Harrison, Kathryn A.; Wright, Daniel G.; Trembath, Philip. 2017. **Implementation of a workflow for publishing citeable environmental data: successes, challenges and opportunities from a data centre perspective**. *International Journal on Digital Libraries*, 18 (2). 133-143. 10.1007/s00799-016-0175-5

Implementation of a workflow for publishing citeable environmental data: successes, challenges and opportunities from a data centre perspective.

Kathryn A. Harrison, Daniel G. Wright, Philip Trembath

Affiliation: Centre for Ecology & Hydrology, Lancaster Environmental Centre, Library Avenue, Bailrigg, Lancaster, LA1 4AP, UK

Corresponding author: Kathryn A. Harrison

Email: kath@ceh.ac.uk

Abstract

In recent years, the development and implementation of a robust way to cite data has encouraged many previously sceptical environmental researchers to publish the data they create, thus ensuring more data than ever are now open and available for re-use within and between research communities. Here we describe a workflow for publishing citeable data in the context of the environmental sciences – an area spanning many domains and generating a vast array of heterogeneous data products. The processes and tools we have developed have enabled rapid publication of quality data products including datasets, models and model outputs which can be accessed, re-used and subsequently cited. However, there are still many challenges that need to be addressed before researchers in the environmental sciences fully accept the notion that datasets are valued outputs and time should be spent in properly describing, storing and citing them. Here we identify current challenges such as citation of dynamic datasets and issues of recording and presenting citation metrics. In conclusion, whilst data centres may have the infrastructure, tools, resources and processes available to publish citeable datasets, further work is required before large-scale uptake of the services offered is achieved. We believe that once current challenges are met, data resources will be viewed similarly to journal publications, as valued outputs in a researcher's portfolio, and therefore both the quality and quantity of data published will increase.

Keywords:

Citation, publication workflow, DOI, dynamic data, metrics, data centre

1.0 Introduction

Historically, there has been resistance from some researchers in the environmental sciences to publishing data, other than referring to it in articles in recognised scientific journals. The act of making data openly available for the public to view, access and re-use is an unfamiliar concept to many, although, for some scientific communities (e.g. bioinformatics and 'omics) data archival is a cultural norm [1]. Inability to access scientific data is an obstacle to interdisciplinary research [2, 3] which is key in the area of the environmental sciences as they cover a broad range of disciplines and often aim to answer complex questions requiring input from a range of specialists. Whilst each year large amounts of research funds (including tax payers' money) are spent generating new data, existing data remain inaccessible, unidentified and therefore underutilised [4].

In recent years there has been increasing pressure on scientists to make the data they generate openly available. Regulatory pressures such as the EU's INSPIRE[1] directive and compliance with research funders' policies (e.g. RCUK'[2] data policy) are compelling researchers to publish their data. Nonetheless, this regulatory approach has done little to prompt a significant change in cultural practices. It is clear that in order for a shift in behaviour to occur, researchers must feel confident that making the data they create available will not adversely impact on their career. Mayernik [5] and Assante et al [6] make reference to the cultural barriers which make scientists unwilling to share results and document the fears that researchers have of being 'scooped', their data being used improperly and the difficulties they face in producing data in a shareable form.

[1] http://inspire.ec.europa.eu/

[2] http://www.rcuk.ac.uk/research/datapolicy/

If a published data resource is regarded as a citeable publication it can impact positively on a researcher's reputation [2] and this in turn will encourage the publication of more data. Generation and subsequent publication of data should be recognised as valuable activities but currently lacks an essential pre-requisite – accepted metrics of significance [7]. For example, it should be possible to collect information on who has re-used the data, what it has been re-used for and how many times has it been re-used. Metrics such as these could ultimately bear on the academic reputation of a researcher amongst their peers in a similar way that metrics on citations of journal papers currently do. Provision of this service alone will not solve all the problems, however, and it will take time to establish. Data centres and research institutions must also consider providing support to researchers, increasing awareness of the issues and developing simple workflows in order that time-limited researchers can engage in the process of making the data they create publicly available and gain credit for doing so.

Providing a means of citing data allows data creators to be perpetually linked to the datasets they produce. However, for researchers to gain credit for their work a formal, community-recognised structure must be set in place [2]. DataCite[3] has been instrumental in developing and supporting the standards behind persistent identifiers for data. They provide a means by which researchers can find, identify and cite research data and other research objects. DataCite currently use the Digital Object Identifier (DOI) system[4] as a persistent identifier for data resources, although other permanent identifiers could be used in a similar way [8].Through this system, DataCite is able to provide a robust mechanism for allowing citation of data resources. The DOI system is one of the more suitable candidates for permanently identifying research data the system is well-established and widely used for identifying research articles and are therefore a familiar entity to researchers [9, 10] and publishers

[3] https://www.datacite.org/
[4] http://www.doi.org/

alike. Whilst a suitable system for identifying data and making it citeable is in place there still exists a gap between the technical ability to cite data and the cultural behaviour of researchers within the environmental sciences (see above). This gap can only be narrowed by researchers interacting with the system in a positive manner and gaining reward for doing so, for example, a raised awareness of a researcher's work within the community leading to increased collaborative or funding opportunities or improved promotion prospects [3].

The Environmental Information Data Centre (EIDC[5]) is a Natural Environment Research Council (NERC) environmental data centre hosted by the Centre for Ecology and Hydrology (CEH). The data centre primarily accepts data resources from NERC-funded research covering a wide spectrum of disciplines including the terrestrial and freshwater sciences and hydrology. Data held by the EIDC is usually 'complete' end of project life data, although the data centre also holds data collected from long-term environmental monitoring programmes – normally deposited in discrete time slices. The EIDC offers researchers the opportunity to obtain a DOI for data they have created and therefore the ability to cite the resource in literature. DOIs are used as a permanent identifier for data held by the EIDC as this is the identifier initially chosen by NERC for use in its data centres. NERC works with The British Library who is the allocation agent for DataCite in the UK. By assigning a DOI to a resource, the EIDC are signifying that the data are complete, stable, in a useable format, have appropriate metadata, have passed the quality control checks within the domain expertise of the data centre and have guaranteed long-term curation [9]. Whilst there is nothing inherent in a DOI that guarantees the data it identifies will remain permanently available and stable, the EIDC holds a form of 'social contract' between itself and the registry (DataCite and the British Library) to ensure that this is the case [10, 11]. The EIDC uses checksums to ensure data remain unchanged once they have been deposited with the data centre and data

[5] http://eidc.ceh.ac.uk/

depositors receive a copy of the checksum so that they may verify this at any given time. As a data centre, the EIDC has been offering DOIs for datasets that it holds since 2011. Here we outline the processes established to provide this service and describe initial community use and acceptance of the system. We explore the impact that this service has had on the data centre, the datasets published by the data centre and the subsequent exposure of those datasets. Further, we discuss future challenges for the EIDC, specifically, citation of dynamic datasets and the collection of citation metrics. Both these issues have the ability to further influence the volume and quality of data published within the environmental sciences community.

