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

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

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

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

Corresponding Author:

Email: dgwr@ceh.ac.uk

Acknowledgements

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

Abstract

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

Keywords

data, publication workflow, infrastructure, data centre

1.0 Introduction

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

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

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

2.0 Issues in Data Publication

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

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

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

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

[2] http://www.nerc.ac.uk/research/sites/data/policy/data-policy/
[3] http://www.rcuk.ac.uk/research/datapolicy/

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

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

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

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

3.0 The Infrastructure

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

3.1 Tracking System

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

3.2 Content Management System (CMS)

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

[5] https://www.atlassian.com/software/jira

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

3.3 The Data Store

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

[6] http://eidc.ceh.ac.uk/
[7] https://plone.org/

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

3.4 Order Manager

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

[8] 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[9], 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[10] and the UK Government's data portal[11], whose records in turn can
239 be harvested by other portals, such as the European Union's INSPIRE geoportal[12]. 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

[9] https://catalogue.ceh.ac.uk
[10] http://data-search.nerc.ac.uk/
[11] https://data.gov.uk/
[12] http://inspire-geoportal.ec.europa.eu/

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

4.0 The Publishing Workflow

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

[13] http://www.dcc.ac.uk/resources/curation-lifecycle-model

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

4.1 Identification

The point of entry to the workflow is identical for all data resources submitted to the data centre – identification of the resource to be published. An initial discussion is held via phone, email or in person, with depositors of the resource to ascertain exactly what the resource constitutes, including the current file format, number of files the resource consists of and resource type (dataset or model). The EIDC has a list of file formats that it prefers to accept for data resources, and will enter into a dialogue with the depositor to determine the most appropriate format in which to make the data resource available. Wherever possible, non-proprietary formats are preferred e.g. csv files over MS Excel spreadsheets, due to their longevity and their facilitation of interoperability. However, the data centre is always willing to work with depositors of data who can make a strong case as to why a resource should be made available in a specific format, rather than one of the EIDC's preferred formats. Regardless of the format selected, the EIDC makes an annual review of the file formats it holds data in. Should the data centre become aware of changes in the availability of certain formats outside of the review window, it would take steps to ensure the currency of the file formats it uses for data storage. Every resource is assessed against standard criteria, including whether the data are replacing/adding to an existing published resource held by the EIDC and whether the EIDC is the most appropriate data centre for hosting of the data, as NERC currently supports six other domain specific data centres besides the EIDC. Assessments are also made regarding whether the data are unique (no other copies are published elsewhere), repeatable (they could be regenerated), underpin a published peer-reviewed paper, and can be provided with sufficient supporting documentation to be re-usable by non-domain specialists. Consideration is also made for the volume of the

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

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

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

4.2 Ingestion Management

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

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

4.3 Preparation

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

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

[14] https://www.nationalarchives.gov.uk/doc/open-government-licence/

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

4.4 Data Transfer

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

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

4.5 Data Storage

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

4.6 Publication

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

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

4.7 Online Ordering

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

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

4.8 Assign DOI

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

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

4.9 Managing Series

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

5.0 Service Management

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

5.1 Web Service Creation

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

6.0 Conclusions

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

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

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

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

Many areas of business, government and research are data driven, so it is clear that in future, the area of data publication is one that will only become of increasing importance. Whilst this should be regarded as good news, given that it will ensure data publication is 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. 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

