.g. an Excel spreadsheet, flat file or a web form
- Templates provide
  - clearly defined format
  - precise semantics / ontology
  - community agreed minimal items
- Templates come with
  - sufficient user documentation with examples
  - data entry validation

Progress so far

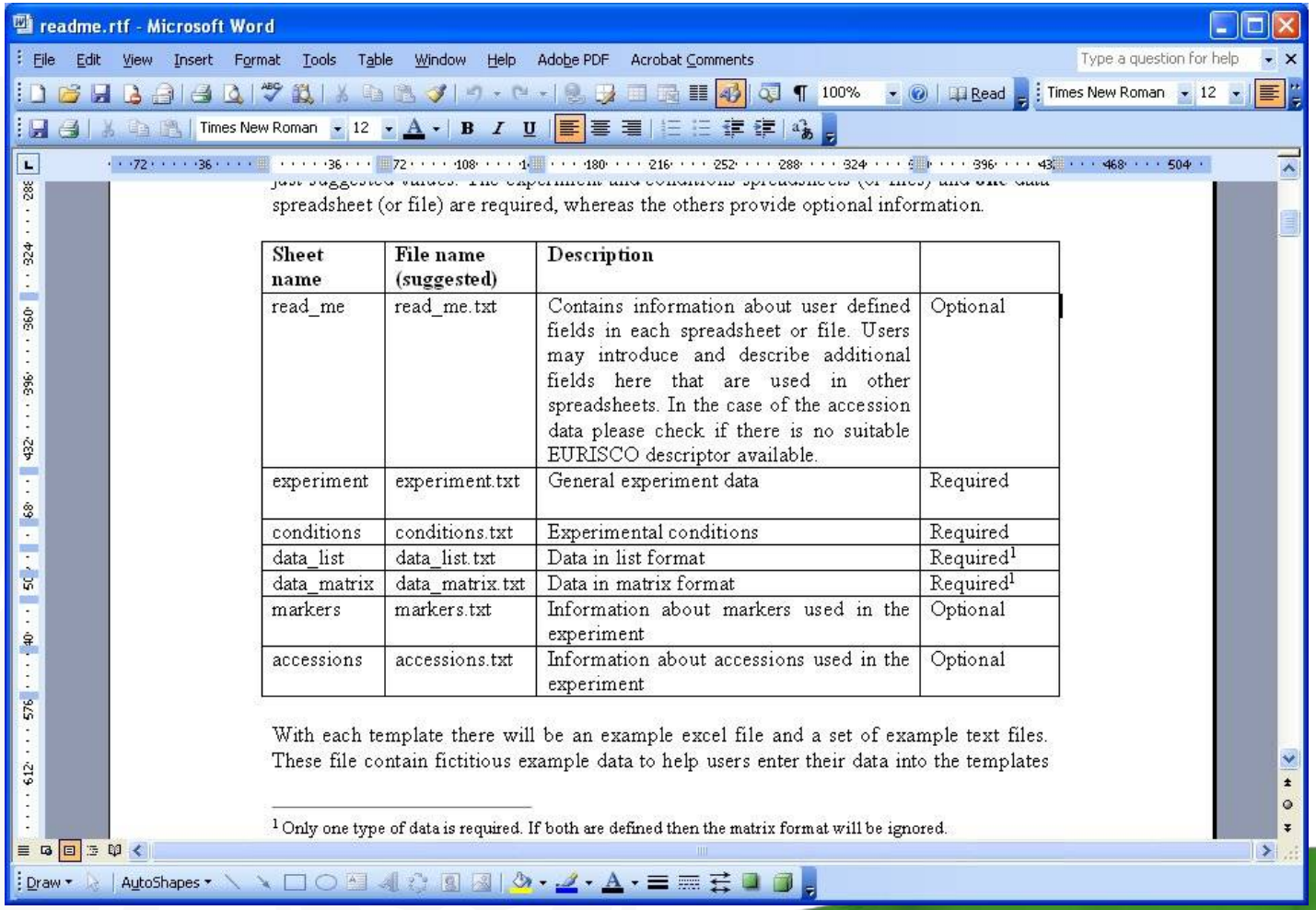
- We have developed templates for
  - SSR Genotyping data
  - Passport
- We will develop
  - Map and QTL data (June 2006)
  - Phenotypic data (June 2006)
  - SNP data (Late 2006)
  - Microarray data (based on MIAME)

