Open data and digital morphology

Thomas G. Davies[1], Imran A. Rahman[1,2], Stephan Lautenschlager[1,3], John A. Cunningham[1], Robert J. Asher[4], Paul M. Barrett[5], Karl T. Bates[6], Stefan Bengtson[7], Roger B. J. Benson[8], Doug M. Boyer[9], José Braga[10,11], Jen A. Bright[12,13], Leon P.A.M. Claessens[14], Philip G. Cox[15], Xi-Ping Dong[16], Alistair R. Evans[17], Peter L. Falkingham[18], Matt Friedman[19], Russell J. Garwood[5,20], Anjali Goswami[21], John R. Hutchinson[22], Nathan S. Jeffery[6], Zerina Johanson[5], Renaud Lebrun[23], Carlos Martínez-Pérez[1,24], Jesús Marugán-Lobón[25], Paul M. O'Higgins[15], Brian Metscher[26], Maëva Orliac[23], Timothy B. Rowe[27], Martin Rücklin[1,28], Marcelo R. Sánchez-Villagra[29], Neil H. Shubin[30], Selena Y. Smith[19], J. Matthias Starck[31], Chris Stringer[5], Adam P. Summers[32], Mark D. Sutton[33], Stig A. Walsh[34], Vera Weisbecker[35], Lawrence M. Witmer[36], Stephen Wroe[37], Zongjun Yin[1,38], Emily J. Rayfield[1]* and Philip C.J. Donoghue[1]*

[1]School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

[2]Oxford University Museum of Natural History, Parks Road, Oxford OX1 3PW, UK

[3]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, UK

[4]Museum of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

[5]Department of Earth Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK

[6]Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, L7 8TX , UK

[7]Department of Palaeobiology, Swedish Museum of Natural History, Box 50007, SE-104 05 Stockholm, Sweden

[8]Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3AN, UK

[9]Department of Evolutionary Anthropology, Duke University, Box 90383, Biological Sciences Building, 130 Science Drive, Durham, NC 27708, USA

[10]Computer-assisted Palaeoanthropology Team, UMR 5288 CNRS-Université de Toulouse (Paul Sabatier), Toulouse, France

[11]Evolutionary Studies Institute, University of Witwatersrand, Johannesburg, South Africa

[12]School of Geosciences, University of South Florida, Tampa, FL 33620, USA

[13]Center for Virtualization and Applied Spatial Technologies, University of South Florida, Tampa, FL 33620, USA

[14]Department of Biology, College of the Holy Cross, Worcester, MA 01610, USA

[15]Department of Archaeology and Hull York Medical School, University of York, York, YO10 5DD, UK

[16]School of Earth and Space Science, Peking University, Beijing 100871, China
[17]School of Biological Sciences, Monash University, VIC 3800, Australia
[18]School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, UK
[19] Department of Earth & Environmental Sciences and Museum of Paleontology, University of Michigan, Ann Arbor, MI 48109, USA
[20]School of Earth and Environmental Sciences, University of Manchester, Manchester, M13 9PL, UK
[21]Department of Genetics, Evolution & Environment and Department of Earth Sciences, University College London, Gower Street, London SW17 7PL, UK
[22]Structure & Motion Lab, Department of Comparative Biomedical Sciences, The Royal Veterinary College, Hawkshead Lane, Hatfield, Hertfordshire AL9 7TA, UK
[23]Institut des Sciences de l'Evolution de Montpellier, CC64, Université de Montpellier, campus Triolet, Place Eugène Bataillon, 34095, Montpellier cedex 5, France
[24]Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de Valencia, 46980 Paterna (València), Spain.
[25]Unidad de Paleontología, Dpto. Biología. Universidad Autónoma de Madrid. 28049 Cantoblanco (Madrid), Spain.
[26]Department of Theoretical Biology, University of Vienna, Althanstrasse 14, 1090 Austria
[27]Jackson School of GeoSciences C1100, The University of Texas at Austin, Austin, Texas 78712, USA
[28]Naturalis Biodiversity Center, Postbus 9517, 2300 RA Leiden, The Netherlands
[29]Paläontologisches Institut und Museum der Universität Zürich, Karl Schmid Strasse 4, 8006 Zürich, Switzerland
[30]Department of Organismal Biology and Anatomy, University of Chicago, 1027 E. 57th Street, Chicago, IL 60637, USA
[31]Department of Biology II, Ludwig-Maximilians University Munich (LMU), Großhadernerstr. 2, D-82152 Planegg-Martinsried, Germany
[32]University of Washington, Friday Harbor Labs, Friday Harbor, WA 98250, USA
[33]Department of Earth Science and Engineering, Imperial College, London SW7 2AZ, UK
[34]National Museums Scotland, Chambers Street, Edinburgh, EH1 1JF, UK
[35]School of Biological Sciences, The University of Queensland, St. Lucia QLD 4072, Australia
[36]Department of Biomedical Sciences, Ohio University Heritage College of Osteopathic Medicine, Athens, Ohio, 45701 USA
[37]School of Environmental and Rural Science, University of New England, Armidale, NSW, Australia, 2351

[38]State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology, Chinese Academy of Sciences, Nanjing 210008, China

*Authors for correspondence: Emily J. Rayfield and Philip C. J. Donoghue

Abstract

Over the past two decades, the development of methods for visualizing and analysing specimens digitally, in three and even four dimensions, has transformed the study of living and fossil organisms. However, the initial promise, that the widespread application of such methods would facilitate access to the underlying digital data, has not been fully achieved. The underlying datasets for many published studies are not readily or freely available, introducing a barrier to verification and reproducibility, and the reuse of data. There is no current agreement or policy on the amount and type of data that should be made available alongside studies that use, and in some cases are wholly reliant on, digital morphology. Here, we propose a set of recommendations for minimum standards and additional best practice for 3D digital data publication, and review the issues around data storage, management and accessibility.

