final report

Project code: B.CCH.1048

Prepared by: Brian Dalrymple and Yutao Li

CSIRO Livestock Industries

Date published: November 2011

PUBLISHED BY
Meat & Livestock Australia Limited
Locked Bag 991
NORTH SYDNEY NSW 2059

Scoping study for RELRP database

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

Abstract

Large amounts of research data have been, or are anticipated to be, generated during the course of the current RELRP (Reducing Emissions of Livestock Research Program) projects. However, there is no systematic infrastructure in place to effectively store the data, or to allow the access and interchange of data between institutes or the researchers within the program. A scoping study of the issues around the development of the framework (including a metadata database) for all the projects within the RELRP has been undertaken. A schema for the storage of the metadata has been designed and implemented in Microsoft Access to facilitate the process of gathering and storing the metadata about the datasets generated and held by the component projects of the RELRP. Two rounds of metadata collection have been conducted to collect information on the experimental designs, the parameters and variables relating to the measurement of green house gas emissions etc., the data format and structures, and the data storage and management systems used by the research groups. Collation of the metadata information from individual projects identified a range of issues relating to the accuracy of the definition of datasets and experiments, data formats and concerns about IP. Through redefining the system structure and quality editing of metadata information (including applying a controlled vocabulary list developed for this project), a metadata database illustrating the basic framework of a future database and data summary sheets showing the current status of data attributes for each project have been produced. A number of options for future database implementation phases and potential benefits of the database to a wide range of communities (scientists, policy makers, consultants and farms) are discussed. The completion of the metadata database with a web interface is highly recommended. An implementation of the actual database focussed on key datasets should be seriously considered. In addition, a number of recommendations for future development of the database if new funding allowing an extension of the current research activities of the RELRP projects were to become available are also provided.

Executive Summary

The MLA/DAFF RELRP aims, through an integrated program of R&D and Demonstration projects, to provide the knowledge, tools and strategies which primary producers will require to manage the Green House Gas (GHG) emissions of their livestock production systems. Large amounts of research data have been, or are anticipated to be, generated during the course of the current RELRP projects. However, there is no systematic infrastructure in place to effectively store the data, or to allow the access and interchange of data between institutes or the researchers within the program. The RELRP project B.CCH.1048 aimed to complete a scoping study on the system requirements of developing a framework (including a metadata database) for all the projects within the RELRP. The specific goal was to understand the attributes of the data, coding protocols, parameters and variables, output formats and data security and storage protocols including Intellectual Property from each individual RELRP project.

In order to fulfil the goal of the project, a range of activities was carried out by the project team to develop the framework (including a metadata database) for gathering and storing the metadata required to complete the scoping study. These include:

- Designing a metadata database in Microsoft Access to facilitate the process of gathering and storing the metadata about the datasets generated and held by the component projects of the RELRP;
- Designing a metadata survey form distributed to all the leaders of RELRP projects to gather the metadata information needed to populate the database;
- Conducting face-to-face meetings and teleconferences to inform, debrief, explain and assist the project leaders and/or experimental leaders to complete the survey forms;
- Collating, quality editing and summarising metadata information collected from initial survey forms to identify the issues or crucial missing information required to populate the database;
- Reformatting the structures of different tables in the metadata database and developing a controlled vocabulary list to standardise the variable names;

- Conducting a second round of data collection with a specific questionnaire targeting the missing information;
- Consolidating the metadata information from the second round of data collection and producing an updated data summary sheet for all the projects.

The results of these activities reveal the following data attributes of the program:

- There are over 200 different types of measurements taken by the RELRP projects, ranging from greenhouse gas emission (CH_4, NH_3, CO_2, NO_2 and NO_3) for various livestock and plant species, feed components, pasture composition, climate parameters, milk composition, animal productivity to molecular sequences;
- At least 32GB of data (including sequence data) is held in various institutes. The actual final size of the datasets could be much higher due to incomplete data information;
- The majority of the datasets are in MS excel format files stored on personal computers of individual institutes. Three projects have molecular sequence data saved on the individual institute servers. Two projects have GPS data saved as PDF format.
- The responses from the survey forms clearly indicate that all projects share the same concern of potential data access by unauthorised third parties. All projects agree that the access to the data information generated by the RELRP projects should be subject to the agreements between MLA/DAFF and host institutes (as per the contracts). However, the Information on animal pedigree, animal production measurements and SNP markers from B.CCH.1010 will not be available for the RELRP due to the IP held by the sheep CRC.

The key outputs of the scoping study can be summarized as follows**:**

- A metadata database illustrating the basic framework structure for a future database and containing detailed descriptions of the project, experiments, treatments and the datasets (i.e. metadata) from individual RELRP projects

has been produced. Although the metadata database currently has limited functionality it provides the fundamental infrastructure for establishing a future workable database if the next phase of database implementation project is to proceed.

- A controlled vocabulary list with standardised terminologies for the metadata fields has been generated. In a full implementation of the metadata database this will used to generate standards terms for the queries to extract the information from the database.

- A data summary sheet showing current status of data attributes from all the RELRP projects has been produced. This information is of great importance to MLA/ DAFF RELRP:
    - it can serve as key IP background information held by the program should they develop new R&D projects;
    - it can be used to provide verification data sets for future modelling projects.

Despite the above achievements, there are still a couple of issues/problems remaining unresolved. These include:

- Project or data complexity causes concerns about the amount of time required to properly document methods or datasets (e.g. some of the project leaders not willing to spend large amount of time, a question frequently asked was “who should be the one responsible doing the tasks?”).
- We are still unclear about the total number of datasets generated and the size of total datasets due to slow response from some of the RELRP projects. This has hampered our effort to accurately document the attributes of datasets from these projects.

Summary of the major recommendations

- Recommendations for the next phase of the RELRP project
    - Complete the implementation of the metadata database for all projects and datasets up to the end of the current RELRP project.

  - Provide access to the metadata as a web-enabled public database.
  - Consider a targeted implementation of the actual database with data from a subset of the RELRP research projects.

- Recommendations for future development of database as part of an extended RELRP or RELRP-aligned activity.
  - Database and project development and execution are part of in integrated process.
  - Determine data to be deposited, access rules milestones etc. up front
  - After the initial development phase separate management of the database from users of the database

Table of contents

1. Background

The MLA/DAFF RELRP aims, through an integrated program R&D and Demonstration projects, to provide the knowledge, tools and strategies required by the Primary producers to manage their emissions including the ability to respond to the commercial imperatives arising from emissions trading. In order to achieve the outcome, the program requires the crucial information such as data concerning methane emissions from ruminant livestock species from multiple themes of the program. Large amounts of research data have been, or are anticipated to be, generated during the course of the current RELRP Projects. However, there is no systematic infrastructure in place to effectively store the data, or to allow the access and interchange of data between institutes or the researchers within the program.

2. Project objectives

The initial phase of this project (Objective 1) is a scoping study that is designed to assess the systems requirements for quantitative and qualitative data streams from the whole of RELRP. The study will also identify data storage and management systems used by the research groups, the standardized data format and structure, as well as the ability of the data storage systems to incorporate time step information for individual animals. The main objective is to understand the attributes of data, coding protocols, parameters and variables, output formats and data security and storage protocols (including access and IP).

3. Methodology

The RELRP comprises 22 projects forming four themes, namely, methane measurement method development; rumen functions; genetic improvement of livestock species for reduced emission; and industry demonstration. Amongst the 22 projects, there are two projects with which the scoping study project aimed to

collaborate closely to determine the attributes of the database. They are B.CCH.1043 (Title: Further development of the FarmGAS calculator), and B.CCH.1047 (Title: Development of historic database and meta-analysis facility for livestock emissions). The remaining 20 projects contain the experimental data that need to be fully documented. Due to the program's complexity, a series of activities were conducted to assess the system requirements for establishing a database. The first step to unravel this complexity was to get an overview of the types of experiments undertaken in the projects and then to design a database system to gather the information and collate the datasets.

