

## Article (refereed) - postprint

---

Wright, Daniel G.; Trembath, Philip; Harrison, Kathryn A. 2017. **Meeting the challenge of environmental data publication: an operational infrastructure and workflow for publishing data.** *International Journal on Digital Libraries*, 18 (2). 123-132. [10.1007/s00799-016-0176-4](https://doi.org/10.1007/s00799-016-0176-4)

© 2016

This version available <http://nora.nerc.ac.uk/513868/>

NERC has developed NORA to enable users to access research outputs wholly or partially funded by NERC. Copyright and other rights for material on this site are retained by the rights owners. Users should read the terms and conditions of use of this material at <http://nora.nerc.ac.uk/policies.html#access>

**This document is the author's final manuscript version of the journal article, incorporating any revisions agreed during the peer review process. There may be differences between this and the publisher's version. You are advised to consult the publisher's version if you wish to cite from this article.**

**The final publication is available at Springer via**  
<http://dx.doi.org/10.1007/s00799-016-0176-4>

Contact CEH NORA team at  
[noraceh@ceh.ac.uk](mailto:noraceh@ceh.ac.uk)

1

2 **Authors:** Daniel G. Wright, Philip Trembath, Kathryn A. Harrison

3

4 **Title:** Meeting the Challenge of Environmental Data Publication: An Operational  
5 Infrastructure and Workflow for Publishing Data

6

7 **Affiliation:** Centre for Ecology & Hydrology, Lancaster Environment Centre, Library Avenue,  
8 Bailrigg, Lancaster, LA1 4AP, UK

9

10 **Corresponding Author:**

11 **Email:** dgwr@ceh.ac.uk

12

13

14

15 **Acknowledgements**

16 The authors would like to thank Rick Stuart, Peter Vodden and Simon Wright for their  
17 assistance in production of the workflow processes, Mike Wilson, Sabera Adam, Evgeniya  
18 Vetchinkina, Rod Scott, Chris Johnson and Jon Cooper for creation of the infrastructure  
19 components described and two anonymous reviewers for their comments.

20

21

22

## 23 Abstract

24 Here we describe the defined workflow and its supporting infrastructure, which are used by  
25 the Natural Environment Research Council's (NERC) Environmental Information Data  
26 Centre (EIDC)<sup>1</sup> to enable publication of environmental data in the fields of ecology and  
27 hydrology. The methods employed and issues discussed are also relevant to publication in  
28 other domains. By utilising a clearly defined workflow for data publication, we operate a fully  
29 auditable, quality controlled series of steps permitting publication of environmental data. The  
30 described methodology meets the needs of both data producers and data users, whose  
31 requirements are not always aligned. A stable, logically created infrastructure supporting  
32 data publication allows the process to occur in a well-managed and secure fashion, while  
33 remaining flexible enough to deal with a range of data types and user requirements. We  
34 discuss the primary issues arising from data publication, and describe how many of them  
35 have been resolved by the methods we have employed, with demonstrable results. In  
36 conclusion, we expand on future directions we wish to develop to aid data publication by  
37 both solving problems for data generators and improving the end-user experience.

38

## 39 Keywords

40 data, publication workflow, infrastructure, data centre

41

## 42 1.0 Introduction

43 Initially, it can appear that publication of data is relatively straightforward to achieve – identify  
44 the data to publish, and make it available [1]. However, this alone will not ensure that the  
45 published data are permanently and openly available [2]. With further consideration, several  
46 issues become evident, which must be addressed before successful publication of data can

---

<sup>1</sup> <http://eidc.ceh.ac.uk/>

47 be achieved. These are discussed in greater detail below, but include identification of which  
48 data to publish, where to publish and to which community, and how to ensure that the data  
49 are both discoverable and reusable. It is important to recognise that the needs of data  
50 producers and data users are not always aligned - the best solution for one party will not  
51 always result in a satisfactory outcome for the other. Data users may want access to the  
52 data they need as quickly as possible, whereas data providers may seek to produce as  
53 many publications as possible using the data before it becomes publicly available [1].  
54 Publication can therefore sometimes be a compromise and data publishers should aim to  
55 ensure that a successful publication has a satisfactory, if not optimum, outcome for both data  
56 producers and end-users. Further, there are significant restrictions placed on the publisher of  
57 data, with which they must comply, for example, the responsibility to describe metadata and  
58 data using national and/or international standards. Here, we describe the main issues  
59 affecting data publishing and how they have helped to shape a functioning workflow and its  
60 supporting infrastructure, enabling publication of environmental data resources via the  
61 Environmental Information Data Centre (EIDC). The EIDC is a Natural Environment  
62 Research Council (NERC) Data Centre specialising in terrestrial and freshwater  
63 environmental data, and as such has responsibility for publishing a broad spectrum of  
64 environmental data in a variety of different formats. We shall conclude by examining the  
65 evidence that this approach works and expanding on future areas for development.

66

## 67 2.0 Issues in Data Publication

68 The first issue to be addressed is selection of the data to publish. Does all data have value,  
69 or should only a selection be made available? The rate of data generation has shown rapid  
70 increases in recent years [3]. To publish all data generated would be both impractical for  
71 data publishers in terms of storing, cataloguing and dissemination of data, and inefficient for  
72 end-users, who would have to spend more time searching for useful data. It is therefore  
73 apparent that, given the finite resources available to data centres such as the EIDC, a form

74 of selection for data must be made, but what criteria should be used to identify the data  
75 which are suitable for publication? To assist with this decision, NERC has produced some  
76 guidelines for identifying suitable data [4]. These include ensuring that the data are within the  
77 scope of the data centre's remit (for the EIDC this is the terrestrial and freshwater  
78 environmental sciences), consideration of whether the data support a publication, whether  
79 the data are repeatable reusable and that no other copies are stored in another data centre.  
80 The EIDC utilises these general guidelines when deciding on the suitability of resources for  
81 publication, as well as incorporating some practical considerations, such as the volume of  
82 the data to be published and whether suitable supporting documentation can be provided.

