



A joint Research Councils Programme co-sponsored by Defra and SEERAD

Data resources for rural sustainability research: realising their combined potential.

Final report for reludata website

**Helen M^cKay, Nigel Boatman, James Aegerter, Naomi Jones, Robert
Stones, Alistair Murray, Chris Short* and Les Firbank⁺**

h.mckay@csl.gov.uk

Central Science Laboratory, Sand Hutton, York YO41 1LZ

***Countryside and Community Research Unit, University of Gloucestershire,
Cheltenham, GL50 4AZ**

**⁺Centre for Ecology and Hydrology, Lancaster Environment Centre,
Lancaster, LA1 4AP**

7 November, 2005



Contents

Executive summary.....	1
Introduction and aims	3
Methods.....	3
Results.....	7
Data access and availability	7
Background: legislation and data access initiatives.....	7
Questionnaire responses: researchers’ needs for data access and availability.....	8
Consultations: trends and issues in data access and availability.....	9
Interdisciplinary working.....	10
Background: interdisciplinary research and implications for data needs	10
Questionnaire responses: interdisciplinary working among respondents.....	11
Consultations: under-representation of social scientists in the questionnaire survey.....	12
Data integration.....	13
Background: definition, methods and initiatives	13
Questionnaire responses: data integration undertaken by respondents.....	14
Consultations: data integration issues, tools and advice	15
Data management models	18
Background: data policies and metadata standards	18
Questionnaire responses: experience and service requirements of respondents..	19
Consultations: research programmes and data management	19
Recommendations.....	22
Conclusion	22
Acknowledgements.....	22
Annex A	Questionnaire survey of the RELU community
Annex B	Review of legislation relevant to the RELU programme
Annex C	Review of data sources relevant to the RELU programme.
Annex D	Data integration workshop
Annex E	Review of key Metadata Standards
Annex F	Data requirements of RELU first call award holders

Executive summary

This report seeks to present data management as accessible, relevant, interesting and above all, useful, to a wide range of scientists, by using a common language. In it, we explore generic and interdisciplinary issues of data management and integration relevant to the aims of the RELU programme, and provide a wider perspective on the policy and organisation of rural economy and land use data in the UK. We did this by means of a questionnaire survey of the RELU community, consultation with specialists and by hosting a workshop on data integration (jointly with the RELU Data Support Service).

Data access and availability are addressed by a number of UK and EU laws and directives, the most relevant being the Freedom of Information Act, the Environmental Information Regulations and the Data Protection Act (see Annex B for a review). Information can be protected by intellectual property rights (IPR).

Currently, the most widely used types of data by respondents to the questionnaire survey of the RELU Community (see Annex A for the full report on the survey) were: land use/land cover, agriculture/horticulture and administrative boundaries. Compared to current use, there is a high demand for socio-economic datasets (social attitudes, social structure/social exclusion, recreation, energy, plant and animal disease and waste). Most respondents discovered data through colleagues or directly from the owning organisation's website. Fewer respondents used portals, gateways or data catalogues. Data portals, gateways and hubs provide direct access to data and are reviewed along with organisations providing access to data, in Annex B. Half of the respondents requiring access to external datasets had difficulties gaining access. The most common difficulties were related to cost (44%), confidentiality (21%) and expertise/data structure (17.5%). Ownership difficulties were relatively infrequent (5%).

Consultation with specialists indicated a trend towards greater awareness of the value of data within organisations, research programmes and the government. There are current moves towards providing more data over the web. Access to government-held data is a particular area of contention among researchers (but is improving). The current charging policy of some organisations appears to be holding back progress in research, perhaps resulting in the under-use of some important datasets.

Interdisciplinary working makes particular demands on data availability and use; information on the existence of datasets, as well as documentation and metadata (data about data), are generally organised by discipline. Language is a major barrier to data discovery. Once identified, datasets from different disciplines frequently use different scales or frameworks and require integration prior to joint analyses.

A post-survey classification of respondents to the questionnaire survey of the RELU community indicated a bias against social scientists: the majority of respondents were natural scientists (58%) rather than social scientists (23%) or economists (19%). The most frequently used types of data and particular datasets, were environmental rather than socio-economic (Annex A). Nearly a third of respondents integrated interdisciplinary datasets but fewer intended to do so in the future (12%).

Targeted consultation with social scientists indicated that the language used in the questionnaire was the main difficulty. The terminologies, particularly in relation to data discovery and integration, caused confusion and made the questionnaire difficult and time-consuming to complete. In addition, its relevance to social science research was not always clear.

Data integration refers to merging, or combining, two different data sets, to allow joint analyses. A variety of tools are available, the simplest and most widespread being spreadsheets and databases. The more complex include statistical and mathematical software packages and Geographical Information Systems (GIS). A number of the initiatives reviewed in Annex C relate to data integration.

A quarter of respondents to the questionnaire survey of the RELU community integrated datasets, and a greater number (30%) intended to do so in the future. The main difficulties were related to integrating interdisciplinary datasets. There appeared to be a certain level of ignorance about issues and potential pitfalls, prompting us to organise a workshop on data integration.

Considerations when integrating datasets include (in order of importance): resolution (including scales, accuracy and sampling strategy), the nature of the data (point, line, area or surface) and the distribution of the data (uniform, patchy or continuously varying). Other issues include: data and metadata availability, boundary changes, differences in data format and how to integrate qualitative data.

The **data management models** used by UK Research Councils, research programmes and government departments are described in the report, and metadata standards briefly reviewed in Annex E.

Nearly a half of the respondents to the questionnaire survey of the RELU community had experience of other research programmes, not only funded by the UK Research Councils, but other government and non-government programmes and overseas programmes. Most respondents favoured a proactive approach to data management, including provision of information on data ownership and access, and collation of data and metadata produced by the programme. Specific services which respondents would clearly like to be provided by RELU included: information on data sources/availability, help with data access and acquisition, a communication facility and a collaborative facility for data sharing. There was also a need for legal advice.

From review, consultation and the workshop, the following recommendations are made in this report:

- data management must be adequately resourced, both by the Research Councils and by Government
- data management needs to begin before the projects begin, and needs to start with applicants rather than award-holders
- Successful data management requires a certain level of stakeholder engagement, service provision and training as well as effective enforcement.
- Successful data management for interdisciplinary research requires facilitation of communication between the disciplines and possibly specific training

The Research Councils can learn from each other, in particular the balance between enforcement and training/outreach.

Introduction and aims

The Rural Economy and Land Use programme aims to advance understanding of the social, economic, environmental and technological challenges faced by rural areas, by funding interdisciplinary research to inform policy and practices of how to manage the countryside and rural economies. The RELU programme is a collaboration between the Economic and Social Research Council (ESRC), the Biotechnology and Biological Sciences Research Council (BBSRC) and the Natural Environment Research Council (NERC). This interdisciplinary approach is particularly relevant with the current emphasis on Evidence-Based Policy-Making in the UK government (<http://www.defra.gov.uk/science/how/evidence.htm>).