Keywords:
digital data, 3D models, phenotype, computed tomography, visualization, functional analysis

1. Introduction

Three-dimensional (3D) digital morphological data are commonly employed by palaeontologists and biologists in research. In palaeontology and anthropology, the widespread application of tomography (especially X-ray computed tomography, CT), laser and structured light scanning and photogrammetry, has revolutionized the study of morphology [1-4]. In biology, optical microscopy, magnetic resonance imaging (MRI) and contrast-enhanced CT are important tools for investigating soft-tissue anatomy [5-11]. The revolution brought about by these technologies has increased the amount and detail of anatomical information recovered from fossil and living organisms, transforming the nature of scientific enquiry in related fields (Figure 1). The resulting datasets are often reconstructed and presented as 3D digital models, which are themselves sometimes used in downstream analyses, including geometric morphometrics [12, 13], finite element analysis [14], multibody dynamics analysis [15], and computational fluid dynamics [16], thereby facilitating quantitative tests of functional and evolutionary hypotheses [3]. These types of studies have yielded important advances in our understanding of the anatomy of living and fossil organisms, e.g.,[10, 17, 18], as well as fundamental aspects of their biology, from feeding mode [19-21] to mobility [22, 23], development [24, 25], and physiology [26-28], as well as developments in taxonomic practise [29, 30]. Barriers to data sharing and access to specimens can be eroded because data exist as digital files that can be easily copied and readily distributed, allowing simultaneous analysis by multiple researchers [31]. These attributes should also enhance the verifiability and reproducibility of studies, facilitating the reuse of data and metadata, more in-depth interrogation of any given dataset, and broader-scale comparative analyses through the assembly of large datasets of multiple specimens or taxa.

However, authors of studies involving 3D digital datasets of biological and palaeontological specimens often do not publish their supporting data, meaning that results and conclusions cannot easily be verified or replicated, and that this potentially valuable source of novel data cannot be further explored [31]. Ultimately, digital data collected but unpublished are likely to be lost to science [2, 29]. This also represents a substantial waste of financial and other resources, and places vulnerable original specimens at greater risk of damage or loss, as the same specimens are likely to be reimaged repeatedly to enable different groups of workers to reproduce the data [29, 32]. Consequently, the promise of 3D digital data has not yet been fully realized.

This is not news [2, 29, 31]. However, most national and international funders have imposed regulations on data access and sharing that are forcing researchers and institutions to finally confront this challenge [33]. These regulations range from funder-mandated full release of all data [33], through declarations that the data are available from authors on request, to no release of supporting data [33]. When data are released, they are deposited in a diversity of online databases (e.g. BIRN, Dataverse, Dryad, EOL, Figshare, GigaDB, Github, MorphoBank, MorphoDBase, MorphoMuseuM, MorphoSource, Phenome10K, Zenodo), institutional and funder repositories, physical museums, and research group websites. At least in part, this diversity of approaches reflects uncertainty about the available repositories for data deposition and the cost of storing the comparatively large files associated with digital imaging-based research. Researchers can also be reluctant to share data that remain part of an active research program [34], or to share a subset of data that is part of a larger, unpublished package. There is also a lack of consensus and widespread confusion over issues of data ownership and copyright, and conflict that emerges between institutional policies asserting copyright ownership (e.g. public museum or even private collections) and the regulations of funding bodies and publishers with regard to open data. Consequently, sharing or publishing supporting data is often a low priority and has effectively been considered optional when not prescribed by a journal. Partial datasets (e.g. low-resolution visualizations or external surfaces) can be insufficient for reproducibility or even verification. As digital morphology has evolved, most of us in the research community have failed to achieve what might now be considered best practise of open data.

The academic world has already taken important steps towards overcoming some of these motivational and practical obstacles. Platforms for both archiving and sharing data online are becoming more commonplace, and can handle large file sizes. The standard in molecular biology is Genbank (https://www.ncbi.nlm.nih.gov/genbank/) where sequence data underpinning studies are accessioned before publication. For other data formats, journals and publishers offer a mixed landscape of policies on data publishing that is in need of standardization [35, 36], but many not only mandate data deposition, some are even prepared to bear the associated costs, making data deposition easier and ultimately improving science, both in terms of practice and accessibility. There are also initiatives to integrate data submission with submissions to peer-reviewed journals, requiring, or at least allowing, the submission of data in the article submission process and enabling reviewers to examine supporting data as part of the review process [37]. However, collectively, these initiatives have not been integrated [35] and they have not yet translated into common practice within many subdisciplines in biology, palaeontology and anthropology.

If a consensus can be established among authors, repositories, journal editors, peer reviewers and funding agencies, there is the prospect of finally realizing the potential of digital morphology in the open-data era. Here, we make recommendations on the nature and extent of essential and recommended best practice datasets that should be made available to support scientific publications using 3D digital datasets across biological sciences (summarised in tables 1 and 2). We review the requirements of associated metadata, discuss the current range of repositories available for such studies, and comment on issues affecting their utility.

2. Publishing tomographic data

A range of methods exist for studying 3D specimens through the creation of 2-D image stacks (i.e. tomography), including X-ray CT (encompassing medical CT, micro-CT and synchrotron tomography), MRI, neutron tomography, optical tomography, histological microtomy and physical tomography [1, 3, 4, 38-40]. All of these techniques generate datasets consisting of up to several thousand parallel sections or slices (tomograms) through a specimen, with each tomogram represented by an image file. Various techniques exist for the construction of 3D digital models from sets of tomograms [1].

(a) Data essential for scientific verification

The image stack: Image stacks are the starting point for most tomographic studies. These provide immediate insight into internal and external features, and form the basis for any subsequent construction of 3D models. Image stacks exist in a range of non-proprietary file formats, but the most common include DICOM, TIFF, JPEG, PNG, vol, RAW, and BMP [41]. All such files can be opened and viewed in free software such as ImageJ, Drishti, SPIERS, Horos, and 3D slicer [42], and can be converted into different formats, although this can be more difficult with DICOM files which exist in a multitude of sub-formats, not all of which can be handled by all software. For most purposes, TIFFs (16- or 8-bit) provide the best balance of accessibility, file size, and data quality (lossless compression), but any lossless, standard image file-types are sufficient. Most JPEG formats enforce a lossy compression scheme that may degrade over multiple save operations; lossless JPEG formats do exist (JPEG-LS, JPEG 2000), but they are not widely used. These differences underlie the importance of specifying the file standard used [41]. Minimally, image stacks should retain the contrast resolution (bit-depth) and spatial resolution used in the study. In cases where the image stack is derived from K-space filling (e.g. MRI) or a series of angular projections (e.g. X-ray CT), the process of generating the image stack is largely automated and we do not consider it necessary to publish the raw projections.

Metadata: An image stack alone will not contain all the information necessary to make full use of the data. For example, scale is only preserved if the resolution (e.g. voxel size or slice spacing) is encoded in the files, and for some datasets slice spacing is not constant and requires per-slice documentation. In the case of DICOMs, this information is typically retained within the file or can be added to the file with a header tag editor (e.g. ImageJ). Otherwise, a text file detailing the voxel or pixel size and slice spacing is the minimum necessary information that must accompany publication of any image stacks. Additionally, metadata information should include full details of how the images were acquired (including scan settings) and further information on data copyright, repository and accession of specimens scanned and, if appropriate, comments on preparation or specimen storage for biological specimens; see table 1). This information is necessary to reproduce studies, as well as to evaluate if better quality data could be obtained with a different set of parameters [43]. Minimally, these data should be provided in a simple text file (e.g. .txt or .vgi) associated with the dataset, regardless of whether the information is provided in any study based on the data.