83

84 Further, a decision needs to be made regarding whether raw or derived values should be  
85 published. Generally, raw values are preferred, as this enables new users to interpret the  
86 data without introducing bias from the data producers' own analysis. However, sometimes  
87 data producers are only able or willing to publish derived values. Where this is the case,  
88 detailed supporting documents detailing how derived values were obtained must be provided  
89 alongside the data. The formats to be used for publishing the data should also be  
90 considered. Proprietary file formats have a greater likelihood of becoming obsolete over time  
91 than non-proprietary formats. Therefore, to ensure the longevity of the resource, non-  
92 proprietary formats should be used to make resources available.

93

94 Decisions also must be made regarding who should be able to access a resource, and how  
95 they will find it. In the UK, for most publicly-funded data, it is now a requirement, that the  
96 data are made publicly available following completion of data generation<sup>2,3</sup> [5]. This must be  
97 within a reasonable period of time, although NERC does sanction embargoes on release of

---

<sup>2</sup> <http://www.nerc.ac.uk/research/sites/data/policy/data-policy/>

<sup>3</sup> <http://www.rcuk.ac.uk/research/datapolicy/>

98 data in order to enable the researchers who generated the data to publish scientific papers  
99 based on their analyses<sup>2</sup>. Data centres should also provide searchable catalogues of their  
100 data holdings to enable users to find resources. If the records held in catalogues conform to  
101 metadata standards, they can be harvested by other catalogues. Being publicly available  
102 does not necessarily mean that end-users are entirely free to use data without limitations or  
103 crediting the data providers, as data centres frequently only make resources available under  
104 licence. Licence terms may include conditions regarding use of the data and also require  
105 users to cite the original creators of the resource.

106

107 One mechanism to enable the ability to refer to a data resource is the allocation of a Digital  
108 Object Identifier (DOI) to a resource. The EIDC uses DOIs to identify the data resources it  
109 holds, and this is discussed in greater detail below. The use of DOIs is not necessarily  
110 suitable for all datasets, and they are best used to represent static resources or ‘snapshots’  
111 of dynamic datasets. Citation of dynamic datasets is more problematical, and the EIDC has  
112 representation on, and has hosted, the Data Citation Working Group of the Research Data  
113 Alliance (RDA)<sup>4</sup> to attempt to provide long-term solutions to this problem. To enable other  
114 users who are unfamiliar with the data resource, to be able to use it, detailed supporting  
115 documents should be provided [6]. Supporting documents should cover specific areas,  
116 including how data are structured, the nature and units of the recorded values, how data  
117 were collected/analysed (including details of instrumentation used and calibration values)  
118 and any quality control measures employed. Not all of these areas will be relevant to every  
119 data resource. For example, biodiversity data may not require information on laboratory  
120 instrumentation, if none was used. The published resources will require a delivery  
121 mechanism that enables users to obtain a copy of the resource. As stated above, this will  
122 require users to agree to licensing conditions before they are granted access. Providers of

---

<sup>4</sup> <https://rd-alliance.org/>

123 data for publication need to be confident that the resource being made available contains the  
124 same data that they provided to the data centre, and similarly, users requesting data want to  
125 know that they are receiving uncorrupted data. To solve this problem, the EIDC uses  
126 checksums to verify the condition of the resources it holds - the mechanism for doing so is  
127 detailed in a subsequent section. Publishers are also required to comply with  
128 national/international legal requirements, such as the Infrastructure for Spatial Information in  
129 Europe (INSPIRE) European directive [7]. Ensuring that their data are published via  
130 recognised data centres relieves data originators of the responsibility to meet these  
131 conditions, which passes to the data centre when it becomes the custodian of the data  
132 resource. As an additional incentive to publish, an increasing number of journals require that  
133 data which underpin a research paper are deposited in a suitable data repository, so that  
134 users may access the data to verify the conclusions of the researchers. This has become of  
135 greater importance following incidents such as the Climatic Research Unit email controversy  
136 [8]. The data centre must take into account all of these considerations in developing robust  
137 processes and infrastructure to enable publication of environmental data.

138

### 139 3.0 The Infrastructure

140 To enable the publication of high quality, reusable environmental data, it is crucial that a  
141 stable, defined infrastructure is in place to provide the various required services. Detailed  
142 below are the components of the infrastructure assembled by the EIDC to enable publication  
143 of data submitted to the data centre.

144

#### 145 3.1 Tracking System

146 All work to be undertaken by the data centre is captured by an issue tracking system. The  
147 EIDC uses JIRA from Atlassian<sup>5</sup> to manage its workload. JIRA delivers an extremely flexible  
148 task management and work allocation system. It provides creation of custom dashboards,  
149 allowing users to create their own view of the issues within the system, or to share a pre-  
150 existing dashboard so that data centre staff can all work from a standard view of the issues  
151 when required. Further, a range of standard and bespoke issue types can be created and  
152 progressed through a configurable status workflow. This enables users to quickly identify  
153 what type of work an issue describes and how far particular issues have progressed within  
154 the workflow. The tracking system provides an audit trail of comments from users conducting  
155 the work on an issue and is also able to record time spent working on individual issues, thus  
156 enabling management and reporting of human resources. Issues can be passed easily  
157 between colleagues for individuals to carry out specific parts of the publishing workflow.  
158 JIRA is also configured to send and receive emails to notify users of changes to issues.  
159 Export of data from JIRA is possible, in a range of non-proprietary formats such as XML or  
160 HTML. This means that if in future the EIDC were to switch to use an alternative issue  
161 tracking system, the audit trail of work undertaken would be retained. Exported data could be  
162 imported to a new system, or compressed and stored for long-term storage if it was decided  
163 that immediate access was not required.