This scoping study explores generic and interdisciplinary issues of data management and integration relevant to the aims of the RELU programme, and provides a wider perspective on the policy and organisation of rural economy and land use data in the UK. It is intended to complement the work of the RELU Data Support service, which provides direct support and information for RELU award holders on data collection, access, management, storage and archiving (<http://relu.esds.ac.uk/>).

Researchers in rural economy and land use require access to a wide range of data and information as well as a range of computational tools and research methods (for example, developing a new way to use both qualitative and quantitative information) to carry out integrated analyses. In this area there is a greater than usual potential for an interdisciplinary approach, in which scientists from different disciplines work closely together, developing novel methods. There is therefore a need to identify the data requirements of researchers and review the approach of research councils to interdisciplinary programmes such as RELU.

This report is organised into four sections: data access and availability, interdisciplinary working, data integration and data management models. It concludes with recommendations on data management solutions, for the RELU office and for the Research Councils regarding future programmes. Further supporting information is presented in the annexes.

This report seeks to present data management as accessible, relevant, interesting and above all, useful, to a wide range of scientists, by using a common language.

Methods

At the beginning of this scoping study a website reludata.csl.gov.uk was set up to provide background information, access to a questionnaire form, information on events and feedback/outputs.

Data access, management and information requirements were evaluated through a questionnaire survey of the RELU research community (Annex A). An e-mail list of those who had registered an interest in the RELU programme through the <http://www.relu.ac.uk/> website was obtained from the RELU Programme Director's office, and an e-mail request sent to the list asking those involved in research to complete the survey form. An electronic facility was provided via the project website, enabling data to be captured directly into a database for processing and

analysis. A version of the form in MicroSoft Word was provided for those who did not wish to complete the web-enabled version (see Annex A for the complete questionnaire). Table 1 summarises the questions on the survey form.

Table 1. Summary of questions on the form sent to the RELU research community.

1	Main research interest(s)?
2	Type of organisation (categories given)?
3	Have you experience of any other research programmes?
4	Do you require access to external datasets (i.e. those held outside your organisation)?
5	Select data types relevant to you (list provided), and say whether you currently use them or would like to in the future.
6	Give examples of specific datasets you currently access.
7	How do you discover the datasets that you use (categories given)?
8	Have you had any difficulty accessing them?
9	If 'yes', please give details (which datasets, what problem, if and how resolved).
10	Would you like to obtain access to additional datasets for your research?
11	List up to five examples.
12	Do you anticipate any difficulty in obtaining access to any of them?
13	If 'yes', please give details (what dataset, anticipated difficulty from list).
14	Do you currently integrate or use integrated datasets?
15	If 'yes', please give examples (two datasets and the scale of integration).
16	Do you intend to integrate datasets in future?
17	If 'yes', please give examples (two datasets and the scale of integration).
18	Do you anticipate any difficulties?
19	If 'yes', please indicate the difficulties anticipated.
20	What tools do you use for data management and integration?
21	Are there any you would like to use but they are not available to you?
22	What are they?
23	What do you find most frustrating in terms of data availability/access/compatibility?
24	What aspects of data management were managed well (or not) in other programmes?
25	What data services would you like to see provided by RELU (from list)?
26	Any other developments you would like to see implemented?

Supplementary information was extracted from the Data Management Plans completed by award holders and supplied to us by the RELU Office (Annex F). Current legislation relevant to data access and sharing was reviewed (Annex B), as well as relevant data sources (Annex C), using web searches as the primary sources of information.

The results of the questionnaire survey suggested that social scientists were not well represented (see Annex A). Interdisciplinarity, data access and availability issues, trends and technological advances were investigated further through (a) consultation with a small sample of social scientists and specialists, and (b) a data integration workshop held jointly with the RELU Data Support Service on 19 May, 2005, in York.

The consultation was carried out by firstly identifying individuals in five categories (see Table 2). Each was then contacted by e-mail to ask if they would be prepared to take part, and providing a list of questions that would be asked (see Table 3). Social scientists were asked about the questionnaire survey (Table 4). If they agreed, an interview (by telephone or visit) was then carried out at a mutually agreed time. The interview was structured around the pre-prepared questions, but not limited to these, so that the interviews were allowed to extend to wider areas where the interviewee

had relevant information to impart. Notes were made during the interview, and written up shortly afterwards. These notes were used as the basis for reporting, by categorising and collating comments on particular themes. Table 2 lists the individuals consulted and their specialisms.

Table 2. Individuals consulted.

		Social Scientist	Computer Scientist	GIS/statistician	Data Manager	Data Provider
1	Anne Owen, York University.			1		
2	Bill Frogatt, Defra.			1		
3	Colin McClean, York University.		1	1		
4	Dav Stott, CSL.		1	1	1	
5	Mustafa Ahmet and Phil Holden, UK Data Archive.		1		1	
6	Isabella Tindall, CEH and RELU DSS.				1	1
7	Louise Corti, ESRC and RELU DSS.	1			1	1
8	Mark Thorley, NERC.				1	
9	Richard Baker, CSL.			1	1	
10	Richard Budgey, CSL.			1		
11	Ruth Swetenham, CEH.			1	1	
12	Steve Langton, Defra.			1	1	1
13	James Aegerter, CSL.			1		
14	Gavin Parker, University of Reading	1				
15	Matt Lobley, University of Exeter	1				
16	Nick Evans, University College Worcester	1				
17	Susanne Seymour, Nottingham University	1				
	Total	5	3	9	8	3

Table 3. Questions used in consultations with specialists.

GIS:

1	What methods have you used for integrating spatial data? What problems have you encountered, and how have you tackled them?
2	What do you consider are the three major issues with integrating spatial data? (describe each, and explore minor issues within them, how would you set about tackling them). An example of an issue would be: 'consideration of error in mapping'.
3	How would you prioritise these issues?

Data management:

1	What research programmes have you experience of?
2	In each of these, were award-holders required to provide information on their data requirements and the data they will be collecting, in a plan at the beginning of a research programme?
3	In each programme, were researchers helped with data location, or negotiating access?
4	Were researchers aided in data interchange/sharing?
5	Did any of these programmes provide a forum for communication between researchers?
6	Did any of these programmes require researchers to archive data and metadata at the end of

	projects? How was this process influenced/aided? How was it enforced?
7	What aspects of data management (in the research programmes you have experience of) worked well and which did not? Why?

Data provision

1	What proportion/amount of time do you spend in providing data to others?
2	Are you/your organisation financially recompensed adequately for this?
3	Does the system you use for providing data work well? What sort of problems have you encountered when providing data and how have you tackled them?
4	Do you have plans to further develop your data provision interface(s) and if so, how?

Computer science:

1	What new technologies are you aware of which could have a significant impact on data access, management and integration over the next few years?
2	Describe the breadth of access to these technologies i.e. how widely are they available and how widely are they used?
3	What infrastructure is needed to use them e.g. hardware, software, level of knowledge/experience?
4	What impacts do you anticipate for each of these and over what timescale?

Table 4. Questions used in consultations with social scientists.