2.0 Data centre process for obtaining a DOI

Data resources are taken into the EIDC following a defined workflow, which includes strict process and quality control measures. Data resources which are identified as suitable for deposit are curated by the data centre in order that they may be viewed and accessed over the long-term. For a data resource to be deemed suitable it must meet a number of criteria such as subject area, funder, repeatability and uniqueness – data held elsewhere would not be considered for deposit. The EIDC first began using a defined workflow in 2011 and to-date holds a total of nearly 400 data resources including datasets, models, model outputs and web services. Only datasets that have passed through the workflow and been formally 'ingested' into the EIDC are eligible for a DOI. Each of the seven NERC data centres (of which EIDC is one) has a representative who can register DOIs for NERC datasets. Whilst the act of registering a DOI with DataCite is the same for all data centres, the manner in which datasets are prepared to a form which is acceptable for allocation of a DOI varies.

2.1. Support for researchers

The EIDC is hosted by the Centre for Ecology and Hydrology (CEH) and as such, the data centre accepts data from both ‘internal’ depositors (i.e. researchers from CEH) and ‘external’ depositors (i.e. researchers employed elsewhere such as universities and other research institutes). The process for ingesting data is identical for both internal and external depositors; the support given to researchers prior to submission of the data is also broadly similar and will be described here briefly. CEH employ a team of Informatics Liaison (IL) staff whose role it is to support researchers with data management and all that it entails. Members of the IL team will work with researchers ideally from the very start of a project to ensure a data management plan is created and regularly reviewed and updated. Likewise, this support is also available for ‘external’ researchers whose data will ultimately be considered for deposit with the EIDC. Data management plans identify the data resources that will be offered to EIDC and also list the supporting documentation which will accompany the deposit. The IL staff will initiate a deposit once the researcher is ready, and support them through the process – for example, helping to complete discovery metadata records, giving advice on formatting the data for deposit, creating any supporting documentation and discussing issues such as licensing and citation. The workflow whereby the EIDC registers DOIs for data it holds is described below. However, this workflow does not solely include actions carried out automatically by the data centre with the researcher in isolation, support from IL and data centre staff is provided throughout. A full description of the complete workflow for ingesting data into the data centre is not within the scope of this article but has been described elsewhere within this special issue.

2.2. Discovery metadata

At the EIDC, the process for obtaining a DOI begins with the collection and storage of discovery metadata. The EIDC uses the UK GEMINI[6] metadata specification for describing

[6] http://www.agi.org.uk/join-us/agi-groups/standards-committee/uk-gemini

the data resources for discovery purposes. This standard has a set of mandatory requirements and includes elements such as title, abstract, lineage and keywords. The EIDCs discovery-level metadata is stored in a metadata file store based on Git[7], a distributed revision control system, which ensures a complete history of all changes made to metadata is maintained. Metadata are stored as JSON[8], an open data-interchange format that records data as attribute-value pairs. The JSON format allows the data centre to transform the data and present them in a number of different formats targeted at distinct audiences – being both human- and machine- readable. For example, the metadata can be presented as a human readable HTML page, as GEMINI-compliant XML for data exchange to data.gov or as XML in the DataCite schema[9] for registering DOIs and populating the DataCite catalogue. Metadata records are created by the researcher depositing data with help from data centre staff, who enter the information using a bespoke metadata editing tool; the metadata is accessed from the CEH catalogue[10]. This catalogue was developed in-house to provide the public with a user-friendly interface for finding, viewing and accessing data (Fig 1).

The discovery metadata record for the data resource also acts as the landing page for the DOI, once registered, and it was designed with this in mind. Although much of the information about a resource is captured using the GEMINI metadata standard, how it should be presented to function as a DOI landing page was key to the decisions made about the way the page was fashioned. As stated by Ball and Duke [12] a landing page should “enable readers to ensure they have located the right dataset, to (re-)familiarize themselves with the research context and supporting documentation, to consider licence terms prior to downloading and to switch to a more recent version of the data if required” (pg 12). The EIDC is keen to promote the use of data citations, therefore, once a resource has a DOI, this

[7] http://www.git-scm.com/
[8] http://www.json.org/
[9] http://schema.datacite.org/
[10] https://catalogue.ceh.ac.uk/

appears, together with the reference to be quoted with any subsequent re-use, at the very top of the page, immediately below the title. An abstract describing the resource follows the DOI and to the right, in a 'Get the data' panel, information on how to order the data, access to supporting documentation and another full citation for the data is presented with the clear instruction 'If you reuse this data you must cite:' (Fig. 1). In designing the landing page, particular care was taken to use accessible language rather than adopt the somewhat opaque language of the metadata standard. For example, 'resource locator' is labelled 'online resources' and 'responsible organisation' is labelled 'contacts'. The GEMINI XML view of the metadata retains the standard terms, it is solely the landing page/catalogue view that presents the more user-friendly version.

2.3. DOI registration

The CEH Catalogue generates DataCite metadata directly from the GEMINI metadata using a simple mapping (Table 1). To register a new DOI, the designated DOI administrator makes a request by clicking a hyperlink on the data resource's record in the data catalogue. This hyperlink only appears on the record if a number of conditions are met. First, only a DOI administrator has access to the link – it does not appear if a user without the necessary permissions is logged in. Second, all the key pre-requisite elements of DataCite metadata must be present within the record – namely: at least one author; a date of publication; a title; and a publisher (other information is also included in the DataCite metadata but these are the only mandatory fields). Thirdly, the landing page must be publicly accessible. Fourth and finally, there must not already be a DOI registered for that resource. By clicking the hyperlink, this triggers a series of actions which occur programmatically without the need for further user intervention. The metadata is posted to DataCite's REST API[11], this creates an entry in DataCite's metadata store. A second request is then immediately posted to the

[11] https://mds.datacite.org/static/apidoc

same API which registers the DOI and specifies its landing page (the page in the CEH catalogue from which the administrator made the request). Next, a request is made to the shortDOI service[12] which creates a more practical, shorter DOI alias. Both the new DOI and the shortDOI are then automatically added to the metadata record in the data catalogue, along with information about how to cite the data resource (Fig 1). Once a DOI has been registered for a data resource, subsequent updates or amendments to the metadata which affect the DataCite metadata are automatically submitted to the DataCite API. This ensures that the DataCite metadata is always representative of the GEMINI metadata held in the CEH catalogue.

The researcher who deposited the data is emailed to inform them that a DOI has been given to the data they created. The email contains details of the DOI, the shortDOI and recommendations on how to use the DOI and cite the data. This notification is currently carried out manually by a member of staff at the data centre. The EIDC also maintains an inventory of all the datasets it holds that have a DOI. This DOI inventory is also manually updated upon the registration of a new DOI. Whilst both these actions are currently carried out manually, the EIDC hopes to automate them in future in order to reduce staff time spent carrying out the processes and provide a more efficient service to depositors.

