finalreport

Secure web delivery of research data

Abstract

The Sustainable Grazing Systems (SGS) Key Program ran from 1997 to 2002. One of the important outcomes of that Program was the SGS national experiment database which was created as six separate but compatible MS Access databases for use on desktop computers.

This Proof of Concept project has successfully demonstrated that two of the above relational databases from the SGS key program could readily be combined into a single large database which was then successfully migrated to a SQL server implementation. Queries of the SQL server combined database can be carried out with simple 'point and click' operations via a Web portal developed within this project. User-selected subsets of data can be exported in formats which clients commonly use to further explore data.

Provision of access to these datasets will add value to the prior SGS investments long into the future.

It is recommended that Phase 2 of this project be immediately embarked upon to ensure that all of the SGS data are made available via the Web portal within the next year.

Executive Summary

The Sustainable Grazing Systems (SGS) project (1997-2002) resulted in a wide array of outcomes including a comprehensive relational database of all data collected at each of the six national experiment sites.   At the conclusion of that project, all 6 sites received their own compatible database whilst MLA was sent all 6 databases as separate database files. At that time, no permissions were given to merge them into a single national SGS database. Since then, permission has been granted to explore the merging of more than one SGS database within this project.

This project was developed as a Proof of Concept to provide authorised end users with secure access via the Internet to the SGS national experiment database (represented by the combined datasets of two site databases). This project aimed to allow users to view and/or download raw or summarised data through a range of tables of data and graphical queries using a new web-based tool. This tool has similar functionality to the FlexiGraph tool already available to scientists with the MS Access version of the SGS national database developed during SGS.

The primary aim was to enable greater value to be gained from quality research data beyond the completion of an important national project such as Sustainable Grazing Systems. For example, it may be that in the future, other researchers and/or grazing system modelers may be provided with authorisation to access some of the unique data sets contained in this national compendium of biophysical data, to query the data, explore it and potentially to download specific sub-sets of data for any authorised purposes.

We believe that the results achieved in this project demonstrate that all of the objectives of this project have been met in full measure.

The capacity to run large queries with the SQL Server implementation of the SGS database adds considerable value and power to the original separate databases. Interrogation of the database is clearly feasible via the Web portal. There is great potential to expand the capacity of the SGS portal to deliver more sophisticated and more complete datasets across the entire SGS national database.

The outcomes from this project are relevant to MLA's other interests in relational databases of research data including Sustainable Grazing on Saline Lands, Better Fertiliser Decisions, Grain & Graze, EverGraze, etc. in that similar delivery of national research data could be achieved for these other databases.

The development of a central national research data repository for MLA would lend itself to the development of additional technologies such as data mining and advanced statistical exploration of complex datasets. These further developments would enable MLA to gain greater benefit from both its prior investments in research as well as provide the means to make better research investment decisions into the future.

Making historic research data available, such as that from the SGS Key Program, is clearly an efficient use of resources as these datasets, which cost MLA and its partners at least $13 m to accumulate over 5 years, can now be made available to any authorised users in the world through a convenient and robust web-based solution.

If Phase 2 of this project were to be approved, MLA could see all of the SGS data made available through this means in a short time frame (say over the next year).

Within 5 years, given adequate resourcing, this concept could be 'rolled out' to allow for other MLA research datasets to be made available to authorised users in a similar fashion. This could enable new research projects to build upon already completed research, to enable linkages to be made between different experiments and to greatly enhance the capacity of computer model developers to test their models against extensive datasets. Ultimately, this will enable the provision of more reliable and comprehensive information and tools to meat producers.

This Proof of Concept project has been successfully completed on time and within budget. It is clear that providing access to complex research databases such as the SGS national database can be readily provided via convenient web access.

It is therefore recommended that the proposed Phase 2 of this project (Delivery) be commenced immediately so that the momentum of this project can be maintained and especially so that the skills of the database staff who have carried out this work (Mr. Colin Lord and Mr. Dion Gallagher) can be retained, thus ensuring efficient delivery of Phase 2.

Simultaneously, it is recommended that MLA work with its collaborating research and funding partners to secure agreements within one year to provide access to these SGS national research datasets via the web portal. As part of this effort, MLA and its partners should consider the legal and intellectual property implications of the provision of research data so that suitable arrangements can be made over the next year to provide access to all of the SGS national experiment data if and when Phase 2 is completed.

Contents

1 Background

The Sustainable Grazing Systems (SGS) project (1997-2002) resulted in a wide array of outcomes including a comprehensive relational database of all data collected at each of the six national experiment sites. At the conclusion of that project, all 6 databases were delivered to MLA as separate databases as, at that time, no permissions were given to merge them into a single national SGS database.

This project was developed as a Proof of Concept to provide authorised end users with secure access via the Internet to the SGS national experiment database (represented by the combined datasets of two site databases). This project aimed to allow users to be able to view and/or download raw or summarised data through a range of tables of data and graphical queries using a new web-based tool. This tool has similar functionality to the FlexiGraph tool already available to scientists with the MS Access version of the SGS national database developed partly before and partly during SGS.

The primary aim is to enable greater value to be gained from quality research data beyond the completion of an important national project such as Sustainable Grazing Systems. For example, it may be that in the future, other researchers and/or grazing system modelers may be provided with authorisation to access some of the unique data sets contained in this national compendium of biophysical data, to query the data, explore it and potentially to download specific sub-sets of data for any authorised purposes.

2 Project Objectives

By 30 December 2006 to have demonstrated a functional web version of FlexiGraph, so that authorised users can view user defined graphs using elements from pre-defined queries. The web interface will also allow limited subsets of data to be downloaded in suitable generic formats (e.g. MS Excel or CSV files) for exploration with other software tools or statistical programs.