1	Open discussion on questionnaire. Did you complete it? If not, why not? If you did, what did you find difficult about it? What particular questions caused you difficulty? Why was that?
2	If you didn't complete the questionnaire, would you be willing to allow me to go through it with you now? If not, could you just answer a few more questions:
3	What official or formal data sets do you use to develop your research areas?
4	How does this link in with the other methods of data collection that you use?
5	Do you or will you have need to access rural/environmental data?
6	How do you get hold of datasets you use and how do you expect to use them?
7	What software/tools do you use?
8	Once the project is complete how is the data you generated recorded, analysed and stored?
9	Have you ever re-visited your own or someone else's research for the purpose of reworking or developing this research area?
10	Archiving– does this pose a problem? Would you want to access data that has been archived?
11	The QA issue – how does it impact on you?
12	What does data integration mean to you?
13	Are there aspects of interdisciplinary research between social and natural sciences that you find difficult with respect to data issues? If so, please describe.

Data integration was chosen as the subject of the workshop because it was identified as a topic of wide interest to researchers in the RELU programme, and one where there was a perceived need for further information. The workshop had four aims: (i) to provide information to RELU researchers on issues relating to data integration, (ii) to provide feedback to respondents on the results of the questionnaire survey, (iii) to provide a forum for the Data Support service to interact with researchers; (iv) to obtain additional information and views on topics which were either not suited to inclusion in, or were suggested by the responses to, the questionnaire. The workshop was advertised on the RELU website, in the RELU newsletter, and in addition all those who responded to the questionnaire were sent an e-mail invitation. A number of speakers were invited to give presentations on a range of topics concerned with data integration, and in addition, a breakout session was held, based on a pre-selected set of issues. All presentations given at the workshop were mounted on the project website after the event, by kind agreement of the authors. Further information is provided in Annex D.

Results

DATA ACCESS AND AVAILABILITY

For the purposes of this scoping study, data have a wide definition. They can be spatial and/or temporally structured, quantitative¹ or qualitative², and context-specific. They have been collected for a purpose. They can be observations on a process/information, evidence, explanatory (to answer a question), and therefore have a function (from data integration workshop break-out sessions, Annex D).

By adopting this definition, all researchers can be described as data users (either their own data or collected by others), so data access and availability have a wide relevance.

Background: legislation and data access initiatives

A number of UK Laws and EU Directives relate to data access (reviewed in Annex B). The Freedom of Information Act 2000 gives the public the general right to see recorded information held by public authorities. Both Environmental and Personal information are exempt, as they are covered by other legislation. The relevance to RELU is that researchers can: (a) request data held by public authorities, and (b) expect information on data holdings to exist. The requested information must be communicated if the information exists and if it doesn't fall within an exemption category; Commercial/In Confidence-labelled information is not automatically exempt.

The Data Protection Act 1998 covers all personal records and data held in paper and electronic systems and provides a right of access to the public. Sharing personal data is unlawful if it would breach confidence or break a law. The Environmental Information Regulations 1992: 1998 allow people to request environmental information from public authorities and those bodies carrying out a public function. Confidentiality clauses do not generally prevent disclosure or sharing of data.

Information can be protected by intellectual property rights (IPR). Intellectual property law enables people and organisations to own the fruits of their creativity and

¹ **Quantitative:** relating to numbers or amounts. Cambridge Dictionaries Online
<http://dictionary.cambridge.org/>

² **Qualitative data** analysis is a term for a very wide range of methods for handling rich data records (text, images or sound), without merely reducing them to numbers.

These methods are used across disciplines and professions, including all social and health sciences, market and business research, information, legal, political and historical studies, life histories and policy evaluations.

Analysis of qualitative data requires sensitivity to detail and context, as well as accurate access to information. The researcher aims to create new understanding of a situation by exploring and interpreting complex data from interviews, group discussions, field notes, archival documents or other records.

Extract from QSR International (supplier of qualitative data analysis software) website.

<http://www.qsr.com.au/>

innovation in the same way that they can own physical property. The owner can control and be rewarded for its use.

A number of initiatives exist which relate to data availability (UK and EU), and these are reviewed in Annex C. They range from the Environment Research Funders' Forum (a new focus group bringing together the UK's major public sector sponsors of environmental science) to the Great Britain Historical Geographical Information System (a unique digital collection of information about Britain's localities as they have changed over time, information coming from census reports, historical gazetteers, travellers' tales and historic maps).

Questionnaire responses: researchers' needs for data access and availability

The questionnaire survey of the RELU community indicated researchers require access to a wide range of data (Annex A). Three quarters of respondents indicated that they required access to external datasets (i.e. those held outside of their organisation) to carry out their research.

The most widely used types of data were land use/land cover, agriculture/horticulture and administrative boundaries (>45% of respondents replying to this question). The demand for these was much lower when respondents were asked what they would like access to in the future, suggesting that these areas, plus soil, topography and food and drink, are the only areas suggested where 'supply' in terms of data meets 'demand'. Compared to current availability, there is a high demand for datasets on social attitudes, social structure/social exclusion, recreation, energy, plant and animal disease and waste. It is worth noting that at least three of these are largely socio-economic issues. See Annex A for more detailed analysis.

Regarding specific datasets, Digimap (Ordnance Survey via EDINA) was the most frequently accessed (21% of respondents), closely followed by Countryside Survey (CEH) with 20 respondents (20%). Population and agricultural census, river flow and meteorological data were also widely used. Apart from the population census, the majority of datasets identified were land use or environmental.

Most respondents discovered datasets from colleagues (68%), and/or the owning organisation website (55%). Data portals were used by 40%, almost a third used a library (31%) and only 10% used a data catalogue. Data portals, gateways and hubs provide direct access to data and are reviewed along with organisations providing access to data, in Annex B.

Half of the respondents requiring access to external datasets had difficulties gaining access. Specific datasets mentioned were: Defra Agricultural Census (due to confidentiality) and National Soil Research Institute Soil survey (due to cost). Environmental and economist respondents were more likely (41%) than social science respondents (23%) to have difficulties with access. The most common difficulties with data access were related to cost (44%), confidentiality (21%) and expertise/data structure (17.5%). Ownership difficulties were relatively infrequent (5%). The majority of respondents indicated they would like to gain access to additional datasets to undertake their research. The datasets most frequently needed were soil, climate and land use/cover.

Consultations: trends and issues in data access and availability

All the specialists consulted felt that there was a general trend towards greater awareness of data and information issues, and appreciation of the value of data, within organisations, research programmes and the UK government. This trend is perhaps a consequence of changes in both legislation (see Annex B for a review) and Government policy (Evidence-Based Policy Making³). The ways in which data are now being used are more and more integrated, because the questions that Defra is asking are more holistic.

There are current moves towards providing more data over the web, using gateways, hubs and portals (reviewed in Annex C). Technological advances, such as grid technology⁴, will enable greater access to data in the future (see also the presentation by Mark Birkin of NCeSS at the Data Integration Workshop <http://reludata/>). The United States has a long tradition of data sharing and ensuring that publicly funded data are in the public domain. Within Europe, there are the beginnings of a move towards open access to and unrestricted use of data⁵.

