Open Data and Palaeontology

Ross Mounce
University of Bath, Bath, UK

Introduction

Palaeontology is the study of ancient life in all its forms: vertebrates, arthropods, plants and many other weird and wonderful types of organism. As an academic discipline, it suffers from a perception in some quarters that it is a less quantitative, less analytical, 'soft science'—a kind of Rutherfordian-view that the study of fossils is just 'stamp collecting'. Yet modern palaeontology is often highly computational, generating lots of data with which to test and form hypotheses. In the digital age, once published, if provided in the right format, data can be easily reused by further

How to cite this book chapter:
Mounce, R. 2014. Open Data and Palaeontology. In: Moore, S. A. (ed.) *Issues in Open Research Data*. Pp. 151–164. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/ban.j

studies to advance the sum of all human knowledge. This chapter examines the availability of palaeontology-related research data online and the reuse conditions under which it is made available.

Example Data Generating Studies in Modern Palaeontology

A typical study in systematic palaeontology may attempt to retrace the relationships between extinct life forms using an evolutionary tree (phylogeny). The source data in this instance may be a matrix of many thousands of observations of the morphology of fossil forms, codified into discrete states for analysis. These observations often come from comparative examination of specimens or, more likely, high-resolution photographs of these specimens that enable features to be examined side-by-side even if the physical specimens themselves are kept continents-apart in different museums.

Other palaeontological studies go one further and aim to create 'virtual fossils'—accurate three-dimensional interactive vizualisations of specimens to aid their interpretation, with the aid of tomographic methods. Methods such as X-ray imaging and magnetic resonance imaging (MRI) generate data non-destructively, so the original fossil is preserved undamaged. Both these types of palaeontological study represent just a small subset of the full range of palaeontological studies but what they have in common is that they heavily rely on imaging data; either photographs of specimens in the first instance, or the creation of three-dimensional image data. Much of palaeontology thus relies on the interpretation of morphology and thus image data, and the online sharing of image data is crucial to advancing palaeontological science.

Infrastructure Enabling Data Sharing in Palaeontology

There are many specialist sites specifically catering for or allowing palaeontological data, some of which incorporate helpful data management, collaboration and analysis tools that further incentivise use of their platform. I do not pretend to provide an exhaustive listing here—there are no doubt many more, the projects discussed herein reflect my own personal biases towards vertebrate palaeontology and systematic palaeontology. The main point of this selection is to highlight the variance in approach to data licencing that each of these projects has adopted. See 'From card catalogs to computers: Databases in Vertebrate Paleontology' for a review with a different focus (Uhen et al. 2013).

The Paleobiology Database *http://paleobiodb.org/*

This project collates taxonomic and collection-based occurrence data for all fossil groups, of all geological ages. It is widely supported and contributed to by the palaeontological research community.

Towards the end of 2013 (Kishnor & Peters 2013), it set a great example by uniformly re-licencing all the data it contains under the Creative Commons Attribution (CC BY) 4.0 International License to ensure that it provides open, reusable data.

Their frequently asked questions (FAQs) (Alroy, adapted by Uhen 2013) suggest that for large (how large is left undefined) dataset analyses, data reusers should download an accompanying 'secondary bibliography' to provide evidence of data provenance for subsequent journal publication as a supplementary material

file. Whilst this strategy certainly fulfils the legal requirements of the CC BY licence, such a request is extremely unlikely to provide counted citations, which help researchers demonstrate their academic impact. Most of the traditional bibliometric data indexers, e.g. Thompson Reuters Web of Knowledge and Google Scholar, only index the main paper for citations. Citations provided in supplementary files are typically ignored (Kueffer et al. 2011).

Ancient Human Occupation of Britain Database (AHOB)
http://www.ahobproject.org/database/

This project documents data on British and European Quaternary dig sites: geographical co-ordinates, photographs, stable-isotope data, faunal lists and more. It has received funding from three Leverhulme Trust programme grants.

Access is entirely restricted to project members-only for the life of the project. According to Uhen et al. (2013) the data '… will be made publicly available at the end of the project in 2013.' Yet in 2014 the database is still access-restricted, project member login-only. Licencing of the data contained in this database is unknown. Even if some of the data cannot be shared openly because it might be sensitive, it strikes me that at least some of the data, e.g. faunal lists and stable-isotope data, is clearly non-sensitive and therefore can without doubt be reasonably made publicly available.