3D models: Typically, tomographic studies involve the reconstruction of 3D models from image stacks, in some cases after image segmentation or other preparation (see below). 3D models are normally triangle-mesh geometries generated via isosurfacing (usually known as surface models) [1]. Publication of the 3D models resulting from isosurfacing allows for the interactive examination of specimen morphology in three dimensions; a wide range of free software is available for this task [1, 3], although no ideal general-purpose file-format exists for complex models (see below). 3D models may have been modified after initial isosurface-construction, for example through smoothing, island removal or hole-filling. Consequently, the most appropriate model to publish to enable verification is the final model (or models) on which the results of the study are based, or which is used in downstream analyses.

The 3D models generated using tomographic data are available in a range of different file formats [1, 44]. The choice of file type may be influenced by various factors including file size and whether colour/texture information is required; it is essential that openly accessible, standard formats are used (e.g. STL, PLY or OBJ), but there is no single ‘ideal’ file format. The Stereolithography (STL) format is the most widely used standard for publishing 3D triangle meshes derived from tomographic techniques, and it is simple and supported by the vast majority of 3D visualization programs, including freely available software [1]. STL files are also compatible with most modern 3D printers, offering potential for wider applications in specimen conservation, public outreach or teaching [3, 45]. However, STL files cannot store data on colour, texture, or scale. Where these are an essential part of the study, an alternative format such as PLY, OBJ with MTL, or VAXML [1, 41, 44] will be required. These formats are also recommended for meshes with a high number of triangles, which can result in very large file sizes in the STL format.

(b) Additional data required for best practice

Prepared datasets. While some tomographic datasets are reconstructed as 3D models without any modification or mark-up, this is unusual. Most datasets are subjected at least to segmentation, the semi-automated or manual differentiation of voxels (3D pixels) into distinct regions-of-interest (using, for example, ‘label fields’ in Avizo, or ‘masks’ in SPIERS). Some datasets also require semi-automated or manual modification of the data (e.g. through brightness modifications) to better separate specimen from background (we term this ‘editing’). These processes involve a degree of subjective interpretation; this is especially true for palaeontological datasets, which are often very noisy and can require extensive manual intervention to extract maximal information from the original data. Thus, publication of the original tomographic dataset and final 3D model may not be sufficient to enable other researchers to assess the association between the two. Segmenting and/or editing a tomographic dataset can be very time-consuming and therefore difficult to reproduce in practice; without access to prepared datasets, most secondary users would not be able to fully interrogate the data underlying a 3D model. In such instances, prepared datasets should be released. No standard file-format exists, but labels and masks can be released in the native formats by the software used to generate them, or as binary image stacks, which can then be readily reconstructed as a 3D model in a variety of software packages [1, 44].

Development of back-projection algorithms can improve signal to noise ratio in generated image stacks and, hence, recent open data mandates at synchrotron facilities require archiving of the radiograph projections, not the resulting slice data [46]. Thus, it may be sensible for authors to archive the raw projection libraries themselves. This is especially important where access to the same specimen may be problematic, or as a precaution in case unique specimens are damaged, lost or destroyed.

Image registration: For physically destructive and optical tomography, tomograms need to be registered (aligned relatively and absolutely in the X, Y, and Z planes, either manually or semi-automatically) prior to any reconstruction of 3D models. This adds a potentially subjective step that may have a bearing on downstream analyses, and so we recommend publishing both the original (unregistered) and registered image stacks as best practice.

3. Publishing 3D data from surface-based methods

Alternative surface-based methods exist for digitizing only the exterior features of specimens in 3D, most notably laser or structured light scanning [47] and photogrammetry [1, 48, 49]. For photogrammetry, data begin as 2-D photographs, whereas in surface-scanning techniques, the 3D shape is usually directly captured as 3D point clouds, with or without texture capture (colour) for each point. In photogrammetry, a 3D polygonal mesh with texture data is generated and warped onto the 3D surface (typically automatically), giving each triangle a colour value. Scanning methodologies may directly visualize point clouds, or may generate and visualize a 3D triangle mesh, with or without texture mapped onto triangles or vertices.

(a) Data essential for verification:

3D models: The production of the initial 3D surface from photographs or surface-scans is largely automated. The most critical data are the final 3D surface file(s) (which may be fused from the original component meshes), e.g. in STL, PLY or OBJ format(s) [41]. In cases where the surface texture (i.e. colour information) is directly relevant to the outcomes of a study, the published 3D models must retain this information (i.e. should be provided in PLY or OBJ formats). Surface models are not normally segmented into multiple geometric objects, so single-file models in PLY or STL format are practical.

Metadata: A text file of metadata should be provided that documents details of the imaging settings and techniques used to generate the 3D model (Table 1). Preparation of 3D meshes may involve a range of operations, including trimming irrelevant data, realigning or reorienting components of the mesh, fusion into a single mesh, smoothing, hole-filling, and/or manual manipulation of the location of individual point coordinates or surfaces. These operations should be detailed in the metadata file. Where such operations are non-trivial and/or involve interpretation, those data (photographs, raw point clouds) are an essential provision, in open and widely accessible formats, where possible.

(b) Additional data required for best practice

Models including texture information: Colour data from the surface can provide useful information to help interpret the specimen (e.g. taphonomic preservation). As best practice, this should be included if available, in PLY or OBJ format.

Original capture data: The photographs or data captured by the scanner or the 3D data generated by the photogrammetry software allow verification of the processes used to generate the model and should be included as best practice. For 3D scanning, in some cases it may only be feasible to release the raw data in proprietary formats but, where possible, widely compatible (e.g. STL) surfaces should be exported. For methods that involve the digital alignment of different aspects of a specimen, or significant manual intervention in the model construction, the unfused data should be released as the accuracy of the original alignment may be of variable quality.

4. Downstream analyses (morphometric and functional analyses)

It is important to consider not only the generation of 3D models, but also the data that may be produced in the course of downstream analyses to which these data are subjected. Common types of analysis include: (1) Size and shape analyses through topological and landmark-based techniques such as geometric morphometrics; and (2) assessment of the functional performance of specimens through computer modelling approaches, such as finite element analysis (FEA), multibody dynamics analysis (MDA), or computational fluid dynamics (CFD). These studies are often based on 3D models with the data subsequently analysed in specialist software packages [1].