Access to UK government data is a particular area of contention among academics. In the past, access to data such as the population Census and Agricultural Census has been restricted (due to confidentiality) but the situation is improving, due both to recent changes in legislation and also the adoption of Evidence-Based Policy. New methods for anonymisation of data should also help to improve access still further.

³ <http://www.defra.gov.uk/science/how/evidence.htm>

⁴ The idea behind **Grid technology** is to link up computers around the world over the Internet to create a new generation of enormously powerful machines. BBC News website <http://news.bbc.co.uk/1/hi/technology/3152724.stm>

Grids allow organisations to share geographically-distributed applications, data and computing power over the internet. It means one company's data could be used in another's application running on someone else's hardware, regardless of the technology.

Computer active website <http://www.computeractive.co.uk/computing/news/2068874/ibm-locks-future-grid-technology>

⁵ Extract from a document on the **OECD** website: 'Towards International Guidelines for Access to Research Data from Public Funding' by Peter Schröder <http://www.oecd.org/> :

January 2004, science ministers endorsed a Declaration inviting the Committee to draft formal OECD guidelines on the subject. Considering the above issues, the ministers proposed that the technical, institutional, financial, legal and cultural aspects of data access regimes be addressed along the lines of the following:

THE PREMISE: OPEN ACCESS

Scientific progress relies on open access to research data. An open exchange of information and knowledge is indispensable to the advancement of scientific research. Open access to and unrestricted use of data outside their initial use promotes open scientific inquiry, diversity of analysis and opinion, scrutiny and testing of hypotheses and results, methods and techniques of analysis and facilitating teaching of research. Non exclusive access to data increases the efficiency of research by avoiding unnecessary duplication of data collection and permitting the creation of new data sets by combining data from multiple sources. Openness should balance the interests of open access to data to increase the quality and efficiency of research and innovation with the need for restriction of access in some instances to protect social, scientific and economic interests.

The core principle of data access regimes should be open access as the default: publicly-funded research data should be openly available to society, subject only to legitimate restrictions. Open access has a price and requires robust funding arrangements.

The message from Defra and other Government Departments seems to be: ‘don’t assume you can’t get access to data, because you couldn’t in the past’. Having said that, barriers to data access and availability still exist. Overall, resources within Defra are getting tighter, and there is a tendency to concentrate on providing data to administrators (EU) rather than to researchers, perhaps due to ignorance (among Government data holders) of the value of the data to other departments and organisations. Consequently, the need for maintaining good time series of data is not fully appreciated, and insufficient effort is made to address compatibility issues, such as changes in boundaries (see the following section on Data Integration). There is also a danger that some data will not be available because they will no longer exist (when administrative requirements change, datasets could be discontinued). It is conceivable that changes in government data policy will threaten the integrity and value of datasets that are key to the RELU community, and to researchers in land use and rural economy in general, both now and in the future. One of the recommendations of this scoping study is that the Research Councils need to work with the Government and Defra in particular, to put adequate resources into data provision (which includes documentation, maintenance and archiving) as well as data collection.

Having said that, there is a general trend towards greater access to government held data. For example, within Defra, the ‘Whole Farm Approach’ initiative will enable farmers to have greater and direct access to government information (see Annex C). There is also a trend towards taking an overall view of data, away from the ‘silo-mentality’ of the past, in which government departments rarely considered their data in relation to data held in other departments.

Where intellectual property (IP) is an issue, data access usually involves extra time and effort to negotiate an agreement to share or exchange data (this is the general experience of the specialists consulted). Free access to data is regarded by some as giving them away to competitors, and some organisations seem to charge commercial rates for data as a default. It takes time and effort to negotiate discounts for academic, non-commercial use. Open access may therefore not be achievable for some datasets unless appropriate conditions can be negotiated to protect owners of IP, or compensation provided.

There seems to be an issue with overcharging for national data by data holders. With developments in technology, researchers are now able to deal with large datasets, yet some organisations’ charging policy still involves multiplying up the cost of regional or sub-regional datasets (e.g. NSRI soils data, where problems had been encountered by a number of questionnaire respondents and specialist consultees), resulting in prohibitive charges, and the consequent under-use of national datasets. Where datasets contribute to policy indicators, government intervention is required: organisations must be adequately funded for data maintenance and provision, as well as data collection. This should help reduce costs of data acquisition for non-commercial gain.

INTERDISCIPLINARY WORKING

Background: interdisciplinary research and implications for data needs

The rural economy and land use programme enables selected researchers to work together to investigate the social, economic, environmental and technological challenges faced by rural areas (RELU website <http://www.relu.ac.uk/about/>). This is

of particular relevance for Defra. In the overview of the First Report Of The Sustainable Farming And Food Research Priorities Group⁶ it states: *There was a very strong emphasis on socio-economic issues. There is relatively little of this type of research in the current Defra, SEERAD, DARD5 and BBSRC sustainable farming and food programmes.*

Interdisciplinary research, through developing novel approaches, can sometimes provide benefits over more traditional disciplinary approaches. There are, however, barriers to overcome. Data access and availability, information on the existence of datasets, as well as documentation and metadata (data about data), are generally organised by discipline (see Annex C).

For example, the Gigateway is aimed at increasing awareness of and access to geospatial information in the UK. The Data Locator can be used to find out what geographic datasets exist. The Data Directory will help locate organisations which supply geographic data, products and services. The keywords and directories which can be used to search for data are mainly derived from natural sciences.

Taking another example, the UK Data Archive (UKDA) is a centre of expertise in data acquisition, preservation, dissemination and promotion. It is curator of the largest collection of digital data in the social sciences in the UK and houses a major collection of computerised historical material. The subject categories used to browse for data, in contrast to Gigateway, are social science-orientated.

Following on from these examples it is also true to say that language is a major barrier to discovery and identification of relevant datasets for interdisciplinary research.

Questionnaire responses: interdisciplinary working among respondents

A post-survey classification suggested that the majority of respondents to the questionnaire survey of the RELU community were natural scientists (58%) rather than social scientists (23%) or economists (19%). Currently, the most widely used types of data are environmental (land use/land cover and agriculture/horticulture). Specific datasets most frequently accessed are Digimap (Ordnance Survey) and Countryside Survey.

Respondents were asked to give examples of integrated datasets that they use, and these datasets were categorised according to discipline ('natural science' or 'socio-economic'). The majority of examples given were of integration between natural science datasets and only four of the examples were of integration between socio-economic datasets. Most were integrated using landscape rather than political spatial units. Nearly a third of the examples given (29%) involved interdisciplinarity (between natural science and social data). Respondents were asked to give examples of datasets they intended to integrate in the future. A lower percentage of the examples given were interdisciplinary (12%). The reason for this apparent difference was not clear. The difficulties encountered in integration, however, may discourage researchers.

These results suggested that socio-economic scientists were under-represented in the survey. We decided to investigate the reasons for this by targeted consultation of social scientists.