MorphoBank
http://www.morphobank.org/

MorphoBank (O'Leary & Kaufman, 2011) is a website primarily used by researchers concerned with morphology-based phylogenetics or cladistics research (reconstructing evolutionary trees).

It has strong features that help researchers build, version control, annotate, manage and enable effective collaboration around their phylogenetic research data, as well as providing a web-space in which to make all that data publicly available after publication of the associated research paper. As of early April 2014, there are over 300 publicly accessible projects on MorphoBank as well as over 600 non-public projects in progress. The *Journal of Vertebrate Paleontology* should be congratulated as one of the first journals to publicly support the use of MorphoBank (Berta & Barrett 2011); as a result of this, there are more MorphoBank projects with data from *Journal of Vertebrate Paleontology*-published studies than any other journal.

Initially, data uploaded to MorphoBank is private, until researchers are ready to choose to make it public. When making their data public, MorphoBank allows researchers to choose from the full range of Creative Commons licences available. MorphoBank guides users towards choosing open licences on their FAQ but does not enforce their preference:

> MorphoBank would prefer for content providers to choose CC0 or CC BY reuse policies because they (and only they) are Open Data licenses. Please be aware that choosing an NC (non-commercial usage only) license may prevent your data submission from being used on open-content only websites such as Wikipedia.
>
> (MorphoBank 2014)

It is difficult to search media by licence, but I estimate (supporting data on figshare; Mounce 2014) that of the >27,000 publicly viewable images hosted on MorphoBank, less than half are made available under Open Knowledge Definition (OKD)-conformant open licences (see **Figure 1**). Over 77% of projects share less than 10 images, with most (modal) sharing only one image—MorphoBank forces users to upload at least one image.

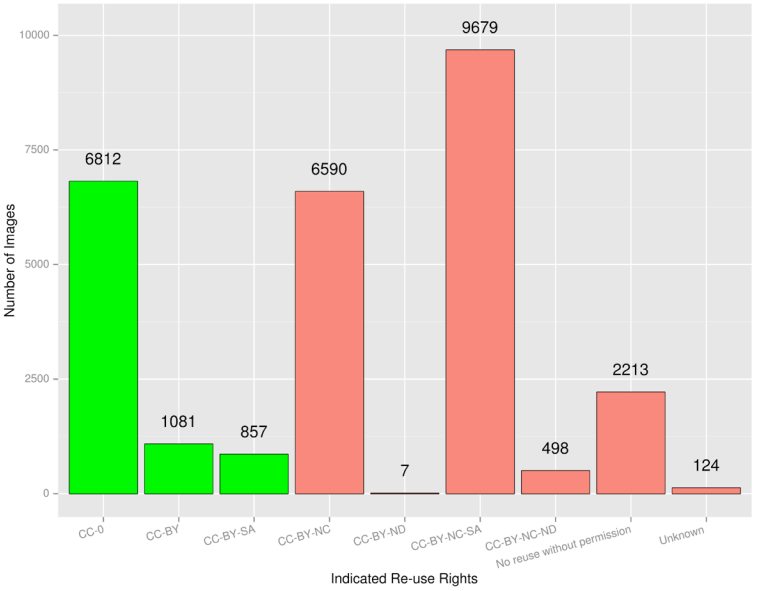

Figure 1: Images in MorphoBank by re-use rights. The three leftmost columns in green indicate OKD-conformant open licences. Figure generated in R (R Core Team, 2014) with the package ggplot2 (Wickham, 2009).

Morphbank
http://www.morphbank.net/

Not to be confused with its close namesake, Morphbank is an earlier project that specifically focuses on biological specimen image data sharing. As of early April 2014, this database makes publicly available over 372,000 images of biological specimens. By default, images are licenced under Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA; not an OKD-conformant open licence) but contributors may opt to change that for their uploads to a less restrictive Creative

Commons license, including even Public Domain Dedication. As with MorphoBank, it does not appear possible at this point to easily filter or search images by reuse licence so I am unable to determine the distribution of licences chosen by contributors to the site.