3 Methodology

Two Microsoft Access databases (NW Slopes and Central Tablelands) were merged into a single database.

The resulting single MS Access database was migrated to a Microsoft SQL Server platform.
The Access SQL queries were modified to run as SQL Server queries.
Access was provided to a database interface via a web portal.
A system for limiting access to data based upon the user's status was implemented.
A basic recording system for logging data accessed was implemented.

A flexible interface to the database has been provided with the following components:
- tree view of data
- datasheet view of data
- facility to download limited subsets of data in generic formats (eg Excel, CSV)
- basic flexible graphing

4 Results and Discussion

4.1 Results - Phase 1 (proof of concept)

There are two phases of work which were envisaged for this project (1. Proof of Concept and 2. Delivery); only Phase 1 has been addressed by this contract.

The Contractor has developed a web portal which allows access to a functional web version of the SGS national database (two sites combined) and allows users to explore data using a web-based version of FlexiGraph, so that authorised users can view user defined graphs using elements from pre-defined queries.

The web interface allows limited subsets of data to be downloaded in suitable generic formats for use with tools such as Excel or statistical programs.

In detail, this has included:
The migration of the two core datasets from two SGS Access databases (NW Slopes and Central Tablelands) into one database (approximately of 1 Gigabyte size) including:

- Rainfall (manual recorded data)
- Meteorological weather station (automatically recorded data)
- Pasture Composition
- Pasture Biomass
- Pasture Growth Rates
- Runoff
- Soil Nutrient Sampling Data
- Soil Moisture data
- Stocking rates
- Weights/Condition Scores

In fact, *all* tables from both databases have been merged into the one database; this is more than was anticipated in the contract. This combined database was then migrated to run on SQL Server.

The datasets can be queried down to plot level where appropriate.

The system limits access to data based upon the user's status.

The web portal allows for a basic recording system for logging data accessed.

The flexible interface to the database has the following components:
- tree view of data
- datasheet view of data
- facility to download limited subsets of data in generic formats (eg Excel, CSV, etc.)
- basic flexible graphing

Documentation attached to each dataset, which lists field types, units, collectors and collection methodologies is still to be added.

The data ownership table/s and links are still to be added to the SQL version.

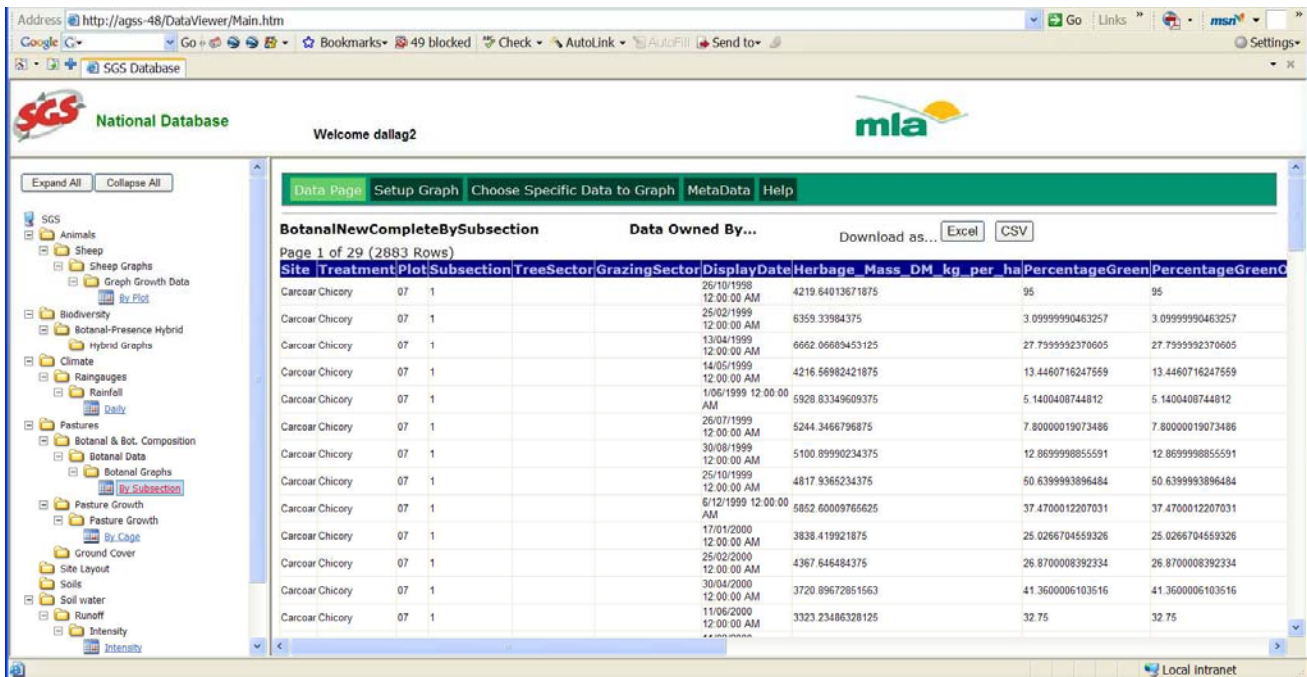The 'screen grabs' shown below demonstrate the various components listed above.

