Data mashups can help answer the world's biggest questions

April 25 2014, by Lora Fleming

We collect climate data. We collect health data. What if we combined the two? Credit: Kevin, CC BY-SA

As the world wakes up to the power of data, we need to start working out how to join up all this information. We need to turn it into meaningful findings that will help us to make changes to the way we live. A new technique is emerging as part of this quest – the data mashup. This approach to linking data could help us shed light on phenomena such as the health impacts of climate change.

Data comes in all shapes and sizes. It can record boundless sets of characteristics over different time scales and geographic areas. But this diversity means that individual databases are often created for specific areas, such as health research, and are rarely shared or combined with others.

Yet it is becoming increasingly apparent that by joining these disparate sources of data, academia, governments and businesses may be able to access information that is currently hidden within closed systems. So researchers are now turning to techniques developed by computer scientists in order to access this Aladdin's cave of information.

The Medical and Environmental Data Mashup Infrastructure (MEDMI) project is one of the initiatives doing this. We're hoping to enable research into the links between climate, weather, environment and health. By bringing databases from each of these areas together and allowing access through one web-based portal, we're aiming to create a shared resource for medical, environmental, and public health researchers.

The collection of health and environment data over the past 20 years has provided a growing resource of information. It includes detailed monitoring of weather and climate variables like temperature and rainfall and digital health records, among other useful additions.

With this information, you could combine temperature and air pollution data to predict when people with chronic lung disease might have respiratory problems if they go outside. The UK Met Office did this and now provides an early warning service for patients, their families and healthcare providers.

But joining such varied forms of data presents some significant hurdles. For a start, in many cases we are at the mercy of the way data has been

collected historically. Pollen data, for example, traditionally suffers from a lack of resolution. Only a few measurement locations cover the whole of the UK but pollen moves rapidly in the air all over the place. These differences in resolution over time and space make it difficult to identify links with other more finely recorded factors, such as individuals with certain types of skin cancer and radon levels in a [particular area](#).

The huge disparities in data collection are even more apparent when considering other environment and health variables. Take rainfall or cloud cover data for example, which are measured on an hourly basis, at very high resolution, over the whole of the UK.

It might be interesting to combine large scale environmental data with the [Avon Longitudinal Study of Parents and Children](#), which followed the pregnancies of 14,000 mothers in the Avon Valley, to see if solar irradiance (a measure of vitamin D levels) exposure is related to the [development of allergic diseases](#). But this is difficult to do because the Avon study only collects data every couple of years and the participants predominantly live in a small geographic area.

Bridging the gaps

Merging data types ranging from a description of a person's mental health to measurements of ocean currents, requires some serious head scratching. Fortunately, statistical techniques and methods such as Geographic Information Systems (GIS) provide us with a really good start, and the standardisation of spatial data services by the [Open Geospatial Consortium](#) has begun to create a common international language between databases. There is also a growing interest from the private sector, with companies like Google dedicating resources to connecting data and enabling access over the web.

Perversely (to health researchers at least), the link between the changing

climate and human health has received little scientific attention, particularly when compared to investigating how climate affects the weather and environment. We're hoping that MEDMI will begin to redress this trend by allowing us to investigate where climate and health data overlap.

For example, we want to identify risk hot spots – places where climate and other environmental factors converge to affect vulnerable populations – early enough to both mitigate the consequences and study these interventions.

The sheer number of partners working on the project highlights the dizzying complexity of any mashup endeavour. And of course there is the veritable minefield of protecting confidential and sensitive health data. The importance of that cannot be overstated.

But scientists like me, already committed to the data mashup cause, aren't fazed by these challenges. We're already looking towards a future where these linked databases can be queried in real time.

We're imagining a world where a regional cold snap can be associated with flu cases and hospital admissions as it happens. That would mean local resources could be quickly and efficiently deployed. We're hoping that long-term predictions about climate and [human health](#) hot spots can help us to plan our cities so that they are more resilient.

Living in a world undergoing rapid environmental change will increasingly require this kind of vision. We're not there yet, not even close, but just like television on your mobile phone, we may get there sooner than you think.

This story is published courtesy of [The Conversation](#) *(under Creative Commons-Attribution/No derivatives).*