⁶ <http://www.defra.gov.uk/science/documents/RPG/Papers/FinalRPGreport.pdf>

Consultations: under-representation of social scientists in the questionnaire survey

The most common reason given for not responding to the questionnaire survey was 'lack of time' (from consultations, as well as non-respondents, Annex A). However, the social scientists consulted highlighted a related issue: the questionnaire at first sight appeared to be 'quite demanding', and they were put off by 'talk of large datasets'. The social scientists consulted often worked with small datasets, supported by summary data.

Other difficulties with the terminologies used in the questionnaire were indicated (see Annex A for the exact wording used on the form). Taking some examples: the term 'discovery of datasets' had caused confusion as it was regarded as being part of the research process. 'Quality Assurance' generally elicited limited response. The 'third party data' term was not used by social scientists. 'Data ownership' categories were not clear to social scientists. 'Data linking' was used rather than 'data integration', and market research used the term 'data layering'.

Other meanings for data integration given by social scientists included:

folding in different types of data to give a more balanced picture
integrating qualitative and quantitative data trying to get sensible results from different spatial units

Examples given by social scientists of data integration included:

integrating Countryside Survey data with farm survey data
analysing aerial photos with agricultural census data.

These meanings and examples seem to indicate that once the differences in terminologies are considered, the practical tasks are much the same as for environmental scientists. Examples provided by respondents to the questionnaire survey (from Annex A) included integrating:

agricultural census data with administrative boundary data
bird census data with Land Cover Map 2000 data
climate data with soils data

The tools used by the social scientists consulted were similar to those used by respondents to the questionnaire survey: Excel, Access, SPSS, MapPoint. However, NUD*IST was also mentioned (software for analysing qualitative data, now called N6, see footnote 2).

When asked about difficulties with interdisciplinary research, it was suggested that common language (for example, an ontology⁷) was the first pre-requisite, and data less important. Issues of: scale, sample size, sampling methodology and representativeness were also cited, as well as qualitative data (getting other scientists to appreciate its value). Complex statistical tests were a major barrier.

When asked about data access, it was clear that most social scientists did need access to external data, but not necessarily large datasets (e.g. Government Department summaries of data, non-electronic data). These data were often used as a background to inform the researcher's own study. Defra data were commonly cited in regard to difficulties with access. In the data integration workshop, it was noted that different

⁷ the term 'ontology' has been used very loosely to label almost any conceptual classification scheme. Extract from Wikipedia (free encyclopedia) web site
http://en.wikipedia.org/wiki/Ontology_%28computer_science%29

disciplines view data differently. There seems to be a tight disciplinary division, e.g. GI data from Gigateway, social science data from ESDS DA. Social scientists who use qualitative data often don't look for, or need, data collected by others because they are trying to understand a particular view or behaviour, rather than measure in a way that requires a national baseline for comparison. Discovery metadata⁸ are also discipline-specific. Naming conventions and common units are important in provision of metadata useful to interdisciplinary research. Descriptions of metadata fields are often not self-explanatory to non-specialist users, and examples can be helpful in understanding their use and application

In the wider consultation, it was the experience of some data specialists that geospatial information is often absent from social science datasets. When data are collected, it may not appear to be necessary, or useful, to attach a spatial element to observations, even when it is possible to do so. If all datasets were geo-spatially referenced (if possible), this would enable subsequent re-use and integration with other datasets. This might be possible retrospectively, if the raw data are archived.

DATA INTEGRATION

Background: definition, methods and initiatives

Data integration means different things to different disciplines. In its simplest form it refers to merging, or combining, two different data sets, to allow joint analyses. It can include combining data of different types, e.g. qualitative with quantitative data, or data of different spatial units, e.g. aerial photography with agricultural census data. Geographically-referenced datasets can differ in terms of: boundaries, scale units and resolution (e.g. 1 km², parish, NUTS⁹, SOAs¹⁰), geography (e.g. point, line, area or surface), and distribution (uniform or patchy). A variety of tools are available to

⁸ What is **Discovery Metadata**?

There are three different levels of Metadata: Discovery metadata - which answers the question, "What datasets hold the sort of data I am interested in?"; Exploration metadata - "Do the identified datasets contain sufficient information to enable a sensible analysis to be made for my purposes?"; and Exploitation metadata - the process of obtaining and using the data that are required.

Extract from gigateway web site <http://www.gigateway.org.uk/metadata/default.html>

⁹ NUTS nomenclature (the official regional breakdown for all EU countries).

For the latest status of the NUTS nomenclature, please refer to the Eurostat Internet site:

http://europa.eu.int/comm/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC

Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS)

Country level: 15 countries;

Level 1: 72 regions;

Level 2: 213 regions;

Level 3: 1091 regions;

Level 4: 1904 regions (for 6 countries only);

Level 5: 98433 communes or equivalent.(situation of 1991)

¹⁰ Super Output Areas (SOAs) are a new geographic hierarchy designed to improve the reporting of small area statistics in England and Wales. Their first statistical application was for the Indices of Deprivation 2004, giving them instant publicity and usage across the local government sector. They have been increasingly used for datasets on the Neighbourhood Statistics (NeSS) website and it is envisaged that they will eventually become a standard across National Statistics and beyond.

National statistics website <http://www.statistics.gov.uk/geography/soa.asp>

integrate data. The simplest and most widely available tools include spreadsheets and databases. The more complex include statistical and mathematical software packages and Geographical Information Systems (GIS).

Geo-referenced and map data can be integrated using GIS. In the 1930s and 40s geographical analysis was conducted by overlaying different types of maps of the same area and based around the discipline of cartography. Since the 1950s systems have evolved to convert this mapping into digital form and more recently to use these data for analysis and problem solving. Nowadays GIS is widely used. The GIS must be able to store information about:

- the geometry: the shape and location of the objects
- the attributes: the descriptive information known about the objects, normally displayed on a map through symbology and annotation¹¹

Extract from Ordnance Survey website

(<http://www.ordnancesurvey.co.uk/oswebsite/>).

A number of initiatives exist which relate to data integration (Annex C). The Great Britain Historical GIS project provides a spatial framework for historical information about Britain. "Historical" means it is focused on documentary evidence, not on archaeological finds, and it deals with extensive "places" and administrative units, not precise points on the ground - and dates are usually quite precise. They work only with sources that cover the whole or large parts of Britain, but these include statistics, boundary information, historical maps and even travel writing. The project has collaborative relationships with the Office of National Statistics, the National Archives, the British Library, English Heritage and the Environment Agency.

The Defra SPIRE (SPatial Information Repository) project will create a geographical information repository. The EU INSPIRE (INfrastructure for SPatial InfoRmation in Europe) initiative aims to make available 'relevant, harmonised and quality geographic information for the purpose of formulation, implementation, monitoring and evaluation of Community policy-making'. These initiatives will need to address data integration in future stages of the projects.

The NERC DataGrid Programme aims to enable users to compare and contrast data from an extensive range of (US, European, UK, NERC) datasets from within one specific context. The project initially involves datasets within the meteorological and oceanographic community, however it eventually it will allow appropriate data held across all the NERC disciplines to be available via the NERC DataGrid interface. Data integration is an integral part of the programme.