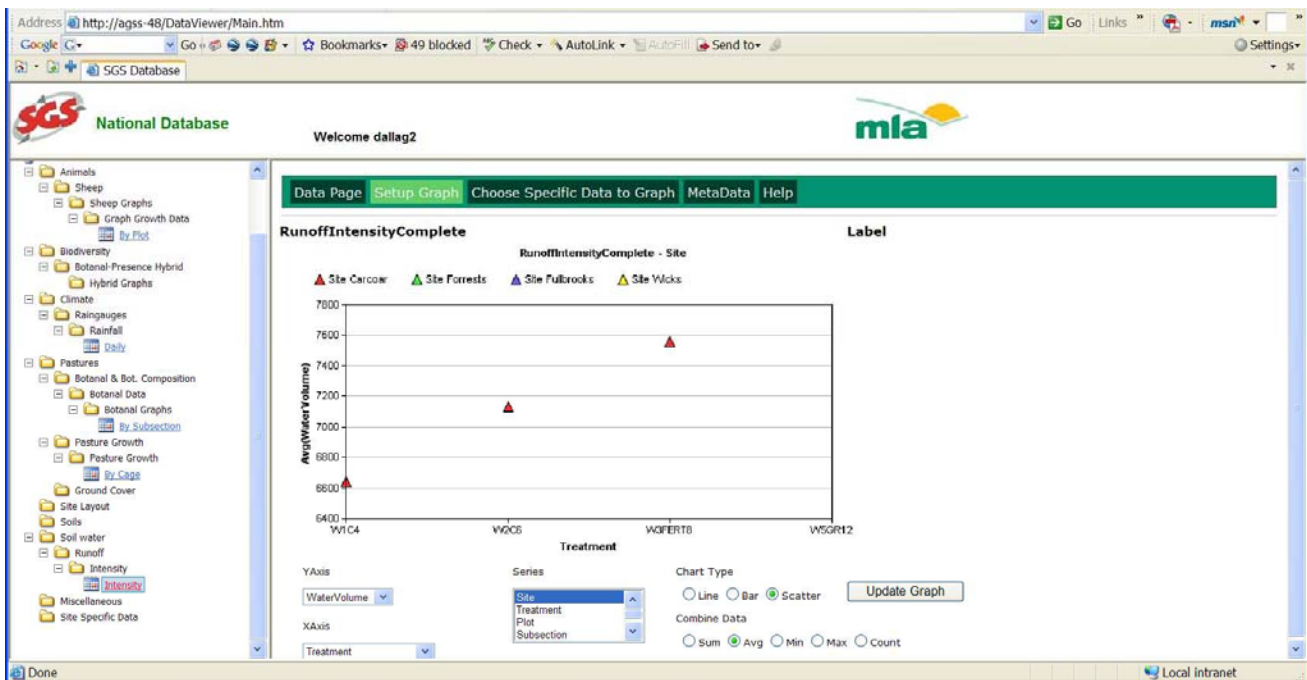**Figure 1. Data page showing part of 2883 rows of Botanal data.**

Figure 2. Scatter plot of runoff water volume vs. treatment for the Carcoar site.
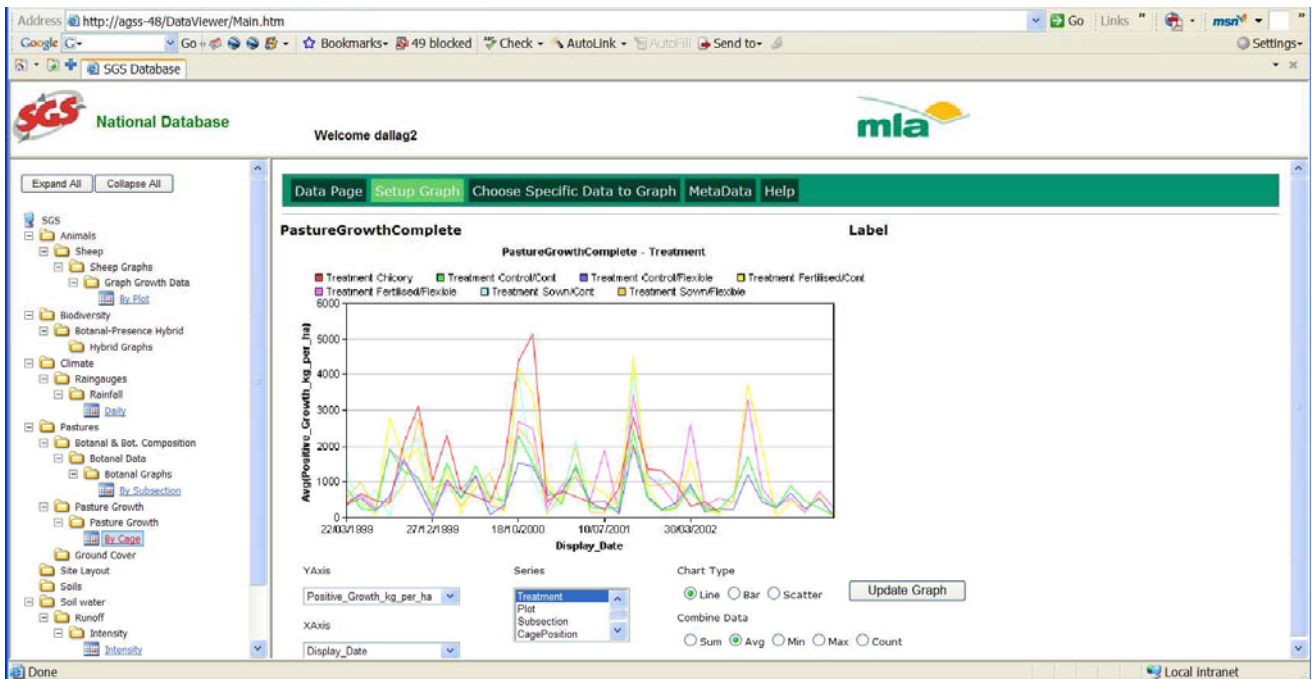
Figure 3. Line graph of average pasture growth by treatment for the Carcoar site.
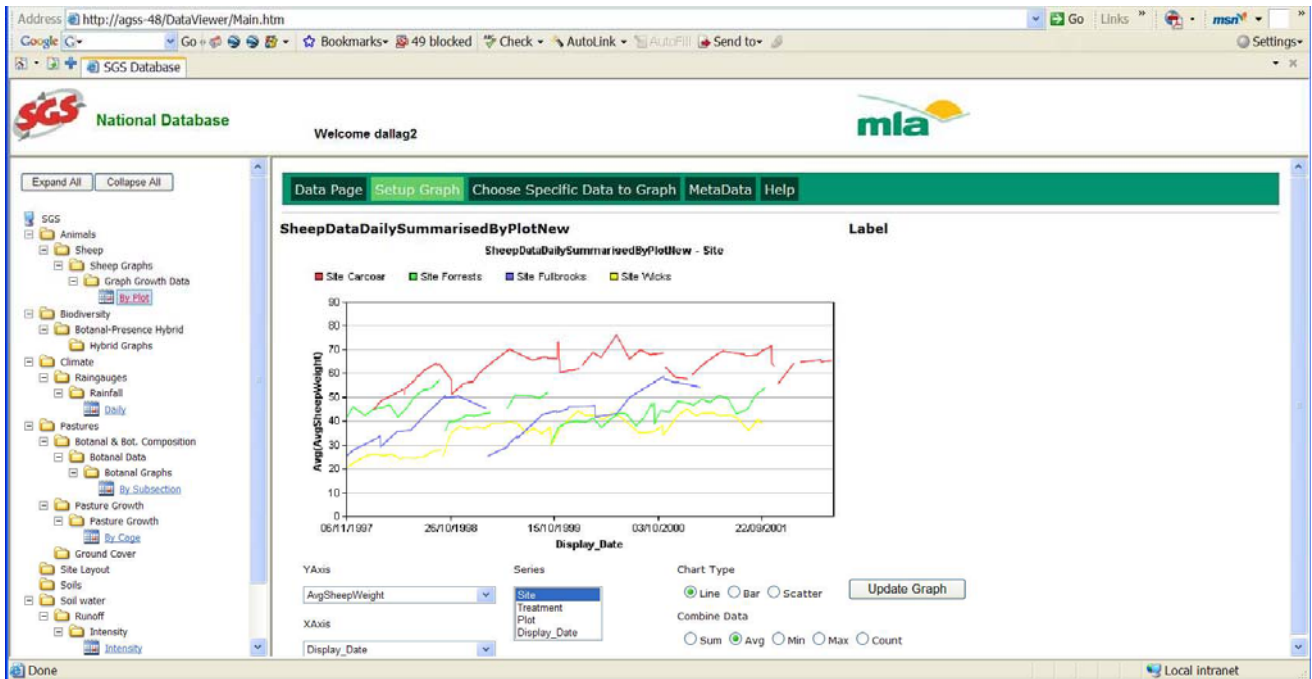