3.1 Initial development of the structural attributes of the database.

Database system design. A scoping study metadata database was designed in Microsoft Access to facilitate the process of gathering and storing the information about the datasets generated and held by the component projects of RELRP. Metadata is data that provides summary descriptions of projects, experiments, measurements, and datasets. The metadata database (see **Appendix A)** is anticipated to sit on top of the actual data database and access systems in the implementation phase (see **Appendix B)**. **Appendix C** lists the tables and fields in the metadata database showing the different attributes of the projects, experiments, treatments, measurements and datasets collected.

Project Survey Design. Initially a survey form was designed to gather the metadata information needed to populate the database. The survey form was partly prepopulated based on the information from the MLA/DAFF RELRP one-page individual project descriptions, individual project progress reports and the presentations in RELRP Workshop in Melbourne (23-24 March 2011). The form contains three major sections for each project, they are, Project, Experiment and Dataset, as shown in **Appendix D**.

Contacting Project Leaders. In order to make the survey responses more efficient and less painful for the project leaders, prior to sending the survey forms, we filled in as much information as we could from the information sheets provided by MLA. The partially completed survey forms were then distributed to 21 project leaders (See

Appendix E) by email except for Ian Bland (B.CCH.1047), to whom a special visit was made.

Metadata Collection. After dispatching the survey forms, extensive efforts were then undertaken to ensure that the project leaders understood the importance of the database and the requirement to complete the forms. The actions taken include:

- Eleven face-to-face information meetings were conducted with the staff from 14 projects (either project leaders or experimental leaders) in Armidale, Brisbane, Melbourne and Perth respectively. The meetings (9 of them were held individually and 2 were as groups) involved the debriefing, explanation and assistance in completing the forms.
- For other geographic locations (Townsville, South Australia and Wollongong), telephone conferences were conducted with three project leaders to debrief and answer the queries.
- These were then followed by the reminder emails and telephone calls.

3.2 Collaboration with the projects B. CCH.1036 and B. CCH. 1047 to determine attributes of data.

Interaction with B.CCH.1043. During the initial round of the database survey, Sally Davison (B.CCH.1043) provided very specific data requirements for her online FarmGAS calculator. These include field measurement of actual emissions (for comparison between results obtained from the calculator and those obtained in the field), measurement of methane mitigation from changed management practices such as feed additives, field measurement of nitrous oxide levels and changes as a result of mitigation activities, where the research was done, who the research leader was, and some of their contact details. These requirements have since been incorporated into the metadata database structure fields.

Interaction with B.CCH.1047. A special visit was made by Mr. Bryce Little (CSIRO Software Developer) to Dr Ian Bland (the project Leader of B.CCH.1047) in Melbourne University on 14 July to explore the possible links between Ian's database and the database proposed in this project. With RELPP project B.CCH 1047, Ian had performed a literature search for experiments that record the emissions of

greenhouse gases from livestock. His intention was to summarise these into a form that would allow easy comparison of values to be made. However, the principal problem was that there were a wide variety of experiments that produced results under a wide variety of conditions. General conclusions were drawn, but detailed data was not provided, only means with an error range. He noted that although his review considered peer-reviewed papers, some experiments relied upon data from un-published sources, via previously published non-peer reviewed reviews. These factors made any direct comparisons difficult within the available resources of his project.

Due to these limitations, both Ian and our team have agreed that building a historic database is unachievable at this stage of the B.CCH.1048 project. Rather than ignoring this literature, we would propose to store the list of literature identified by Ian into the data repository until such time as the future utility of the information was determined,

3.3 Refining the metadata database structures to improve data query outcomes

Quality examination. After receiving the initial database survey responses and collating the information in the metadata database, a range of issues/problems relating to dataset and experimental definitions were identified.

Refining metadata database structure. Given that most confusion occurred at the level of definition of treatment and dataset, we decided to add two more tables (“Treatment” and “Measurement”) to the original three tables (“Project”, “Experiment” and “Dataset”) of the metadata database. Each individual table was modified to contain much more specific and comprehensive information (**see Appendix F**). To address some confusion around descriptions of the experimental datasets, an extensive effort was undertaken to create new data fields to clarify what is considered an “experiment” and what a “dataset” should contain.

Creating a controlled vocabulary list. To ensure that the individual projects are consistently represented and to aid in the cross referencing data in the metadata database, we generated a new controlled vocabulary list to standardise the names or

terms to be used for each field in the metadata database (**See Appendix G**). The list was generated and refined in consultation with the research groups and using reference information in particular to check plant names and groupings.

The controlled vocabulary will be used in the implementation phase to harmonize all of the data received so that the use of terms was consistent across the whole metadata database. This process is essential to ensure that correct operation of queries in the database.

Quality editing the individual project metadata information and generating a summary data sheet for all projects. In order to get a clear picture of the number of measurements, the number of datasets and the dataset size from the RELRP projects, an extensive effort was undertaken to summarise the data information from the individual projects. The data summary sheet (**See Appendix H**) shows all the information collated from the first round RELRP metadata database survey forms. It spells out what sort of information we are after and gives some example projects for which we have most or all of the information that we require for the scoping study. Most importantly the file shows the status of the metadata for individual projects and what information is still missing.

3.4 Collaboration with all projects to determine attributes of data

Constructing a questionnaire and conducting the second round of data collection. Given that there was still crucial data information missing for most of the projects, we decided to conduct a second round of metadata collection. A set of specific questions was constructed for the data collection. These were:

1. How many datasets each experiment generated?
2. How many records were in each dataset?
3. How big each dataset in terms of kB or MB?
4. Can you give us an estimate about the total number of datasets from your project and the total size of the dataset?
5. Have we interpreted the information from your survey response correctly? If not, could you please update if needed.

Conducting the second round of data collection. The data summary sheet **(Appendix H**) and the questionnaire were sent to 20 RELRP project leaders for their inputs. For the two projects (B.CCH.1012 and B.CCH.1034) which did not respond to the initial database survey forms, we sent the project leaders (Phil Vercoe and Frances Phillips) the initial survey form, the controlled vocabulary list, second round questionnaire and the data summary sheet for them to complete.

4. Results

4.1 Initial survey responses and the issues identified.

For the first round of database survey, of the 21 projects receiving the survey forms, 18 projects responded and provided information specific to their respective projects (**Appendix E**). After collating the metadata from the initial individual project survey responses, several issues/problems were identified. The issues and some of the actions taken include:

- Of the 18 initial responses, after collation and assessment we found that only 5 of the responses contained all of the information that we required to complete the project.

  Action: a second round of data collection was carried out.
- In three cases (B.CCH.1007, B.CCH.1011 and B.CCH.1013) the data collection forms referred to their milestone reports for data information;

  Action: the MLA was contacted to seek permission to access the individual project milestone reports. The latest milestone reports for B.CCH.1007 and 1011 were provided, but not for B.CCH.1013 due to an IP concern. The reports were examined for additional data information to be added into the metadata database.

- There was confusion about the definition of experiments and datasets, which was not helped by the fact that we were building our view of the data as the collection progressed;

Action: this distinction was clarified through refining the metadata database structures

- We still remain unclear about the total number of datasets generated in many of the projects and thus the total size of the datasets;

  Action: this was partly addressed by the second round of data gathering

- We did not provide a standard or controlled vocabulary for data attributes as part of the initial data collection survey. Thus a range of different names/formats were used for the same information in a number of the data attributes such as the "Measurement type", "Site", "Plant species";

  Action: a controlled vocabulary was constructed, applied to existing data and supplied as part of the second round of data collection

- Project or Data complexity caused concerns about the amount of time required to properly document methods or datasets (a question frequently asked was "who should be the one responsible doing the tasks?"). This has led to the expectation that any database implementation project will have to cover all additional costs/activities required to organise their data sets for incorporation. In one project alone this was estimated as one person month (Roger Hegarty pers. comm.)