To date, just over 70% of the data resources held by EIDC have a DOI allocated to them. Currently, researchers are asked upon deposit whether they would like a DOI for the data they have created – they are not minted automatically for every data resource taken in. The reason researchers don’t always request a DOI is usually due to the data being ‘legacy’ data i.e. data that was generated a long time ago (on the scale of decades) and has already been discussed in the scientific literature, therefore researchers feel they have nothing to gain

[12] http://shortdoi.org/

from obtaining a DOI for them. When a DOI for a data resource is resolved using a web browser, the user sees a landing page which is the discovery metadata record for that resource. The landing page includes information on how to obtain the resource and how to cite it in future publications (see above). DOIs can only be allocated to data resources that have been formally deposited with the EIDC; this normally occurs towards the end of a project or section of work. Data must have passed documented quality checks and be held within the data centre itself. DOIs are allocated prior to the data being made publicly available (although this usually happens immediately after). The EIDC supports NERC's option of allowing researchers a two year embargo on the release of the data they created. In the case of embargoed data resources, DOIs are registered when the data are deposited, as this allows researchers to use the DOI in any publications they have planned. The DOI is documented on the landing page for the data resource along with details of the embargo and a date when the data are to be made available.

3.0 Uptake and use of DOIs for data from a data centre perspective

The motivation for requesting a DOI for data deposited with EIDC has varied over time. At first, requests came in solely because it was now a service offered by the data centre and this had been communicated to depositors by the IL staff. DOIs were initially requested even though some researchers were not fully aware of what they could be used for. This is not unsurprising, as it has been noted previously that there is a lack of clear recommendations on how to cite data within scientific literature. The Data Citation Guidelines for Data Providers and Archives [10] state that among Federation of Earth Science Information Partner (ESIP) members, current recommendations for citing data range from casual acknowledgement within the text of a paper to formal and specific citations within the references section of the paper. Mayernik [5] also stated that even when data is widely shared, users do not commonly cite datasets in formal ways. Rather than formally citing datasets, data users typically acknowledge data use in the text of an article in the acknowledgement section.

One of the first DOIs assigned by the EIDC was for data created by Beresford et al [13] which was subsequently quoted in a journal paper [14]. However, the authors failed to include the recommended DataCite citation in the reference list and merely added a statement to the paper, “All data associated with this study are available from the CEH Information Gateway (https://gateway.ceh.ac.uk/) and the data have been allocated a digital object identifier (http://dx.doi.org/10.5285/1a91c7d1-ec44-4858-9af2-98d80f169bbd)"

This indicates they did not regard it as a reference in the same way as they would a journal paper.

Other researchers requested a DOI as they were publishing in a data journal and it was a mandatory requirement of submission. Data journals, especially in the field of the environmental or natural sciences are a relatively new concept, however, they are increasing in number. Journals such as Earth System Science Data (ESSD), Geoscience Data Journal, Scientific Data and Data in Brief publish peer-reviewed data papers – papers that describe datasets [3]. The majority of data journals require data to be stored in an approved repository with a permanent identifier assigned enabling reviewers to access the data. At the EIDC, one of the first datasets referred to in a data paper was from Haxton et al. [15]. This dataset was deposited with the EIDC and given a DOI which was subsequently cited in an ESSD paper by Prudhomme et al. [16]. Furthermore, the ESSD paper has since been cited by at least five other journal publications (as recorded by CrossRef[13]) including one the author co-authored [17]. It should be noted that each of these outputs (the data paper and

[13] http://www.crossref.org/

the dataset itself) is a publication in its own right - there is no requirement for the data resource to have the same lead author as the data paper. They are separate entities with their own individual reference and can be referred to as such. When re-using data or tracking where data have been re-used it is important to use the citation for the dataset itself, rather than the reference for the data paper. If simply referring to the work carried out by a group of authors, citing the data paper would be appropriate. By publishing a data paper based on a dataset, authors are adding value to the dataset for the future consumers of the data [9] as the data they created has undergone a scientific peer review process. Datasets published by the data centre have reached a certain level of quality as required to obtain a DOI, but they are not peer reviewed.

As a case study, the above example has encouraged other researchers within the organisation to engage with the data centre which has further increased the number of datasets being offered for deposit. In the financial year 2012-2013, the EIDC had 35 deposit requests i.e. researchers contacting the data centre wishing to deposit data. These figures contrast with those of the financial year 2014-2015, where the EIDC had 83 deposit requests (it should be noted that one deposit request may lead to the deposit of one dataset, or many which is often the case). In 2015-2016, the EIDC had 61 deposit requests in the first 6 months of the financial year, therefore it is likely that the number of deposit requests this year will exceed those of the previous year. The reason for this increase in engagement with the EIDC could be due to case-studies such as the one above being advertised keenly throughout the organisation (CEH), however, it is more likely that pressure from publishing houses, as discussed below, has had a greater impact on these figures.

The final reason researchers are now offering their data to the data centre to publish (and requesting a DOI) is that increasingly scientific journals are recommending, or even mandating, that data referred to in an article must be archived in an appropriate public archive [3]. The archive must provide public access and guarantee long-term preservation of the data resource. Some journals also require that the data have been assigned a permanent identifier (e.g. a DOI). The pressure from publishing houses (e.g. British Ecological Society, Ecological Society of America, Nature and Science) is urging those researchers in the environmental science community previously resistant to the idea of publishing data to actively participate. Whilst this is encouraging it is often done in an untimely manner. Despite the support and advice available, some researchers are still unaware of the importance of data management and citation, or it fails to make the list of their priorities for reasons discussed above [3]. Many researchers are currently offering data resources to the data centre for publication only after a journal paper has been written and submitted, and hence require the deposit process to take place hurriedly. This is often not possible as the EIDC processes mandate that data coming into the data centre be accompanied by sufficient supporting information which depositors have usually not prepared in advance. The EIDC is bound by NERC to take in data of long-term value so that it may be stored securely in perpetuity and have the potential to be re-used where suitable. It is therefore not possible for the data centre to 'fast-track' data deposits with the aim of meeting the requirement from depositors that they must have a DOI for data referred to in a journal paper. Data accepted into the EIDC must be complete, be in a non-proprietary format and have sufficient supporting information so that it may be understood and re-used by others without the need to contact the creator. It is therefore critical that researchers engage with data centre staff as early as practically possible in their projects, to develop data management plans and ensure the correct documentation will be provided upon deposit of the data. In cases where researchers have taken advantage of the support provided and deposit of data has occurred in a timely manner, the process of obtaining a DOI and publishing the data can occur rapidly as the workflow operated within the datacentre is automated, where appropriate, and can be completed in a matter of seconds. If researchers have not planned in advance and approach the data centre requesting a DOI as a matter of urgency, the process can take somewhat longer. This is because time has to be spent preparing the data and supporting information. Therefore, whilst the pressure from publishing houses has prompted increased awareness of the requirement to publish data, it may take some time before researchers realise they must engage with this process at an early stage, before a project or grant is completed and prior to preparing articles for submission.