164

### 165 3.2 Content Management System (CMS)

166 The EIDC uses a CMS in a number of crucial roles. First, an administrative area is required,  
167 for keeping all official data centre documentation, such as the standard processes followed  
168 by data centre staff, the checklists used for quality assurance and documentation relating to  
169 ingestion of data resources, such as Service Agreements. The CMS also contains  
170 inventories for data, web services and DOIs the EIDC has issued, and also contains a

---

<sup>5</sup> <https://www.atlassian.com/software/jira>



171 Licence Store for storage of copies of the licences to be used when users are placing orders  
172 for copies of resources. The administrative area is only viewable by data centre staff, and  
173 requires users to sign in. The remainder of the CMS is used as the data centre's website,  
174 and is publicly available<sup>6</sup>. These public facing pages contain information about the data  
175 resources held by the data centre, including supporting documents available to assist users  
176 in re-use of the data, as well as information on the services provided to people wishing to  
177 deposit their data with the EIDC. The CMS that the data centre has selected to fulfil these  
178 purposes is Plone<sup>7</sup>, which is freely available and Open Source. Export of content from Plone  
179 is possible, thus enabling all existing content to be imported to a new CMS should the need  
180 to use an alternative product arise in future. There would therefore be no loss of the audit  
181 trail.

182

### 183 3.3 The Data Store

184 The EIDC needs secure storage locations to hold the data it is responsible for. Data  
185 deposited with the data centre is stored primarily in two places: the file store and the spatial  
186 database. The file store contains both a staging area, for deposits which haven't been  
187 checked against the EIDC's standard acceptance checks, and an area for accepted data  
188 resources which have successfully passed the checks. Everything stored in the file store is  
189 backed up on a daily basis, so could be quickly retrieved if any resources were ever to be  
190 deleted in error. Spatial data, in addition to being stored in the file store, has a copy stored in  
191 the data centre's spatial database, which is a version of Oracle. This permits users ordering  
192 spatial data to select from a range of file formats, co-ordinate reference systems and  
193 coverages. As the EIDC is hosted by the Centre for Ecology and Hydrology (CEH), all data  
194 is stored on disk, using CEH's Storage Area Network (SAN). These are backed-up to tapes,

---

<sup>6</sup> <http://eidc.ceh.ac.uk/>

<sup>7</sup> <https://plone.org/>

195 stored on-site inside a fire safe daily, with further back-ups being stored in an off-site fire  
196 safe on a weekly basis.

197

### 198 3.4 Order Manager

199 The Order Manager is a bespoke java web application developed in-house by the EIDC. It  
200 allows users to order copies of files from the EIDC. In order to enable ordering of data  
201 resources, data centre staff must first configure the Order Manager with the relevant details.  
202 A key aspect of the Order Manager is that before an order can be placed, users must  
203 indicate their acceptance of the licensing conditions under which the resource is being made  
204 available. Licences for a resource are selected during configuration. For flat files, delivery of  
205 data resources is via an email to users, containing a link to download the file they have  
206 ordered. The download link is operational for 30 days. For spatial data, Order Manager  
207 operates in conjunction with the Feature Manipulation Engine (FME), a proprietary piece of  
208 software from Safe Software<sup>8</sup>, allowing creation of workflows for data manipulation. Using  
209 FME alongside Order Manager allows users to select the file format, co-ordinate reference  
210 system and coverage they want when they place their order for data. This is particularly  
211 helpful for large datasets, where download of the whole resource may take hours. The ability  
212 to select file formats and co-ordinate reference systems also facilitates interoperability  
213 between disparate data resources, and hence data re-use. For users to be able to place  
214 orders for data using Order Manager, they must first register with the EIDC. This consists of  
215 simply providing an email address, a password and a display name. This information is used  
216 only to provide an email address to which the data centre can send emails containing  
217 download links for any resources ordered and to create an account so that users can review  
218 the history of any orders they have placed. The history includes details of any polygons used  
219 for subsetting the data, time periods, spatial reference system and file formats, so that users

---

<sup>8</sup> <http://www.safe.com/>

220 can recreate an order if required, and details of the licensing conditions under which the  
221 order was agreed. The EIDC does not use the information provided for any other purpose, or  
222 forward users' details to any other parties.

223

### 224 3.5 Catalogue

225 The EIDC has a catalogue<sup>9</sup>, containing discovery metadata records for the resources it  
226 curates. The catalogue is another bespoke java web application created specifically for use  
227 by the data centre. It contains a metadata editor, permitting data centre staff to create  
228 metadata records and verify them against a selected metadata standard, such as GEMINI  
229 2.2 [9], (a UK discovery metadata standard compatible with INSPIRE [10]), or ISO 19115  
230 [11], meaning the metadata records contained in our catalogue are compatible with those  
231 contained in other data catalogues, and can therefore be harvested by other catalogues as  
232 described below. Users can search the catalogue by entering search terms, selecting facets,  
233 spatial search, or any combination of these methods. Metadata records are presented as  
234 human-readable HTML web pages, with DCAT [12] compliant XML or JSON representations  
235 also being available if required. In addition, the catalogue is available as a Web Accessible  
236 Folder (WAF) containing GEMINI XML records for the EIDC's published resources, which  
237 can be accessed by other data catalogues in order to harvest the records, such as NERC's  
238 data catalogue service<sup>10</sup> and the UK Government's data portal<sup>11</sup>, whose records in turn can  
239 be harvested by other portals, such as the European Union's INSPIRE geoportal<sup>12</sup>. This  
240 ensures that simply by publishing a record publicly via the EIDC's catalogue, the resource  
241 will be discoverable by a much larger user community than would otherwise be possible if it  
242 were published in only a single catalogue (Fig. 1). The vast majority of metadata records  
243 held by the data centre are viewable by the public, because depositors of resources want