For some reason, however, few palaeontologists seem to have adopted the use of Morphbank to share their image data. Alberto Prieto-Marquez, a vertebrate palaeontologist, is one notable exception in that regard—he has made over 1700 images relating to his research available via this site.[1]

Dryad
http://datadryad.org/

Another more recent initiative to encourage data sharing that is open to palaeontologists is Dryad. All data submitted to Dryad is released to the public domain under the Creative Commons Zero waiver (CC0). The Paleontology Society journals (*Journal of Paleontology*, *Paleobiology*) were the first significant palaeontological adopters of Dryad, and now the palaeo-relevant journals *Palaeontology*, *ZooKeys* and *Zoological Systematics and Evolution* also make use of it to share supplementary, publication-associated data. The journal *Evolution* deserves special praise for being one of the first well-respected evolutionary biology journals to mandate data archiving for all its articles (Fairbairn 2011), something that many journals still just weakly 'encourage'. Key to the popularity of Dryad is probably its assignment of a digital object identifier (DOI) to each and every dataset contributed, which allows easier citation and tracking of the reuse of data. Of course, data does not actually need a DOI to be 'citable' but, for many, a DOI certainly

[1] User record available at http://www.morphbank.net/?id=78418

does seem to encourage formal citation. This may explain why some authors have even gone to the trouble of uploading datasets relating to long-ago published papers—something I would imagine they would not do if they saw no benefit to themselves in this service.

Figshare
http://figshare.com/

Figshare, similar to Dryad, is a 'generalist' data sharing website that is open to palaeontology but also contains data relating to a much wider array of subjects. Like Dryad, they also assign DOIs to datasets but they go one further in assigning each and every file within your data upload a separate DOI if you so wish. Unlike Dryad, figshare also allows the upload of data not related to publications, so it is ideal for uploading 'work-in-progress' data and data from projects that would otherwise be left in a file-drawer unfinished forever. I estimate at least 2000 research objects (figures, images, data, posters, manuscripts, code) relating to palaeontology have so far been made available at figshare. From a reuse rights perspective, figshare by default makes uploaded figures, media, posters, papers and filesets available under CC BY. Datasets are made available under CC0, and code under the MIT License. All these are OKD-conformant open licences.

Summary of Data Sharing Infrastructure for Palaeontology

As you can see from this small selection of palaeo-relevant databases, there is huge variance between them in terms of reuse rights. Some make nothing publicly available (e.g. AHOB),

whilst many allow users to initially upload data privately and then make it publicly available at a later date (e.g. figshare, Dryad, MorphoBank, the Paleobiology Database). When data is made publicly available at these sites, some allow a wide choice of reuse rights options and content uploaders do typically make use of all of these options if options are provided (e.g. Morphbank and MorphoBank). Others such as figshare, Dryad and the Paleobiology Database have made a conscious and reasoned decision to not allow a choice of licences when making data available; all these three enforce OKD-conformant licenses—either CC BY or CC0.

Interestingly, prior to the late 2013 licencing change by the Paleobiology Database committee, PaleoDB (as it was then known) used to allow data contributors to upload data under a variety of different Creative Commons licences. Many contributors chose different licences, and some of these licences were incompatible with each other! This along with many other reasons (given in Hagedorn et al. 2011; Klimpel 2012) is why PaleobioDB opted to adopt CC BY only.

Is licence choice really a good thing?

Having content available in a variety of different licences in projects such as at Morphbank and MorphoBank creates a lot of additional complexity for bulk reusers of content. Having to accommodate this variability is hard, especially if some of those different terms and conditions are incompatible with each other. Databases such as Dryad that use CC0 impose no legal restraint on data reuse, and instead trust academic cultural norms to ensure that data is cited appropriately if reused. I am confident that in science we do not need to resort to copyright-led enforcement of

citation, and that academic cultural norms and the self-policing nature of academia are enough to ensure citation from data reuse. As testament to this, I know of no instances in which data made available at Dryad or figshare has been reused without appropriate citation.