Figure 4. Line graph of average sheep weight data from Central Tablelands and NW Slopes sites.
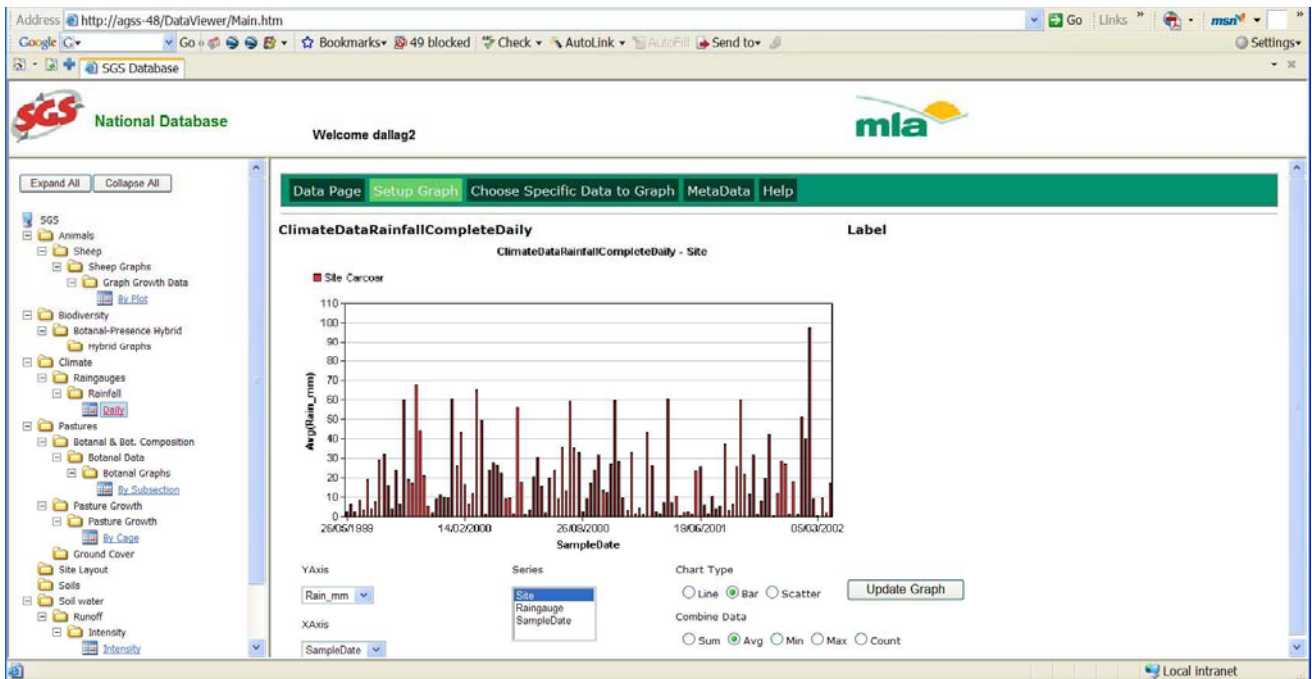
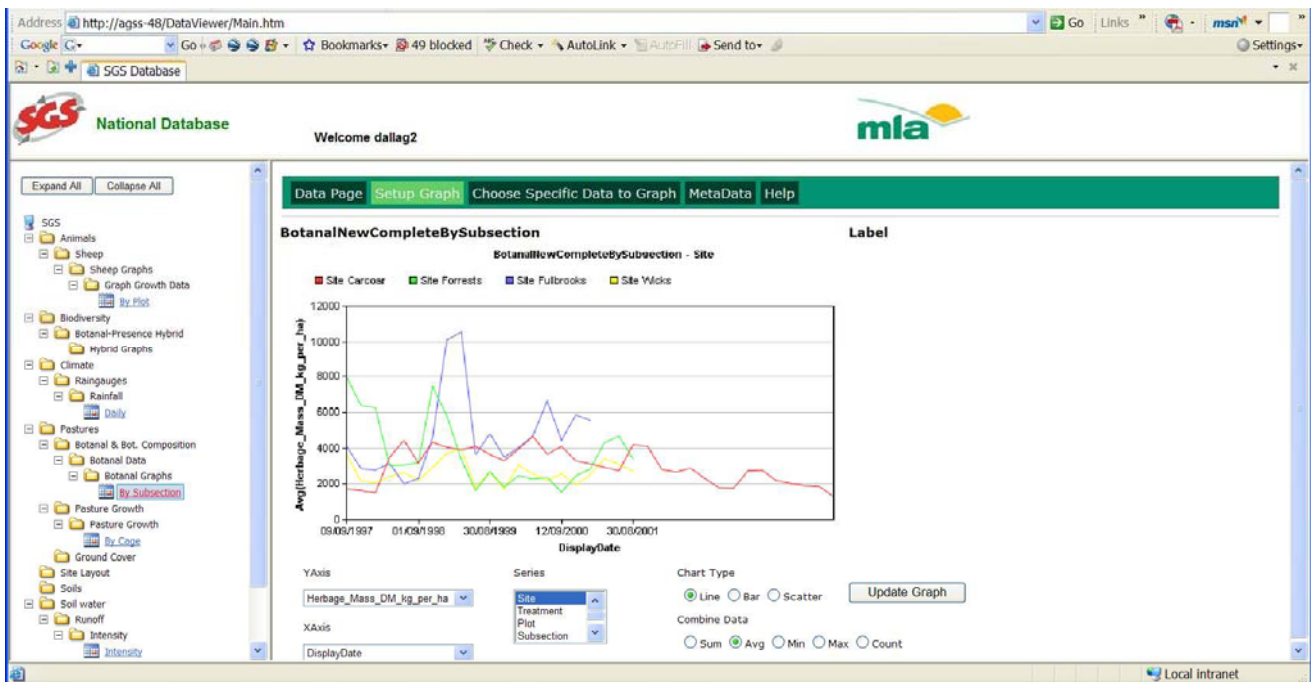**Figure 5. Bar chart of rainfall by date for the Carcoar site.**