Since the EIDC began issuing DOIs for data resources we have seen an increase in researchers' awareness of the requirement to make data available, predominantly driven by data journals and journal publications. The EIDC is receiving an increasing number of enquiries about depositing data from scientists interested in submitting data papers and research articles as they realise that this is a mechanism whereby they can gain academic credit for a body of work which was previously unacknowledged. Our ability to identify and cite data resources in a reliable manner is largely down to the system put in place by DataCite and the use of DOIs (although it is possible that other permanent identifiers could work in an equally successful way [8]) as it offers researchers an incentive for releasing the data they have created. Without this incentive, we believe many data resources available today through the EIDC would not have been deposited with the data centre and therefore be inaccessible.

4.0 Future challenges for the data centre

The advent of a robust method for making data resources citeable has gone some way in addressing the lack of published data available in the field of environmental sciences but there are still areas where improvements could be made to further increase openness and re-use of data. Many of the data resources archived by the EIDC are created from long-term environmental monitoring programmes and therefore data are being regularly updated. The challenge of making this type of dynamic dataset citeable is well documented, as data such as these do not fit the commonly used DOI system well [5, 11, 12]. In line with DataCite recommendations, once a dataset held by the EIDC has been given a DOI, it will not be changed, updated or corrected [18]. If any of these alterations are required, a new DOI is issued. This is so users can identify and retrieve the exact same data identified by a DOI irrespective of how long it has been since it was registered. The EIDC currently offers researchers two choices when depositing dynamic data, based on the approaches outlined by Ball and Duke [12]; either a new time-slice can be deposited into the data centre and a new DOI issued, or the whole dataset can be taken in including the previous data and any new data (a new snap-shot). In the latter case, the previous version is deprecated and a new DOI is issued for the whole resource. An example of this is data from the UK Butterfly Monitoring Scheme (UKBMS) deposited into the EIDC. The UKBMS deposits data annually on collated indices and species trends. The first deposit was made in 2011 and the data ran from 1976 to 2011 [19, 20]. In 2012, UKBMS submitted a new snap-shot of the data, this time running from 1976 to 2012 [21, 22]. The addition of the new data not only added extra data values but as a consequence also changed the values of the previous years' data. Once a new snap-shot is published and has a DOI, the old snap-shot is deprecated by labelling it an 'Historical archive' in the discovery metadata record. The catalogue is configured so that for records labelled as such, a banner automatically appears at the top of the record stating 'This dataset has been withdrawn' (Fig 2). In this way, the DOI still resolves to the correct landing page so remains a permanent identifier and the user can clearly see that this is not the most current version of the dataset (a link to a record for this collection of data resources is available from the deprecated dataset landing page so users can easily find the most up-to-date version, should they wish to).

However, some researchers are unhappy with the current system and indeed, from a data centre's perspective, snap-shots can become unwieldy for regularly updated time-series data which are common in the Earth Sciences [10]. Instead, researchers would prefer one identifier for the whole resource that never changed regardless of how many updates or additions of data were made. Such a system would ensure citation metrics for the resource were not diluted with new citations generated each time an update was made. However, this is a service we are currently unable to offer using the system we have in place. The Research Data Alliance[14] has a working group dedicated to exploring solutions to the problem of citing dynamic datasets and a position paper by Andreas Rauber and Stephan Pröll has been produced describing a conceptual model for scalable dynamic data citation [23]. However, this paper addresses the problem from a data re-user's perspective so may not solve the issues that researchers depositing to the EIDC have raised. Rauber and Pröll propose using timestamped, versioned data that can be assembled into specific subsets by using queries which subsequently have permanent identifiers assigned to them. This system enables authors to cite only the query, rather than the whole dataset, ensuring users can access exactly the same data referred to by the identifier for perpetuity [23].Whilst this addresses the issue of ensuring users are able to precisely identify specific subsets of data that may have changed over time it does not solve the issue of citation dilution raised by researchers depositing to EIDC. Also, the DOI system, as currently implemented by DataCite, does not support Template Handles, thus a parameterized DOI would not resolve to a particular subset but to the whole dataset [11]. It is clear that attempts are being made to address the issue of citing dynamic datasets but also that one size does not fit all [3, 5], therefore systems may have to adapt in future to accommodate researchers requirements.

Another issue which, if addressed, could further promote data publication in the environmental sciences is that of citation metrics. For the production and publication of data to be recognised as valuable scholarship it requires accepted metrics of significance [7].

[14] https://rd-alliance.org/

Researchers are more likely to publish the data they create if they can measure its impact, track its use and receive credit for creating it [3, 7]. A researcher's academic success is frequently measured by the journal publications they produce, specifically in the impact factor of the journals in which they publish and the number of times articles are subsequently cited. If mechanisms were put in place to provide similar information for datasets, researchers would be able to measure the impact of the data they produce which could input into the professional reward process [5]. Tracking data use is difficult as datasets are inconsistently cited by data users [5]. However, respondents to a survey carried out by Kratz and Strasser [7] found that citation and download counts were more useful than search rank or altmetrics. Therefore, a method for measuring data impact based on data citation counts, though difficult to implement would be desirable to researchers. Data papers can go some way to providing this type of information. For example, the data journal ESSD provides metrics on views and citations of the data papers they publish (Fig. 3). Crucially, however, this is not tracking the citation of the data itself which has its own DOI and mechanism for citation. Thomson Reuters[15] Web of Knowledge now provide a service called the Data Citation Index (DCI) which provides access to data, links data to the articles it supports and tracks citation of datasets. Unfortunately, the DCI is currently not open and free to use (a subscription is required) and repositories have to agree to have information about the data they hold harvested by Thomson Reuters. The EIDC is working with Thomson Reuters to ensure that the data it holds can be included in the DCI and this has recently been achieved by allowing Thomson Reuters to harvest metadata held by DataCite about data held by the EIDC. This is an important first step, although, as CEH is not a subscriber to the DCI, the data centre is unable to obtain information on the citation counts for data it holds.