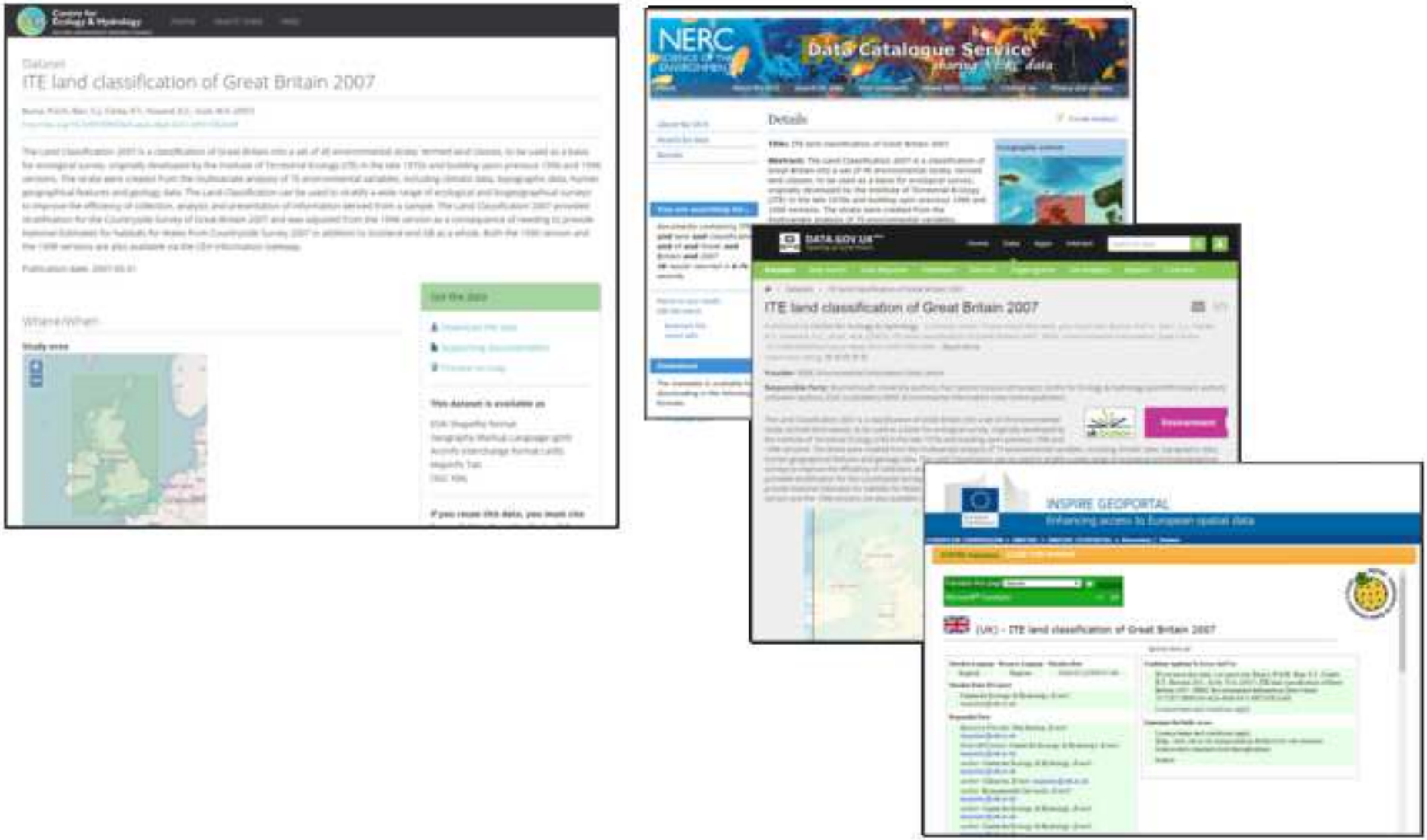
633

Figure captions

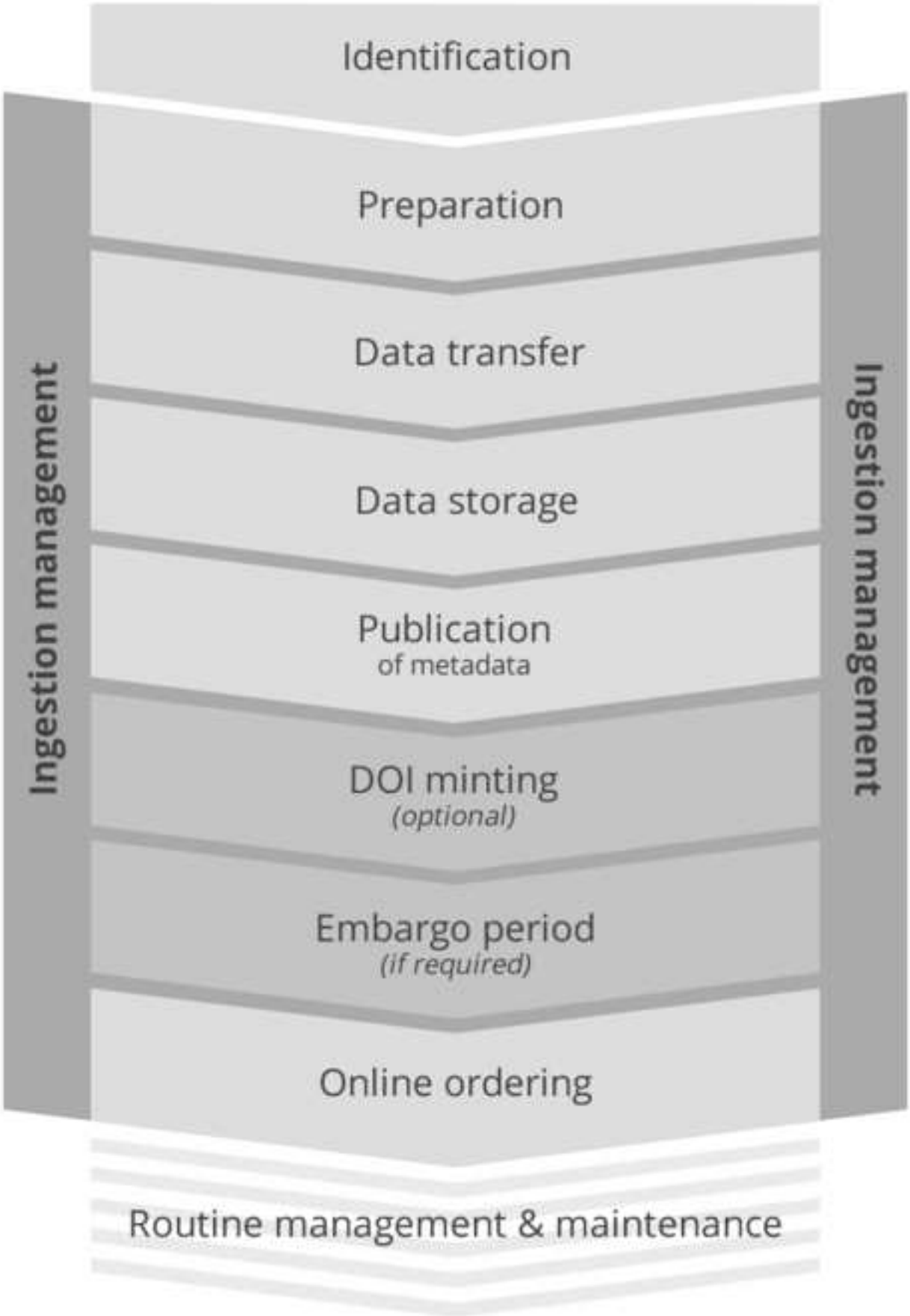
Fig. 1 Illustrating how a discovery metadata record from the EIDC's data catalogue, (on the left), has been harvested by three other data portals: the NERC data catalogue service, UK Government's data portal and the European Union's INSPIRE geoportal.

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

Figure


Centre for Ecology & Hydrology
Dataset
ITE land classification of Great Britain 2007
Study area
Get the data
This dataset is available as
NERC
Data Catalogue Service
Details
DATA.GOV.UK
ITE land classification of Great Britain 2007
INSPIRE GEOPORTAL
Enhancing access to European spatial data
(UK) - ITE land classification of Great Britain 2007

Figure


Identification
Preparation
Data transfer
Data storage
Publication
of metadata
DOI minting
(optional)
Embargo period
(if required)
Online ordering
Routine management & maintenance
Ingestion management
Ingestion management