Figure 6. Line graph of the average herbage mass for all Central Tablelands and NW Slopes sites.
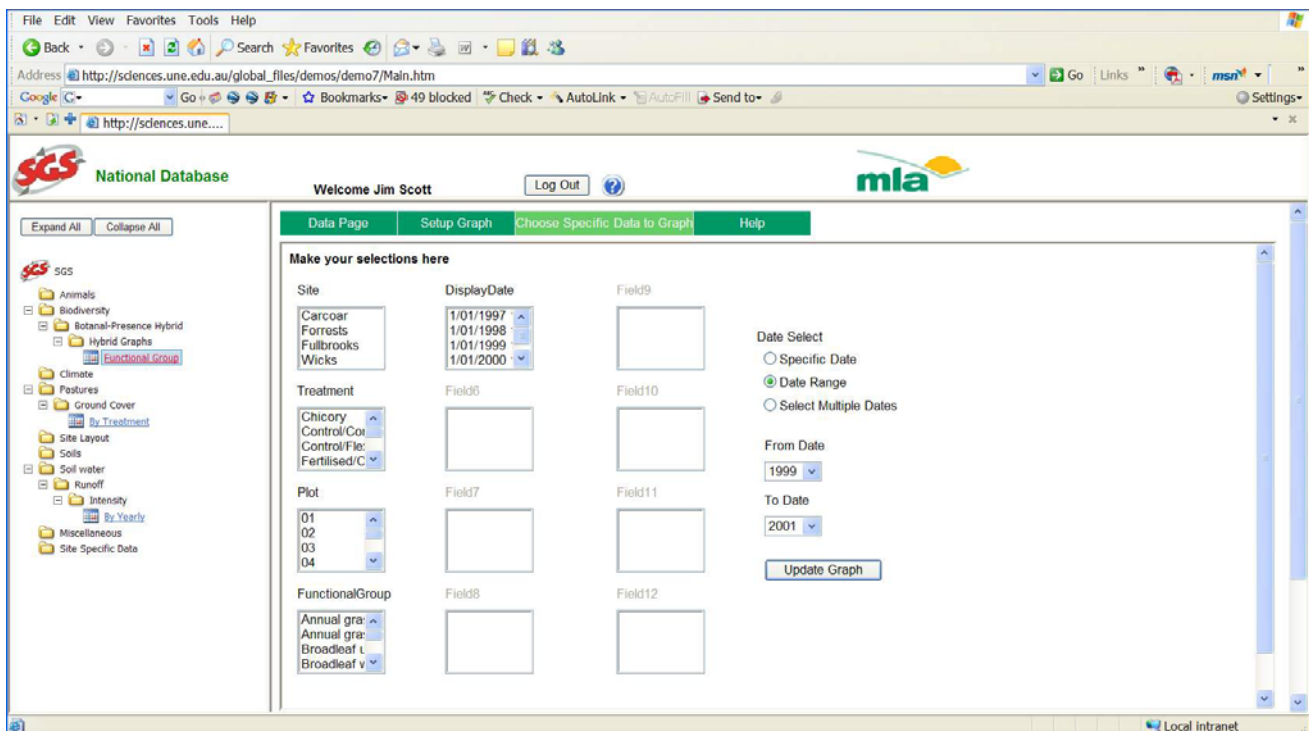
Figure 7. Layout of "Choose specific data to graph" tab.
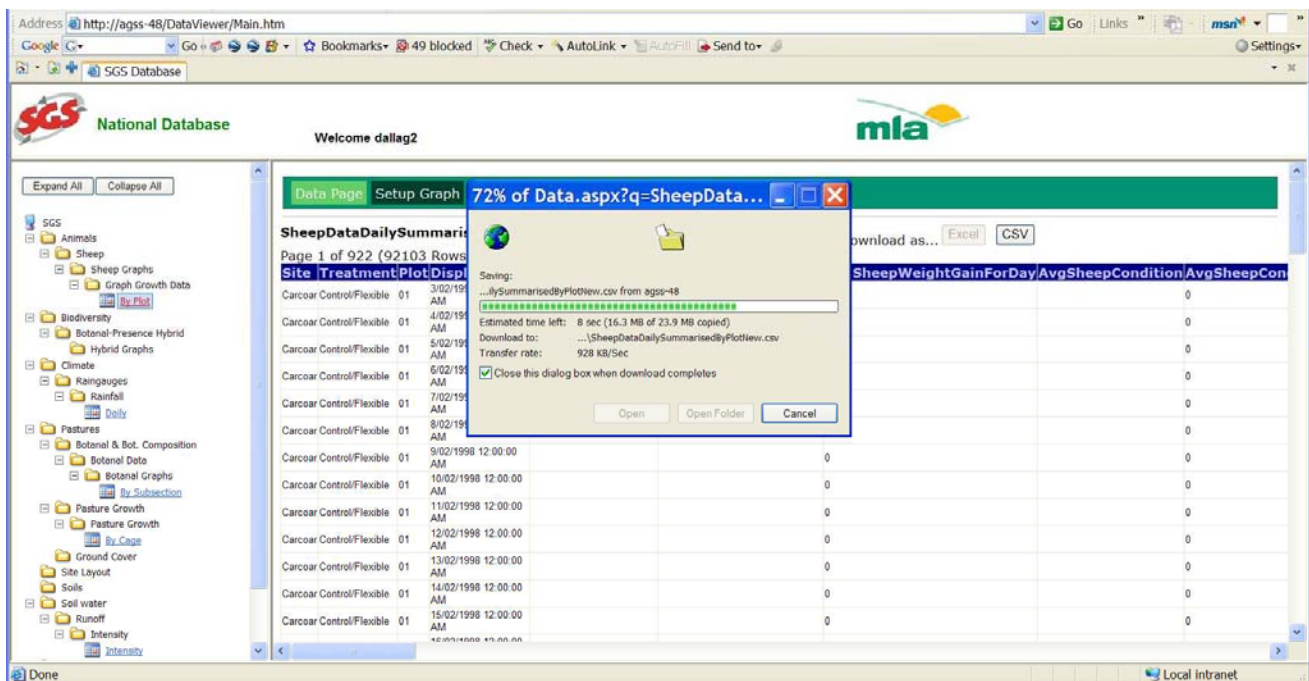