---

<sup>9</sup> <https://catalogue.ceh.ac.uk>

<sup>10</sup> <http://data-search.nerc.ac.uk/>

<sup>11</sup> <https://data.gov.uk/>

<sup>12</sup> <http://inspire-geoportal.ec.europa.eu/>

244 their data to be discoverable, because this promotes its re-use and therefore the likelihood  
245 that they will gain credit for creation of the data resources. It is also a requirement for issue  
246 of a Digital Object Identifier (DOI) that a publicly available metadata landing page for the  
247 DOI, is available. Issue of DOIs by the EIDC is discussed below. However, the design of the  
248 catalogue also allows users registering with the data centre to be assigned to specific  
249 groups, and as such, it is possible to create catalogue records for resources which are  
250 restricted to specific groups of users. This feature helps in facilitating work between different  
251 academic institutions, or groups within an institution.

252

253

#### 254 4.0 The Publishing Workflow

255 All data resources submitted for publication by the EIDC pass through the same, proven  
256 workflow (Fig. 2), developed to provide solutions to the issues outlined above. Many of the  
257 elements of the workflow developed by the EIDC have parallels within the Curation Lifecycle  
258 Model proposed by the Digital Curation Centre (DCC)<sup>13</sup>, though not necessarily performed in  
259 the same order. The EIDC is also gradually adding to the list of services it can provide,  
260 though most of the transformation services offered are currently only available for spatial  
261 data. The process by which resources are transferred from the researchers who generated  
262 the data to the EIDC is termed 'ingestion'. Any resources which the data centre publishes  
263 will therefore have been ingested by the EIDC prior to their publication. The majority of the  
264 data centre's data holdings are datasets, but models, web services and other data-related  
265 applications are also considered for curation. All processes used by the EIDC as part of the  
266 ingestion workflow have been designed to be as generic as possible, using general names  
267 for infrastructure components, rather than specific names of applications (e.g. tracking  
268 system rather than JIRA). This was done to make the processes as 'future-proof' as

---

<sup>13</sup> <http://www.dcc.ac.uk/resources/curation-lifecycle-model>

269 possible, meaning if an infrastructure component changes, it does not necessitate alterations  
270 to the processes.

271

#### 272 4.1 Identification

273 The point of entry to the workflow is identical for all data resources submitted to the data  
274 centre – identification of the resource to be published. An initial discussion is held via phone,  
275 email or in person, with depositors of the resource to ascertain exactly what the resource  
276 constitutes, including the current file format, number of files the resource consists of and  
277 resource type (dataset or model). The EIDC has a list of file formats that it prefers to accept  
278 for data resources, and will enter into a dialogue with the depositor to determine the most  
279 appropriate format in which to make the data resource available. Wherever possible, non-  
280 proprietary formats are preferred e.g. csv files over MS Excel spreadsheets, due to their  
281 longevity and their facilitation of interoperability. However, the data centre is always willing to  
282 work with depositors of data who can make a strong case as to why a resource should be  
283 made available in a specific format, rather than one of the EIDC's preferred formats.

284 Regardless of the format selected, the EIDC makes an annual review of the file formats it  
285 holds data in. Should the data centre become aware of changes in the availability of certain  
286 formats outside of the review window, it would take steps to ensure the currency of the file  
287 formats it uses for data storage. Every resource is assessed against standard criteria,  
288 including whether the data are replacing/adding to an existing published resource held by  
289 the EIDC and whether the EIDC is the most appropriate data centre for hosting of the data,  
290 as NERC currently supports six other domain specific data centres besides the EIDC.

291 Assessments are also made regarding whether the data are unique (no other copies are  
292 published elsewhere), repeatable (they could be regenerated), underpin a published peer-  
293 reviewed paper, and can be provided with sufficient supporting documentation to be re-  
294 usable by non-domain specialists. Consideration is also made for the volume of the

295 resource, as large resources may incur a charge for their curation, although this is not the  
296 primary criterion used for assessment of suitability.

297

298 If, after this assessment, the resource is considered to be suitable for deposit, the depositor  
299 is notified of the positive identification outcome and the request for deposit becomes a full  
300 ingestion 'job' in the EIDC's tracking system. The ingestion job is assigned to a member of  
301 data centre staff who will manage the ingestion of the resource/s to the data centre, ensuring  
302 that all appropriate tasks are completed.

303

304 For resources that are deemed unsuitable for deposit, the depositor is notified of the  
305 outcome and the reasons why. If it is considered that the data being offered for deposit  
306 would be more suitable for deposit at one of NERC's other data centres, then the depositor  
307 is advised to contact the relevant data centre. No further action is taken, unless the depositor  
308 disagrees with the reasons given for rejection of the resource, in which case the issue is  
309 referred to the manager of data centre operations, who will consider the case and make a  
310 final decision.

