THE IMPORTANCE OF OPEN ACCESS TO CHRONOLOGICAL INFORMATION: THE INTCHRON INITIATIVE

Christopher Bronk Ramsey*[1], Maarten Blaauw[2], Rebecca Kearney[1], Richard Staff[3]

* Corresponding author: christopher.ramsey@rlaha.ox.ac.uk

[1] Oxford Radiocarbon Accelerator Unit, School of Archaeology, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, United Kingdom.

[2] ^{14}CHRONO Centre, Queen's University Belfast, 42 Fitzwilliam Street, Belfast BT9 6AX, United Kingdom.

[3] Scottish Universities Environmental Research Centre, Rankine Avenue, Scottish Enterprise Technology Park, East Kilbride G75 0QF, United Kingdom.

ABSTRACT. The development of chronologies relies on integrating information from a number of different sources. In addition to direct dating evidence, such as radiocarbon dates, researchers will have contextual information which might be an environmental sequence or the context in an archaeological site. This information can be combined through Bayesian or other types of age-model. Once a chronology has been developed, this information can be used to estimate, for example, chronological uncertainties, rates of change, or the age of material which has not been directly dated.

Dealing with the information associated with chronology building is complicated and re-evaluation of chronologies often requires structured information which is hard to access. Although there are many databases with primary dating information, these often do not contain all of the information needed for a chronology. The Chronological Query Language (CQL) developed for OxCal was intended to be a convenient way of pulling such information together for Bayesian analysis. However, even this does not include much of the associated information required for reusing data in other analyses.

The IntChron initiative builds on the framework set up for the INTIMATE (Integrating Ice core, Marine and Terrestrial Records) chronological database (Bronk Ramsey et al. 2014) and is primarily an information exchange format and data visualization tool which enables users to pull together the types of information needed for chronological analysis. It is intended for use with multiple dating methodologies and while it will be integrated with OxCal, is intended to be an open format suitable for use with other software tools. The file format is JSON which is easily readable in software such as R, Python and MatLab. IntChron is not primarily intended to be a data depository but rather an index of sites where information is stored in the relevant format. As an initial step, databases of radiocarbon dates from the Oxford Radiocarbon Accelerator Unit (including those for the NERC radiocarbon facility), the RESET tephra database, the INTIMATE chronology database and regional radiocarbon databases for Egypt and Southern Africa are all linked. The intention is that users of OxCal will also be able to make published data accessible to others and to store working data, visible only to the user, to be used with the associated analysis tools. The IntChron site allows data from third party sources to be accessed through a representational state transfer (REST) application programming interface (API) in a number of different formats (JSON, csv, txt, oxcal) and associated bibliographic information in BibTeX format.

The aim of the IntChron initiative is to make it easy for users to provide data (in the single JSON format with limited minimum requirements) as well as to access data and tools, while promoting robust chronologies including realistic estimates of uncertainties. It is hoped that this will help to bring the chronological research communities to a point where data access is as easy as it is in some other fields. This is particularly important for Early Career Researchers and for those seeking to use large datasets in novel ways.

INTRODUCTION

Throughout the scientific disciplines there is a revolution taking place in how data is disseminated and used. In some fields such as genetics there are major initiatives like GenBank[1] which ensure that data is available to other researchers and in a form which enables further use. In the area of high-energy physics data sharing has been so critical that it was responsible for the creation of the WorldWideWeb[2] by Tim Berners-Lee in 1990. In this context it is reasonable to ask ourselves where we are within the field of chronological data in the area of usable data sharing.