Figure 8. Screen grab of process of saving large CSV output file (29 megabytes) of sheep data from two SGS national sites.
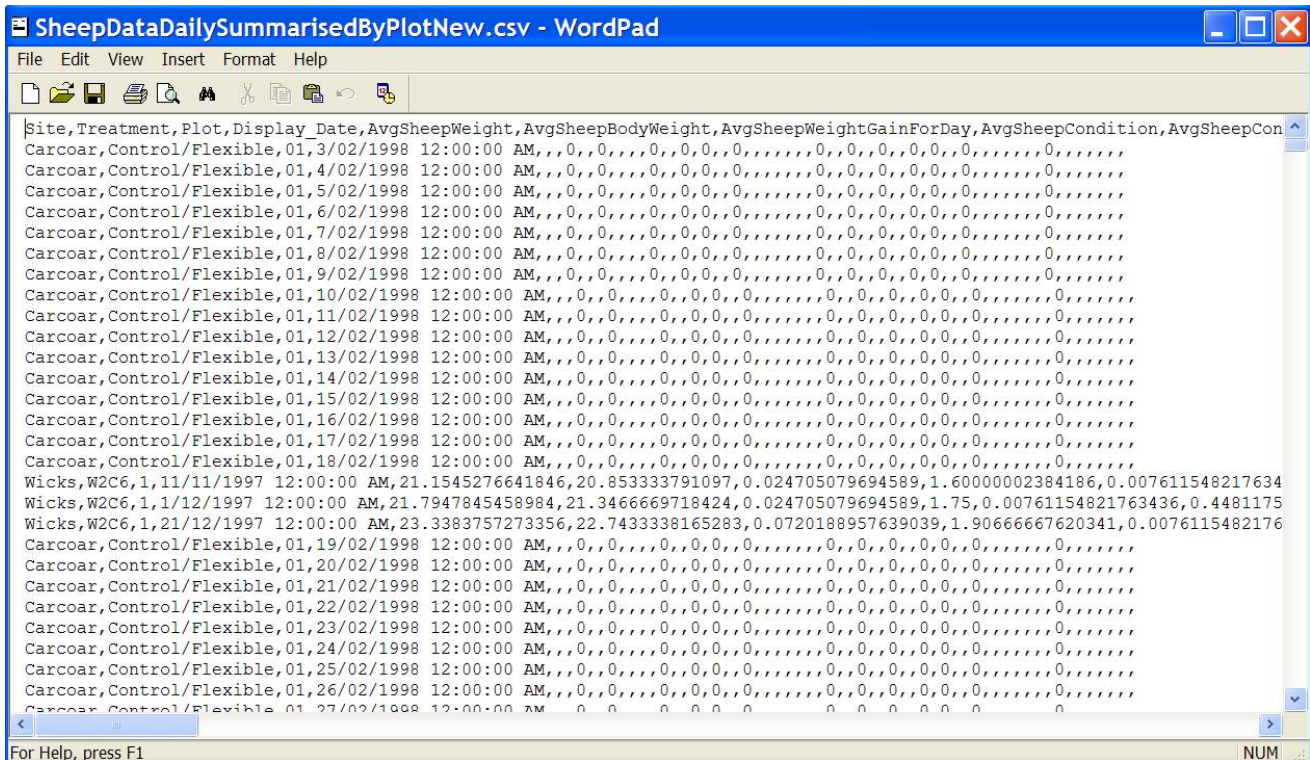