Another troubling aspect is the seemingly widespread adoption of the 'non-commercial' (-NC) Creative Commons licences where they are allowed. I suspect this is based upon misunderstanding of the type of reuse(s) that these licences prevent. Many assume that non-commercial licences only prevent for-profit businesses from reusing content for profit. But non-commercial is about commerce, not profit, and that is an important difference. In my experience, few realise that these non-commercial licences are far more restrictive: -NC content cannot be reused in most educational settings in schools or universities, likewise -NC content cannot be uploaded to Wikimedia for use on Wikimedia projects like Wikipedia (Klimpel 2012). Indeed, a recent ruling in Germany shows that -NC content is only 'safe' for personal use (Haddouti 2014): any other use, even by a non-profit organisation, may get the content reuser sued many years later. Myself and many others would not want to expose ourselves to this risk and thus -NC licenced content is unusable for us.

The Role of Journals in Encouraging Data Sharing

In my opinion I see journal policy as key to encouraging and enforcing data sharing. There are the beginnings of a trend to be observed in which the better journals mandate the archiving of all publication-related supporting data to encourage its examination and reuse (Fairbairn 2011). This is in both the authors' and journals' interests because sharing data is known to be associated

with an increased citation rate (Piwowar, Day & Fridsma 2007; Piwowar & Vision 2013), as well as being cost-effective (Piwowar, Vision & Whitlock 2011). I would like to think these advantages alone would facilitate spontaneous data sharing, but I do not see that happening in the palaeontological community, so research-funder and journal policies are still needed to encourage and enforce data sharing.

The journals *Evolution*, *Journal of Paleontology*, *Paleobiology* and *ZooKeys* clearly mandate that all data should be shared. Then there are a lot of journals like the *Journal of Vertebrate Paleontology* (Berta & Barrett 2011) that merely encourage full data archiving. Even within the same society there is policy variance: of the Linnean Society journals, the *Biological Journal of the Linnean Society* requires Dryad data archiving, whilst the *Zoological Journal of the Linnean Society* does not mandate data archiving, anywhere. I have had to contact the editor of the *Zoological Journal of the Linnean Society* many times with regards to data issues in that journal. It would help my research, and presumably many others, if *Zoological Journal of the Linnean Society* took a stronger approach with regards to its data sharing guidelines.

Conclusions

Palaeontological data and its availability in the digital era is an interesting subject with many ongoing developments. For many types of data that would concern palaeontologists, there are no unsolved technical barriers in the way of sharing data openly anymore; the only barrier is social adoption, willingness to share. For phylogenetic data there are well-established data standards such as Nexus and 'hennig' with which to exchange data in small plain

text files, as well as specialist databases for it, e.g. MorphoBank. This phylogenetic data is increasingly being uploaded online. But for images and photographs the trend is different. Despite a much wider selection of databases available, I detect a certain reluctance from palaeontologists to upload their specimen research photographs in their entirety.

Palaeontology, and indeed all morphology-based biological research, is utterly dependent upon the interpretation of specimen morphology, so it is vital that photographic imagery of these specimens and their attributes are made available for all to see and use (Ramírez et al. 2007; Cranston et al 2014). Until full, high-resolution images of specimens are abundantly and openly available online, systematic palaeontology will continue to be an expensive endeavour, often requiring researchers to travel to museums all across the world to view and take photos of specimens they need for their comparative research. Thus, even despite the Internet, much of palaeontological research still operates in a kind of pre-Gutenberg manner akin to the age where scholars had to travel to each of the best libraries in the world to read books of which there were no copies anywhere else. The Internet has revolutionised the dissemination of written works, enabling their free and easy copying. But, for palaeontological specimens and research-quality images of them, the digital revolution has really yet to begin. For three-dimensional imaging, the many hundreds of gigabytes of raw tomographic data required for each specimen may seem to be a valid barrier for not sharing them openly online. However, I see no such good excuse as to why there are not more openly available high-resolution photographic images of palaeontological research specimens. The infrastructure is certainly in place and cost-efficient, if not 'free', for researchers—it just needs to be used!

References

Alroy, J adapted by Uhen, M D 2013 *Frequently Asked Questions* (Paleobiodb.org). Available at http://paleobiodb.org/#/faq/citations [Last accessed 14 August 2014].

Berta, A and Barrett, P M 2011 Editorial. *Journal of Vertebrate Paleontology*, 31(1): 1. DOI: http://dx.doi.org/10.1080/02724634.2011.546742.

