



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

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

Annex D.

**Data integration workshop
held on 19 May, 2005 at King's Manor, York.**

**Helen M^cKay and James Aegerter, Central Science Laboratory, Sand Hutton,
York YO41 1LZ, U.K.
h.mckay@csl.gov.uk**

23 May, 2005.



Annex D. Data Integration Workshop

Contents

Attendance	1
Programme.....	1
Discussion topics for breakout sessions.....	2
Outputs from breakout sessions	2

Attendance

Approximately 30 delegates attended, from a wide range of disciplines (economists, biologists, social and environmental scientists and geographers) and institutions (including Defra statistics, Sport England and the Countryside Agency), data specialists (from Forest Research and EDINA) and the RELU office.

Programme

10:00-10:15	Registration and coffee
10:15-10:20	Welcome/introduction <i>Nigel Boatman, CSL</i>
10:20-10:50	E-social science <i>Mark Birkin, NCeSS</i>
10:50-11:20	Rural Data Hub <i>Miles Templeton, Defra</i>
11:20-11:50	Case Study 1: Developing Spatial Data for the Classification of Rural Areas According to Socio-Economic and Environmental Sustainability Factors <i>Meg Huby/Anne Owen/Steve Cinderby, University of York</i>
11:50-12:20	Case Study 2: Integrating Spatial Data on the Rural Economy, Land Use and Biodiversity <i>Colin Maclean/Piran White, University of York</i>
12:20-12:40	RELU Data Support Service <i>Isabella Tindall, DSS</i>
12:40-13:40	Lunch
13:40-14:10	Questionnaire survey: Data Resources for Rural Sustainability Research: Realising Their Combined Potential <i>Helen McKay, CSL</i>
14:10-15:10	Breakout Sessions <i>Nigel Boatman, CSL</i>
15:10-15:25	Tea/Coffee
15:25-16:10	Report back and discussion
16:10-16:15	Concluding remarks <i>Nigel Boatman, CSL</i>

Discussion topics for breakout sessions

The aim of the breakout sessions was to tease out and explore differences between disciplines in their approaches to data management and integration, and how these differences can be resolved to enhance interdisciplinary working.

The discussion was structured around the following issues. Other issues were introduced if relevant, but we asked group chairs and rapporteurs to try to capture the views of the group on each of the issues listed below.

1. What do you regard as 'data' (list some examples linked to the discipline of the contributor)
2. What do you understand by data quality? How important is it and how is it assessed?
3. How much effort is made in your area to measure/quantify errors and uncertainties in the data? Are current practices adequate or should they be improved?
4. How is your data stored (e.g. spreadsheet, database, hard copy etc). At what stage in projects is method of storage and retrieval, metadata and archiving method determined? Are improvements needed?
5. What are the major technical issues in data integration? Do you believe the implications of these are sufficiently well understood to interpret the resulting output? If not, how can this be addressed?
6. What technical developments and trends in data management and integration do you foresee becoming important in the next few years?
7. Are your data needs being adequately addressed by the RELU Data Support Service? If not, what additional services would you like to see in this and similar research programmes?
8. There were fewer respondents from social scientists than from natural scientists to our questionnaire survey. Why do you think this occurred?
9. The answers to one of the questions in our survey suggested that respondents anticipated doing less integration between natural science and socio-economic data in the future than at present. Do you believe that the need for this type of research will grow or decline in the future?

Outputs from breakout sessions

GROUP A

- Group A was composed of mainly environmental scientists who used quantitative data, but some social scientists who used qualitative data were also present.
- **What are data?** spatial and/or temporally structured quantitative or qualitative and context-specific pieces of information.

Annex D. Data Integration Workshop

- **Data collection:** Social scientists tend to spend some time formulating the problem and then collect data. Environmental scientists often use previously collected data. Different disciplines view data differently, and this has implications for the metadata collected and made available
- **How do you find data?** Web links, data portals, portal of portals. Tight disciplinary division e.g. GI data via GI gateway, social science data through ESDS DA. Discipline-specific discovery meta-data.
- **Social scientists' use of data:** Social scientists who use qualitative data often don't look for, or need, data collected by others (different context, different questions). Social scientists may under-use qualitative data sources. Often unaware of data and availability of data. Methodological processes can be obscured in context and approach. Many could make better use of a greater range of data.
- **RELU data requests.** 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 be useful to alert award holders to those datasets for which the DSS could help with access, to avoid duplication of effort.
- **Data quality:** without gaps (complete), includes issues of scale and definition. Issues of interpretation at inappropriate resolution – how can this be avoided? These are often due to data collection issues and information being unavailable to the eventual users of the data. End users should not accept data at face value e.g. issues of uncertainty. This is certainly true of large data sets – it is often assumed that they are of high quality, without supporting evidence. Quality assurance is often the first casualty when resources are limited.
- **Quality of data collection.** Important for subsequent user to know how data collected. Good metadata important. It is particularly important for social scientists where data is context-specific. Anonymisation can subsequently reduce value and may partly explain why a low proportion of social science data is archived, despite high submission rates. It would be useful to compare retrieval rates for different types of data.
- **Data integration.** Problems with divergent scales (effectively eroding the quality of the data at the highest resolution, as you always ought to work at the resolution of the coarsest data set). Includes not only temporal (time series) and spatial integration, but conceptual integration (data can be mapped together to reach a common interpretation rather than formally mathematically combined).
- **Technical issues.** Unintentional abuse of spatial data, difficulties in finding the appropriate data. When deciding on metadata for archiving, naming conventions and common units are important. Devolution is therefore a potential threat. Even basic lexicography can confound non-specialists. Provision of metadata useful to inter-disciplinary studies can be challenging.

Annex D. Data Integration Workshop

- **Training.** RELU could undertake training to encourage an improved understanding of other disciplinary fields, methodologies and terminologies, and thus help to foster interdisciplinarity.

GROUP B

- **Group composition:** Group B involved: an economist, data manager (forestry), geographers, an environmental scientist, PhD researcher, recreation, social policy research, a few modellers, human processes/biodiversity, a few statisticians.
- **Variety of data:** the starting point is the processes being considered/ formulation of the problem, units of analyses (spatial vs. non-geographically referenced data), information from an interview. Vertical/horizontal analogy is useful (human dimension vs. spatial dimension).
- **Definition of data:** observations on a process/information, evidence, explanatory (to answer a question), data are relevant (collected for a purpose), data have a certain level of quality, are descriptive, are a resource, are multifaceted (text, photos, video)
- **Categories of data:** structural (what form, how collected) vs. functional (what you want them for, how you could use them, models)
- **Quality:** suitability, appropriateness (inc. scale issue), reliability. Regarding qualitative data, analytical techniques have been well documented. Assessment of quality: reliability, validity and coverage.
- **Data collection plan:** not always relevant e.g. if data have already been collected. Often plans are informal, varying level of thought applied, have been cavalier in the past.
- **RELU framework:** novel. But, can be viewed as another bit of bureaucracy, an extra form to fill in. Purpose unclear. DSS likened to Social Security.
- **Technical issues:** non-geo-referenced data. Temporal frameworks can differ between datasets as well as spatial units.
- **Technical developments:** grid, e-science
- **Are the DSS addressing needs?** No. Timing (too late), communication and feedback (lack of). Other programmes have done it better.