Excel template example



	A	B	C	D	E	F	G	H	I	J	K
1	SampleID	Accession	Marker	Gel/Run	Dye	Allele	Size	Quality	Height	Volume	Amount
2	1	1	gpsb014	Sb_test48_group24	700	289	288.5	200		190509	1
3	2	2	gpsb014	Sb_test48_group24	700	283	282.9	200		281960	1
4	3	3	gpsb014	Sb_test48_group24	700	273	272.3	200		122231	1
5	4	4	gpsb014	Sb_test48_group24	700	283	283.3	200		328691	1
6	5	5	gpsb014	Sb_test48_group24	700	287	287.2	83		324693	1
7	6	123	gpsb014	Sb_test48_group24	700	277	276.6	200		365617	1
8	7	328	gpsb014	Sb_test48_group24	700	283	282.7	200		345350	1
9	8	1723	gpsb014	Sb_test48_group24	700	283	283.2	200		308883	1
10	9	4122	gpsb014	Sb_test48_group24	700	281	273.0	200		62408	1
11	10	5418	gpsb014	Sb_test48_group24	700	279	279.3	200		349492	1
12	11	6264	gpsb014	Sb_test48_group24	700	283	283.3	200		347196	1
13	12	6426	gpsb014	Sb_test48_group24	700	273	273.1	200		359491	1
14	13	7755	gpsb014	Sb_test48_group24	700	279	278.6	200		334057	1
15	14	8195	gpsb014	Sb_test48_group24	700	277	277.4	200		357859	1
16	15	8234	gpsb014	Sb_test48_group24	700	283	283.4	200		329829	1
17	16	8948	gpsb014	Sb_test48_group24	700	275	275.0	200		358045	1
18	17	10964	gpsb014	Sb_test48_group24	700	273	272.6	200		365090	1
19	18	10984	gpsb014	Sb_test48_group24	700	273	272.7	200		350716	1
20	19	12048	gpsb014	Sb_test48_group24	700	281	281.3	200		340524	1
21	20	12386	gpsb014	Sb_test48_group24	700	281	281.3	200		341518	1
22	21	12731	gpsb014	Sb_test48_group24	700	281	282.0	200		308065	1
23	22	16071	gpsb014	Sb_test48_group24	700	273	272.6	200		357545	1
24	23	23423	gpsb014	Sb_test48_group24	700	283	283.4	200		327346	1
25	24	26289	gpsb014	Sb_test48_group24	700	273	272.9	200		366096	1
26	25	27516	gpsb014	Sb_test48_group24	700	281	280.8	200		301466	1
27	26	27748	gpsb014	Sb_test48_group24	700	283	283.0	200		355440	1
28	27	27762	gpsb014	Sb_test48_group24	700	283	283.7	200		251322	1
29	28	31524	gpsb014	Sb_test48_group24	700	283	283.8	200		344476	1
30	29	32301	gpsb014	Sb_test48_group24	700	283	283.5	200		242332	1
31	30	32368	gpsb014	Sb_test48_group24	700	265	264.9	200		366009	1
32	31	32561	gpsb014	Sb_test48_group24	700	287	287.9	200		329023	1
33	32	32571	gpsb014	Sb_test48_group24	700	273	272.6	200		364239	1

Readme – instructions on use



readme.rtf - Microsoft Word

spreadsheet (or file) are required, whereas the others provide optional information.

Sheet name	File name (suggested)	Description	
read_me	read_me.txt	Contains information about user defined fields in each spreadsheet or file. Users may introduce and describe additional fields here that are used in other spreadsheets. In the case of the accession data please check if there is no suitable EURISCO descriptor available.	Optional
experiment	experiment.txt	General experiment data	Required
conditions	conditions.txt	Experimental conditions	Required
data_list	data_list.txt	Data in list format	Required[1]
data_matrix	data_matrix.txt	Data in matrix format	Required[1]
markers	markers.txt	Information about markers used in the experiment	Optional
accessions	accessions.txt	Information about accessions used in the experiment	Optional

