Keeping the lights on

Important information may be lost after funding for maintaining research-data resources runs out. **Craig Nicholson** hunts for long-term solutions.

When Lora Fleming, an epidemiologist at the University of Exeter in the UK, applied for funding for a project to link environmental and human health data called Medmi, she didn't think much about what would happen after the funding ended.

"I think I had a naive idea that somehow, if Medmi was seen as something really positive, the research councils and research users would keep it going," she says. "I think I had no idea how much resource it takes to do that."

Fleming and her colleagues won a total of about £1 million (€1.1m) from two UK public funders to combine existing datasets and make them available through a secure analysis and visualisation platform. When the three-year project ended in 2016, about 30 researchers were signed up to access the data. When she spoke to Research Europe, the number of users had increased to 80.

Two of the project partners—the Met Office, which provides the UK's weather service, and the government agency Public Health England—are "trying to keep Medmi going", Fleming says. Her colleague Cristophe Sarran, a researcher at the Met Office, is still refreshing the data and responding to queries from researchers. But there's "very little funding" for those efforts, she says.

Medmi exemplifies a problem with data management that affects numerous other projects. Another researcher at the University of Exeter, data scientist Sabina Leonelli, covered the issue in a report in 2017.

"Many research projects in big data and human health are typically set up at most for five years, with no possibility to extend funding further so as to maintain and update the datasets and related infrastructure," the report said. "When the funding ends, access to data deteriorates and is sometimes lost entirely, leading to a loss of knowledge resources."

The UK's Medical Research Council, Medmi's main funder, didn't want to comment on that individual case. Fleming suggests the MRC's view would be that her group should have submitted more grant proposals. "We did and weren't successful, so that's on us," she says.

But she adds that in her opinion the MRC and the Natural Environment Research Council, the other funder, "didn't take any responsibility for or ownership of the project". And extra project grants "wouldn't have addressed the long-term issues around data infrastructure", she says.

The MRC was happy to talk about its data-management work more generally. Rachel Knowles, the council's programme manager for clinical sciences, wrote its open-data policy. She says that all researchers who apply for funding are asked to submit a data-management plan. The MRC is "particularly interested" in ensuring that the value of the data generated by its funding is maximised through sharing and reuse, she says.

But the MRC and other funders are "really uncertain" about which data should be kept for reuse in the long term, Knowles says. It is costly to keep data up to date, and budgets for this are limited.

In January, the MRC, along with the research charities the Wellcome Trust, Cancer Research UK and the Bill and Melinda Gates Foundation, joined Clinical Study Data Request, a website set up by drug companies to share clinical data. One of their aims, Knowles says, is to better understand which data researchers are interested in.

Louise Corti is associate director of the UK Data Archive, a department of the University of Essex that gets public funding to maintain research datasets from across the UK. It does short and long-term archiving, and Corti says it is the latter that comes with high costs and requires decisions on what is worth keeping. "If you're going to curate things for the next 100 or 200 years, you need to migrate the dataset formats forward," she says. "That's very expensive to do."

Deciding what to keep is "not as clear cut as you might imagine", she says. Data from expensive large studies and long-term studies with a track record of usefulness are generally preserved, but predicting which smaller studies might be useful decades later is trickier.

The UK Data Archive has criteria to help it decide what to preserve, including that the data "must be high quality, authoritative and reliable". The data should also be "important resources for current research purposes, meet new demands in research, or supplement areas of the collection that the service seeks to expand".

Another source of data loss, Corti says, is research projects—such as Medmi—that set up their own data resources, rather than depositing their data in a repository with a high likelihood of long-term support. "When the principal investigator leaves, often that stuff gets stuck and has to be resurrected," she says. These kinds of projects leave a legacy when their funding runs out, she warns. "The earlier that can be integrated into a workable model, the better."

Something to add? Email comment@ResearchResearch.com

'Deciding what data to keep is not as clear cut as you might imagine.'