311

## 312 4.2 Ingestion Management

313 Ingestion Management is the process whereby the tasks required to ingest the data resource  
314 to the EIDC are controlled. The individual responsible for completion of ingestion  
315 management is designated the 'Ingestion Manager'. Ingestion Managers are responsible for  
316 ensuring that all the tasks required for ingestion and subsequent curation of the data are  
317 performed successfully, and that they are undertaken in the correct order. The first task for  
318 the Ingestion Manager is to review the information collected during Identification. They will  
319 then create tasks in the tracking system to manage the ingestion of resources, the first of

320 which is 'Preparation', with one task being created for each identified resource. Once a  
321 Preparation task is complete, it is the Ingestion Manager's job to quality assure the work.  
322 This is achieved by completing a checklist to confirm that critical actions have been  
323 completed appropriately. If the work undertaken is satisfactory, the Ingestion Manager will  
324 then create tasks for 'Data Transfer', 'Data Storage', 'Online Ordering', 'Publication' and, if  
325 required, 'DOI Minting'. The objectives of these tasks are detailed below. As with the  
326 Preparation task, the Ingestion Manager assures all work undertaken in these tasks by  
327 completing quality checklists. Completed checklists are stored in the administrative area of  
328 the CMS, thus providing an audit trail of quality checks for each resource ingested by the  
329 data centre.

330

#### 331 4.3 Preparation

332 Every resource which is to be ingested to the EIDC will have a Preparation task created for  
333 it, the primary purpose of which is to create a document called the Service Agreement (SA)  
334 via liaison with the depositor of the data resource. The SA is critical to the whole process of  
335 ingestion, as it clearly defines what services the data depositor can expect from the EIDC  
336 and similarly, details of the resource and supporting information that the data centre can  
337 expect from the depositor. A completed SA will include a definitive title for the resource, the  
338 file format/s in which it will be provided, the data volume, details of supporting documents,  
339 licensing information and whether an embargo on the availability of the resource and  
340 supporting documents is required. The supporting documentation is required to enable re-  
341 use of the data and provide details of the resource's provenance – a list of the topics about  
342 which information should be supplied is provided by NERC [4]. Both the data resource itself  
343 and the supporting documentation are, in isolation, of limited use, but when used together,  
344 should provide data which can be used without further recourse to the generator of the data.  
345 As with the data resource itself, supporting documents should be provided in non-proprietary  
346 formats, as this will help to ensure the currency of the documents and facilitate their use by

347 parties wishing to utilise data resources. The licence stipulates the conditions under which  
348 the data may be accessed and used. Most of the data resources held by the EIDC are made  
349 available under the UK Open Government Licence (OGL)<sup>14</sup>, in-line with NERC guidance [4].  
350 Sometimes depositors and/or funders require an alternative licence to be used, though  
351 depositors are advised that the EIDC's default position is to make resources available under  
352 the OGL unless there are valid reasons not to do so. This is easily accommodated, but  
353 depositors must liaise with the EIDC's data licensing team to ensure that the alternative  
354 licence is acceptable, and a copy of the licence is provided and added to the licence store of  
355 the data centre's CMS. The SA also captures the details of whether a DOI is required by the  
356 depositor and the authors of the resource, to enable citation of the resource. It also identifies  
357 whether the resource is covered by the INSPIRE (Infrastructure for Spatial Information in  
358 Europe) directive, designed to enable interoperability between European spatial datasets [7],  
359 and if so, by which theme it is covered. The data centre staff will negotiate a date for transfer  
360 of the resource to the EIDC and discuss what type of data is being provided: raw data or  
361 derived values. Ideally, raw data is preferred, to allow different users to analyse the data  
362 using their preferred methods without any existing bias. However, in some instances only  
363 derived values are provided, and where this is the case, the data centre strives to ensure  
364 that the supporting documentation contains details of how derived values were obtained  
365 from raw values. An area for the resource is created in the EIDC's CMS to store documents,  
366 including a 'Private' folder for administrative documents relating to the ingestion and a  
367 'Public' area for holding supporting documents for the data resource. An incomplete 'stub'  
368 entry is created in the data centre's data catalogue to enable recording of discovery  
369 metadata, including details of the provenance of the resource via the 'lineage' statement.  
370 The initial, draft version of the SA is checked by the Ingestion Manager to ensure the content  
371 is appropriate, before being sent to the depositor for their agreement. If satisfied with the

---

<sup>14</sup> <https://www.nationalarchives.gov.uk/doc/open-government-licence/>



372 details, the depositor emails the data centre to confirm their agreement, and the ingestion of  
373 the data resource can proceed.

374

#### 375 4.4 Data Transfer

376 The Data Transfer task follows that of Preparation. The objective of Data Transfer is to  
377 ensure the transfer of the data resource and all supporting documents from the depositor to  
378 the EIDC. This can occur via several methods, though the most common route for transfer is  
379 by email to the data centre's email account. This generates a notification in the tracking  
380 system to advise the data centre that the transfer has occurred. Alternative means of  
381 transfer, often employed for resources too large for email transfer, can include ftp or, very  
382 rarely, even via physical media (hard-drive or DVD) sent in the mail. On receipt of the data  
383 resource, the depositor is sent a 'Goods Received Note' (GRN) to indicate that the data have  
384 been received. The data are moved to the data centre's staging area – a folder in the  
385 filestore, which is backed up on a daily basis. The resource is also checksummed, with the  
386 resulting checksum being sent to the depositor. The primary reason for checksum creation is  
387 to provide the depositor with the opportunity to verify that the correct resource has been  
388 received by the data centre, and no corruption of files has occurred during transit. The  
389 checksum also permits data centre staff to move the resource between locations and quickly  
390 verify that no alterations to the resource have occurred. During Data Transfer, the 'stub'  
391 discovery metadata record is completed for the resource and validated against metadata  
392 standards. This will enable users to find the resource by searching the data centre's  
393 catalogue. An entry for each transferred resource is created in the Data Inventory, logging  
394 exactly what the resource is and its current location. Some basic 'Resource Acceptance  
395 Checks' are then performed on the resource to ensure that the data centre are satisfied that  
396 the resource is appropriate. These include checks that the resource name, format and size  
397 match that agreed in the SA, the resource opens using an industry standard application and  
398 contains the correct type of data. If these are passed, the task is passed back to the

399 Ingestion Manager for quality assurance, who will also send a 'Data Deposit Completion  
400 Notice' (DDCN) to the depositor, informing them that the deposit meets the agreed criteria.  
401 This ends the stage of resource deposit involving input from the depositor - all other steps  
402 will now completed solely by data centre staff, although the depositor will be notified when  
403 key milestones are reached.

