Open data in studies of the water-energy-food nexus

Simon Lambert, Vasily Bunakov, Scott J. McGrane, and E. Marian Scott

Published version information

Citation: S Lambert et al. "Open data in studies of the water-energy-food nexus." In: Benoît Otjacques, Patrik Hitzelberger, Stefan Naumann, Volker Wohlgemuth (Eds.) **From Science to Society: The Bridge provided by Environmental Informatics**. Adjunct Proceedings of the 31st EnviroInfo conference. Shaker (2017): 221-226.

https://www.shaker.eu/en/content/catalogue/index.asp?lang=en&ID=8&ISBN=978-3-8440-5495-8

Open data in studies of the water–energy–food nexus

Simon Lambert[1], Vasily Bunakov[2], Scott J. McGrane[3], E. Marian Scott[4]

1. Introduction

Since the 2011 Bonn Conference, there has been a growing recognition that our critical water, energy and food (WEF) resources are interlinked, and that strong interdependencies occur across the three sectors [1]. Yet studies which seek to map WEF resources and flows in particular geographies, sectors or scenarios have been slow to emerge. An overarching aim of nexus research has been the eradication of silo mentality and vertical management of the WEF sectors. However, the vast majority of nexus studies have not yet freed themselves from the pitfalls of undertaking analyses through a specific sectoral lens. In this paper, we examine some of the issues confronting quantitative studies of the nexus, opening the debate around open data in this sphere.

Technological advances have significantly enhanced our capabilities to record, store and analyse data at particularly fine resolutions, but despite the increased capacity for data acquisition, difficulties remain for researchers working within WEF nexus spheres. At the national scale, data can provide a robust overview of resource stocks and consumption, but without modelling approaches (e.g. downscaling) do not enable researchers to drill down to more local geographies or processes. In countries with complex administrations such as the United Kingdom—with its devolved governments and civil service bodies—further problems arise with regard to provenance and comparability. For example, the Environment Agency is responsible for generating and reporting environmental data across England, while the Scottish Environment Protection Agency and Natural Resources Wales undertake the same function for Scotland and Wales respectively. Often monitoring regimes are different, with variance in purpose, instrumentation, units and even monitoring frequencies. As a result, comparisons across areas are often not possible without some considerable data pre-processing and post-modelling. At the opposite end of the scale, technology that facilitates the monitoring of WEF resource flows at increasingly local levels (e.g., household, institution, building) has enabled the acquisition of high-resolution data. However, much of these data remain unavailable to external users, as issues of corporate and customer sensitivity preclude release of these data to third party users.

In Open Science the ideal of FAIR data (Findable, Accessible, Interoperable, Reusable) is gaining traction [2]. With reference to nexus studies, all of these desirable ideals are to some degree lacking. In

[1] Science and technology Facilities Council, UK, simon.lambert@stfc.ac.uk

[2] Science and Technology Facilities Council, UK, vasily.bunakov@stfc.ac.uk

[3] School of Mathematics & Statistics, University of Glasgow, Glasgow, UK, scott.mcgrane@glasgow.ac.uk

[4] School of Mathematics & Statistics, University of Glasgow, Glasgow, UK, marian.scott@glasgow.ac.uk

this paper we will examine the reasons for that, and propose a perspective on data that offers a principled approach to data reuse.

2. Problems of data in the WEF nexus: open data

Critical to nexus studies is the availability and accessibility of data sources across the WEF sectors. Data in nexus studies can be classified into three types: i) data that are collected via monitoring, ii) data that are collated into open access databases (e.g., such as government or transnational institutions (e.g., EuroStat, EEA, United Nations)), and iii) private, often 'closed' data, that require payment for access, or preclude access in its entirety due to commercial sensitivity. Within the sphere of WEF nexus research, often all three data types are necessary to quantify flows across the three sectors at contrasting spatial scales. Yet data access itself is often a barrier to the successful implementation of nexus assessments, especially at more local scales where closed data dominate.

