JULY 2016

Options for a New Integrated Natural Resource Monitoring Framework for Wales

Project Document

Briefing note: The Requirements and Opportunities for Data and Informatics in an Integrated National Monitoring Programme

Welsh Government Contract No. C147/2010/2011

Agreed Additional Work Requirement Dated 8th March 2016

NERC CEH Project: NEC05945

How to cite this document: Henrys, P., Chadwick, D., Chapman, C., Guest, P., Letheren, B., Neil, S., Siriwardena, G. (2016) **Options for a New Integrated Natural Resource Monitoring Framework for Wales; Phase 1, Project Document - Briefing note: Requirements and Opportunities for Data and Informatics in an Integrated National Monitoring Programme;** Report to Welsh Government (Contract reference: C147/2010/11; Agreed Additional Work Requirement Dated 8th March 2016). NERC/Centre for Ecology & Hydrology (NERC CEH Project: NEC05945)

Options for a New Integrated Natural Resources Monitoring Framework for Wales

Project Document - Briefing note:

Requirements and Opportunities for Data and Informatics in an Integrated National Monitoring Programme

Peter Henrys (CEH)

Colin Chapman (WG)

Stuart Neil (WG)

Paul Guest (WG)

Barnaby Letheren (NRW)

David Chadwick (Bangor University)

Gavin Siriwardena (BTO)

July 2016

Intentionally blank

Briefing note:
Requirements and Opportunities for Data and Informatics in an Integrated National Monitoring Programme

For any future monitoring programme it is important to have a clear and comprehensive strategy to steer the collection, management, use and dissemination of its data, information and evidence.

Rationale

The Welsh Government recognises the value of the data and information collected and used within environmental monitoring programmes. Increasing the accessibility of data and evidence and encouraging its re-use can lead to improvements in efficiency, promote transparency and dialogue and raise the level of quality of a shared environmental evidence base. This can also lead to novel uses and give individuals and organisations the ability to access the data and combine it with other data sources in new ways. The Welsh Government's Open Data Plan sets out its commitment to publish data, where it is appropriate to do so, under the Open Government Licence (OGL) and make it accessible to as wide an audience as possible without restriction. In addition to this, it is acknowledged that to fully understand and report on key aspects of the environment, there is a growing need for integrated data analysis. These key issues of openness of data and data integration highlight the need for a formalised approach to data and informatics within any future natural resources monitoring programme.

The complexity and high data requirement of a Natural Resource Management monitoring programme requires a clear strategy sensitive to the diverse nature of the content, quality and ownership of datasets in order to maximise appropriate usage and exploit the possibilities of data sharing and integration. A number of key themes must be considered when developing an appropriate strategy. These include:

1. Data strategy and governance
2. Appropriate consideration and management of data accessibility
3. Utilise and promote existing Data standards
4. Preserve and expose auditability and provenance of evidence
5. Make use of data capture technologies

Data strategy and governance

Before any monitoring takes place it is important to define the strategy and governance arrangements for capturing, storing, managing, quality controlling, disseminating and using the data. This will include considerations such as data flow, ownership, access permissions, roles and responsibilities, policies and procedures, retention etc. Fundamental to this is clearly defining the purpose of the monitoring and the expected use of the data and information gathered. In other words, being clear about the question that the monitoring is designed to answer.

Data Accessibility

Wherever it is practicable and appropriate, a future national monitoring programme would aim to make datasets and data products available for re-use in an appropriate format. However there are a range of situations where data may not be publishable without restriction. For example, there are many data sets that have usage restrictions that legally must be adhered to due to regulation, confidentiality, licensing or compliance and they often have very different constraints that impact their use. It is therefore important to consider these separately and for each case to be assessed in its own right. For some, usage restriction limits the ability to disseminate derived outputs, whereas for others restrictions can be such that the raw data itself is concealed and is only available for analysis in an aggregated form. Such data sets, however, can often be central to analyses that underpin the evidence base one wishes to present. It is therefore important that, whilst usage restrictions are maintained, the potential of the data is maximised and that evidence is not compromised. As an example, it may be necessary for a particular data set to be hosted by a particular institution and copies cannot be shared. The data, however, is needed for integrated analyses and provides key evidence in its own right, as such it is important that any natural resources report utilises this data. In such cases consideration will be needed to make metadata available so that the data can be identified and information about how to access the data and the appropriate restrictions are clear.