Figure 9. Part of 29 megabyte CSV output file saved of sheep data from 2 SGS national sites.
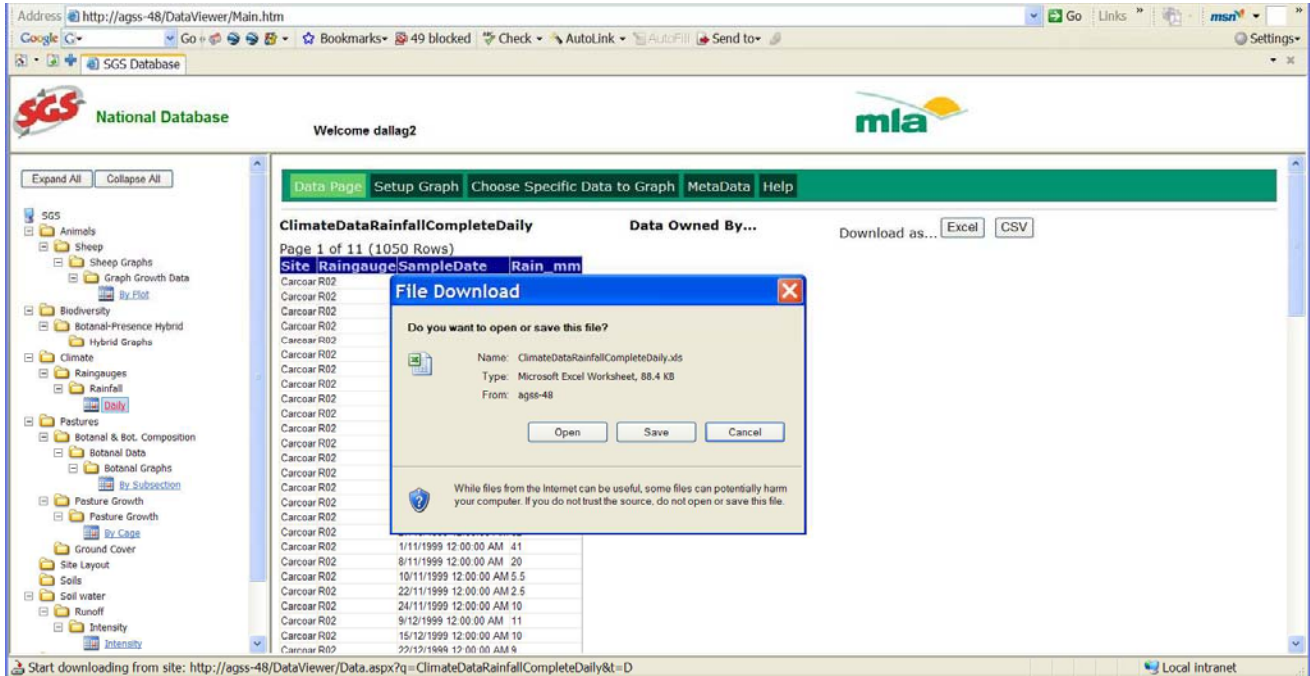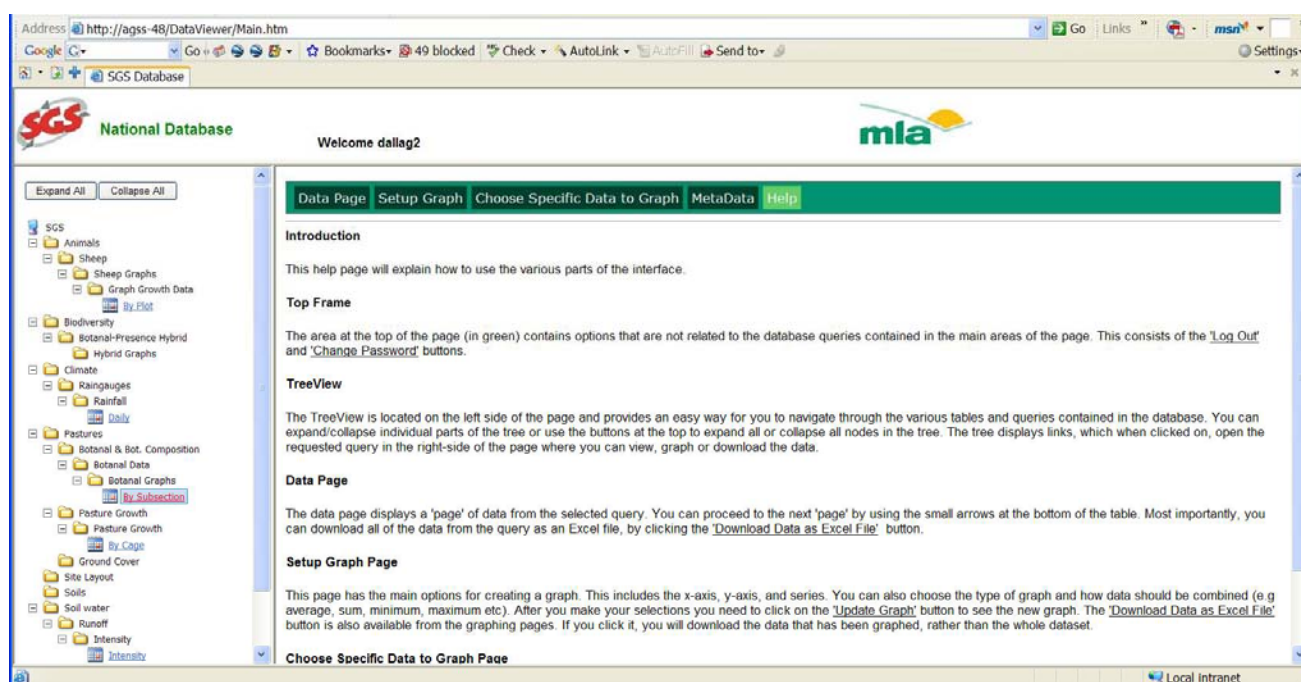