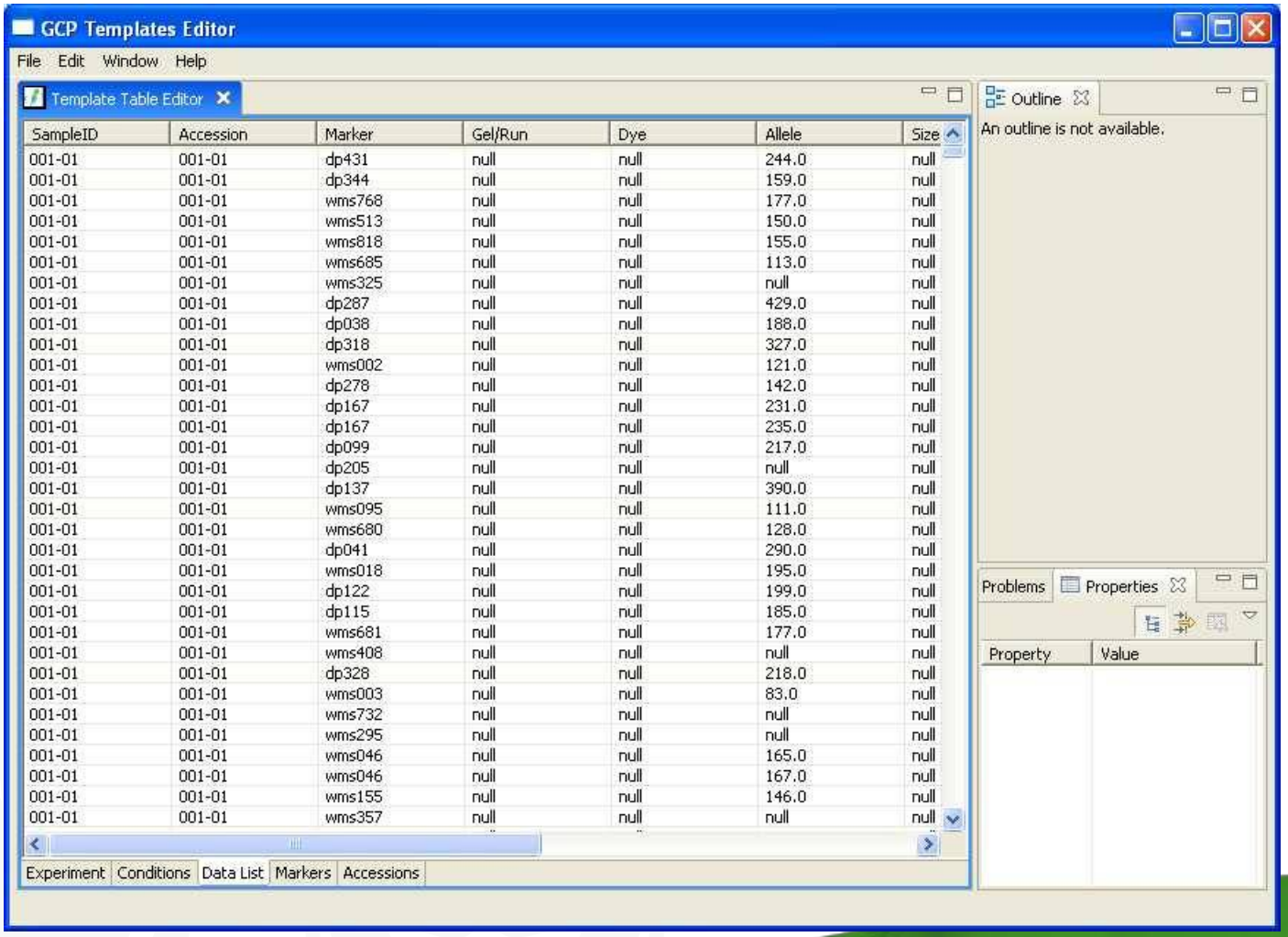
With each template there will be an example excel file and a set of example text files. These file contain fictitious example data to help users enter their data into the templates

[1] Only one type of data is required. If both are defined then the matrix format will be ignored.

Each template has

- A defined format for Excel and text files
  - We recommend text files due to limitations with Excel
- A read me file
  - Explaining how to format the data in Excel or text
  - Describes fields and valid entries
- A parser to convert the data to XML
  - A generic validater/parser has been developed to support several different data models for each template
- A transformer to convert the XML to other formats

Template Editor



CIMMYT MR

Crop Research Informatics Laborator

2005 Template Team

- CIMMYT — Marilyn Warburton & Miguel Anducho
- CIRAD — Brigitte Courtois & Manuel Ruiz
- IITA — Sarah Hearne
- IPGRI — Tom Hazekamp
- IRRI — Thomas Metz
- SCRI — David Marshall

2006 Template Team

- CIMMYT — Marilyn Warburton & Miguel Anducho
- IRRI — Richard Bruskiewich
- ACGT — Jane Morris and Ayton Meintjes
- IPGRI — Tom Hazekamp
- GrainGenes — Dave Matthews
- MaizeGDB
- Gramene — Junjian Ni and Noel Yap