Data standards

To maximise the reuse of data and to understand the potential of integrating data together, certain standards should be adhered to. Standards are created so that attributes and associated meta-data of data sets are exposed and an understanding of the underlying data structure is made as simple as possible and common across different sources. An example of such data standards is the EU INSPIRE Directive, which Welsh Government is committed to implement by 2019. This directive aims to enhance the sharing of environmental spatial information and better facilitate public access to spatial information. Data standards are also used to ensure consistency across data, which can be crucial for integrated analyses or presentation of evidence across multiple data layers. Both geographic and temporal consistencies, as well as consistency of terminology, measurements and data tags via the use of controlled ontologies and thesauruses can be particularly important. There are existing examples of good use of data standards within environmental science, thesauruses established and ontologies published (e.g. Darwin Core) that should be used wherever possible.

When designing any new monitoring activity or data collection task, the data collection should consider how the data will be archived and what associated meta-data is required. This can lead to efficiencies in post processing. Sorting primary data and metadata into the correct formats for archiving can take a huge amount of time post collection, so careful planning is needed to avoid this.

The challenge of controlling data standards is intensified with third party data sets that are used in integrated analyses or to supplement the evidence base. For these “independent” data sets it may be ok to insist on minimum data quality limits, but to insist on or to encourage changes in practices by external organizations would require funding and could be expensive. Ultimately, good communication across data providers is key to understanding where compromise is needed and where strict codes of practice are needed.

Auditability and Provenance

When assessing the suitability of data for a particular use or presenting evidence, it is important that the quality of the evidence can be assessed in a systematic fashion. A key element of evidence quality is having a clear audit trail such that it can be traced back to the point of data collection. This requires exposing the workflow and data sets that contributed to the evidence. If this is clearly described in associated meta-data, then the user has an increased confidence in the evidence presented. In this sense, the provenance of data or evidence links back to data standards. Further, if derived output and evidence is tested and challenged by comparison against existing models, expert knowledge, controlled studies or published research, then confidence in the product is increased and robustness satisfied. Based on these principles, sufficient resource should be allocated to support conversion of data into robust evidence products.

It is also important that a publishing workflow is established and potentially presented with any dataset. For example what checks were made on the data to ensure it was appropriate for publication before being signed off for public dissemination. These checks would typically include quality assessment, and environmental or legal sensitivity. Critical to this is the transparent recording of data characteristics, which allow appropriate controls and caveats to be applied to the raw data. In such circumstances, it is better to provide sound outputs in which such caveats have already been taken into account with the link back to the raw data made clear.

Data Capture

Over the last 10-20 years there have been huge gains made in the field of informatics relating to data capture. A driving principle behind much of the development has been to find increased efficiencies via a reduction in any post-processing and improved data quality. This has led to an increasing move towards electronic data capture, whereby surveyors themselves input the data either out in the field via computer software or post-hoc via web-based forms. There are many examples of such systems in place in environmental recording each with varying degrees of success. One particular example of this move to electronic data capture was the 2007 Countryside Survey, where a GIS oriented solution was adopted utilising both a strong database design and capture software facility. It was estimated that the move to electronic data capture saved the survey in excess of £700,000. The Breeding Bird Survey also successfully utilises a system of web forms to allow the participants to fill in their own records online. This has helped improve data standards and reduce post-processing of paper-based data entry.