- Concerns on data IP access, most of the people hope that special security measurement can be put into place before others to access "their" data and potentially publish manuscripts on "their" data;

  Action: see the recommendation for next phase of database implementation project

- What are the data storage formats we should use for GPS and sequence data?

  Action: currently the PDF format is being used by a number of groups

  Action: there are a number of standard formats used for sequence data, any of these can be readily implemented as required.

4.2 Second round of data collection and data attributes

As of 15th October 2011, of the 20 projects with experimental data information, 14 projects had responded to the second round and 13 of them had provided the update on the questionnaire. They are: B.CCHs.1003, 1004, 1005, 1007, 1008, 1009, 1012, 1013, 1014, 1015, 1018, 1032, 1034 and 1036. Despite reminder emails and follow-up telephone calls were made, the remaining 6 projects still have not responded.

After consolidating the information from the second round of data collection, a updated data summary sheet showing current status of data attributes from all the RELRP projects was produced (**Appendix I**).

Size of existing datasets. Based on the data information gathered from the responding projects, it is estimated that at least 32GB of data (including sequence data) is held in various institutes. However, the actual final size of the datasets from all the projects is expected to be significantly larger than the estimated value (could be 3~4 times). There are a couple of reasons for the expectation:

- Among 13 projects responded, three projects were unable to provide the expected actual size for the datasets to be generated from the experiments to be undertaken in the near future ;
- The slow responses from some of the RELRP projects has hampered our effort to accurately document the attributes of datasets from these projects;
- Having large datasets generated, but lack of financial, time and personnel resources, the leader of one project, B.CCH.1010 (Roger Hegarty), has estimated that to properly document the experimental datasets from the project one full time person would be required for one month to complete the tasks.

Because of the incomplete data at this stage, it is impossible for us to accurately estimate the number of datasets available from all the RELRP projects and the scale of datasets which need to be consolidated in the next phase of the database implementation. However, by most standards the estimated size of the datasets is small and if it was 10 or even 100 times larger would not pose a storage problem on current data storage systems.

Data variables and parameters. In total there were over 200 different types of measurements taken by the RELRP projects (**Appendix I**). The variables range from greenhouse gas emission (CH_4, NH_3, CO_2, NO_2 and NO_3) for various livestock and plant species, feed components, pasture composition, climate parameters, milk composition, animal productivity to molecular sequences.

The number of variables is a significant contributor to the complexity of the database and is very large compared to the size of the datasets.

Data formats, storage and management systems used by the research groups. The majority of the datasets are in MS excel format files stored on the personal computers of individual researchers in their respective institutes. Three projects (B.CCH.1005, 1007 and 1018) have molecular sequence data. Two projects (B.CCH.1003 and 1033) have GPS data saved in the PDF format. Among all the projects, only five (B.CCH.1003, 1005, 1011, 1018 and 1036) have stored their data on a secured server of host institutes.

Data IP access. The responses from the survey forms clearly indicate that all projects share the same concern of potential data access by unauthorised third parties. All projects agree that the access to the data information generated by the RELRP projects should be subject to the agreements between MLA/DAFF and host institutes (as per the contracts). However, the information on animal pedigree, animal production measurements and SNP markers from B.CCH.1010 will not be available to most of the RELRP researchers due to the IP held by the sheep CRC.

Consolidated database or federated database system? A small number of the research groups have access to CRC or institutional databases for the storage of data. The majority of the data is however stored on desktop computers of institutional servers as files (frequently excel spreadsheets) and folders of files. In the main these are not well organised, very dependant on the knowledge of individual researchers and certainly not readily accessible from outside of the host institution, let alone in a form suitable for remote querying.

Therefore, with the possible exception of data entered into the Sheep CRC and Beef CRC databases importation and consolidation of RELRP data into a single data repository and database appears to be the simplest option.

The Sheep and Beef CRC database curators have not been approached to discuss remote access, but in principle this should be possible once the appropriate IP arrangements have been put in place.

Standard data formats/structures v. user specific data formats/structures. Across the current RELRP projects there are a range of different formats used for equivalent types of data which have been developed for the requirements of individual researchers and research groups. Within the database a single data format/structure is used. It is unlikely that this format/structure will suit the needs of all, and possibly none, of the research groups.

Currently our preferred approach is not to impose a standard format for datasets across the projects, rather work with the project leaders and research staff to ensure that we understand the data and then use programming tools to automatically reformat and process the data into the required format.

By taking this approach researchers would not be unnecessarily constrained to fit into a one size fits all approach.

Standard units. Across the current RELRP projects there are a number of cases where different units used for equivalent types of data. As above the preferred approach is to convert incoming data to the database standard allowing researchers to use whatever system they are most comfortable with.

Again discussion would be held with researchers so that we and they are well informed of the processes required.

Phenotypic and genotypic data from animals used in RELRP, but not accessible to all researchers in the RELRP. The animals used in a number of the projects have other genotypic and phenotypic data which is held by the Sheep and Beef CRCs in particular. This data may well be valuable to projects aiming at integrative analysis across a number of datasets. Consideration should be given to how access to this data by the full group of RELRP researchers can be facilitated. This could be controlled using indirect access to data held in the Sheep and Beef CRC databases.

Issues remaining unresolved. Despite the above achievements, there are still a couple of issues/problems remaining unresolved. These include:

- Project or data complexity causes concerns about the amount of time required to properly document methods or datasets (e.g. some of the project leaders not willing to spend large amount of time, a question frequently asked was “who should be the one responsible doing the tasks?”).
- We are still unclear about the total number of datasets generated and the size of total datasets due to slow response from some of the RELRP projects. This has hampered our effort to accurately document the attributes of datasets from these projects.

5. Benefits of a database to wide range of communities

5.1 Research community

Examples of research questions which could be investigated

- Interaction between animal feed, microbial populations, GHG emission and growth rates.
- List all the Geographic locations for which methane measurement data is available
- Focus on collation of research about specific microbes
- Focus on collation of research about specific breeds
- Focus on specific feed types across different projects (i.e. Pasture types, grains, native shrubs
- Focus on experiments that used a specific technology – or compare technologies

- Specific Suggestion for a Proof of Concept**.**
    - Data across three “seemingly independent” projects should be linked providing a proof of concept of the value of the proposed resource. Our suggestion is to exploit a pathway with logical physiological continuum by linking data and results from Feed Supplementation (project B.CCH.1014) to Rumen Microbiota (project B.CCH.1008) to Manure Management (project B.CCH.1020),. **Appendix J**.

5.2 Policy makers

The database will provide a source of raw data to enable the modeling of the impact of uptake of different approaches to the reduction of GHG emissions on the actual emissions. This will help inform the debate about what is realistic and inform policy decisions appropriately.

5.3 Consultants

A good example will be consultants from the Australian Farm Institute. They will be able to use the information from the database to apply their FarmGAS calculator to provide the individual farms the information on how their farms perform in terms of GHG management.

5.4 Primary Producers

At this stage it is not envisaged that the primary producers will be accessing information directly from this database. Rather they will access tools such as the FarmGAS calculator which in turn use data from the database.

6. Key outputs of the study

The major outputs of the scoping study can be summarized as follows:

- A metadata database illustrating the basic framework structure for the metadata database component of the database and containing detailed descriptions of the project, experiments, treatments and the datasets (i.e. metadata) from individual RELRP projects has been produced. It provides

fundamental infrastructure for establishing a future workable database if the next phase of database implementation project is to proceed.

- A controlled vocabulary list with standardised terminologies for the metadata fields has been generated. This can also be used as a query list to extract the information from the database.
- A data summary sheet showing current status of data attributes from all the RELRP projects has been produced. This information is of great importance to MLA/ DAFF RELRP: 1) it can serve as key IP background information held by the program should they develop new R&D projects; 2) it can be used to provide verification data sets for future modelling exercises.
- A draft design for the full database.

7. Key outcomes of the study

The members of the scoping study team, the RELRP project leaders and many of their staff, and MLA/DAFF etc. have a much better understanding of the state of the datasets generated by the RELRP projects and the options for metadata and full database in the immediate future and for the longer term.