404

#### 405 4.5 Data Storage

406 Following successful completion of Data Transfer, the Ingestion Manager will assign a Data  
407 Storage task to a member of the data centre staff. The EIDC's data store is regularly  
408 backed-up, but recovery from accidental deletions is time-consuming, so for security issues,  
409 the number of staff able to access the data store (and therefore complete Data Storage  
410 tasks) is limited. The resource will be located using the location stored in the Data Inventory,  
411 and moved to the data store. The checksum is verified to ensure no corruption has occurred  
412 to the file during the move, and the location of the resource is updated in the Data Inventory.  
413 Further, if the resource is in a spatial data format, such as personal geodatabase or  
414 shapefile, a copy is added to the data centre's spatial data store. This permits the data to be  
415 sliced by location, and also to be used in Web Services if required. Where appropriate, the  
416 data centre may also store extremely large datasets consisting of multiple files on an ftp site,  
417 which permits users who have requested the access details from the data centre to  
418 download individual files quickly, as opposed to attempting to download one extremely large  
419 file. On completion, the task is quality assured by the Ingestion Manager.

420

#### 421 4.6 Publication

422 Publication tasks cover the publication of one or more data centre objects, such as a  
423 metadata catalogue record for a data resource (which also functions as the landing page for  
424 a DOI), supporting documentation, or web services, such as Web Map Services (WMS). The

425 Ingestion Manager will specify exactly which resources are to be published, to what  
426 audience (public or a specified group, as detailed in the Service Agreement) and the date for  
427 publication. Many of the publication dates for data centre resources are determined by  
428 embargo, which is a period between transfer of a resource to the data centre, and the date  
429 of its public availability, during which time the depositor of a data resource has opportunity to  
430 make use of the data. Embargoes typically last up to two years after the last data of data  
431 generation, though can be shortened on instruction from the depositor for any reason, for  
432 example to coincide with the publication of an academic paper. Timing of publication is also  
433 dependent on whether the depositor of the resource has requested a DOI for their resource,  
434 in order to enable other users to cite it. If a DOI has been requested, then the landing page  
435 for the data resource is required to be publicly available prior to issue of the DOI. In this  
436 instance, the landing page is made available to the public, but the data resource itself is not,  
437 in order to ensure that all users are only able to access the resource once the mechanism to  
438 enable its citation is in place. However, if no DOI is requested, then publication of the  
439 discovery metadata record does not occur until after the resource has been made publicly  
440 available, via the process of 'Online Ordering', detailed below. On completion of the task, the  
441 work undertaken is quality assured, and a 'Publication Notice' is sent to the depositor,  
442 notifying them that publication has now occurred.

443

#### 444 4.7 Online Ordering

445 Online Ordering is the process whereby a data resource is made available so that users can  
446 order a copy, by clicking a link in the discovery metadata record for the resource. This is  
447 achieved by configuring the 'Order Manager' application, a component of the EIDC's  
448 infrastructure. Configuration involves specifying what type of resource is to be made  
449 available (flat file or spatial data), the licences which users placing an order for the data must  
450 agree to, name of the file to be delivered and, if it is spatial data, any specific options  
451 requested, such as user choice of file format and coverage required. Once this has been

452 successfully completed and tested, the discovery metadata record held in the data centre  
453 catalogue is updated to enable users to order a copy of the resource. If an embargo has  
454 been requested by a depositor, Order Manager will not be configured until expiry of the  
455 embargo period. In the interim, users attempting to order a copy of the data are instead  
456 directed to the data centre's 'embargo' page, which explains the reasons why the resource is  
457 not currently available. As with other tasks, the completed work is quality assured by the  
458 Ingestion Manager.

459

#### 460 4.8 Assign DOI

461 The process for assigning a DOI to a data resource is undertaken only for those where the  
462 depositor has requested a DOI for their deposited resource. The required information (list of  
463 authors, title and publication year) is extracted from the SA and entered into the discovery  
464 metadata record, if not already present. The data centre staff member undertaking the work  
465 clicks a button in the catalogue record to create an XML document in DataCite's required  
466 schema [11]. This is automatically sent to DataCite's Application Programming Interface  
467 (API), which mints the DOI. Details of the DOI are automatically entered into the discovery  
468 metadata record, which becomes the landing page for the DOI. An entry is created in the  
469 'DOI Inventory' area of the data centre's CMS, thus allowing the data centre to track all DOIs  
470 it has issued. The depositor is then sent a 'DOI Issued Notification' email, informing them  
471 that the DOI has been issued and explaining how to use the DOI to cite the resource. The  
472 work is subsequently quality assured by the Ingestion Manager. The EIDC strongly advises  
473 depositors to obtain a DOI for their deposited resource to enable its citation, but does not  
474 mandate it. Minting of DOIs is not free and there is a small, but real, financial cost to the data  
475 centre for their issue. For a small minority of depositors, there may be valid reasons why  
476 they do not wish to obtain a DOI. For example, users may wish to deposit an early version of  
477 a resource for sharing with a specific group of users, knowing in advance that the resource  
478 may be subject to change, or will be replaced after a period of time. Once a DOI has been