Open data sources can be categorised as either centrally derived (i.e., central government or transnational bodies such as the European Union or United Nations), or data that are available from other research projects. Large volumes of data across a sphere of sectors are collected by national governments every year in the preparation for national accountancy exercises, providing insight into resource availability, flows of resources, economic performance and international trade data (to name but a few). These data are ordinarily spatially and temporally coarse (national, or large regions, annual), but are useful in providing an overview of how countries are performing. However, care must be taken in undertaking comparative assessments, as variations in units, monitoring strategies in energy and food production systems can result in incomparable data. In some cases, results from other research projects can also provide a readily minable source of data. However, this is subject to the data being published or made accessible, as many projects will produce or collate data that are subject to embargoes or non-disclosure agreements that can restrict or prohibit reuse of particular datasets.

Standardisation initiatives aim to address some of these problems. The INSPIRE Directive [3] is aimed at a spatial data infrastructure for the EU, facilitating the sharing of environmental spatial information. One of its principles is that it should be possible to combine seamless spatial information from different sources and share it with many users and applications, and for information collected at one level/scale to be shared with all levels/scales. This is certainly of great value, but nexus-related data is more diverse than that envisaged by INSPIRE and much falls outside its scope of environmental concerns.

While data generated from other research may provide some utility, individual data management plans may mean that the underpinning data are unavailable to other research groups. For example, Tassou et al. [4] highlighted the usage of energy in cold storage and supply chains across the supermarket industry. The data behind this study were protected by a non-disclosure agreement and were subsequently unavailable to other research projects. Furthermore, many other research publications and outputs are very case or geography specific, resulting in a lack of certainty about transferability.

There have been efforts, particularly for data that is perceived to be of high reuse value, to develop services for unifying disparate data sources. A recent example [5] provides the UK research community

with fully processed quasi-realistic GIS maps of spatial population distribution, interpolated over census area statistics.

3. Problems of data in the WEF nexus: closed data

A major barrier to nexus research is the prevalence of 'closed data', which are data that are owned by private business, exist behind a paywall or are deemed commercially sensitive, and are therefore unavailable to nexus researchers. There are several key examples of where closed data may provide crucial insight into nexus practices, but remain unavailable in the public domain. The Farm Business Survey (FBS) collates detailed responses from every farm in England, highlighting their business setup, resource use and expenditure. At the most detailed level, the FBS can identify every single farm within England, and as a result, access to it is strictly protected. Yet agricultural landscapes are often a focal point of nexus research, as they represent areas where all three sectors come together on a daily basis, using water and energy to produce food crops, and often diversifying to produce biofuel crops or generating renewable energy via on-farm renewables. There is no such contemporaneous assessment undertaken for farms in Scotland, Wales or Northern Ireland, which makes comparisons of farm practice across the UK near impossible. In some instances, data from other research projects or from research institutes can also be 'closed', which is to the detriment of continual advancement of our understanding of the nexus, and more widely, our capability of tackling grand societal issues. An example of this is the Crop Map that is produced by the Centre for Ecology and Hydrology (a NERC funded research institute in the United Kingdom). This GIS map, produced in conjunction with a consultancy firm, details cropland across England (once again, this dataset does not extend to Scotland, Wales or Northern Ireland) at a 2 ha resolution. While this dataset is considerable in its size and detail, the cost of accessing the entire map at the scale of England is in the order of £35,000, and therefore prohibits most researchers being able to utilise this resource. While it is possible to request smaller areas and pay for those accordingly, the ability to interpret some of these data at even a county or regional level comes with a considerable financial cost attached.

Much of the data that are collected and protected by private firms or public utilities are very local scale data, often at the household or institutional level. For example, utility companies that provide energy and water to households are increasingly able to monitor usage to a detailed level, including individual appliances. While water meters lag behind the progress observed in gas and electricity meters, new households are increasingly being fitted with devices that enable monitoring. Companies that collect these data are often responsible for maintaining the security of their customer information, and access to such data is often intensely protected. Supermarkets represent a meeting point of producers and consumers, and their infrastructure is often both water and energy intensive. Energy is used for lighting, cooking and refrigeration of food produce, whilst water is used to maintain hygiene and cleanliness standards across estates. The emergence of new monitoring technologies and digital infrastructure allows most supermarkets to monitor their utility usage in real-time, enabling them to derive strategic plans to reduce their resource consumption and increase their sustainability footprints. While summary data (e.g., total carbon footprint, water usage or energy consumption) are often published in annual stakeholder

reports, more detailed data remain unavailable beyond these companies, who store and analyse their own data accordingly.