[15] http://thomsonreuters.com/en.html

In contrast, ResearchGate is a free service, which enables researchers to share their publications and access citation metrics. ResearchGate allows registered users to add articles, book chapters, conference papers, datasets and unpublished work to their home page - once added, metrics are collected on the publication. This would seem like a suitable solution to the problem of collecting citation metrics for the data held by the EIDC, at least from an individual researcher's point of view, as they should theoretically be able to include information about datasets they have deposited with the datacentre (e.g. title, DOI) and obtain information of citation metrics over time. However, when registering a dataset with ResearchGate, users are required to attach the data as a file and are therefore uploading a copy of the data to the ResearchGate site. This is not something the EIDC can recommend for a number of reasons. First, additional copies of data would be unnecessarily generated and stored. Second, ResearchGate mandate that any data uploaded to its site is free from any Intellectual Property Rights which in the majority of cases is not true for data generated though public or private funding. Third, uploading data to ResearchGate is often impractical as the volumes of data in question are often very large (500GB or more)

It is clear that whilst some solutions are available, further work is still needed to implement an openly accessible tool to capture and present metrics for datasets. Until researchers can quantify the impact data resources they have created have on the academic community as a whole they may not receive the full scholarly credit they deserve. In the meantime, the EIDC plans to include information on download counts for each dataset on its landing page. Whilst not ideal, it provides researchers with a highly regarded 'second-choice metric' [7] and can be used as an interim measure until a more informative system is put in place.

5.0 Conclusions

Whilst there is still a long way to go before data resources are viewed as valued outputs from a researcher's work in the same way journal publications have always been, data centres, such as the EIDC, are facilitating a cultural shift in practices with regard to data publications. By providing a robust workflow enabling the identification of datasets and providing a means for data to be cited, data centres are providing the building blocks on which more wholesale changes in attitude and behaviour can occur. Working in conjunction with publishing houses, data centres are beginning to convince researchers that publishing the data they have generated can be beneficial to their research careers. Data centres can further improve on the volume of data published in the environmental sciences by enabling the citation of dynamic datasets, ensuring long-term environmental monitoring experiments can be cited as a single entity, rather than having to generate a new DOI and citation after each new addition of data. In addition, the generation and publication of citation metrics that provide an indication of the impact a dataset has had on the academic community could also, encourage more researchers to publish the data they have created. Much has been accomplished in the last few years but there are still many issues left to address. It will take time for a cultural shift to occur, but by putting flexible robust systems in place and by seeking to illustrate to researchers the benefits of publishing the data they produce, in time data resources and those that generate them will receive the credit and standing they deserve.

Acknowledgements

The authors would like to thank Rick Stuart and Pete Vodden for their work in helping to create the workflows and processes and Rod Scott, Chris Johnson, Jon Cooper, Evgeniya Vetchinkina and Mike Wilson for developing technical solutions.

References

1. Lawrence B, Jones C, Matthews B, Pepler S, Callaghan S (2011) Citation and peer review of data: Moving towards formal data publication. Int J Digital Curation. 6(2): 4-37

2. Klump J, Bertelmann R, Brase J, Diepenbroek M, Grobe H, Hock H, Lautenschalger M, Schindler U, Sens I, Wachter J (2006) Data publication in the Open Access Initiative. Data Science Journal 5: 79-83.

3. CODATA-ICSTI. (2013) ed. Y Socha. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal, 12(0): CIDCR1-CIDCR7. DOI: http://doi.org/10.2481/dsj.OSOM13-043

4. Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen L, Moorman D, Uhlir P, Wouters P (2004) Promoting access to public research data for scientific, economic, and social development. Data Science Journal 3: 135-152

5. Mayernik, M, (2013) Bridging data lifecycles: Tracking data use via data citations workshop report. NCAR Technical Note NCAR/TN-494+PROC, http://dx.doi.org/10.5065/D6PZ56TX.

6. Assante M, Candela L, Castelli D, Manghi P, Pagano P (2015) Science 2.0 repositories: Time for a change in scholarly communication. D-Lib Magazine 21:(1/2) doi:10.1045/january2015-assante. http://dx.doi.org/10.1045/january2015-assante

7. Kratz JE, Strasser C (2015) Making Data Count. Scientific Data 2:150039. Doi: 10.1038/sdata.2015.39

8. Duerr R, Downs R, Tilmes C, Barkstrom B, Lenhardt W, Glassy J, Bermudez L, Slaughter P (2011) On the utility of identification schemes for digital earth science data: An assessment and recommendations. Earth Science Informatics 4:139-60. http://dx.doi.org/10.1007/s12145-011-0083-6

9. Callaghan S, Donegan S, Pepler S,*et al.* (2012) Making data a first class scientific output: Data citation and publication by NERC's environmental data centres. Int J Digital Curation*.* 7(1): 107–113.

10. ESIP (Federation of Earth Science Information Partners) (2012) Data Citation Guidelines for Data Providers and Archives. Ed. Parsons MA, Barkstrom B, Downs RR, Duerr R, Tilmes C and ESIP Preservation and Stewardship Committee. ESIP Commons. http://dx.doi.org/10.7269/P34F1NNJ

11. Klump J, Huber R, Diepenbroek M (2015) DOI for geoscience data-how early practices shape present perceptions. Earth Science Informatics 1-14. http://dx.doi.org/10.1007/s12145-015-0231-5

12. Ball A, Duke M (2015) How to Cite Datasets and Link to Publications. DCC How-to Guides. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides

13. Beresford NA, Barnett CL, Howard BJ, Howard DC, Tyler AN, Bradley S, Copplestone D (2011). Observations of Fukushima fallout in Great Britain. NERC Environmental Information Data Centre. doi: 10.5285/1a91c7d1-ec44-4858-9af2-98d80f169bbd

14. Beresford NA, Barnett CL, Howard BJ, Howard DC, Wells C, Tyler AN, Bradley S, Copplestone D (2012) Observations of Fukushima fallout in Great Britain. J Environ Radioact. 114:48-53. doi: 10.1016/j.jenvrad.2011.12.008.

15. Haxton T, Crooks S, Jackson CR, Barkwith AKAP, Kelvin J, Williamson J, Mackay JD, Wang L, Davies, H, Young A, Prudhomme C (2012). Future flows hydrology data. NERC Environmental Information Data Centre. doi:10.5285/f3723162-4fed-4d9d-92c6-dd17412fa37b

16. Prudhomme C, Haxton T, Crooks S, Jackson C, Barkwith A, Williamson J, Kelvin J, Mackay J. Wang L, Young A, Watts G (2013) Future Flows Hydrology: and ensemble of a daily river flow and monthly groundwater levels for use for climate change impact assessment across Great Britain. Eath Syst. Sci. Data 5:101-107. doi:10.5194/essd-5-101-2013

17. Royan A, Prudhomme C, Hannah DM, Reynolds SJ, Noble DG, Sadler JP (2015) Climate-induced changes in river flow regimes will alter future bird distributions. Ecosphere 6(4): 50. doi: 10.1890/ES14-00245.1

18. British Library, DataCite (2013) Working with the British Library and DataCite – A guide for Higher Education Institutions in the UK. British Library http://www.bl.uk/aboutus/stratpolprog/digi/datasets/WorkingWithDataCite_2013.pdf