8. Options and recommendations

8.1 Options for the further development of the metadata base/full database within the current RELRP, or in the absence of further funding for activities coordinated with the current RELRP activities

Option 1. Distribute the current metadata database as an MSAccess file.

Option 2. Complete the metadata database and make available to RELRP.

Option 3. Complete the metadata database and make available to RELRP projects via a web interface.

Option 4. Construct a web enabled database with a subset of data from key projects.

Option 5. Construct the full web enabled database.

	benefit	cost	Benefit to cost
1	Partially completed limited functionality metadata database of limited value to general users	No further cost	
2	Additional value for current RELRP program	Small additional cost	
3	Additional value for current RELRP program. Easier to maintain and for users to access	Small additional cost	Recommended solution
4	Little additional value for data generators, value to data users unclear	Substantial extra cost, including potentially complex negotiation of IP issues	Possible solution, see below
5	Little additional value for data generators, value to data users unclear	Substantial extra cost, inclusion of unnecessary data, including potentially complex negotiation of IP issues	Not recommended

8.2 Recommendations for the further development of the metadata base/full database within the current RELRP, or in the absence of further funding for activities coordinated with the current RELRP activities.

R1 The metadata database should be completed (including all datasets generated during the course of the project) and made available to all RELRP project researchers via a web interface

a. This is required as the first step in the completion of the full database implementation.

R2 The interface of the metadata database should link the key outputs of the RELRP projects to the underlying datasets used to generate the results.

R3 Consideration should be given to making the metadata database available publicly.

R4 Careful consideration should be given towards also building the database itself, but only including data from a selected subset of projects.

a. *Since provision of data to a database has not been explicitly funded in the current project allocations, the data organisation and reformatting etc. should be funded through the database project*

b. All of the data sets not included in the database proper will be stored in a file based data repository; this will include the historical data from data B.CCH.1047.

c. data will made available from the database to the Australian Farm Institute (AFI) for use in the FarmGAS calculator as required

8.3 Options for the further development of the full database in a future RELRP, or activities coordinated with the current RELRP activities.

Option 1. Construction of a comprehensive database

Option 2. Construction of a targeted database developed in coordination with the research activities and vice versa, i.e. there is a clear understanding about what data will go into the database, why it is going into the database and how it will be used in the future.

The construction of a database should not be seen as an end in itself, the world is full of underused databases. Whilst there is value in being able to point to the database and say “all of the data is here” this is a relatively limited value compared to being able to say “look at what we have done with all of this data”. To achieve the latter construction of the database must be tightly linked to it’s use. However, trying to be all things to all people is also a recipe for substantial cost for limited extra value.

8.4 Recommendations for the further development of the full database in a future RELRP, or activities coordinated with the current RELRP activities.

R5 The database is a central component of the coordinated activities.

- *Database development is undertaken alongside the development of projects*
- *Development of data generation projects is undertaken alongside the development of data analysis projects and the database itself.*

R6 Contribution (or not) of data to the database by projects is clearly defined during the drafting of project plans and contracts.

- *For many early stage research projects there is very limited value in contribution of the data to a database.*

- *For some activities there may be little value in contributing data to the database.*

R7 For those projects contributing data to the database the costs of project specific data management and provision of data to the database in the agreed format should be included as part of the project funding.

R8 That the conditions of access to the data and the protocols for the access and use of data are agreed by the parties up front and form part of the project contracts.

R9 That the RELRP generated datasets should be imported into a single database – a federated database approach is not practical.

R10 Careful consideration should be given to the value of ancillary data about animals with RELRP supported methane measurements.

- *For example access to associated genotypic and other phenotypic data currently limited to a single project may have significant value to other projects in the RELRP.*

R11 That for projects contributing data to the database deposition of the data in the database is a milestoned activity with payment attached.

R12 That to manage concerns about IP issues the construction of the database, data management etc. be conducted at arms length from the use of the database, i.e. the project team managing the database does not use the database for research.

- *However there is clearly value in having the initial stages of development and implementation of the database undertaken by a team which understands the data and the research objectives.*
- *In the longer term, once the database has been set up and running smoothly consider transferring the day to day management to a specialist database group/service provider.*

9. Planned Database Implementation Project Phases

9.1 General recommendations

To address concerns about access and IP concerns

- Implement standard password authentication and control of access to actual datasets at the level of identification of individuals
- Implement email notification of access to data identifying the query, data accessed and of the individual accessing the data to the owner of the data
- Develop agreed protocol for access to and use of data upfront

General database

- Implement in mySQL or similar

General web interface

- Implement in php or similar to high visual standard and ease of use
- Interface available 24x7

9.2 Web-enabled completed metadata database RELRP datasets

There are a few important characteristics of a good web-based database. These are: allows global access, easy to use (familiar interface that is user-friendly and easily learned), available across multiple platforms, cost effective.

Completion of the metadata database and web interface. A number of tasks are required to complete the implementation of the metadata database.

- Finalization of the metadata database design
- Collect any remaining/new metadata
- Complete the development of the standard vocabulary
- Implement the standard vocabulary across the data received
- Data normalization
- Finalize the querying ability required
- Design the interface
- Link the interface to the metadata database

We estimate that depending on the exact scope of the project as much as the equivalent of 1 person for 6 months may be required to complete the project. Some of the activities listed above could be undertaken in parallel, which would reduce the actual elapsed time required to complete the work.

This estimate is for the development work and does not cover the hosting of the completed database or the maintenance of access, password generation etc.

9.3 Web-enabled database with a subset of data from key projects

Once the metadata database and the web interface have been completed there are two separate stages for the completion of the full database.

Stage I (at least 3 months, assuming IP issues have all been resolved).

- Gathering of datasets into the data repository – decisions made on which data is to be put in the working database
    - 1 month alone will required to sort out the data held be one of the groups (Roger Hegarty pers. comm.)
- Database design to be finalized (according to the types of data to be imported and types of queries required)
- Code to be written to convert the many different datasets to enable them to be imported into the database (requires communications with the data owners)
- Database to be build by importing the converted Datasets

Stage II (3 months)