In the early days of radiocarbon dating, laboratories made considerable efforts in making data available to researchers through 'date-lists'. This was a particular focus of this journal at the time. However, the published radiocarbon 'date-list' has become a rarity now due to the large number of dates being generated and because paper publication of such information makes the data difficult to use and provides insufficient space for associated information. In many ways we have now regressed to a state where accessing chronological information has become more difficult. For radiocarbon dates themselves, we usually now rely on the data being published in journal articles and books, in a myriad of different formats, and often missing key information such as the laboratory code (which potentially allows further information to be found) and associated stable isotopes. In most instances the dates themselves are published in the form released by the associated laboratory and so conforming to the standard formats (Stuiver and Polach 1977, Reimer et al. 2004, Millard 2014) but some older data may not comply. For stable isotope measurements it is important that only values measured by Isotope Ratio Mass Spectrometry (IRMS) are widely distributed because values measured by AMS cannot be used to infer the original isotope ratios within the samples themselves due to isotopic fractionation through the combustion and graphitization stages of sample preparation. Additional problems exist with dated environmental records. For example, the proxy data against depth and age-depth models are rarely available in any open-access form, making reanalysis, or analysis for different purposes impossible without specific requests to authors who may be too busy to respond, unwilling to share the data, or no longer in the field. This does not show the commitment to open science to which we should all be aspiring.

There are a number of competing pressures which help to perpetuate this situation. The first is that some of these data are quite complex in their nature and therefore difficult to make available in a consistent format. However, this is not usually the case for radiocarbon dates. In some cases very complex schema or requirements create a barrier to sharing data amongst the scientific community. If we want to encourage data sharing we should make it relatively simple. This may not include all possible associated metadata, but it is a significant advance if key data are available and referenced to places where more metadata are available. In the case of radiocarbon dates, for example, pre-treatment details and calculation methods are probably better explained within peer-reviewed publications rather than trying to encapsulate all of this information into a single data source.

---

[1] https://www.ncbi.nlm.nih.gov/genbank/

[2] https://www.w3.org/People/Berners-Lee/WorldWideWeb.html

The purpose of this paper is to suggest a way forward which resolves some of these data presentation and access problems. One important element of this is that it allows for a distributed data model with data being held in different databases, files and archives, and compiled by both users and producers of data. Another crucial aspect is that the data format is easy and concise, so that basic information is available in addition to more extensive information, where applicable. The final key component is that all of the data should be associated with a publication so that further information can be found if it is required.

Before looking at the proposed solution, this paper will look at some partial solutions which have helped to make sharing of chronological information easier already.

DATABASES

Quite reasonably the provision of radiocarbon dates in published date-lists has been replaced by availability in databases. In general, however, these databases tend to have particular uses of the data in mind and the format, methods of access, and associated data are all very different. In some cases the databases also have a limited lifetime due to funding constraints. Because radiocarbon dates are normally commissioned by users of radiocarbon labs, these labs can only make the data available if they know that they have been published. This makes it harder for labs to make their data public even if it has been funded by organisations that require open data access.

Here we highlight three examples of radiocarbon databases which illustrate some of the advantages and limitations of such databases. The specific examples are chosen because they are also relevant to the discussion of the IntChron initiative below.

The Oxford Radiocarbon Accelerator Unit Database[3] (ORAUD) is a laboratory-based database which allows public access to published radiocarbon dates measured at the lab. Originally it was set up to provide online access to radiocarbon dates published in the laboratory's date-lists periodically released in the journal Archaeometry. Since then, this has been expanded to include any dates for which the lab knows the publication – principally those where the lab has been involved in the publication, or for dates funded through the UK's Natural Environment Research Council (NERC) national radiocarbon facility and measured at Oxford, for which publication data are collected. This has over 10,000 radiocarbon dates in it. Because the data are directly drawn from internal records, the integrity of the radiocarbon data itself is assured. However, alternative data, in particular site locations and names, species, etc. are all supplied by submitters and so vary somewhat in format, precision of location and detail available. This type of database can have considerable advantages in making a relatively large overall dataset available, with links to publications for further details and from a wide range of submitters. It is unlikely that these data would be so easily accessible if each submitter had to make their own arrangements for making them available.

The Egyptian Radiocarbon Database[4] (ERD) was set up for a specific project aimed at assessing the relationship between radiocarbon dates and the Egyptian dynastic chronologies (Rowland and Bronk Ramsey 2011). This database has much more information on individual samples including categorization into reigns, and museum

[3] https://c14.arch.ox.ac.uk/database

[4] https://c14.arch.ox.ac.uk/egyptdb

acquisition numbers. It also includes radiocarbon dates measured at a wide range of different laboratories. This type of information is very specific to this project and not suitable for a more generic radiocarbon database. In this case there are a smaller number of radiocarbon dates (~1650) and the details about context and site location have been checked as part of the compilation. On the other hand, the radiocarbon data themselves are largely taken from the literature in some cases without some useful information, for example, the $\delta^{13}C$ values. Because this database was developed for a specific project, it is effectively a static resource unless someone can manage and regularly update the information.