Engaging businesses in the nexus process is crucial for advancing the sphere of research. Highlighting the benefits and incentives of becoming involved with such research is essential to enable full collaboration between industrial partners and academics. As mentioned with regard to supermarkets, many private companies employ their own analysts and planning managers to oversee their sustainability footprints and highlight areas where they, or their customers, can make savings in terms of both finance and resources. An important challenge is communicating with private business, elucidating how their involvement with academic research projects can subsequently assist their daily practices.

4. A framework for the role of datasets in understanding the WEF nexus

From the foregoing discussion we have established the importance of data in understanding the WEF nexus, and that there are a great many possible sources of data, whose use is hampered for various reasons, whether related to the accessibility of the data itself or its suitability for reuse. We next propose a framework for thinking about and reasoning about the role of datasets. The starting point is to examine in what ways a dataset might be able to contribute to understanding the nexus. That is, to adopt a 'data-centric' view in which datasets are considered primary, and the question is whether and how they may be utilized in a nexus study. By the very nature of the nexus, connections or relationships are key, so we may think not simply of what a particular dataset measures (for example, abstraction of water from rivers) but of how it can shed light on connections. These connections are not necessarily formulated within a "model" (of any degree of formality), though they might be.

Arising from the work of the WEFWEBs project [9], three general types of relationship have been identified, in terms of the role that is played by data in their elucidation.

- Quantifying, calibrating or scoping a relationship
  Example: Sankey diagrams. Sankey diagrams are a type of flow diagram, allowing representation of the flow of conserved quantities within a system. For the WEF nexus, Sankey diagrams may be derived by informed analysis of the elements of the system and the flows between them, but the quantification of the flows obviously requires data as its origin.
- Providing evidence for a relationship
  Example: informal 'influence relations' arising from stakeholder workshops, where it may be desired to test the proposed relations, determine any limits on their validity, etc.
- Providing input to a relationship
  Example: hydrological models of flooding, requiring very specific data on terrain profile and land cover in order to run the models.

Having distinguished these types of contribution, we can then ask what are the necessary conditions for a dataset to be able to shed light on such relationships. In the case of "providing inputs", the dataset must relate to the appropriate quantity (terrain altimetry for example), have suitable geographical and temporal scope, precision, and units of measurement. Similar sets of conditions apply in the other cases,

though for example for “providing evidence” or “scoping” there will be a need to have a scope that allows falsification of the relationship.

At this point it might seem that all that can be said is that there is a need for adequate metadata associated with datasets to allow the checking of the necessary conditions, a judgement (whether by human or automatically) of whether the conditions are met. This is indeed an interesting issue in general, but it is possible to go further. First, on this basis it is possible to conduct a gap analysis relative to the particular relationship under consideration, based on whether or not there are accessible datasets that meet the necessary conditions.

However, we can also ask what would have to change in order to meet the necessary conditions. If we cannot find a dataset that precisely matches the conditions for our particular area of study, are there other datasets that can be adopted or adapted? There are two ways this could be possible: by finding a dataset with a different scope, that can nonetheless be mapped, possibly with caveats, to the desired scope; or finding a dataset that can serve as a proxy for the required quantity. These can be seen as an attribute of *transferability* of a dataset. It is clear that some types of data are more transferable than others. A terrain model for a particular area, used as input for hydrological modelling, is completely specific to that geographical area and would be useless for hydrological modelling of any other area. But data on food consumption by the population of a particular area A within country C, or indeed aggregated for country C as a whole, might well be usable as an adequate substitute for another area B. At least it would be worthy of consideration for that purpose, in the absence of the specific dataset relating to area B. The relevance may be subject to critique, and the assumption should be made explicit.

Therefore thinking in terms of necessary conditions allows not only identification of gaps in data but some reasoning about what is needed to fill the gaps, where assumptions or estimates could be used, what other data sources might be applicable, …—in general, seeking a dataset that allows transfer, interpolation or extrapolation so as to satisfy the conditions, while making explicit that this is what has been done and opening it up for critique.

