



Temples 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.



Temples task: The problem

- Researchers, analytical tools and databases all want data in **different formats**
- Ideally, we only want one way to **store** the definitive version of the data
- The medium to long term solution to this problem is to develop data standards
 - Through a collaborative process
 - Domain modeling task
- However we need a short term solution
 - Since the data is being produced now!
 - This is the goal of the template task



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 – syntax / semantic

Progress so far

- We have developed templates for
 - Genotyping data
 - Passport
- We will develop by the end of the end year templates for
 - Map and QTL data (Nov 2005)
 - Phenotypic data (prototype) (Early 2006)

Excel template example



Microsoft Excel - example.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF

Look for: Find Now Search in Outlook & File System folders

SampleID

| | 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 | 8196 | 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 | | 364299 | 1 |

Ready

Readme – instructions on use



just suggested values. The experiment and conditions spreadsheets (or files) and the data 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
 - Describing fields and valid entries
- An XML schema to define the XML form of the data
- An Eclipse RCP application
 - Converting the data to XML and
 - Editing the XML file

Template Editor



file:/C:/Documents%20and%20Settings/Gdavenport/My%20Docun

Resource Set

- Allele 204
- Locus gpsb023
 - Allele 152
- Locus gpsb023
 - Allele 172
- Locus gpsb023
 - Allele 220
- Locus gpsb023
 - Allele 176
- Locus gpsb023
 - Allele 196
- Locus gpsb023
 - Allele 178
- Locus gpsb023
 - Allele 188
- Locus gpsb023
 - Allele 178
- Locus gpsb023
 - Allele 200
- Locus gpsb023
 - Allele 164
- Locus gpsb023
 - Allele 210
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027
- Locus gpsb027

Outline

- file:/C:/Documents%20and%20Settings/Gdavenport/My%20Docun
 - Dataset
 - Content Metadata Example
 - Experimental Metadata Individuals
 - Loci Type
 - Markers Type
 - Accessions Type

| Property | Value |
|----------|-------|
| Amount | 1.0 |
| Name | 200 |
| Quality | 200.0 |

Selection Parent List Tree Table Tree with Columns

Selected Object: Allele 200

Generation Challenge Programme - Microsoft Internet Explorer


File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites AutoLink AutoFill Options

Address <http://www.generationcp.org/bioinformatics.php?da=0526023>

Google Search New! Check AutoLink AutoFill Options

A CGIAR CHALLENGE PROGRAMME



Generation Challenge Programme

CULTIVATING PLANT DIVERSITY FOR THE RESOURCE POOR

home | about us | subprogramme 1 2 3 4 5 | resources | news | publications | virtual workspace | contact us | site map

Bioinformatics

Data Submission Templates

The Generation Challenge Program (GCP) is generating a wide spectrum of data. Part of these data will go into public databases, part will be included in GCP or institutional databases that will be available as web services. However, there will always be data sets that do not -- or not yet -- fit in one of the available structures. Therefore, a facility has to be maintained that allows access to these data sets in downloadable files in a consistent but flexible format.

The main objective of the **Templates** task is to provide simple templates for the temporary storing or distributing of the different data sets that are being produced within SP1, SP2, and SP3, for which there is no current provision in public or institutional databases. These templates must contain consistent but sufficient explanatory notes on how they should be used. The completed data sets should contain the necessary information to be stand-alone and should be simple enough to be understood. For example, enough description of material and methods used, and no use of acronyms or a coding system that only the data provider can understand. The templates will be consistent with the [GCP Domain models](#) being produced and will be used to store data in the central repository.