479 issued, the EIDC will continue to make the resource that the DOI has been assigned to  
480 publicly available, even if this is only via email request. This is because the data centre  
481 believes that where a data resource has been made available to be used and cited in a  
482 piece of research, then that exact same resource should be available for anyone wishing to  
483 replicate or verify the results of the study. By not obtaining a DOI, the EIDC does not commit  
484 to continuing to make a resource available and so the data centre is able to replace or  
485 withdraw a dataset without maintaining access to it. For data resources which do not have a  
486 DOI, individual resources can be identified using a unique identifier which all resources are  
487 assigned when they enter the data centre, though this should not be considered a substitute  
488 for a DOI. Users are able to cite the URL of the data catalogue entry for a resource, though  
489 should be aware that the EIDC has no responsibility to maintain this in perpetuity. As such, if  
490 citation of the resource is important to depositors, then they would be advised to obtain a  
491 DOI.

492

#### 493 4.9 Managing Series

494 Some data resources form part of a series, for example where a new year of data has been  
495 generated. Where this is the case, the discovery metadata records are collected together as  
496 child records of a Series record, thus enabling a user to quickly identify all related datasets.  
497 This approach can also be used to relate a series of versions of a data resource, such as  
498 models, which may undergo several iterations during their lifetime. This is achieved via  
499 creation of a 'Manage Series' task by the Ingestion Manager. The member of staff assigned  
500 to complete this task must ensure that the Series record complies with the relevant metadata  
501 standard, and that all required child records are associated with it. This work is then quality  
502 assured by the Ingestion Manager.

503

#### 504 5.0 Service Management

505 Creation of Web Services, such as WMS, are managed in a similar manner to the ingestion  
506 of data resources. A 'job' is created in the data centre's tracking system, which enables the  
507 Service Manager to co-ordinate the activities required to create and publish a web service.  
508 This consists of creating a 'Web Service Creation' task, to oversee the production of the  
509 service, and a 'Publication' task, as described above, to enable publication of the service.

510

## 511 5.1 Web Service Creation

512 The service manager assigns the task for creation of a view service to a member of the  
513 EIDC staff with the required technical skills. They will create a conceptual design for the  
514 service. Where possible, this is reviewed with the original depositor of the resource to ensure  
515 they are satisfied with the representation of the data. The service is then created, the  
516 technical details of which are not discussed here. As with datasets, a discovery metadata  
517 record for the service is created in the EIDC's data catalogue, to enable users to find the  
518 service. An entry for the service is also created in the Service Inventory of the CMS to act as  
519 a record of services for which the EIDC has responsibility. The service is then thoroughly  
520 tested, prior to publication. The Service Manager quality assures the finished product before  
521 its release.

522

523

## 524 6.0 Conclusions

525 The field of data publication is not as straightforward as it may at first appear, but as the  
526 areas detailed above have demonstrated, many of these issues can be resolved through a  
527 combination of constructing the publication workflow correctly and utilising a robust and  
528 stable infrastructure for publication. This is evidenced by the successful publication of over  
529 300 datasets, over 200 DOIs issued, and 20 web services published, all using the workflow

530 and infrastructure detailed above. The EIDC has also been recognised as an accepted  
531 repository for data by the British Ecological Society, the Nature Publishing Group and the  
532 Earth System Science Data journal. It has been shown that many researchers' primary  
533 concern over data publication is failure to receive credit for their work [2]. The workflow and  
534 infrastructure utilised by the EIDC has therefore enabled producers of environmental data to  
535 publish the data they have generated in the public domain, safe in the knowledge that the  
536 data are secure and that, by ensuring the data are citeable, they will receive credit for their  
537 work. The EIDC has witnessed an increase in the number of requests to deposit, and a  
538 corresponding increase in the number of published data resources. For the financial year  
539 2013-2014, 35 deposit requests were made, increasing to 83 for the year 2014-2015. Not all  
540 of these requests were granted, but the same time period saw an increase in the number of  
541 resources published from 25 in 2013-2014 to 92 in 2014-2015. Based on figures for the first  
542 half of 2015-2016, the total requests and published resources this year will exceed those in  
543 previous years. Dealing with this increase in both requests and published resources can  
544 easily be accommodated by the infrastructure and workflow that the EIDC has put in place,  
545 with the primary limit on processing of deposit requests being resource.

546

547 Even so, there are still some outstanding issues which remain. No citation mechanism for  
548 fluid datasets, where the content is updated regularly, but users wish to always cite the most  
549 recent version of the dataset currently exists, or to cite only a specific subset of a dynamic  
550 data resource [13]. This problem is recognised within the data publishing community, but so  
551 far no robust solution has been determined. Duerr et al [14] reviewed many of the different  
552 available identification schemes, and recognised one of the key criteria in using identifiers is  
553 that users want to know they are referring to the exact same dataset as other users who  
554 have cited the resource, but also acknowledged that resources, such as time-series, can be  
555 subject to alterations. Whilst many of the identifiers reviewed were capable of identifying a  
556 unique resource, none was able to provide an identifier for a resource in a state of flux. The