(a) Data essential for verification

Morphometric data: For morphometric approaches, the original landmark coordinates, or the rules defining landmark location should be provided as these constitute the raw data for the morphometric analyses. For 2D landmark data, a .tps file or similar format links landmarks to their constituent images. Where 3D landmark data points are collected via a 3D digitizer, it is common practice to tabulate the specimen number of the digitized specimen. Where the analyses are based on 3D surfaces or digital models, it is desirable that the models (surface or volume) used in the analysis should be published in an accessible format (following the guidelines outlined above).

Downstream functional data: Functional analyses typically convert 3D digital datasets into proprietary formats for specific methodologies, such as FEA, CFD and MDA. Free software packages do exist, but typically industry standard commercial packages are employed. These have the advantage of reliability and standardized algorithms underpinning the computational analysis.

Project files or metadata: Specialist software has the disadvantage that it outputs data in proprietary file formats that may not be widely accessible to many potential users. For morphometrics, a text file detailing any corrections or transformations applied to the data and an explanation of the analyses should be published. If the morphometric analysis is conducted in the R environment, an annotated .R script is a convenient solution. For 3D functional analyses, the (usually proprietary) files containing the analysis set-up and parameters, either with or without the results files, are required for model verification. This addition enables a user with access to the appropriate software to replicate the analyses. Full metadata should be provided with details of processing techniques used to generate the final model, as well as a description of any parameters specified by the user in the analysis (Table 1).

(b) Data required for best practice.

Project and results files: Analytical techniques used to investigate the function and biomechanical performance of 3D modelled taxa will produce a range of additional digital data, which should also be made available in order to replicate studies. In the case of FEA, programs use volumetric meshes consisting of a finite number of elements. For MDA and CFD, formats such as the parasolid standard are often essential to perform the analyses. Further parameters and boundary conditions are then defined in specialist software (e.g. Abaqus, Ansys, Strand 7, Adams, Opensim, Gaitsym, COMSOL). Ideally, both the model set-up as well as the result files would be published alongside a study (e.g. [50]). For commercial packages, viewing software is sometimes available which allows the display of models and results files, but no additional analyses. Some industry software packages have text editor readable files that list and detail the location and nature of boundary conditions, e.g. .inp files for Abaqus FE software.

5. Data repositories

Researchers have a responsibility to ensure that all of the data necessary to reproduce a published study are made available. As explained above, for 3D digital datasets these data may include original 2D images, prepared/segmented 3D images, 3D geometries, and relevant metadata. These datasets can be, *in toto*, very large by today's standards; over 100 GB per specimen is possible in some scenarios, and there may be some instances where single publications utilize huge numbers of specimens, the storage of which is in itself a project. Publishers and other institutions hosting repositories must manage and facilitate access to the data they host, with these obligations persisting into the future, ideally indefinitely. Museums and other institutions holding original specimens often consider digital data as an intrinsic aspect of the specimen, and request researchers to deposit these data with them. Many have active programmes of 2D and 3D digital curation and normally make data freely available for research purposes. Data access for commercial use is a source of much needed income, and commercial reuse of data released for research purposes is a genuine concern. However, most museums do not yet have systems, policies, or resources in place for the long-term curation and distribution of digital morphological data [31]. This is not surprising given the paradigm shift in the concept of the accessioned specimen brought about by digital morphology, expanding from the physical specimen to a diversity of avatars.

Digimorph.org pioneered the curation of digital morphological data, and there are now a number of general and specialist repositories facilitating the publication and dissemination of supporting data at a variety of scales (Table 3). Many journals have agreements with such repositories and will cover charges, even for relatively large datasets. In addition, many funding agencies are building in facilities to cover costs of long-term data storage, and many institutions have developed their own data repositories to manage research data generated by their own researchers. Out-moded promises to make data "available on request" should give way to permanent URL links to 3D image data in biology, anthropology, and palaeontology (cf. [36]).

(a) Available data repositories

A range of repositories are available that cater for 3D digital datasets arising from research in biological sciences (Table 3). These can vary greatly in terms of the size and types of data they are willing to accept, as well as the cost of storage. In some cases, the choice of repository may be prescribed by the funding body or journal, but this decision will most often be made by the researcher. Modern facilities for publically sharing datasets include national data centres (typically supported by a research funding body; e.g. RCUK data centres), multidisciplinary (e.g. Dryad; [datadryad.org]; figshare [figshare.com], MorphoMuseuM [morphomuseum.com], MorphoSource [morphosource.org], Phenome10K [phenome10k.org], and Zenodo [zenodo.org] or discipline-specific (e.g. XROMM [xromm.org]) repositories, and institutional repositories for data produced in-house (e.g. Bristol University's Research Data Repository [data.bris.ac.uk/data], Natural History Museum London's Data Portal [http://data.nhm.ac.uk]). It is not entirely clear that all of these are sustainable in the long term. Traditional repositories of physical specimens can also store and disseminate data, and many are moving towards online access to their digital collections.

(b) Necessary standards for data repositories

Digital repositories should have the same qualities as repositories of physical specimens, in that they should ensure the long-term persistence and preservation of datasets in their published form, provide expert curation, stable identifiers for submitted datasets, and facilitate public access to data without unnecessary restrictions. However, by their very nature, they should also ensure that the data are discoverable online, provided with unique, permanent and citable reference codes (e.g. DOIs), associated with relevant metadata (e.g. .readme text file), and have links to relevant publications and funding bodies [2, 29].

The specific license used by the repository should be considered. Many facilities currently use the CC-BY-NC licence, which disallows re-use for commercial activities. This may be desirable where there are concerns over activities such as selling 3D prints of museum specimens with no benefit to the institutions charged with maintaining those collections. Authors may prefer to choose the CC-BY license, which is among the most open creative commons licenses available and has become the standard for open access publication of journal articles. This license lets others distribute, edit and build upon the original data, even commercially, as long as they credit the original creator. The CC-0 license (Dryad default) goes further and allows copyright-owners to waive all rights. CC-BY-ND is less attractive, since it allows sharing but does not allow the end user to publish derivatives of the data.

3D digital datasets associated with published studies should be verifiable and fully traceable from production to publication, and later republication. One option is digital watermarking, which provides a means of achieving verification of the authenticity and integrity of data, is imperceptible to the human eye, but also durable in both digital and printed forms, surviving most image edits, file format conversions, data compression, filtering, and partial data removal, smoothing. Another option would be to require users to register with the repository before data can be downloaded and used, a practice already imposed by some repositories (e.g. Dryad, Morphosource). Registration is usually free and open to everyone, but allows the repository to track data access.