The third example radiocarbon database is the Southern African Radiocarbon Database[5] (SARD). This has been set up to help research into this region (Loftus et al. submitted). This again contains radiocarbon dates from a whole range of different laboratories and has associated data associated with environmental context, relevant archaeological periods and context. This has been set up to be a live collection of data from the region rather than for a specific project. Again the data has been compiled to ensure consistency of contextual information but sometimes lacking details such as the $\delta^{13}C$ values.

All three databases provide information which is potentially very useful for large scale analysis or for identification of relevant publications and further resources. Four key elements are required for these data to be useful:

1. Linked references that can provide further information on the measurements and their context. Where possible this includes the DOI to ensure unequivocal identification of the publication.
2. Uncalibrated radiocarbon dates with their uncertainty and associated laboratory code. This ensures that date information can be tied to a laboratory and that duplicate data from different sources can be identified.
3. Site information including name and country and ideally longitude and latitude to provide a geo-reference. The precision of the location information might vary and so cannot be guaranteed to provide an exact location but should be tight enough to ensure that sites with similar names are distinguished and overall geographic distributions of dated material are correct. In some cases, there may be security reasons not to publish very precise location information.
4. The material being dated and, where known, the species, though in both cases the exact form this takes is likely to vary.

This sets some minimum requirements for data which can be usefully shared. Of course for some applications you might need much more information than this – and in those cases only some databases will contain what you require. Ideally pre-treatment information would also be included, and this is something which radiocarbon laboratories should be encouraged to provide, although the details are likely to be laboratory specific. There are other databases which also include chronological information from other dating methods, and databases which include more complex types of information such as age-depth models and time transfer functions (Bronk Ramsey et al. 2014). These are relevant to many of the same research questions as primary radiocarbon date information.

[5] https://c14.arch.ox.ac.uk/sadb

TOOLS

When considering the importance of open access data, and suitable methods for dissemination, the different types of analytical tools which can be used for chronological analysis should also be considered.

There are a wide range of types of analytical tool which are relevant here. These include very generic tools such as spreadsheets, packages like R and MatLab, and programming languages like Python. These can be used for all kinds of data analysis and the input required for them is typically unstructured data tables (tab delimited ascii text or .csv) or in the case of R, MatLab and Python, more structured data formats such as JSON.

We then also have very specific tools such for radiocarbon specifically for performing a whole variety of tasks from radiocarbon calibration to age depth modelling (Blaauw and Christen, 2005, 2011, Bronk Ramsey 2008, 2009, Bronk Ramsey and Lee, 2013, Lanos and Philippe 2017). These all require, as a minimum the uncalibrated radiocarbon date and its uncertainty. In the case of age-depth models (Blaauw and Christen, 2005, 2011, Bronk Ramsey 2008) we also require depth information and for special analysis (Bronk Ramsey and Lee, 2013) longitude and latitude.

Many of these tools are either embedded within packages (particularly R), or are web-based and so all have the capability of being able to draw on information available via the internet. All of these more specific tools do require specific data fields to be interpreted in a particular way and for this reason providing information for them is better done in a more structured form.

INTEGRATION OF DATABASES AND TOOLS

The need for data to be structured can be addressed in a number of different ways. The most straightforward is to integrate the data sources and tools into one package. This is the approach taken for tools associated with using time transfer functions provide in the pilot INTIMATE database (Bronk Ramsey et al. 2014), and also for the tools for comparing glass chemistries in the RESET tephra database (Bronk Ramsey et al. 2015). The advantage of this approach is that the data have been specifically formatted for the tools in question. The disadvantage is that it is difficult to use the tools with data not stored within these particular databases, and in particular with unpublished data that are the subject of active research.

It seems a better strategy to separate data sources from tools but to structure the data in a way which makes analysis possible without manual manipulation.