An important issue to consider is how further development of data capture technology can be used to provide additional efficiencies and improve data quality. One consideration may be the use of open source software for field data collection that can be shared across providers and modified accordingly. An existing example of which is the COBWEB project that provides a facility to easily generate mobile apps for environmental citizen science. Another consideration for increased efficiencies may be to align the data collection initiatives directly with the database formats. An additional consideration may be whether the same software application could be used across professional surveyors, volunteers and across different environmental domains. Using existing data and/or reference data to suggest confidence is also an area of great potential. Ultimately, the pros and cons of each system should be considered specific to the monitoring activity in question. As an example, one may consider the pros and cons of using open source GIS software such as QGIS or proprietary software such as ESRI ArcGIS for spatial mapping of habitats. In this instance, whilst the free open source QGIS solution offers much the same user functionality as the paid-for ESRI product and provides an easier basis on which to develop bespoke software extensions, the back-end database solution provided by ESRI offers significant advantages in terms of data storage, structure and accessibility. Further, one can expect maintenance and stability in paid-for software as well as a managed program of updates from one version to another, which can be of critical importance for sensitive environmental data. Hence it is important to fully consider the pros and cons of any data capture system prior to deployment and to take advantage of any previous investment.

In many cases base maps and spatial reference aids such as GPS are used to aid data collection and have provided a significant improvement in data capture over recent years. In some situations open base mapping datasets (such as Open Street Map or Welsh Government’s Digital Aerial Imagery) or the use of location enabled data recording devices (e.g. via GPS) can be used to reduce reliance on heavily restricted base mapping products such as Ordnance Survey Mastermap which can hinder wider publication and data sharing. Care must be taken though to ensure that use of different reference datasets does not impede data integration.

Current Initiatives

There are many examples of current applications that present evidence or provide a window to available data. Data catalogues such as lle.gov.wales, catalogue.ceh.ac.uk and data.gov.uk and evidence portals such as the NBN gateway, GMEP reporting portal, UKSO, StatsWales and the future Atlas of Living Wales all provide clear examples that should be utilised wherever possible rather than re-inventing the wheel. Ongoing national and international activities should be exploited where possible and lessons learnt from past experiences. Though the operational functionality of data storage, archiving and tagging of data and the dissemination of key results and summaries are very different, it is important that they are not viewed in isolation.

Relationship with new technologies considered as part of the future monitoring programme

The use of emerging technologies within a monitoring context provide additional opportunities and challenges from a data and informatics perspective. Most notably, many new initiatives collect vast volumes of data and often require a considerable amount of processing prior to analysis. In addition to this, coordination of data capture and adherence to strict data collection protocols and consistency across observations must be maintained.

The use of citizen science to aid data collection introduces a particular set of considerations with regard to data capture and protocols. To save post-processing time and to ensure consistency across the volunteers it is important that some central coordination effort is in place and that data collection exercises are suitable for non-professional surveyors in order to minimise errors and increase efficiencies. Using electronic data capture technologies can help with this. There are current examples already in place such as the iRecord suite of mobile applications used by many biological recording societies and the web-form system used by the BTO for the Breeding Bird Survey. There are also additional open source smart phone apps that could be easily configured to record environmental information, for example COBWEB, Fieldtrip GB and EpiCollect+.

The use of eDNA approaches in environmental monitoring produce a large amount of data that are processed and condensed into environmental indicators of interest. The raw data itself is then of little use except for re-analysis. The issues then centre around how and where the vast quantity of raw data are held. It is important to ensure these new data resources are kept and managed accordingly for data citation, retention period etc.

Finally, the use of EO data requires a considerable amount of processing and storage which can be a challenge to computing infrastructures. For example the 2007 Land Cover Map classified over 8.6 million land parcels. There are, however, EO strategies and commitments in place across the political and administrative spectrum where such considerations are already being addressed (eg Defra CoE). The underlying principle of developing EO as a new technology for Government, including related informatics activity, is therefore through collaboration.

Future Potential

The complexities involved in developing an effective informatics strategy, as listed above, are such that a collaborative approach will be necessary to ensure that all data issues are fully captured. As such, any future monitoring programme should have a data and informatics coordination board to oversee this strategy and to increase the sharing of data and evidence.