Cranston, K, Harmon, L J, O'Leary, M A and Lisle C 2014 Best Practices for Data Sharing in Phylogenetic Research. *PLOS Currents Tree of Life*. 2014 Jun 19. Edition 1. doi: http://dx.doi.org/10.1371/currents.tol.bf01eff4a6b60ca4825c69293dc59645.

Fairbairn, D J 2011 The advent of mandatory data archiving. *Evolution*, 65: 1–2. DOI: http://dx.doi.org/10.1111/j.1558-5646.2010.01182.x.

Haddouti, C S 2014 *Verstoß Gegen CC-Lizenz: Deutschlandradio Muss Zahlen*. Available at http://www.heise.de/newsticker/meldung/Verstoss-gegen-CC-Lizenz-Deutschlandradio-muss-zahlen-2151308.html [Last accessed 14 August 2014].

Hagedorn, G, Mietchen, D, Morris, R, Agosti, D, Penev, L, Berendsohn, W and Hobern, D 2011 Creative Commons licenses and the non-commercial condition: implications for the re-use of biodiversity information. *ZooKeys*, 150: 127-149. DOI: http://dx.doi.org/10.3897/zookeys.150.2189.

Kishnor, P and Peters, S 2013 *Paleobiology Database Now CC BY.* Available at http://creativecommons.org/weblog/entry/41216 [Last accessed 13 August 2014].

Klimpel, P 2013 *Consequences, Risks, and Side-Effects of the License Module Non-Commercial – NC 1-22*. Available at http://openglam.org/files/2013/01/iRights_CC-NC_Guide_English.pdf.

Kueffer, C, Niinemets, Ã, Drenovsky, R E, Kattge, J, Milberg, P, Poorter, H, Reich, P B, Werner, C, Westoby, M and Wright, I J 2011 Fame, glory and neglect in meta-analyses. *Trends in Ecology & Evolution*, 26: 493-494. DOI: http://dx.doi.org/10.1016/j.tree.2011.07.007.

Mounce, R 2014 MorphoBank image content analysis. *Figshare*. DOI: http://dx.doi.org/10.6084/m9.figshare.994172.

O'Leary, M A and Kaufman, S 2011 MorphoBank: phylophenomics in the "cloud". *Cladistics* 27: 529-537. DOI: http://dx.doi.org/10.1111/j.1096-0031.2011.00355.x.

Piwowar, H A, Day, R S and Fridsma, D B 2007 Sharing detailed research data is associated with increased citation rate. *PLOS One*, 2: e308+. DOI: http://dx.doi.org/10.1371/journal.pone.0000308.

Piwowar, H A, Vision, T J and Whitlock, M C 2011 Data archiving is a good investment. *Nature*, 473: 285. DOI: http://dx.doi.org/10.1038/473285a.

Piwowar, H and Vision, T J 2013 Data reuse and the open data citation advantage. *PeerJ*, 1: e175. DOI: http://dx.doi.org/10.7717/peerj.175.

R Core Team 2014 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramírez, M J, Coddington, J A, Maddison, W P, Midford, P E, Prendini, L, Miller, J, Griswold, C E, Hormiga, G, Sierwald, P, Scharff, N, Benjamin, S P and Wheeler, W C 2007 Linking of digital images to phylogenetic data matrices using a morphological ontology. *Systematic Biology*, 56: 283-294. DOI: http://dx.doi.org/10.1080/10635150701313848.

The MorphoBank Project 2012 *FAQ*. Available at http://www.morphobank.org/index.php/FAQ/Index [Last accessed 13 August 2014].

Uhen, M D, Barnosky, A D, Bills, B, Blois, J, Carrano, M T, Carrasco, M A, Erickson, G M, Eronen, J T, Fortelius, M, Graham, R W, Grimm, E C, O'Leary, M A, Mast, A, Piel, W H, Polly, P D and Säilä, L K 2013 From card catalogs to computers: databases in vertebrate paleontology. *Journal of Vertebrate Paleontology*, 33: 13-28. DOI: http://dx.doi.org/10.1080/02724634.2012.716114.

Wickham, H 2009 ggplot2: elegant graphics for data analysis. Springer New York.