19. Botham, M; Roy, D; Brereton, T; Middlebrook, I; Randle, Z (2012). United Kingdom Butterfly Monitoring Scheme: collated indices 2011. NERC Environmental Information Data Centre.doi:10.5285/ff55462e-38a4-4f30-b562-f82ff263d9c3

20. Botham, M; Roy, D; Brereton, T; Middlebrook, I; Randle, Z (2013). United Kingdom Butterfly Monitoring Scheme: species trends 2011. NERC Environmental Information Data Centre.doi:10.5285/cad2af6c-0c97-414c-8d5f-992741b283cf

20

21. Botham, M.; Roy, D.; Brereton, T.; Middlebrook, I.; Randle, Z. (2013). United Kingdom Butterfly Monitoring Scheme: collated indices 2012. NERC Environmental Information Data Centre.doi:10.5285/7949cc99-76c4-4a3e-8c33-41a35b8b7777

22. Botham, M.; Roy, D.; Brereton, T.; Middlebrook, I.; Randle, Z. (2013). United Kingdom Butterfly Monitoring Scheme: species trends 2012. NERC Environmental Information Data Centre.doi:10.5285/5afbbd36-2c63-4aa1-8177-695bed98d7a9

23. Rauber A, Pröll S (2015) Scalable dynamic data citation – RDA-WG-DC Position paper https://rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html Accessed 25 June 2015

Figure captions

Fig 1 Example of a record in the CEH catalogue showing recommended citation and display of DOI.

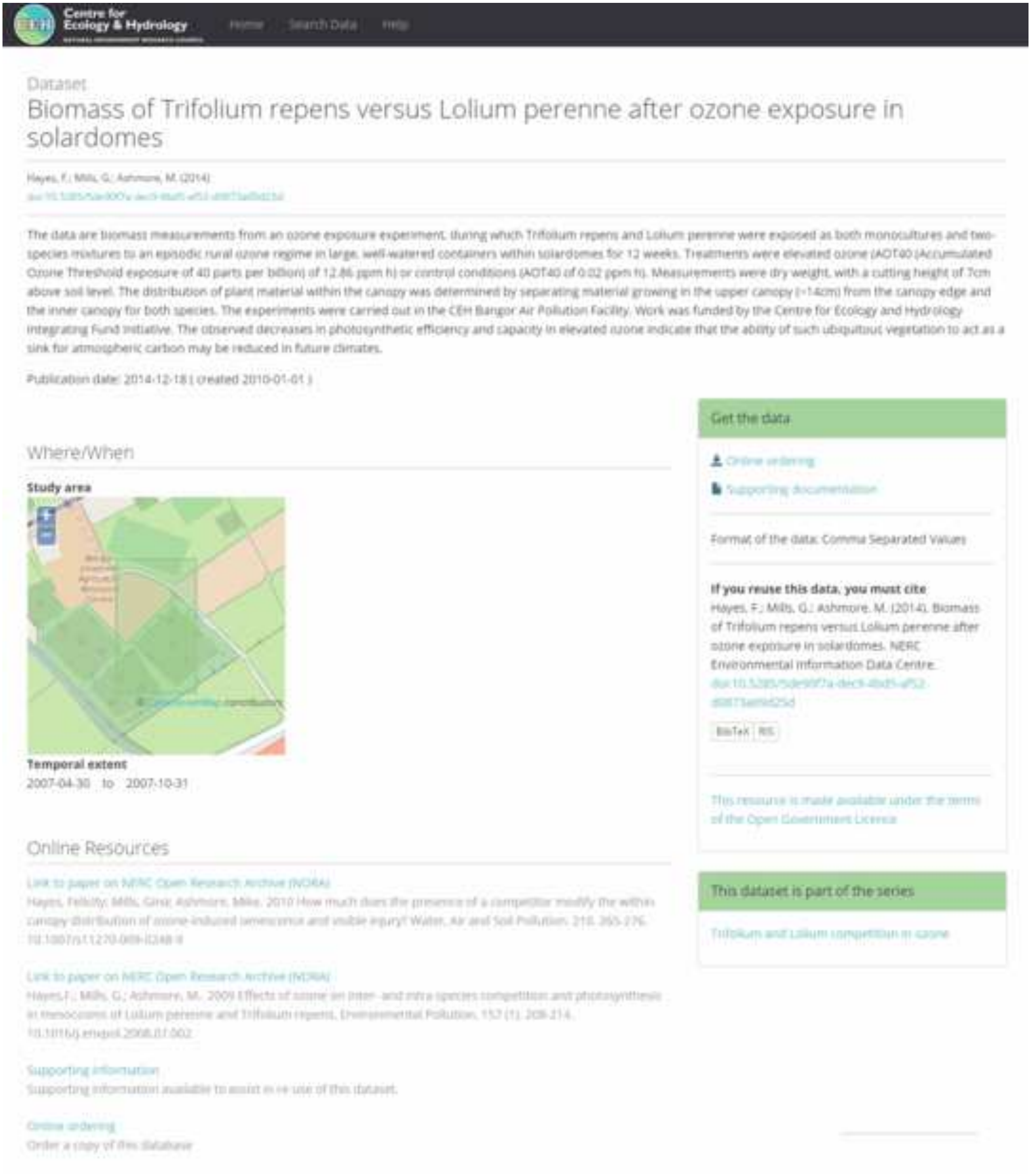
Fig 2 Example of a deprecated metadata record in the CEH catalogue.

Fig 3 Metrics provided by the data journal Earth System Science Data including views and citations.

Table 1. Mapping between GEMINI metadata and DataCite metadata

GEMINI metadata element	**DataCite metadata element**
Based on /MD_Metadata/fileIdentifier	/resource/identifier
/MD_Metadata/identificationInfo/MD_DataIdentification/pointOfContact/CI_ResponsibleParty[role/CI_RoleCode/@codeListValue='author']/individualName	/resource/creators/creator
/MD_Metadata/identificationInfo/MD_DataIdentification/citation/CI_Citation/title	/resource/titles/title
/MD_Metadata/identificationInfo/MD_DataIdentification/pointOfContact/CI_ResponsibleParty[role/CI_RoleCode/@codeListValue=publisher]/individualName	/resource/publisher
/MD_Metadata/identificationInfo/MD_DataIdentification/citation/CI_Citation/date/CI_Date[dateType/CI_DateTypeCode/@codeListValue='publication']/date	/resource/publicationYear
/MD_Metadata/identificationInfo/MD_DataIdentification/descriptiveKeywords/MD_Keywords/keyword	/resource/subjects/subject
-	/resource/dates/date[@dateType='Submitted']
/MD_Metadata/identificationInfo/MD_DataIdentification/language/LanguageCode	/resource/language
/MD_Metadata/MD_ScopeCode/@codeListValue	/resource/resourceType/@resourceTypeGeneral
/MD_Metadata/identificationInfo/MD_DataIdentification/citation/CI_Citation/identifier/RS_Identifier	/resource/alternateIdentifiers/alternateIdentifier
/MD_Metadata/distributionInfo/MD_Distribution/distributionFormat/MD_Format/name	/resource/formats/format
/MD_Metadata/identificationInfo/MD_DataIdentification/resourceConstraints/MD_LegalConstraints/otherConstraints	/resource/rightsList/rights
/MD_Metadata/identificationInfo/MD_DataIdentification/abstract	/resource/descriptions/description[@descriptionType='Abstract']
/MD_Metadata/identificationInfo/MD_DataIdentification/extent/EX_Extent/geographicElement/EX_GeographicBoundingBox	/resource/geoLocations/geoLocation/geoLocationBox