We are currently working **Data Submission Templates** for the following three areas:

- [Passport data](#)

4 BIOINFORMATICS

SUBPROGRAMME LEADER
Theo van Hintum,
Wageningen,
theo.vanhintum@wur.nl

GENERAL LINKS

[Bioinformatics homepage](#)

[GCP Bioinformatics links](#)

COLLABORATIVE DEVELOPMENT

[CropForge](#)

[GCP on CropWiki](#)

DOMAIN MODELING

[Welcome to Domain Modeling](#)

TEMPLATES

[GCP Data Submission Templates](#)

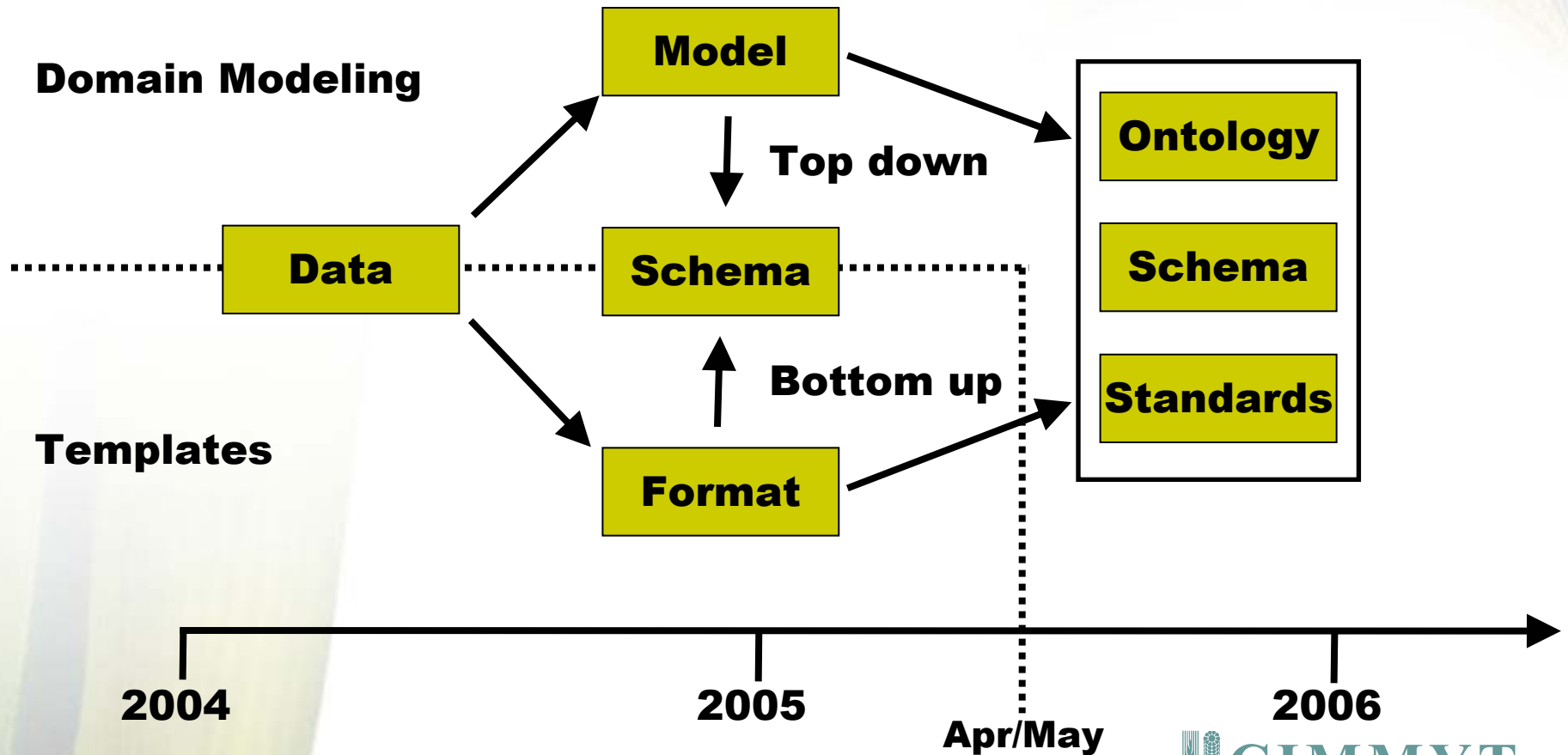


Each template will have

- Eclipse RCP and web applications for
 - Loading and validating the data
 - Improved views of the data
 - Uploading to the central repository
 - Uploading to databases where possible
 - Converting the data to other formats
- Final versions expected early 2006



Templates & Domain modeling



Template Team

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