This line of reasoning is leading us in the direction of *provenance*. Provenance is a concept that arises in a number of contexts. In long-term digital preservation, it refers to the steps of custody or of processing through which a (digital) object has passed [6], and it provides evidence of the authenticity of that object. In scientific workflows, provenance describes the steps in processing scientific data [7] and contributes to reproducibility of scientific research. In the WEF nexus context, provenance would capture the operations applied to datasets to render them suitable for use in the context of the study—operations such as interpolation, geographical transfer (as defined above) and proxy substitution.

Capturing and presenting provenance allows critiquing of nexus studies, an important aspect of openness in science. But there is more than this: the nexus researcher seeking data to bolster their studies should be able to search for prior work that has already utilized the kind of transfer operations mentioned above: that is, to search for provenance patterns to help fill the gaps in their own data landscape. We can see this as the complement of the necessary conditions on data usability: *sufficient* conditions for data reuse based on what has been done previously. Taken literally, there cannot really be sufficient conditions: the usefulness of a dataset will always be subject to interpretation in the context of the study being undertaken. However, if ‘sufficient’ is taken to mean ‘sufficient for consideration’—just as ‘necessary’ really signifies ‘necessary for consideration’—then there is an opportunity, if someone else

has used a dataset for a similar purpose, adapting it in a particular way that can be reused in the new context.

In scientific workflows, two types of provenance have been identified [8]:

- Retrospective: execution traces, what was done to the data
- Prospective: workflow structure, what will be done with the data

In the context of nexus studies, a third type arises:

- Potential: what could be done with the data—that is, how it may be reused outside its original context based on the pattern of operations that made it applicable to that context

5. Conclusions

We have discussed the shortcomings of open and closed data in studies of the WEF nexus. By adopting a data-centric view, and thinking in terms of necessary conditions for applicability of datasets to nexus relationships, it becomes possible to perform a gap analysis, and, beyond that, reason about how alternative data sources might be used to substitute and fill the gaps. Ideas of provenance are applicable to capture what has been done to datasets and therefore what may be done with them. Reproducibility and open critique will be fostered—two of the pillars of Open Science. But not only this lofty ideal is the aim, but more practically the ability to open up structured data behind nexus models and relationships for sharing and reuse.

A number of lines for further investigation have been raised in the course of presenting the framework for the role of data in understanding the nexus:

- What can be said in general about the metadata to be associated with a dataset to allow reasoning about the necessary conditions for its applicability to nexus relationships?
- What types of transferability of datasets are possible, and how may they be represented?
- Are existing provenance models and systems suitable for capturing and and querying the provenance of datasets in nexus studies?
- How can the idea of potential provenance be formalized to enhance data reuse in nexy=us studies?

References

[1] Hoff, H., 2011: Understanding the Nexus. Background paper for the Bonn2011 Nexus Conference. In *The Water, Energy and Food Security Nexus*. Stockholm Environment Institute, Stockholm., pp. 1–52.

[2] The FAIR Data Principles. https://www.force11.org/group/fairgroup/fairprinciples

[3] INSPIRE Infrastructure for Spatial Information in Europe. https://inspire.ec.europa.eu/

[4] Tassou, S.A., Ge, Y., Hadawey, A. and Marriott, D., (2011), Energy consumption and conservation in food retailing, *Applied Thermal Engineering*, **31**(2-3), 147-156

[5] Shi, S., Walford, N. (2010): An automated internet geoinformation service for integrating online geoinformation services and generating quasi-realistic spatial population GIS maps. ISPRS Archives Volume XXXVIII Part 2.

[6] Reference Model for an Open Archival Information System. CCSDS Recommended Practice. https://public.ccsds.org/pubs/650x0m2.pdf

[7] ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance (2014). http://vcvcomputing.com/provone/provone.html

[8] Lim, C. et al. (2010), Prospective and retrospective provenance collection in scientific workflow environments. IEEE International Conference on Services Computing. http://ieeexplore.ieee.org/abstract/document/5557202/

[9] WEFWEBs project. http://www.gla.ac.uk/research/az/wefwebs/

Acknowledgements: The work is supported by WEFWEBs project [9] sponsored by the UK Engineering and Physical Sciences Research Council (EPSRC) and Science and Technology Facilities Council (STFC). The views expressed are those of the authors and not necessarily of the project.