Figure 10. File download dialog box showing capacity to save 1050 rows of rainfall data as an Excel file on the local client machine.

Figure 11. Part of the Help files available on the SGS National Database Web portal.

All of the above 'screen grabs' have been captured from a SQL server installation on our Centre's server. It functions exactly the same as the proposed stand-alone SQL server that we envisage should be used for remote access to the central database.

As we have only just been given access to a stand-alone SQL server implementation on a UNE server at UNE as of December 19, 2006, we are unable at this time to provide real-time access to the SQL server implementation from outside the campus. Nevertheless, should MLA require proof of the above functionality, we should be able to provide that early in 2007.

4.2 Discussion

We believe that the results above speak for themselves and demonstrate that all of the objectives of this project have been met in full measure.

The capacity to run large queries with the SQL Server implementation of the SGS database adds considerable value and power to the original separate databases. Interrogation of the database is clearly feasible via the Web portal. There is great potential to expand the capacity of the SGS portal to deliver more sophisticated and more complete datasets across the entire SGS national database.

The outcomes from this project are relevant to MLA's other interests in relational databases of research data including SGSL, Better Fertiliser Decisions, Grain & Graze, EverGraze, etc. in that similar delivery of national research data could be achieved for these other databases.

The development of a central national research data repository would lend itself to the development of additional technologies such as data mining and advanced statistical exploration of complex datasets. These further developments would enable MLA to gain greater benefit from both its prior investments in research as well as provide the means to make better research investment decisions into the future.

5 Success in Achieving Objectives

It is clear from the above results that we have achieved fully on the Proof of Concept required by this project. We claim therefore that we have fully met the objectives of this project, on time and within budget.

6 Impact on Meat and Livestock Industry – now & in five years time

Making historic research data available, such as that from the SGS Key Program, is clearly an efficient use of resources as these datasets, which cost MLA and its partners at least $13 m to accumulate over 5 years, can now be made available to any authorised users in the world through a convenient and robust web-based solution.

If Phase 2 of this project were to be approved, MLA could see all of the SGS data made available through this means in a short time frame (say over the next year).

Within 5 years, given adequate resourcing, this concept could be 'rolled out' to allow for other MLA research datasets to be made available to authorised users in a similar fashion. This could enable new research projects to build upon already completed research, to enable linkages to be made between different experiments and to greatly enhance the capacity of computer model developers to test their models against extensive datasets. Ultimately, this will enable the provision of more reliable and comprehensive information and tools to meat producers.

For example, recently, permission was given for a Sheep CRC PhD scholar (Mr. Karl Behrendt) to be provided with access to both the SGS Carcoar database and the Cicerone Project database (both in MS Access format) to enable him to conduct bioeconomic research without the need to collect any biophysical data.

This Proof of Concept project has demonstrated how the above approach of providing access to datasets should be considered for a wider range of authorised Australian researchers resulting in the potential for much more value eventually being 'mined' from already completed research, to the ultimate benefit of Australian meat producers.

7 Conclusions and Recommendations

This Proof of Concept project has been successfully completed on time and within budget. It is clear that providing access to complex research databases such as the SGS national database can be readily provided via convenient web access.

It is therefore recommended that the proposed Phase 2 of this project (Delivery) be commenced immediately so that the momentum of this project can be maintained and especially so that the skills of the database staff who have carried out this work (Mr. Colin Lord and Mr. Dion Gallagher) can be retained, thus ensuring efficient delivery of Phase 2.

Simultaneously, it is recommended that MLA work with its collaborating research and funding partners to secure agreements within one year to provide access to these SGS national research datasets via the web portal. As part of this effort, MLA and its partners should consider the legal and intellectual property implications of the provision of research data so that suitable arrangements can be made over the next year to provide access to all of the SGS national experiment data if and when Phase 2 is completed.