Figure



Dataset

Biomass of Trifolium repens versus Lolium perenne after ozone exposure in solardomes

Hayes, F.; Mills, G.; Ashmore, M. (2014)
doi:10.5285/5de90f7a-dec9-4bd5-af52-d0873a69d25d

The data are biomass measurements from an ozone exposure experiment, during which Trifolium repens and Lolium perenne were exposed as both monocultures and two-species mixtures to an episodic rural ozone regime in large, well-watered containers within solardomes for 12 weeks. Treatments were elevated ozone (AOT40 (Accumulated Ozone Threshold exposure of 40 parts per billion) of 12.86 ppm h) or control conditions (AOT40 of 0.02 ppm h). Measurements were dry weight, with a cutting height of 7cm above soil level. The distribution of plant material within the canopy was determined by separating material growing in the upper canopy (~14cm) from the canopy edge and the inner canopy for both species. The experiments were carried out in the CEH Bangor Air Pollution Facility. Work was funded by the Centre for Ecology and Hydrology Integrating Fund Initiative. The observed decreases in photosynthetic efficiency and capacity in elevated ozone indicate that the ability of such ubiquitous vegetation to act as a sink for atmospheric carbon may be reduced in future climates.

Publication date: 2014-12-18 (created 2010-01-01)

Where/When

Study area

Temporal extent
2007-04-30 to 2007-10-31

Online Resources

Link to paper on NERC Open Research Archive (NORA)
Hayes, Felicity; Mills, Gina; Ashmore, Mike. 2010 How much does the presence of a competitor modify the within canopy distribution of ozone-induced senescence and visible injury? Water, Air and Soil Pollution, 210. 265-276. 10.1007/s11270-009-0248-8

Link to paper on NERC Open Research Archive (NORA)
Hayes,F.; Mills, G.; Ashmore, M.. 2009 Effects of ozone on inter- and intra-species competition and photosynthesis in mesocosms of Lolium perenne and Trifolium repens. Environmental Pollution, 157 (1). 208-214. 10.1016/j.envpol.2008.07.002

Supporting information
Supporting information available to assist in re-use of this dataset.

Online ordering
Order a copy of this database

Get the data

Online ordering

Supporting documentation

Format of the data: Comma Separated Values

If you reuse this data, you must cite
Hayes, F.; Mills, G.; Ashmore, M. (2014). Biomass of Trifolium repens versus Lolium perenne after ozone exposure in solardomes. NERC Environmental Information Data Centre. doi:10.5285/5de90f7a-dec9-4bd5-af52-d0873a69d25d

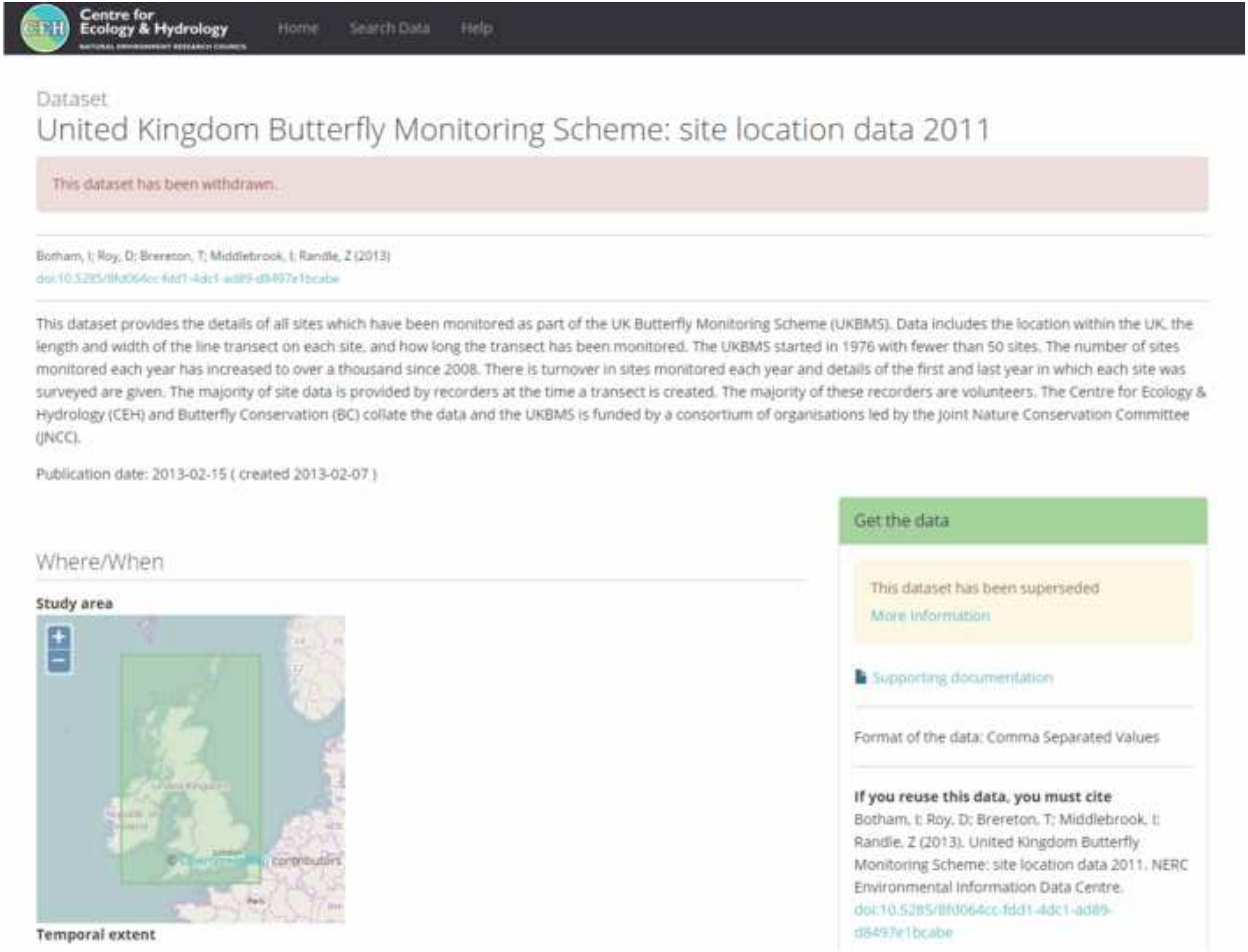
BibTeX | RIS

This resource is made available under the terms of the Open Government Licence

This dataset is part of the series

Trifolium and Lolium competition in ozone

Figure


Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL
Home
Search Data
Help

Dataset

United Kingdom Butterfly Monitoring Scheme: site location data 2011

This dataset has been withdrawn.