(c) Costs

When publishing large (e.g. > 10 GB) 3D digital datasets, it is vital to consider the financial costs, which are typically proportional to the amount of data being stored. Some repositories do not currently charge for accessions (e.g. Morphosource) but, for some, accession charges are not insignificant. The popular online digital repository Dryad [datadryad.org] currently charges $120 per data package of 20 GB plus $50 for each additional 10 GB. Datasets based on synchrotron tomography supporting a single publication can easily run to 100 GB for a relatively small number of scans of individual specimens (e.g. [51]), and it is possible to envisage future projects, especially synthetic papers and large-scale comparative analyses, generating datasets that are orders of magnitude greater in size. Publishing such datasets can quickly become prohibitively expensive; many journals offer to fully or partially cover the costs of depositing digital datasets, but do not have a clear policy for datasets that are 100s GB to TB in size. Applications for research funding are increasingly budgeting for data storage costs, but this does not assist projects making use of pre-existing data, or those where funds for data publication are not available.

One way of minimizing costs is by reducing the total size of data published without compromising the quality. Cropping of redundant space around a volume representing the specimen is an obvious first step. Lossless compression of individual image files is an excellent route to reduce data storage for image stacks in certain formats. For example, LZW compression, both lossless and fully reversible, can provide upwards of 40% reduction in file size on 8-bit TIFFs with no evident effect on data quality, but is often not routinely applied. The PNG image format provides a similar level of lossless compression. As noted above, the JPEG image format enforces lossy compression that degrades data, and should not be used despite appealingly high compression ratios. Placing files into ZIP archives (e.g. one ZIP file per image stack) also reduces disk space through lossless compression and is more convenient for downloading. However, ZIP and .VOL archives are less secure for long-term storage, since, if the single file containing a dataset becomes corrupted, the entire dataset will be lost. Corruption of single files within a large dataset is less serious, and at least some repositories have procedures in place to detect and remediate bitrot [32]. We recommend that unarchived copies of the original data are stored and made available where possible.

In our enthusiasm for recycling 3D digital data and easing reproducibility of morphological studies based on them, the environmental costs of storage should be considered. Most datasets will be accessed infrequently and so there is no need or justification for their storage on spinning disks. Many repositories make use of automated tape storage which is stable and comparatively low in direct costs for the same reasons that make it environmentally low-cost. However, in such cases data will not be available instantly on demand and access will instead have to be requested.

6. Rescuing legacy data and constraints on data use

An increase in the availability and ease of use of data repositories raises the prospect of making data available from previously published studies where the data were not released at the time of publication. Digital datasets can be uploaded to online data repositories and linked to past publications. At present there are no policies or mechanisms we are aware of among journals and publishing houses to link archival publications to newly deposited data. However, there is no material technical barrier to salvaging legacy data in this way. Publishers are likely to welcome such an initiative since it would obviously improve data visibility, facilitate reproducibility, and likely rejuvenate old publications in terms of access, citations and, ultimately, their marketability.

Obtaining digital characterizations of morphology can be time-consuming and expensive, and researchers rarely exhaust their data with the first publication. Funders and publishers are increasingly removing choice over whether to release supporting data, and so it can seem unfair that the researchers who generated datasets have to subsequently compete to exploit them further. This can be particularly difficult for lone early career researchers potentially competing with large experienced research groups [34]. One potential solution to this would be the introduction of time-limited embargos, which can already be facilitated by some data repositories. However, such embargos violate the most basic tenet of open data, that of removing barriers to assessing the reproducibility of research [52]. After the point of publication, it is also effectively impossible to police the release of supporting data and, consequently, we see no alternative to the release of data with publication. A possible compromise may be borrowed from the Bermuda [53], Fort Lauderdale [54], and Toronto [55] agreements of the genomics community. These mandate data release at the time they are obtained but, more germane to morphologists, these agreements provide safeguarding for data generators through published, time-limited, statements of intent of how they propose to exploit the data [55]. Other researchers are free to exploit the data for other purposes, and for any purpose after the stated period of limitation of the statement of intent [56]. Third party users with overlapping research interests are expected to proceed respectfully and in dialogue with the data generators to identify a mutually agreeable publication schedule [55]. Invariably, much more is at stake in such projects, and though these informal agreements are rarely violated, they are generally well-policed by the peer review process [56], and by the reputational damage suffered by those who choose not to observe these agreements.

Practice in the genomics community underscores the point that there is more to gain from open data than the warm glow of altruism [55, 57]. Not only has it led to greater and more rapid scientific advance [52, 55], it can lead to material personal gain, through the proposals for collaborative exploitation of published data, both to achieve stated research objectives, and to achieve new objectives that would not be possible without unforeseen collaborators [55, 57]. Citation and access-tracking of published datasets provides credit to the authors [32]. Attribution of authorship is mandated under CC-BY licenses and is in any case integral to the academic culture. Many journals already mandate citation of published datasets, not (or not merely) the publications describing research based upon them; this must become common practice. Further mechanisms of encouraging researchers to share their data should only add to this motivation, such as explicitly evaluating the open sharing of data as part of CVs in hiring, promotion or other award processes.

Nevertheless, data can be associated with ethical sensitivities that may require the withholding, or restriction on public distribution, of data (e.g. anthropology or medical science [58, 59]). In such instances, the issues that apply should be clearly defined so that beyond these boundaries researchers and publishers can follow an ethos of open data publication. Mechanisms already exist to cope with these constraints while still making data available, such as data anonymization and vetted access [55].

7. Outstanding challenges

While the principle of open data has been mandated by the majority of funders [33], publishers, physical repositories and researchers are all scrambling to meet the resulting challenges. Above all, the competing interests over ownership of digital data need to be resolved between: (i) funders who pay for research, (ii) researchers who collect specimens and create the digital datasets, (iii) research facilities where data are collected, (iv) museums who have a duty of care for the physical specimens, and (v) research publishers. Funders, researchers, and publishers may have converged on an ethos of open data. However, the institutions that are responsible for the physical specimens have not obviously been invited to engage in the development of open data policy, and yet it is museums that will have to change most in terms of their policies on the nature of what they consider intrinsic aspects of the physical specimens that they hold in their care. One solution for museums might be to comply with research funders’ requirements, and waive copyright over digital representations of their collections, along with its associated income stream. Another solution would be for these institutions, which are those best-placed to inform policy on the curation, storage and distribution of data, to develop digital collections with the stability to match that of their physical inventory. Indeed, with the development of cybertypes [29, 30], this may be an inevitable future aspect of the world’s leading museums. However, if this readily realisable vision of data repository quality, stability, and credibility, is to be achieved, it will require the funders who have mandated data deposition to cover the costs of establishing and maintaining such facilities, through block grants, not through piecemeal funding to researchers. If such change is to be achieved, it must not only happen in wealthier countries, but worldwide and, thus, more amply provisioned funders should provide further means to help other countries improve their data-sharing capacities.