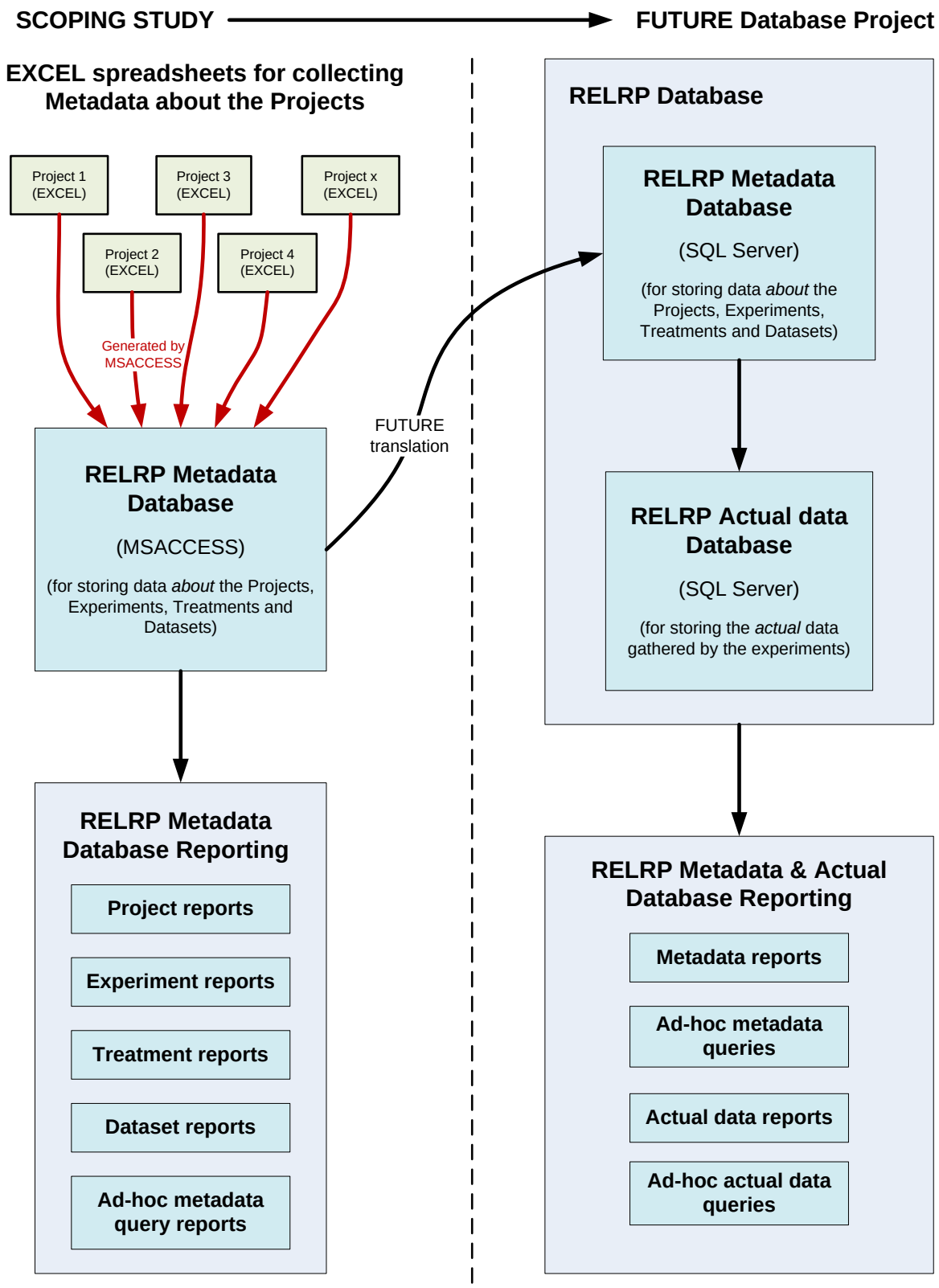
- User interface developed (according to the queries required for different Users – i.e. Researchers, Policy makers, Consultants

Depending on the number of datasets and the size of data and complexity of individual projects, we estimated that up to 2 FTEs with different skills will be required for 6 months for this phase of the database implementation project.

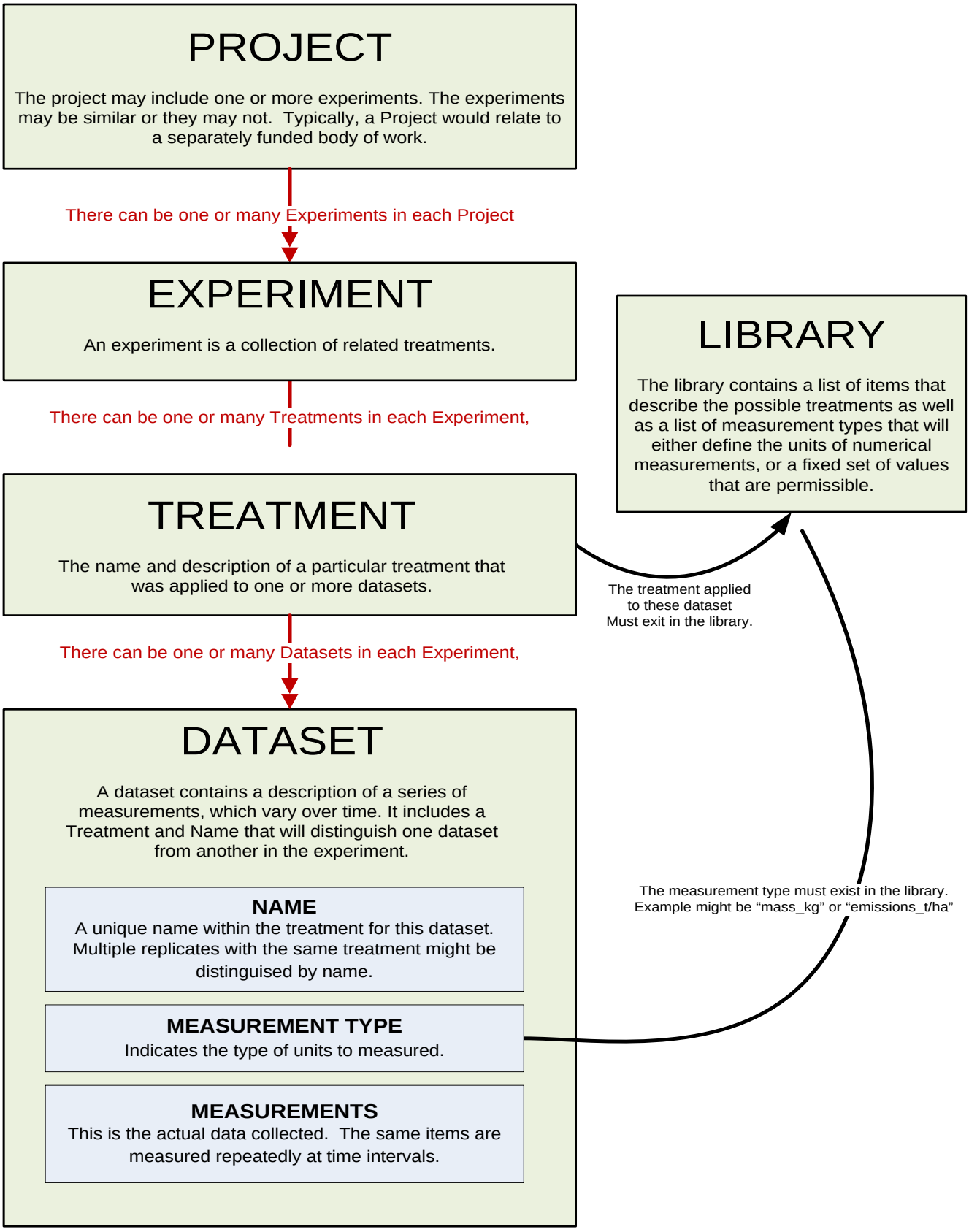
9.4 Potential show stopping issues for a useable and used database v. solely a data repository.

- Issues with the extent of access to data in the data base are not resolved
- Over complex regulation of access such that exploration of the data and the testing of novel ideas and approaches is effectively dissuaded