Questionnaire responses: data integration undertaken by respondents

A quarter of respondents to the questionnaire survey of the RELU community indicated they currently integrate or use integrated datasets (Annex A). Examples of the datasets integrated are listed (Appendix E of Annex A). Most were integrated using landscape (e.g. km², hectare etc.) rather than political (e.g. wards, parishes,

¹¹ See 'GIS files' for further information on the basic concepts of GIS:

<http://www.ordnancesurvey.co.uk/oswebsite/gisfiles/section1/>

counties etc.) spatial units. The main difficulties encountered by respondents were experienced when attempting to integrate social datasets with natural science datasets, and RELU may have an important role to play in enabling researchers to understand key elements of the process required to successfully undertake this.

A greater number of respondents intended to integrate datasets in the future (30%). Most respondents did not know if they were likely to encounter difficulties when integrating datasets (62%), the most frequently anticipated difficulty related to spatial issues.

The most common data integration tool, used by three quarters of all the respondents, was a spreadsheet. Mapping/GIS and statistical software were used by half of the respondents. Databases (36%) and graphical software (24%) were less commonly used. The Countryside Information System (CIS) was used by 10% of respondents. Nearly half of the respondents felt there were no processes that they could not carry out because the appropriate tools were not available (49%), less than half said they didn't know and only 7% thought they needed further tools.

These results seem to indicate firstly that a substantial and growing proportion of researchers integrate datasets. Secondly, there appears to be a degree of ignorance of the difficulties involved, and the tools required. Data integration issues and tools were investigated further in the Data Integration and in consultations with GIS users.

Consultations: data integration issues, tools and advice

Data integration includes not only spatial integration, but temporal (time series) and conceptual (data can be mapped together to reach a common interpretation, rather than formally mathematically combined). Non-geo-referenced datasets, and qualitative data in particular, are seen as a technical and intellectual challenge.

Initial stages of the research process involve formulating the question(s), and then investigating the availability of data and possibly developing a plan for collecting additional data. Datasets, metadata, or documentation are then accessed, or collected, and a decision made on the common framework at which to integrate the data. OS MasterMap, LCM2000 and British National Grid frameworks were frequently mentioned for digital data, SOAs, NUTS and parishes for non-digital data.

A key issue with using geo-referenced data is understanding where problems may occur. This determines what you can do with the data. The RELU project RES-224-25-0062 *Developing spatial data for the classification of rural areas according to socio-economic and environmental conditions* provides a case study. One of the first considerations must be: how the data are distributed (uniform, patchy or continuously varying). The nature of the data (point, line, area or surface) will also indicate how the datasets should be integrated. In addition, resolution should be considered (this includes scales, accuracy and the sampling strategy).

Difference in scales between datasets is an important issue. When integrating a number of datasets, you need to be clear that you are working at the resolution of the coarsest one, effectively eroding the quality of the other datasets. There is a real danger of unintentional abuse of spatial data, which is exacerbated by the trend to more user-friendly GIS software. Advice on the valid use of data, and warnings, should perhaps be given to users along with the metadata, by the data providers.

Datasets may be in different (software-specific) formats. This is becoming less of a problem with the adoption of common formats by the market leaders. The software to deal with differences in formats can be expensive, however.

ArcGIS¹² is the most commonly used commercial software for integration of geo-referenced datasets, and contains tools that can be used to extract information from map data. Spreadsheets such as Excel, databases such as Access and statistics programmes such as SPSS can then be used to manipulate the data. Open source software is also available, for example GRASS¹³ and R¹⁴. The RELU project RES-224-25-0099 *Integrating spatial data on the rural economy, land use and biodiversity* is applying the technique of Genetic Algorithms¹⁵ to integration of environmental with socio-economic data.

Data visualization is a particular area of data integration. Two or more datasets can be displayed on a map, amalgamated into one, or allowed to ‘communicate’ with one another. The simple act of drawing a map prompts questions and helps communication and understanding of data. Recent advances include the Virtual Reality Suite of the Informatics Research Institute at Newcastle University <http://www.ncl.ac.uk/iri/facilities/index.htm>, which uses ViewScape, a multi-input, multi-display control device which allows multiple visualisation sources to be presented as a combined single display in a variety of different formats.

During consultations, GIS users (nine out of the 17 specialists consulted) were asked what they considered to be the most important issues in regard to data integration. The results are shown in Table 5. Difference in resolution/scale was the most important issue, mentioned by six of the nine individuals consulted, and ranked first by three of them. It was closely followed by data availability (mentioned by six and ranked first by two). Nature of the data to be integrated (e.g. point, line, area, also vector or raster) was the third most important issue, and lack of adequate metadata (including information on quality and other documentation) the fourth. Other issues included data distribution (i.e. continuously varying, or patchy), boundary changes (of particular importance when integrating data collected at different points in time), data format (although differences in format are becoming less of an issue with developments in the software) and how to integrate qualitative data.

Raster and Vector are the two fundamental methods of storing map information in digital form (Ordnance Survey GIS website <http://www.ordnancesurvey.co.uk/oswebsite/gisfiles/section1/>). All locations and shapes can be defined in terms of x and y coordinates from a given grid system: it is these numerical values which are used to translate map information into digital form. This applies in both vector and raster formats. In vector data the features are recorded one by one, with shape being defined by the numerical values of the pairs of x y

¹² **ArcGIS** is an integrated collection of GIS software products. ESRI (GIS and Mapping software) website <http://www.esri.com/software/arcgis/about/overview.html>

¹³ Geographic Resources Analysis Support System (**GRASS**) is a GIS <http://grass.itc.it/>

¹⁴ **R** is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. <http://www.r-project.org/>

¹⁵ **Genetic Algorithms** are methods for optimisation. To use a genetic algorithm, you must represent a solution to your problem as a genome (or chromosome). The genetic algorithm then creates a population of solutions and applies genetic operators such as mutation and crossover to evolve the solutions in order to find the best one(s). <http://lancet.mit.edu/~mbwall/presentations/IntroToGAs/>

coordinates. Vector data can be thought of as a list of values. In raster data the entire area of the map is subdivided into a grid of tiny cells. A value is stored in each of these cells to represent the nature of whatever is present at the corresponding location on the ground. Raster data can be thought of as a matrix of values. The major use of raster data involves storing map information as digital images, in which the cell values relate to the pixel colours of the image. Both types of data are very useful, but there are important differences – the characteristics below are broad generalisations which do not necessarily apply in all circumstances.

Vector:	Raster:
relatively low data volume	relatively high data volume
faster display	slower display
can also store attributes	has no attribute information
less pleasing to the eye	more pleasing to the eye
doesn't dictate how features should look in a GIS	inherently stores how features should look in a GIS

The wider consultation revealed differences between disciplines. Computer scientists all thought data format was an issue. Data managers and providers thought metadata was the most important issue, closely followed by data availability.