Data access is not only important post-publication, to aid reproducibility, but during peer review, so that the results of a study and their interpretations can be verified prior to publication. Providing tomographic or 3D data at the point of journal submission is, in our experience, a comparatively rare phenomenon that the publishing infrastructure is not currently well set up to facilitate. Publishers must develop a more homogenous policy on open data [35], along with procedures to ensure data sources are acknowledged and linked electronically to the derivative publications [52]. It is also important that systems are developed to ease the submission of such data, and facilitate secure, anonymised distribution of data to reviewers. Dryad offers an integrated submission system where publishers can coordinate submission of a manuscript with submission of data, which can then be accessed securely by referees and editors. For non-integrated journals, an interim solution may be to host data at a temporary, hidden-URL that can be forwarded to the reviewers via the journal. Authors may be cautious about sharing such data ahead of an article being accepted for publication, and there should be a clear policy governing the restrictions of use for reviewers.

8. Conclusions

Data sharing is essential in order for the benefits of 3D digital data to be fully realized by the scientific community, as well as for the maximum benefit to be gained from the public and private funding that allows these data to be collected. Not only are the benefits of 3D digital data not currently being fully realized, but failure to publish supporting data is rendering many studies based on 3D digital data at least difficult to reproduce. We have presented a series of proposals for open 3D digital data. These outline the minimal standards of verifiability that studies should meet before they are published. We also present more ambitious standards that we hope can be assumed as normal best practice (Table 1). We have all been guilty of failing to meet these standards in the past because of technical and other limitations; however, technology has changed and so must we. There are costs associated with releasing data, both real and in-kind, but these are insignificant in proportion to the real costs of regenerating the data, and the reputational costs to individuals, institutions, journals and editors, of publishing research predicated upon inaccessible data.

Ethics statement: No data were harmed in the formulation of this contribution.

Data accessibility statement: There are no data associated with this manuscript.

Competing interests statement: The authors declare there are no competing interests.

Authors' contributions statement: The project was conceived by TGD, IAR, SL, JAC, EJR, and PCJD, all of whom drafted the original manuscript, to which all others contributed.

Acknowledgements: We thank Zosia Beckles (data.bris), Else-Marie Friis (NRM, Stockholm), Iain Hrynaszkiewicz (Springer Nature), Mark Hahnel (figshare), Elizabeth Hull (Dryad), Phil Hurst (Royal

Society Publishing), Rhiannon Meaden (Royal Society Publishing), Sowmya Swaminathan (Springer Nature), Stuart Taylor (Royal Society Publishing), and Sally Thomas (Palaeontological Association) for discussion.

Funding statement: The authors are funded by BBSRC (PCJD, EJR), The Calleva Foundation and the Human Origins Research Fund (CS), European Research Council (AG, JRH, RBJB), Generalitat Valenciana and MINECO (CMP), Leverhulme Trust (AG, RBJB), NERC (JAC, PCJD, AG, JRH, EJR), NWO (MR), National Science Foundation (AG, APS, SYS), 1851 Royal Commission (IAR), Royal Society Wolfson Merit Award (PCJD), and the Swedish Research Council (SB).

References

1. Sutton M.D., Rahman I.A., Garwood R.J. 2014 *Techniques for Virtual Palaeontology*. London, Wiley.
2. Rowe T., Frank L.R. 2011 The disappearing third dimension. *Science* **331**, 712-714.
3. Cunningham J.A., Rahman I.A., Lautenschlager S., Rayfield E.J., Donoghue P.C.J. 2014 A virtual world of paleontology. *Trends in Ecology & Evolution* **29**(6), 347-357.
4. Weber G.W., Bookstein F.L. 2011 *Virtual anthropology: a guide to a new interdisciplinary field*, Springer.
5. Metscher B.D. 2009 MicroCT for comparative morphology: simple staining methods allow high-contrast 3D imaging of diverse non-mineralized animal tissues. *BMC Physiol* **9**, 11. (doi:10.1186/1472-6793-9-11).
6. Gignac P.M., Kley N.J., Clarke J.A., Colbert M.W., Morhardt A.C., Cerio D., Cost I.N., Cox P.G., Daza J.D., Early C.M., et al. 2016 Diffusible iodine-based contrast-enhanced computed tomography (diceCT): an emerging tool for rapid, high-resolution, 3-D imaging of metazoan soft tissues. *J Anat*. (doi:10.1111/joa.12449).
7. Berquist R.M., Gledhill K.M., Peterson M.W., Doan A.H., Baxter G.T., Yopak K.E., Kang N., Walker H.J., Hastings P.A., Frank L.R. 2012 The Digital Fish Library: using MRI to digitize, database, and document the morphological diversity of fish. *PLoS One* **7**(4), e34499. (doi:10.1371/journal.pone.0034499).
8. Staedler Y.M., Masson D., Schonenberger J. 2013 Plant tissues in 3D via X-ray tomography: simple contrasting methods allow high resolution imaging. *PLoS One* **8**(9), e75295. (doi:10.1371/journal.pone.0075295).

9. Worsaae K., Sterrer W., Kaul-Strehlow S., Hay-Schmidt A., Giribet G. 2012 An anatomical description of a miniaturized acorn worm (hemichordata, enteropneusta) with asexual reproduction by paratomy. *PLoS One* **7**(11), e48529. (doi:10.1371/journal.pone.0048529).

10. Lautenschlager S., Bright J.A., Rayfield E.J. 2014 Digital dissection - using contrast-enhanced computed tomography scanning to elucidate hard- and soft-tissue anatomy in the Common Buzzard Buteo buteo. *J Anat* **224**(4), 412-431. (doi:10.1111/joa.12153).

11. Sharp A.C., Trusler P.W. 2015 Morphology of the jaw-closing musculature in the common wombat (Vombatus ursinus) using digital dissection and magnetic resonance imaging. *PLoS One* **10**(2), e0117730. (doi:10.1371/journal.pone.0117730).

12. Corti M. 1993 Geometric morphometrics: An extension of the revolution. *Trends Ecol Evol* **8**(8), 302-303. (doi:10.1016/0169-5347(93)90261-M).

13. Adams D.C., Rohlf F.J., Slice D.E. 2013 A field comes of age: geometric morphometrics in the 21st century. *Hystrix* **24**, 7-14. (doi:10.4404/hystrix-24.1-6283).

14. Rayfield E.J. 2007 Finite element analysis and understanding the biomechanics and evolution of living and fossil organisms. *Annual Review of Earth and Planetary Sciences* **35**, 541-576.

15. Bates K.T., Falkingham P.L. 2012 Estimating maximum bite performance in *Tyrannosaurus rex* using multi-body dynamics. *Biology Letters* **8**(4), 660-664.