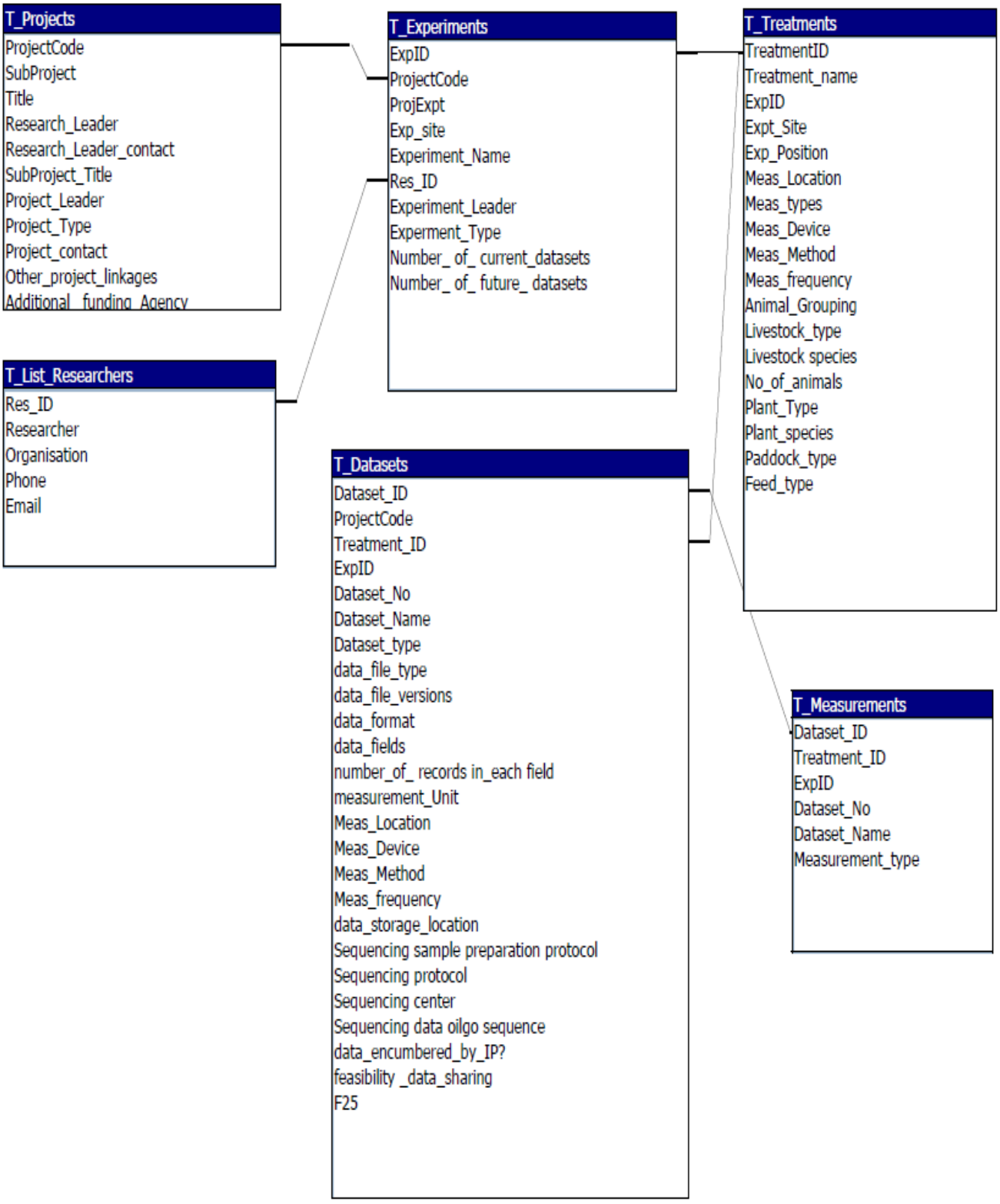
Appendix A. RELRP Metadata Database Design



Appendix B. Draft RELRP Database Design.



Appendix C. Relationships Between the Initial RELRP Metadata Database Query Tables



Appendix D. Initially Designed Metadata Database Survey Form.

FIELDNAME	DESCRIPTION	OPTIONS	VALUES
PROJECTS			
ProjectCode	RELRP Project Code		
SubProject	SubProject Code		
Title	RELRP Project Title		
Research_Leader	Principal Investigator for the RELRP Project		
Research_Leader_contact	contact number and email address		
SubProject_Title	SubProjct Title		
Project_Leader	RELRP Sub Project Leader if several sub projects exist		
Project_Type	Summary of project nature	(i.e. Method development, Industry Demonstration, Rumen functions, Genetic improvement, Rumen functions)	
Project_contact	RELRP Project contact	This can be research leader contact if there is no subproject	
Other_project_linkages	List of other RELRP project(s) linked to this project		

Additional_ funding_Agency	Name of other external funding Agency contributing to project (excluding RELRP and host Institution)	BEEF CRC, Sheep CRC, GRDC etc.	
Other_information	Any other information needed for the project		
Value_of_external_funded_dataset_to_RELRP	Value_of_external_funded_dataset_to_RELRP		
EXPERIMENTS	(REPEAT THIS SECTION IF MORE THEN 1 EXPERIMENTS)		
ExpID	Individual Exp ID		
ProjectCode	RELRP Project Code		
ProjExpt	Exp within the Project Code	A, B, C etc.	
Experiment_Name	Experiment Title		
Experiment_Leader	Experiment Leader		
Experment_Type	Experiment Type		
Number_ of_ current_datasets	Datasets already available		
Number_ of_ future_ datasets	Datasets will be generated		
Expt_Site	Geographic Location of Experiment Site	(i.e. Farm location, Research institute location)	
Exp_Position	Where the measurements were taken on-site	(i.e. Pen in Animal house, cattle yard, paddock pasture, waste water reservoir)	
Meas_Location	Location from which the measurements where taken	(i.e. In-field (paddock), in-vivo (rumen), in-	

Meas_types	Actual types of measurements in the DATASET	vitro (artificial rumen), faeces) (i.e. Gas, Climate, Soil, Production, Microbe profile, Genetic ??) (i.e. FLECK Nano micro-controller, SF6 tracer, Open path laser, Open Path FTIR)	
Meas_Device	Machine used to take measurements		
Meas_Method	Technology, Experiment method used		
Meas_frequency	How often measurements were taken	(i.e. hourly, 5 x daily)	
Animal_Grouping	Whether the measurements are grouped	(i.e. Individual animals, Groups of animals)	
Livestock_type	Livestock involved in the DATASET	(i.e. diary cattle, beef cattle, steer, heifers, sheep, goat)	
Livestock species	Specific Breed or species	(i.e. Fine-wool Merino or strong wool sheep , Angus cattle, Brahman cattle)	
No_of_animals	No. of animals used		
Plant_type	Plants involved in Experiment	(i.e. Forages, Legumes)	
Plant_species	Specific species		

Paddock_type	Paddock grass type	i.e. ryegrass, chickory, lucerne	
Feed_type	Types of Feed involved in the DATASET	i.e. 70% grain-content	
DATASETS	(REPEAT THIS SECTION IF MORE THEN 1 DATASETS)		
Dataset_ID	Individual DATASET ID		
ProjectCode	RELRP Project Code		
ExpID	Individual Exp ID		
Dataset_No	Dataset Number within the Experiment		
Dataset_Name	File name		
Dataset_type	General description of what each DATASET is measuring	(i.e. Genetics, Environment, Microbes, Feed, Gases)	
data_file_type	Specific description of data nature	(i.e. genomic DNA sequence, cDNA sequence, microbial peptides, methane measurement, phenotypic traits including pedigree)	
data_file_versions	Indication of data being raw or processed	i.e. raw data or processed data	
data_format	data file saved format	(i.e. MSExcel, text, MSAccess, database, etc)	

data_fields	Detail variable names		
number_of_ records in_each field	No. of records for each variable		
measurement_Unit	Individual variable measurement unit	(i.e., kg,)	
data_storage_location	Physical location of data files	PC, Server, Portable hard drive, USB, CD etc	
Sequencing sample preparation protocol	sample preparation protocol		
Sequencing protocol	Sequencing protocol		
Sequencing center	Name of sequencing centre used		
Sequencing data oligonucleotide sequence	Oligo sequences if linkers or tags used		
data_encumbered_by_IP?	data_encumbered_by_IP?		
feasibility _data_sharing	feasibility _data_sharing		