In conclusion, the following list may help the user to address the major issues and avoid the common traps in integrating data:

1. Find out which datasets are available which meet your research requirements, and obtain as much information, documentation and metadata as possible about each.
2. Decide on the framework at which to integrate (e.g. the scale, grid system, resolution).
3. Are there historical issues, such as political boundary changes? If so, can this be resolved?
4. Consider the distribution of the data. Is it uniform, patchy, or continuously varying? Consider using weighted averages if uniform, or an alternative if not. Consider using novel methods such as genetic algorithms.
5. Consider how the data are stored. Are they point, line, area or surface? Are they vector or raster? Are they compatible? If not, how can the data be manipulated?
6. Consider scales and resolution. Are they compatible? If not, how can the data be manipulated? Is the resulting integration valid? What assumptions have you made and how can these be made explicit to potential users of the data you create?

Background: data policies and metadata standards

The principles of data policy and management within RELU¹⁶ can be summarised as:

- a) Publicly funded research data are a valuable long-term resource
- b) Data generated by RELU projects are to be well managed
- c) RELU researchers are expected to make these data available for archiving
- d) RELU funds will be available for data management
- e) Long-term data management will be the responsibility of the funding Research Councils

The RELU Data Support Service¹⁷ (DSS) was launched in January 2005 to support award holders and applicants. Data Management Plans (DMPs) were collected from award holders, an advisory service (the DSS) and web-based information portal were set up, a programme of outreach and training embarked upon as well as work on identifying key external data sources for RELU projects.

The NERC data policy handbook¹⁸ describes the nature of the data resource, ownership and custody of data, obligations of data holders and obligations of NERC. Responsibilities for NERC data have been delegated to its seven Designated Data Centres. The handbook also covers access to, and charges for NERC's data as well as legal and contractual obligations. Final payment of awards should not be made until Data Centres agree that award holders have met their obligations (including producing a DMP at the start of the project, and data for archiving at the end).

The ESRC Data Policy¹⁹ includes their definition of datasets, award applicants and holders' obligations (to complete a "data review" before the start of the project, and to offer data for archiving at the end), ESRC support (through the Data Archive or via Qualidata), ESRC obligations, charges (generally nil for academic use) and legal, contractual and ethical issues.

The BBSRC Statement on Safeguarding Good Scientific Practice²⁰ is not prescriptive about individual approaches, but underlines the principles of honesty, openness and guidance from professional bodies. It also describes award holders' obligations (to document results and store primary data, publish results and acknowledge collaborators) and BBSRC obligations (confidentiality, security of data provided etc.).

UK Government Departments generally do not have data policies. Data management has been 'fit for purpose'. Only the elements needed to efficiently carry out government activities are taken on board. Confidentiality, national security and administrative obligations have been paramount. The needs of academic researchers have therefore assumed lower priority. However, with the advent of Evidence Based Policy Making (<http://www.defra.gov.uk/science/how/evidence.htm>), the value of data has come to the fore. Also, public bodies must abide by the Freedom of Information Act and produce Publication Schemes (what information is available and

¹⁶ <http://www.relu.ac.uk/about/data.htm>

¹⁷ <http://relu.esds.ac.uk/>

¹⁸ <http://www.nerc.ac.uk/data/policy.shtml>

¹⁹ http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/DataPolicy2000_tcm6-8233.pdf

²⁰ http://www.bbsrc.ac.uk/funding/overview/good_practice.pdf

how to get it). Defra's accessibility commitment²¹ encourages better information management 'through the creation and storage of records electronically and the implementation of the Code of Practice for Records Management'.

Data management involves the maintenance of metadata. A number of different standards exist and five are reviewed in Annex E (Dublin Core, e-government metadata standard, the geographical metadata standards UK GEMINI and SPIRE and the DDI - Data Documentation Initiative).

Questionnaire responses: experience and service requirements of respondents

Nearly a half of the respondents to the questionnaire survey of the RELU community had experience of other research programmes, not only funded by the UK Research Councils (48% of respondents, specifying 26 programmes), but other government and non-government programmes (68% of respondents) and overseas programmes (7% of respondents).

When asked about other research programmes, most respondents favoured a proactive approach to data management, including provision of information on data ownership and access, and collation of data and metadata produced by the programme.

Specific services which respondents would clearly like to be provided by RELU include:

- information on data sources/availability
- help with data access and acquisition
- communication facility for interaction with other award holders
- a collaborative facility for data sharing with other award holders

There was less support for provision of tools for data integration, possibly due to ignorance (see previous section).

Finally, it seemed clear that respondents were in need of advice on legal issues including Freedom of Information Act and confidentiality and IP issues including data ownership. When asked about other developments in data management they would like to see implemented in RELU, widening the availability of data in an appropriate form seemed to be the overriding message.

Consultations: research programmes and data management

The Data Integration Workshop addressed data management models in the breakout sessions. Delegates felt that RELU could undertake training to enable interdisciplinary working and improved understanding of other disciplinary fields. Delegates considered that the RELU Data Support Service had begun operating too late in the life cycle of the programme, that there had been insufficient communication and feedback to award holders (for example, information from the Data Management Plans, such as datasets required by other projects). Publication of the list of data requests across the RELU programme would be useful. RELU could aim to help with making the most sought-after datasets available to researchers. It would also be useful to alert award holders to those datasets for which the DSS could help with access, to avoid duplication of effort.

²¹ <http://www.defra.gov.uk/corporate/.opengov/accessibility.htm>

RELU is perhaps the most ambitious interdisciplinary Research Council research programme so far. It builds upon the multidisciplinary work of the NERC URGENT and LOIS programmes and the ESRC JAEP programme.

URGENT - Urban Regeneration and the Environment, is a Thematic Programme of research funded by NERC (1997 – 2002). URGENT aimed to stimulate the regeneration of the urban environment through understanding and managing the interaction of natural and man-made processes. URGENT was a test case for developing data management, and a greater proportion of resources were invested than in previous programmes. A Quality Assurance (QA) Guidance Manual was prepared for the Programme, and a completion report details the lessons learnt. Key recommendations relevant to this report were:

- It is essential to get Data Centres to “sign off” completed projects before final payment is made. If this is not done, it is much more difficult to get PIs to co-operate once they have received their final payment.
- It is essential that data management and QA activities for Thematic Programmes commence at, or before, the start of a Programme. Consideration should be given to centralised (top sliced from each Programme?) funding for these activities. This would ensure consistency of approach and continuity across all Thematic Programmes

URGENT developed a web based metadata tool, a search facility and thesaurus, and data management was accomplished by providing help, outreach and generally ‘the carrot rather than the stick’. URGENT was successful in obtaining a high proportion of the data produced in the programme. The key recommendations above addressed perceived shortcomings of the project data management.

LOIS (Land Ocean Interaction Study) was a six year (1992 – 1998) NERC project studying riverine, atmospheric, estuarine, coastal and shelf processes. The programme produced large data sets which were handled by five data centres. Data Centres ensured the maintenance of data standards, and made data available to modellers and other scientists in the LOIS community.