557 data centre currently adopts a policy of directing users to access the most recent version of  
558 updated datasets in the discovery metadata, and only providing offline access to deprecated  
559 resources. This is far from ideal, and the EIDC continues to be involved with the Data  
560 Citation Working Group of the RDA to attempt to provide a practical solution to this problem.  
561 There are also pressures to provide a better experience for users, in terms of ease of use  
562 and greater flexibility in terms of issuing data. Currently, flat files from the data centre can be  
563 ordered only in the format in which they were deposited. Users ordering a copy of spatial  
564 data do have the ability to select from a range of formats and co-ordinate reference systems  
565 when placing an order, provided that the depositor of the data has not specified otherwise in  
566 their SA, and can also select the spatial coverage they are interested in. However, users are  
567 unable to slice the data by time period, meaning that they must frequently order the whole  
568 dataset. This can present problems if the file to be downloaded by the end user is  
569 particularly large, when the required time for complete download can take hours, depending  
570 on internet connection speed. For exceptionally large data resources, approaching a  
571 terabyte in volume, the data centre has made them available from a secure ftp site, to which  
572 registered users can request access. This in itself is problematical, given that no direct  
573 metric of data downloads can be provided – a useful statistic when attempting to measure  
574 impact of a data resource. However, to resolve this issue, the data centre is working on  
575 providing a gridded data store as part of its infrastructure. This would allow users to place  
576 orders for datasets, slicing by time and/or location if desired. The EIDC also undertakes  
577 regular reviews of its processes, and where improvements in efficiency are identified, these  
578 are rapidly incorporated into the current processes.

579

580 Many areas of business, government and research are data driven, so it is clear that in  
581 future, the area of data publication is one that will only become of increasing importance.  
582 Whilst this should be regarded as good news, given that it will ensure data publication is  
583 always treated seriously and should be funded accordingly, it is important to recognise that



584 the challenges faced by data publishers will only grow too. Larger volumes of data are now  
585 being generated more quickly than ever before [3] and therefore the issue of identifying what  
586 to publish and how is becoming ever more acute.

587

## 588 7.0 References

589 [1] Callaghan S, Donegan S, Pepler S, Thorley M, Cunningham N, Kirsch P, Ault L, Bell P,  
590 Bowie R, Leadbetter A, Lowry R, Moncoiffe G, Harrison K, Smith-Haddon B, Weatherby A,  
591 Wright D (2012) Making data a first class scientific output: data citation and publication by  
592 NERC's environmental data centres. *The International Journal of Digital Curation* 7:107-113  
593 doi: 10.2218/ijdc.v7il.218

594 [2] Lawrence B, Jones C, Matthews B, Pepler S, Callaghan S (2011) Citation and peer  
595 review of data: moving towards formal data publication. *The International Journal of Digital*  
596 *Curation* 2:4-37

597 [3] Committee on Archiving and Accessing, Board on Atmospheric Sciences and Climate,  
598 Division on Earth and Life Studies, National Research Council (2007) *Environmental data*  
599 *management at NOAA: archiving, stewardship, and access*. National Academies Press,  
600 Washington D.C.

601 [4] Thorley M (2012) NERC data policy – guidance notes. Natural Environment Research  
602 Council, Swindon. <http://www.nerc.ac.uk/research/sites/data/policy/datapolicy-guidance/>.  
603 Accessed 12 November 2015

604 [5] Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen L, Moorman D,  
605 Uhlir P, Wouters P (2004) Promoting access to public research data for scientific, economic  
606 and social development. *Data Science Journal* 29:135-152

607 [6] Kratz JE, Strasser C (2015) Researcher perspectives on publication and peer review of  
608 data. *PLoS ONE* 10:e0117619 doi:10.1371/journal.pone.0117619

609 [7] European Commission (2007) Directive 2007/2/EC of the European parliament and of the  
610 council of 14 March 2007 establishing an infrastructure for spatial information in the  
611 European community (INSPIRE). Official journal of the European Union 108:1-14

612 [8] Holliman R (2011) Advocacy in the tail: exploring the implications of 'climategate' for  
613 science journalism and public debate in the digital age. Journalism 12:832-846

614 [9] Association for Geographic Information (2012) UK GEMINI: Specification for discovery  
615 metadata for geospatial data resources v2.2. Association for Geographic Information

616 [10] European Commission Joint Research Centre (2013) INSPIRE Metadata implementing  
617 rules: Technical guidelines based on EN IS) 19115 and EN ISO 19119 v1.3. European  
618 Commission Joint Research Centre.  
619 [http://inspire.ec.europa.eu/documents/Metadata/MD\\_IR\\_and\\_ISO\\_20131029.pdf](http://inspire.ec.europa.eu/documents/Metadata/MD_IR_and_ISO_20131029.pdf). Accessed  
620 12 November 2015

621 [11] Technical Committee ISO/TC 211, Geographic information/Geomatics (2003) EN ISO  
622 19115:2003 Geographic information – Metadata. ISO

623 [12] Maali F, Erickson J (2014) Data Catalog Vocabulary (DCAT) W3C Recommendation.  
624 <https://www.w3.org/TR/vocab-dcat/>. Accessed 15 March 2016

625 [13] Pröll S, Rauber A (2014) A scalable framework for dynamic data citation of arbitrary  
626 structured data. 3rd International conference on data management technologies and  
627 applications (DATA2014)

628 [14] Duerr RE, Downs RR, Tilmes C, Barkstrom B, Lenhardt WC, Glassy J, Bermudez LE,  
629 Slaughter P (2011) On the utility of identification schemes for digital earth science data: an  
630 assessment and recommendations. Earth Science Informatics 4: 139-60 doi:  
631 10.1007/s12145-011-0083-6.

632

633

634 Figure captions

635 **Fig. 1** Illustrating how a discovery metadata record from the EIDC's data catalogue, (on the  
636 left), has been harvested by three other data portals: the NERC data catalogue service, UK  
637 Government's data portal and the European Union's INSPIRE geoportal.

638

639 **Fig. 2** A diagram of the publishing workflow designed by the EIDC.



Figure