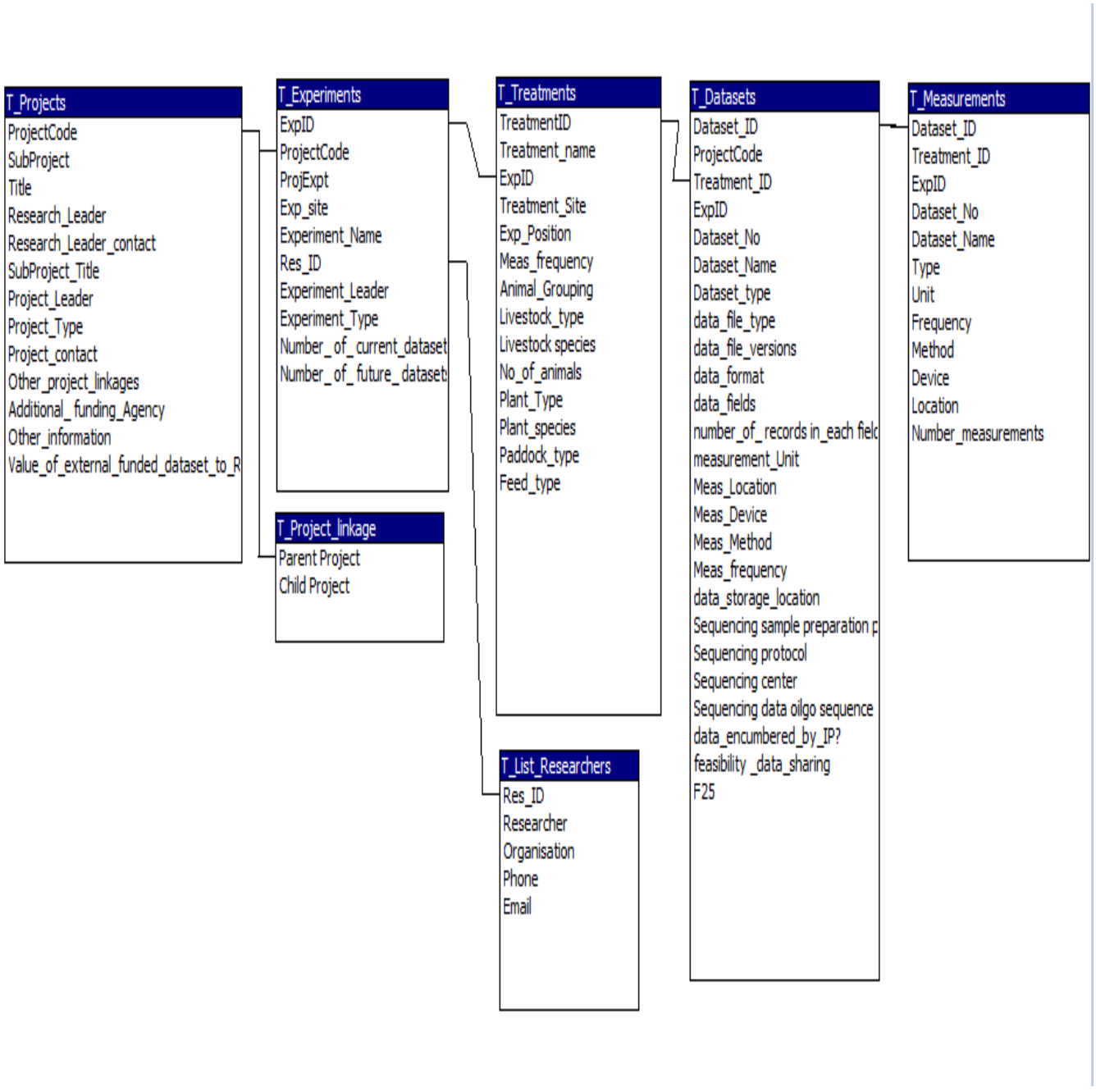
Appendix E. Survey Distribution List and the Response Status

Project name	Title	Research leader	Form Sent	Face to Face Meeting	Telephone conference	Form Return
B.CCH.1003	Novel individual enteric methane measuring system for multiple ruminants	**Chris McSweeney**	√	√		√
B.CCH.1004	Mitigation of methane emissions from the northern beef herd	**Ed Charmley**	√		√	√
B.CCH.1005	Metagenomic analysis of feed utilisation and hydrogen balance in Australian livestock for lower methane emissions	**Mark Morrison**	√	√		√
B.CCH.1006	genetic improvements of beef cattle for greenhouse outcomes	**Robert Herd**	√	√		√
B.CCH.1007	archaeaphage therapy to control rumen methanogens	**Athol Klieve**	√	√		√
B.CCH.1008	rumen microbial profiling- a tool to investigate methane mitigation strategies	**Valeria Torok**	√		√	√
B.CCH.1009	enteric methane abatement strategies for ruminant production systems in SE Australia	**Richard Eckard**	√			√
B.CCH.1010	Novel strategies for enteric methane abatement	**Roger Hegarty**	√	√		√

B.CCH.1011	peptide phage display libraries to discover peptides that are bioactive against rumen methanogens	**Chris McSweeney**	√	√		√
B.CCH.1012	antimethanogenic bioactivity of Australian plants for grazing systems	**Phil Vercoe**	√			
B.CCH.1013	Methanotrophs from natural ecosystems for ruminant methane mitigation	**Athol Klieve**	√	√		√
B.CCH.1014	increasing productivity and reducing methane emissions by supplementing feed with dietary lipids.	**Athol Klieve**	√	√		√
B.CCH.1015	breeding low methane emitting sheep and elucidating the underlying biology	**Phil Vercoe**	√	√		√
B.CCH.1018	microbial ecology of hydrogenotrophic rumen microorganisms in response to methane inhibitors	**Chris McSweeney**	√	√		√
B.CCH.1020	manure management to reduce greenhouse gas emissions from cattle feedlots	**Deli Chen**	√	√		√
B.CCH.1031	demonstration projects for on-farm practical methane management strategies	**Phil Vercoe**	√	√		√

B.CCH.1032	demonstration projects for on-farm practical methane management strategies	**Ed Charmley**	√		√	√
B.CCH.1033	demonstration projects for on-farm practical methane management strategies	**Malcolm McPhee**	√	√		√
B.CCH.1034	demonstration projects for on-farm practical methane management strategies	**Joe Jacobs**	√			√
B.CCH.1036	Open path FTIR project	**Frances Phillips**	√		√	
B.CCH.1043	Further development of the FarmGAS calculator- an online tool assisting farmers and understand emission mitigation potential	**Sally Davison**	√		√	√
B.CCH.1047	Development of historic database and meta-analysis facility for livestock emissions	**Ian Bland**		√		

Appendix F. Relationship between Refined RELRP Metadata Query Tables



Appendix G. A controlled vocabulary list

Factors	Level
age	Value
Australian_native_forage_shrubs	Acacia saligna
Australian_native_forage_shrubs	Acacia mearnsi
Australian_native_forage_shrubs	Acacia sp. (Black wattle)
Australian_native_forage_shrubs	Atriplex nummularia
Australian_native_forage_shrubs	Chamaecytisus palmensis
Australian_native_forage_shrubs	Cullen australasicum
Australian_native_forage_shrubs	Enchylaena tomentosa
Australian_native_forage_shrubs	Eremophila glabra
Australian_native_forage_shrubs	Eremophila longifolia
Australian_native_forage_shrubs	Maireana brevifolia
Breed	Angus cattle
Breed	Brahman cattle
Breed	Friesian cattle
Breed	Friesian cross cattle
Breed	Holstein cattle
Breed	Japanese native goat
Breed	Merino sheep
Breed	Crossbred lamb
Feed component	Hominy meal
Feed component	Brewers grain
Feed component	Cold presses canola
Feed component	Cotton seed
Feed component	Red Grape marc
Feed component	DHAgold
Feed component	Bagasse
Feed type	Dry rations
Feed type	Pasture
Feed type	Green harvest
Feed type	Forage
Frequency	Daily
Frequency	End
Frequency	Every minute
Frequency	Every second
Frequency	Hourly
Frequency	Every two hours

Frequency	Monthly
Frequency	Start
Frequency	Weekly
Frequency	Every season
Gender	Castrated
Gender	Female
Gender	Male
Gas	Ammonia (NH3)
Gas	Carbon Dioxide (CO2)
Gas	Hydrogen (H2)
Gas	Methane (CH4)
Gas	Nitrate (NO3)
Gas	Nitrite (NO2)
GHG measurement place	Animal House
GHG measurement place	Butterbox
GHG measurement place	Chamber
GHG measurement place	Paddock
GHG measurement place	Laboratory
GHG measurement Device	SF6
GHG measurement Device	Open Path
GHG measurement Device	Faecal NIRS
GHG measurement Device	IRD
Livestock species	Bos indicus
Livestock species	Bos taurus
Livestock species	Capra hircus
Livestock species	Ovis aries
Measurement_type	16S sequences
Measurement_type	Accelerometer
Measurement_type	Animal live weight
Measurement_type	Battery voltage
Measurement_type	Biodiversity
Measurement_type	Carbon dioxide
Measurement_type	Device assignment
Measurement_type	Diet composition using F.NIRS
Measurement_type	Diet quality using F.NIRS
Measurement_type	Diet quality using NIRS
Measurement_type	Distance between animals