In contrast to NERC, ESRC funds data management centrally and tends to take a bottom up approach, concentrating on training and outreach rather than enforcement. QA (generally called Good Research Practice) and archiving are important issues. ESRC use the UK Data Archive model, and the DDI²² social science metadata standard (see Annex E).

Data Managers should encourage access to data from outside programmes, with the premise of open access to data. The specialists consulted generally felt that much time and effort was spent negotiating the use of datasets, some key datasets were under-used, and current restrictions were hampering scientific progress.

Programmes with interdisciplinary aspirations need to promote communication between environmental and social researchers. They should also provide information on the other projects within the programme, to encourage collaboration.

Data Managers were very clear about the importance of metadata and detailed and explicit documentation of datasets. The ideal model is a repository of data, with metadata and a system to interrogate it.

²² The Data Documentation Initiative (DDI) is an effort to establish an international XML-based standard for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioral sciences. <http://www.icpsr.umich.edu/DDI/codebook/index.html>

For integration purposes (see previous section) users need to know information on scale and resolution, distribution and the form of data (point, line or surface, vector or raster). Information on Quality Assurance is a pre-requisite for data re-use and sharing, and must be addressed if the full value of datasets is to be realised. Many metadata do not tell the history of the data, nor allow the user to assess confidence (i.e. provide information on variability and/or uncertainty inherent in the data). Information such as confidence intervals, standard errors and uncertainty should be included within metadata. The EU Fifth Framework HarmoniRiB project is addressing the assessment of uncertainty of data, and the metadata elements needed.²³

A number of metadata standards exist (see Annex E for a review of some key standards). Research programme managers have generally selected relevant fields from published standards, rather than adopting one particular standard. There seems also to be a disciplinary divide in the metadata adopted. It would assist interdisciplinary working if environmental and social researchers could agree on the metadata elements to be used.

Data archiving is addressed by both NERC and ESRC, with both services and enforcement encouraging award holders to deposit data at the end of projects. However, researcher motivation remains low, and NERC enforcement does not seem to work particularly well, and is under revision. There need to be incentives for researchers to deposit data. Currently, incentives in educational institutions centre on publications and researchers are not rewarded for archiving or collection of data. Government institutions such as Defra have a poor record for archiving, often use inappropriate formats and tend to 'tidy-up' data after its immediate use is over.

Internationally, the CODATA²⁴ working group is addressing preservation and archiving of scientific data worldwide. The group will deliver a position paper documenting the diversity of best practices and identify the 'best' ones in the area of data archiving and preservation across the science domain. The OECD is developing a handbook on how to present data and metadata²⁵.

Data provision (including management and storage of data) is under-resourced, in both Research Councils and UK government. Data integration usually requires communication and advice from the provider, not just the raw data. It can also involve additional manipulation and integration tasks. Distributed databases and grid technology will enable direct access to data for users and so should reduce the pressure on providers.

Other technological developments and trends relate to data management models. The increase in the volume of data appears to have outstripped hardware available to deal

²³ HarmoniRiB is a research project supported by the European Commission under the Fifth Framework Programme and contributing to the implementation of the Key Action "Sustainable Management and Quality of Water" within the Energy, Environment and Sustainable Development Programme. Contract no: EVK1-CT-2002-00109. <http://www.harmonirib.com/>

The two major expected outputs of the project are (1) a set of methodologies, tools and case studies for assessment of uncertainties and for integration of uncertainties and socio-economic factors into preparation of river basin management plans, and (2) the infrastructure and the datasets containing uncertainties from the network of representative river basins.

²⁴ <http://www.codata.org/> The International Council for Science Committee on Data for Science and Technology.

²⁵ http://www.oecd.org/document/60/0,2340,en_2649_33715_30388604_1_1_1_1.00.html

with it. In academia, this is being temporarily addressed by grid technology, however this has yet to filter down to government and the public. There is a trend to greater data provision over the internet, web based tools (e.g. SPSS on-line facility and e-science tools) and towards greater computer literacy of data users.

Recommendations

Lessons learnt from previous Research Council Programmes have been well documented, and ‘the solution’ appears clear:

- data management must be adequately resourced, both by the Research Councils and by Government
- data management needs to begin before the projects begin, and needs to start with applicants rather than award-holders
- Successful data management requires a certain level of stakeholder engagement, service provision and training as well as effective enforcement.
- Successful data management for interdisciplinary research requires facilitation of communication between the disciplines and possibly specific training

The Research Councils need to address researcher motivation with respect to data management.

The Research Councils can learn from each other and develop best practice in data management. This is particularly important for interdisciplinary programmes, where dataset discovery and metadata availability are holding back progress.

The Government needs to ensure organisations commissioned to collect data are adequately resourced for: (a) data provision and (b) implementing changes to legislation.

Conclusion

Data management is an increasingly important aspect of Research Council programmes, and of interdisciplinary research in particular. Successful data acquisition, maintenance, provision and archiving require not only adequate resources (time and financial), but more importantly, a political will. For these reasons, we as need to take on a coordinated strategy, and continue to move to an organisation-based approach.

Acknowledgements

The project was funded under the Rural Economy and Land Use (RELU) programme, a joint Research Councils Programme co-sponsored by Defra and SEERAD. We are grateful to the staff of RELU Programme Director’s office and the Data Support Service for their help and advice throughout the project.

Thanks to everyone who contributed to this report. In particular to: Louise Corti, Mustafa Ahmet, Phil Holden, John Watkins and Isabella Tindall of the RELU Data Support Service, Meg Huby, Anne Young, Steve Cinderby, Piran White and Colin McClean of York University for useful discussion and involvement in the Data Integration . Thanks also to: Bill Froggatt, Dav Stott, Mark Thorley, Richard Baker, Richard Budgey, Ruth Swetenham, Steve Langton, James Aegerter, Gavin Parker, Matt Loble, Nick Evans and Susanne Seymore for acting as Consultees. And finally, Mark Birkin and Miles Templeton for stimulating presentations at the data integration workshop, as well as all involved in the breakout sessions.

Table 5. Analysis of the most important issues for GIS users.

Numbers within the table refer to the ranking each consultee gave each issue. The score is calculated from the rankings (3 for first, 2 for second and 3 for third or more).

		Issue									
		Distribution of data	Nature of data (point, line, area, vector, raster)	Resolution/Scale	Data availability/access	Boundary changes	Data format	Volume of data	Security	Metadata/standards	Archiving
Consultee	1 Anne Owen	1	2	3							
	2 Bill Frogatt			3	1	4				2	
	3 Colin McClean			2	1		3				
	4 Dav Stott				2		3	1			
	5 Richard Baker		2	1						3	
	6 Richard Budgey		2		3					1	
	7 Ruth Swetenham			1	3						
	8 Steve Langton			1	2					3	
	9 James Aegerter	3	2			1					
Ranking	first	1	0	3	2	1	0	1	0	1	0
	second	2	4	1	2	0	0	0	0	1	0
	third	3	0	2	2	0	2	0	0	2	0
Score	Total	4	8	13	12	4	2	3	0	7	0