16. Rahman I.A., Darroch S.A., Racicot R.A., Laflamme M. 2015 Suspension feeding in the enigmatic Ediacaran organism Tribrachidium demonstrates complexity of Neoproterozoic ecosystems. *Science Advances* **2015**(1), :e1500800.

17. Donoghue P.C.J., Bengtson S., Dong X.-P., Gostling N.J., Huldtgren T., Cunningham J.A., Yin C., Yue Z., Peng F., Stampanoni M. 2006 Synchrotron X-ray tomographic microscopy of fossil embryos. *Nature* **442**(7103), 680-683.

18. Smith S.Y., Collinson M.E., Rudall P.J., Simpson D.A., Marone F., Stampanoni M. 2009 Virtual taphonomy using synchrotron tomographic microscopy reveals cryptic features and internal structure of modern and fossil plants. *Proceedings of the National Academy of Sciences* **106**(29), 12013-12018. (doi:10.1073/pnas.0901468106).

19. Lautenschlager S. 2013 Cranial myology and bite force performance of *Erlikosaurus andrewsi*: a novel approach for digital muscle reconstructions. *Journal of Anatomy* **222**(2), 260-272.

20. Rahman I.A., Zamora S., Falkingham P.L., Phillips J.C. 2015 Cambrian cinctan echinoderms shed light on feeding in the ancestral deuterostome. *Proceedings Biological sciences / The Royal Society* **282**(1818), 20151964. (doi:10.1098/rspb.2015.1964).

21. Wroe S., Ferrara T.L., McHenry C.R., Curnoe D., Chamoli U. 2010 The craniomandibular mechanics of being human. *Proceedings Biological sciences / The Royal Society* **277**(1700), 3579-3586. (doi:10.1098/rspb.2010.0509).

22. Pierce S.E., Clack J.A., Hutchinson J.R. 2012 Three-dimensional limb joint mobility in the early tetrapod *Ichthyostega*. *Nature* **486**(7404), 523-U123.

23. David R., Stoessel A., Berthoz A., Spoor F., Bennequin D. 2016 Assessing morphology and function of the semicircular duct system: introducing new in-situ visualization and software toolbox. *Scientific reports* **6**, 32772. (doi:10.1038/srep32772).

24. Lowe T., Garwood R.J., Simonsen T.J., Bradley R.S., Withers P.J. 2013 Metamorphosis revealed: time-lapse three-dimensional imaging inside a living chrysalis. *J R Soc Interface* **10**(84), 20130304. (doi:10.1098/rsif.2013.0304).

25. Goswami A., Randau M., Polly P.D., Weisbecker V., Bennett C.V., Hautier L., Sanchez-Villagra M.R. 2016 Do Developmental Constraints and High Integration Limit the Evolution of the Marsupial Oral Apparatus? *Integr Comp Biol* **56**(3), 404-415. (doi:10.1093/icb/icw039).

26. Bourke J.M., Porter W.M., Ridgely R.C., Lyson T.R., Schachner E.R., Bell P.R., Witmer L.M. 2014 Breathing life into dinosaurs: tackling challenges of soft-tissue restoration and nasal airflow in extinct species. *Anatomical record* **297**(11), 2148-2186. (doi:10.1002/ar.23046).

27. Porter W.R., Sedlmayr J.C., Witmer L.M. 2016 Vascular patterns in the heads of crocodilians: blood vessels and sites of thermal exchange. *J Anat*. (doi:10.1111/joa.12539).

28. Bourke J.M., Witmer L.M. 2016 Nasal conchae function as aerodynamic baffles: Experimental computational fluid dynamic analysis in a turkey nose (Aves: Galliformes). *Respir Physiol Neurobiol* **234**, 32-46. (doi:10.1016/j.resp.2016.09.005).

29. Faulwetter S., Vasileiadou A., Kouratoras M., Thanos D., Arvanitidis C. 2013 Micro-computed tomography: Introducing new dimensions to taxonomy. *Zookeys* (263), 1-45. (doi:10.3897/zookeys.263.4261).

30. Akkari N., Enghoff H., Metscher B.D. 2015 A new dimension in documenting new species: high-detail imaging for myriapod taxonomy and first 3D cybertype of a new millipede species (Diplopoda, Julida, Julidae). *PLoS One* **10**(8), e0135243. (doi:10.1371/journal.pone.0135243).

31. Hublin J.J. 2013 Free digital scans of human fossils. *Nature* **497**, 183-183.

32. Boyer D.M., Gunnell G., F., Kaufman S., McGeary T.M. in press Morphosource: archiving and sharing 3-D digital specimen data. *Paleontological Society Papers*. (doi:10.101/scp.2016.9).

33. Hahnel M. 2015 Global funders who require data archiving as a condition of grants. (figshare https://dx.doi.org/10.6084/m9.figshare.1281141.v1.

34. Portugal S.J., Pierce S.E. 2014 Who's looking at your data? *Science* **Science Careers**. (doi:10.1126/science.caredit.a1400052).

35. Naughton L., Kernohan D. 2016 Making sense of journal research data policies. *Insights: the UKSG journal* **29**(1), 84-89. (doi:10.1629/uksg.284).

36. Alsheikh-Ali A.A., Qureshi W., Al-Mallah M.H., Ioannidis J.P. 2011 Public availability of published research data in high-impact journals. *PLoS One* **6**(9), e24357. (doi:10.1371/journal.pone.0024357).

37. Anonymous. 2016 Let referees see the data. *Scientific Data* **3**, 160033. (doi:10.1038/sdata.2016.33).

38. Long F., Zhou J., Peng H. 2012 Visualization and analysis of 3D microscopic images. *PLoS computational biology* **8**(6), e1002519. (doi:10.1371/journal.pcbi.1002519).

39. Ziegler A., Kunth M., Mueller S., Bock C., Pohmann R., Schröder L., Faber C., Giribet G. 2011 Application of magnetic resonance imaging in zoology. *Zoomorphology* **130**(4), 227-254. (doi:10.1007/s00435-011-0138-8).

40. Gold M.E., Schulz D., Budassi M., Gignac P.M., Vaska P., Norell M.A. 2016 Flying starlings, PET and the evolution of volant dinosaurs. *Current biology : CB* **26**(7), R265-267. (doi:10.1016/j.cub.2016.02.025).

41. McHenry K., Bajcsy P. 2008 An overview of 3d data content, file formats and viewers. Technical Report: isda08-002. (pp. 1-21. Urbana, IL 61801, Image Spatial Data Analysis Group, National Center for Supercomputing Applications.