Measurement_type	Feed dry matter
Measurement_type	Feed fatty acids
Measurement_type	Feed intake
Measurement_type	Dry matter intake using F.NIRS
Measurement_type	Genetic similarity based on DNA sequence information
Measurement_type	Headspace gas
Measurement_type	hydrogen
Measurement_type	Location
Measurement_type	Methane
Measurement_type	Methane concentration
Measurement_type	Microbes
Measurement_type	Milk composition
Measurement_type	Milk yield
Measurement_type	Min temperature
Measurement_type	Max temperature
Measurement_type	OTU
Measurement_type	Pasture composition
Measurement_type	Presence or absence of methanotroph molecular markers.
Measurement_type	qPCR
Measurement_type	RH
Measurement_type	Rumen pH
Measurement_type	Rumen protozoa
Measurement_type	Sequence
Measurement_type	Temperature
Measurement_type	Treatment assignment
Measurement_type	Wind speed
Measurement_type	Wind direction
Measurement_type	Wind turbulence
Measurement_method	Indirect calorimetry
Measurement_method	Weighing
Measurement_method	Oven drying
Measurement_method	SF6
Measurement_method	Rumenocentisis

Measurement_method	Stomach tube
Pasture type	Improved mixed pastures
Pasture type	Legumes –clover/lucerne
Pasture type	Brassicas
Pasture type	Monocultures – oats/ Italian rygrass
Pasture type	Native grasses
Pasture type	Perennial dairy pasture
Plant type	Tropical forage
Plant type	Herbaceous
Plant type	Tree forage
Plant type	Temperate
Plant type	Australian native shrubs
Plant type	Novel forage
Plant type	Eremophila glabra
Plant type	Algae
Project Type	Farm enterprise model options
Project Type	Genetic improvement
Project Type	Genetic improvement / industry demonstration on farm
Project Type	Improved waster management
Project Type	Industry demonstration on farm
Project Type	Meta-analysis literature on In vivo, In vitro methane studies
Project Type	Methane measurement method development
Project Type	Methane measurement method development / Genetic improvement
Project Type	Method development
Project Type	Method development and Rumen functions
Project Type	Rumen functions
Project Type	Rumen microbial manipulation
Sampling	Faeces
Sampling	Rumen fluid
Sampling	Milk
Sites	Belmont Research Station
Sites	Center for Advanced Animal Science
Sites	Chiswick CSIRO
Sites	Douglas Daly Research Station
Sites	DPI Glen Innes

Sites	DPI Grafton
Sites	EcoSciences Precinct
Sites	Highchester Abattoir
Sites	Lansdown Research Station
Sites	Rockhampton Downs
Sites	Trafalgar Station
Sites	Vic DPI Ellinbank
Sites	Victoria River Research Station
Sites	Wamblana Station
Status	Dry
Status	Lactating
Status	Pregnant
Supplement	Suppplement A
Supplement	Suppplement B
Supplement	Suppplement C
Treatment	Fistulated
Tropical_grasses	Astrebla squarrosa
Tropical_grasses	Astrebla elymoides
Tropical_grasses	Astrebla sp.
Tropical_grasses	Bothriochloa insculpta
Tropical_grasses	Bothriochloa bladhii
Tropical_grasses	Bothriochloa decipiens
Tropical_grasses	Bothriochloa pertusa
Tropical_grasses	Cenchrus ciliaris
Tropical_grasses	Chloris gayana
Tropical_grasses	Chrysopogon fallax
Tropical_grasses	Dichanthium aristatum
Tropical_grasses	Dichanthium sericeum
Tropical_grasses	Digitaria decumbens
Tropical_grasses	Eulalia fulva
Tropical_grasses	Heteropogon contortus
Tropical_grasses	Heteropogon triteceus
Tropical_grasses	Iaesaeum
Tropical_grasses	Iseilema membranacea
Tropical_grasses	Iseilema vaginiflorum
Tropical_grasses	Lablab purpureus
Tropical_grasses	Lolium sp
Tropical_grasses	Lolium perenne (Perennial ryegrass)

Tropical_grasses	Megathyrsus maximum-Panicum maximum
Tropical_grasses	Pennisetum ciliare
Tropical_grasses	Setaria sphacelata
Tropical_grasses	Sorgum plumosum
Tropical_grasses	Themeda triandra
Tropical_grasses	Themeda sp.
Tropical_grasses	Triticum sp
Tropical_grasses	Urochola brizantha-Brachiaria brizantha
Tropical_grasses	Urochola decumbens-Brachiaria decumbens
Tropical_grasses	Urochola humidicola-Brachiaria humidicola
Tropical_grasses	Urochola mosambicensis
Tropical_Legumes	Medicago sativa
Tropical_Legumes	Arachis paraguariensis
Tropical_Legumes	Arachis sp. section Erectoides
Tropical_Legumes	Arachis pintoi
Tropical_Legumes	Macroptilium bracteatum (Burgundy Bean)
Tropical_Legumes	Calliandra calothyrsus
Tropical_Legumes	Centrosema molle
Tropical_Legumes	Centrosema pascuorum
Tropical_Legumes	Centrosema pubescens
Tropical_Legumes	Chamaecrista rotuindifolia
Tropical_Legumes	Clitoria ternatea
Tropical_Legumes	Desmanthus bicornutus
Tropical_Legumes	Desmanthus leptophillus
Tropical_Legumes	Desmanthus virgatus
Tropical_Legumes	Desmodium heterophyllum
Tropical_Legumes	Glycine tabacina
Tropical_Legumes	Glyricidium sepicum
Tropical_Legumes	Kennedia prorepens
Tropical_Legumes	Leucaena leucocephala cv.
Tropical_Legumes	Macroptilium atropurpureum (Siratro)
Tropical_Legumes	Medicago sativa
Tropical_Legumes	Sesbania sesban
Tropical_Legumes	Stylosanthes sp.
Tropical_Legumes	Stylosanthes guianensis var intermedia
Tropical_Legumes	Stylosanthes hamata
Tropical_Legumes	Vigna laceolata
Tropical_Legumes	Desmanthus virgatus
Tropical_Legumes	Stylosanthes scabra
Tropical_Legumes	Stylosanthes seabrana

Appendix H. Data summary sheet showing the status of metadata information from individual projects prior to second round data collection (see the attached excel file).

Appendix I. Data summary sheet showing the status of metadata information from individual projects after second round data collection (see the attached excel file).

Appendix J: Suggestion for a Proof of Concept. Distinct projects will provide information about how to mitigate GHG at individual levels of the overall livestock production system. For instance, B.CCH.1014 will investigate the impact of lipid containing feed additives on the suppression of methane emissions. Parallel to it, B.CCH.1008 will investigate changes in rumen microbial communities in relation to methane production. Finally, B.CCH.1020 will explore options for innovative manure management to reduce methane and N emissions. As a proof of concept, it is expected that a meta-analysis of the data resulting from these "seemingly independent" projects will allow for a holistic view of the entire "ruminant" system. For instance, one such analysis may conclude that canola oil supplements allows for the presence of a substantial population of rumen methanogens without major investments in sophisticated manure management.

Feed Supplementation with Dietary Lipids (B.CCH.1014)	Rumen Microbial Profiling (B.CCH.1008)	Manure Management (B.CCH.1020)
1. Spirulina 2. Algamac 3. Canola oil 4. Safflower oil	1. Bacterial 2. Archea 3. Methanogens 4. Fungi 5. Protozoa	1. Stockpile aeration 2. Urease inhibitors (Agrotain) 3. Non-Treated