In the short term, the emphasis should be focussed on understanding current developments in this area across Welsh Government (and potentially wider) to avoid duplication of effort and to consolidate existing activity. Efforts could concentrate on existing data catalogues such as Lle and data.gov.uk to understand how these could be exploited further and contribute to a future natural resources gateway. As an example, any current environmental data sets that are available via data.gov.uk that could be utilised in a future monitoring programme should be exposed. Utilising existing catalogues ensures that data already conform to certain data standards and hence achieving consistency. Ensuring robustness and consistency across data and evidence should be a clear priority in the short term. The consistency may be in the way that data is stored (eg same file formats), the way that data is collected, the way that data is described (eg species nomenclature) or the way that data is analysed. The robustness of data and evidence may relate to the auditability of the product or whether any conclusions have been challenged via, for example, multiple modelling approaches or expert opinion, to validate inference potentially used in decision making. The goal should be to provide a clear benchmark and guidance for all data providers and analysts for sharing of data and evidence. The key will then be to establish where the data should reside and be disseminated in the long term, for example should biodiversity data go to the NBN.

In the longer term the aim should be to have a single gateway for all Welsh environmental data and evidence. This may cover certain elements of raw data, summary data or the presentation of robust evidence products. This “hub” should provide a window to data products and evidence without necessarily being the one place where all data is stored. Evidence and data may, and in many cases should, exist on other platforms that make the most of existing infrastructures. Some data may be directly accessible, whereas for other data all that is available is a meta-data record and link to a third party site. Similarly for evidence, whilst all available evidence should be clearly presented it may be that this is drawn from 3rd party sites via the use of web services such as WMS for displaying national maps. In reality, this gateway may represent a simple landing page from which other archives and infrastructures can be accessed – building on these existing initiatives will bring the biggest efficiency savings. This would enable clear distinction between raw data and summary results, but provide a single port of call for environmental information across Wales.

Aside from these future priorities, it is perhaps important to recognise that whilst such development can provide efficiency savings a significant amount of resource is required to maintain and develop the infrastructure required. Currently there is precious little infrastructure or skill to manage, analyse and synthesise domain specific data. A significant proportion of resource available should be ring fenced for data and informatics to underpin data coordination and sharing activities, analysis and interpretation both to the end user community and across organisations. It is therefore important to acknowledge the possibility of sharing the available funding resources for this management activity between organizations that may contribute external data sets. An alternative consideration could be that the raw data management and access is outsourced, but consideration would have to be made as to whether this was sustainable or desirable.

Ultimately, it is well recognised now that a well thought out and well-resourced approach to data and informatics can lead to significant efficiencies, increased use of and recycling of data and better engagement with policy makers and public via dissemination mechanisms.

Intentionally blank

NERC SCIENCE OF THE ENVIRONMENT

PEER

INVESTORS IN PEOPLE

Athena SWAN Bronze Award

BANGOR
Centre for Ecology & Hydrology
Environment Centre Wales
Deiniol Road
Bangor
Gwynedd
LL57 2UW
United Kingdom
T: +44 (0)1248 374500
F: +44 (0)1248 362133

EDINBURGH
Centre for Ecology & Hydrology
Bush Estate
Penicuik
Midlothian
EH26 0QB
United Kingdom
T: +44 (0)131 4454343
F: +44 (0)131 4453943

LANCASTER
Centre for Ecology & Hydrology
Lancaster Environment Centre
Library Avenue
Bailrigg
Lancaster
LA1 4AP
United Kingdom
T: +44 (0)1524 595800
F: +44 (0)1524 61536

WALLINGFORD - Headquarters
Centre for Ecology & Hydrology
Maclean Building
Benson Lane
Crowmarsh Gifford
Wallingford
Oxfordshire
OX10 8BB
United Kingdom
T: +44 (0)1491 838800
F: +44 (0)1491 692424

enquiries@ceh.ac.uk
www.ceh.ac.uk