42. Schneider C.A., Rasband W.S., Eliceiri K.W. 2012 NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**(7), 671-675. (doi:10.1038/nmeth.2089).

43. Faulwetter S., Minadakis N., Keklikoglou K., Doerr M., Arvanitidis C. 2015 First steps towards the development of an integrated metadata management system for biodiversity-related micro-CT datasets. In *Bruker microCT User Meeting 2015* (Bruges, Bruker.

44. Sutton M.D., Garwood R.J., Siveter D.J., Siveter D.J. 2012 SPIERS and VAXML: A software toolkit for tomographic visualisation and a format for virtual specimen interchange. *Paleontologica Electronica* **15**(2), 5T (palaeo-electronica.org/content/issue-2-2012-technical-articles/2226-virtual-palaeontology-toolkit).

45. Rahman I.A., Adcock K., Garwood R.J. 2012 Virtual Fossils: a New Resource for Science Communication in Paleontology. *Evolution: Education and Outreach* **5**(4), 635-641. (doi:10.1007/s12052-012-0458-2).

46. ESRF. 2015 The ESRF data policy. (pp. 1-4. Grenoble, ESRF.

47. Cooney C.R., Bright J.A., Capp E.J.R., Chira A.M., Hughes E.C., Moody C.J.A., CaNouri L.O., Varley Z.K., Thomas G.H. 2017 Mega-evolutionary dynamics of the adaptive radiation of birds. *Nature*.

48. Falkingham P.L. 2012 Acquisition of high resolution three-dimensional models using free, open-source, photogrammetric software. *Paleontologica Electronica* **15**(1).

49. Mallison H., Wings O. 2014 Photogrammetry in paleontology – a practical guide. *Journal of Paleontological Techniques* **12**, 1-31.

50. Lautenschlager S. 2015 Estimating cranial musculoskeletal constraints in theropod dinosaurs. *R Soc Open Sci* **2**(11), 150495. (doi:10.1098/rsos.150495).

51. Huldtgren T., Cunningham J.A., Yin C., Stampanoni M., Marone F., Donoghue P.C.J., Bengtson S. 2011 Fossilized nuclei and germination structures identify Ediacaran “animal embryos” as encysting protists. *Science* **334**, 1696-1699.

52. Schofield P.N., Bubela T., Weaver T., Portilla L., Brown S.D., Hancock J.M., Einhorn D., Tocchini-Valentini G., Hrabe de Angelis M., Rosenthal N. 2009 Post-publication sharing of data and tools. *Nature* **461**(7261), 171-173. (doi:http://www.nature.com/nature/journal/v461/n7261/suppinfo/461171a_S1.html).

53. Marshall E. 2001 Bermuda Rules: Community Spirit, With Teeth. *Science* **291**, 1192-1192.

54. Wellcome_Trust. 2003 *Sharing data from large-scale biological research projects: a system of tripartite responsibility. Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA.* London, Wellcome Trust; 6 p.

55. Birney E., Hudson T.J., Green E.D., Gunter C., Eddy S., Rogers J., Harris J.R., Ehrlich S.D., Apweiler R., Austin C.P., et al. 2009 Prepublication data sharing. *Nature* **461**(7261), 168-170. (doi:http://www.nature.com/nature/journal/v461/n7261/suppinfo/461168a_S1.html).

56. Nanda S., Kowalczuk M.K. 2014 Unpublished genomic data-how to share? *BMC genomics* **15**, 5. (doi:10.1186/1471-2164-15-5).

57. Nelson B. 2009 Empty archives. *Nature* **461**, 160-163.

58. Warren E. 2016 Strengthening research through data sharing. *New England Journal of Medicine* **375**(5), 401-403. (doi:doi:10.1056/NEJMp1607282).

59. Hrynaszkiewicz I., Khodiyar V., Hufton A.L., Sansone S.-A. 2016 Publishing descriptions of non-public clinical datasets: proposed guidance for researchers, repositories, editors and funding organisations. *Research Integrity and Peer Review* **1**(1). (doi:10.1186/s41073-016-0015-6).

60. Bright J.A., Marugan-Lobon J., Cobb S.N., Rayfield E.J. 2016 The shapes of bird beaks are highly controlled by nondietary factors. *Proceedings of the National Academy of Sciences of the United States of America* **113**(19), 5352-5357. (doi:10.1073/pnas.1602683113).

Figure and table captions

Table 1. Summary table of recommendations for types of data files that should be published in support of published articles.

Table 2. Summary of the principles of open data for digital morphology.

Table 3. Summary of main online repositories for 3D digital morphological data.

Figure 1. Examples of digital data and downstream uses. (*a*) Medical CT image of the dentary of the holotype of *Tyrannosaurus rex* CM 9380. (*b*) Reconstructed MicroCT dataset of vascular injected green iguana (*Iguana iguana*) skull [OUVC 10677]. (*c*) Slice through braincase region of microCT scanned Iodine-potassium iodide (I_2KI) stained contrast-enhanced grey squirrel (*Sciurus carolinensis*) skull. (*d*) MRI scan midline slice of neonatal white rhino (*Ceratotherium simum*). (*e*) Synchrotron Radiation X-ray Tomographic Microscopy (SRXTM) partial reconstruction of putative red alga from the Ediacaran Weng-an Biota, South China. (*f*) Digital reconstruction of *Offacolus kingi*, a chelicerate from the Silurian of Hertfordshire, UK, reconstructed via serial grinding and optical microscope photography; Inset: digital segmentation of microphotograph. (g) image of stl (stereolithography) file of skull of foetal Tammar wallaby *Macropus eugenii*. (*h*) Optical projection tomography of mouse hindlimb at embryonic stage E19, stained with Alcian blue and Alizarin red and imaged using visible light and fluorescent light to image cartilage and bone respectively (image courtesy of Karen Roddy). (*i*) Photogrammetry reconstruction of guineafowl trackway. (*j*) Surface scan of human subject, with subject-specific skeleton and muscle volumes segmented from MRI scan data and resulting multibody dynamics analysis (MDA) model of same subject. (*k*) SIMM (Software for Interactive Musculoskeletal Modelling) model of *Tyrannosaurus rex* hindlimb. (*l*) MicroCT scan reconstruction of

the skull of the common buzzard, *Buteo buteo*, detailing landmarks and semilandmarks used for geometric morphometrics (GMM) analysis (reproduced with permission from Bright et al. [60]). (*m*) Finite element (FE) model of the skull of *Allosaurus fragilis* (reproduced with permission from Lautenschalger & Rahman in press). (*n*) Results of CFD simulation of water flow around a 3D model of the cinctan echinoderm *Protocinctus mansillaensis*. All images obtained from authors unless stated otherwise.