Botham, I; Roy, D; Brereton, T; Middlebrook, I; Randle, Z (2013)
doi:10.5285/8fd064cc-fdd1-4dc1-ad89-d8497e1bcabe

This dataset provides the details of all sites which have been monitored as part of the UK Butterfly Monitoring Scheme (UKBMS). Data includes the location within the UK, the length and width of the line transect on each site, and how long the transect has been monitored. The UKBMS started in 1976 with fewer than 50 sites. The number of sites monitored each year has increased to over a thousand since 2008. There is turnover in sites monitored each year and details of the first and last year in which each site was surveyed are given. The majority of site data is provided by recorders at the time a transect is created. The majority of these recorders are volunteers. The Centre for Ecology & Hydrology (CEH) and Butterfly Conservation (BC) collate the data and the UKBMS is funded by a consortium of organisations led by the Joint Nature Conservation Committee (JNCC).

Publication date: 2013-02-15 (created 2013-02-07)

Where/When

Study area

Temporal extent

Get the data

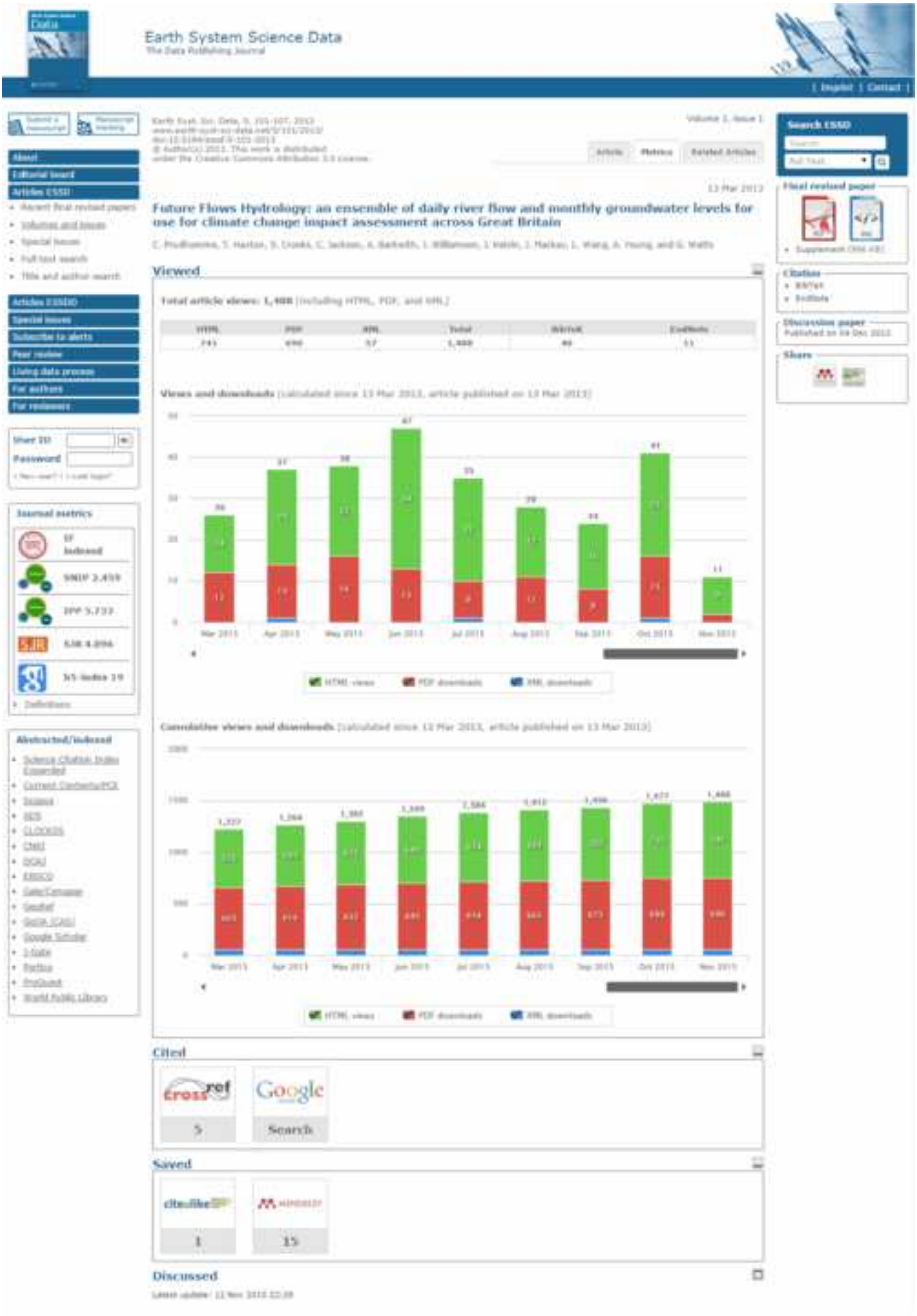
This dataset has been superseded
More information

Supporting documentation

Format of the data: Comma Separated Values

If you reuse this data, you must cite
Botham, I; Roy, D; Brereton, T; Middlebrook, I; Randle, Z (2013). United Kingdom Butterfly Monitoring Scheme: site location data 2011. NERC Environmental Information Data Centre.
doi:10.5285/8fd064cc-fdd1-4dc1-ad89-d8497e1bcabe

Figure



Earth System Science Data
The Data Publishing Journal

| Imprint | Contact |

Earth Syst. Sci. Data, 5, 101-107, 2013
www.earth-syst-sci-data.net/5/101/2013/
doi:10.5194/essd-5-101-2013
© Author(s) 2013. This work is distributed
under the Creative Commons Attribution 3.0 License.

Volume 5, issue 1

13 Mar 2013

Future Flows Hydrology: an ensemble of daily river flow and monthly groundwater levels for use for climate change impact assessment across Great Britain

C. Prudhomme, T. Haxton, S. Crooks, C. Jackson, A. Barkwith, J. Williamson, J. Kelvin, J. Mackay, L. Wang, A. Young and G. Watts

Viewed

Total article views: 1,488 (including HTML, PDF, and XML)

HTML	PDF	XML	Total	BibTeX	EndNote
741	690	57	1,488	49	11

Views and downloads (calculated since 13 Mar 2013, article published on 13 Mar 2013)

Cumulative views and downloads (calculated since 13 Mar 2013, article published on 13 Mar 2013)

Cited

Saved

Discussed

Latest update: 11 Nov 2013 22:28