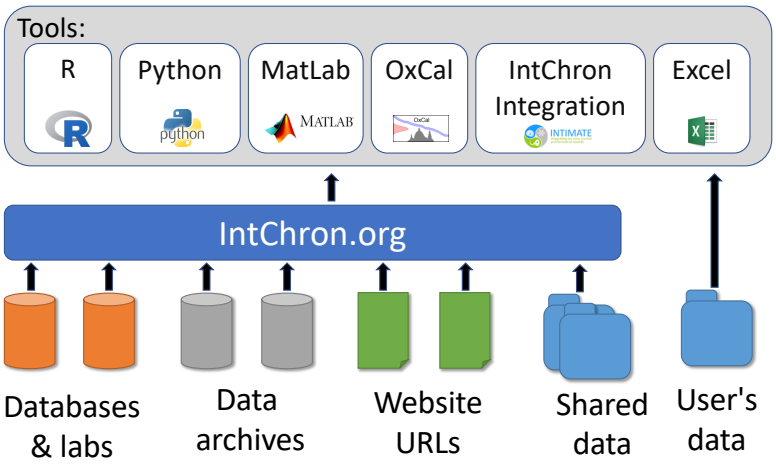

Figure 1: Schematic showing the role of the IntChron site in indexing data from different sources and the IntChron integration tool as one of many possible tools which can use the data from the IntChron site and as compiled by users.

THE INTCHRON INITIATIVE

The IntChron initiative aims to tackle both the issues surrounding easy access to data and those associated with feeding data directly into chronological tools (including generic packages like R and MatLab). The whole approach is ultimately designed around the requirements for chronological analysis. This requires a certain degree of structure, while retaining the flexibility to deal with different types of information, and simple minimum requirements to encourage participation.

It is important to stress that IntChron is not intended to be a database in its own right. What it aims to do is to provide an index of available data, and the tools to convert that data into different formats either for viewing or for input into statistical tools (Figure 1). The main intention is to provide a single point of access to multiple sources and to remove the need for data providers to generate data in more than one format.

A simple analogy might help to explain how IntChron is intended to work. If you have a publication with a DOI then you can always access this through the doi.org website. For example:

https://doi.org/10.1126/science.1189395

will automatically route you to the publication with the DOI of 10.1126/science.1189395. In this case the data for this publication are also available on a couple of the databases discussed above (ORAUD, ERD) and have been indexed by IntChron. This means that:

https://intchron.org/doi/10.1126/science.1189395

will automatically call up the radiocarbon dates from this study as a web-page. Furthermore if you want those data in a specific format (.txt, .csv, .oxcal or .json), you can retrieve this using the same url but with the required extension so for example:

https://intchron.org/doi/10.1126/science.1189395.json

will download a JSON data object with all of the data suitable for reading into a data object in R, MatLab or Python. It is possible to crawl through all of the indexed data which means that it is regularly indexed by search engines such as Google and also means that if you wish to download all of the data this is also possible to do by scripting. By using the .bib file extension it is possible to download all of the bibliographic information in BibTeX format which is an input format used in almost all bibliographic management systems or for direct use in LaTeX.

Methodology

The underlying methodology is quite simple. Each indexed data-source provides a list of publications with associated data as a JSON object. The data-source is also set up to return the data for any of those publications as a JSON object. In practice this can be generated as required live (which is the case with the databases linked so far) or it could be as a static file.

It is also possible for data to be lodged on data servers such as NOAA and then have these directly indexed within IntChron. In this configuration, only the file format (JSON) is being used allowing access to the tools and the search on IntChron. The only disadvantage of this approach is that if at some point the url on the data server changes there will be no mechanism to automatically update that.

To make retrieval of information more efficient from live databases, the system is also set up to retrieve information by site or record name, where the information might be covered by several different publications. However, this is not a requirement for data-sources because any such indexing can be undertaken by crawling the data if required. Further indexes and filters will be provided within the IntChron interface as the number of indexed datasets expands.

The format for the JSON objects exchanged is given in the schema at https://intchron.org/schema but the key elements of the main 'INTCHRON.Project' data type are:

- bibliography: a list of associated publications; as a minimum this contains the DOI. If this is not available, the full bibliographic details for each publication in a format based on BibTeX should be included.
- project_series_list: an array of project level data series.
- records: a list of records containing data
- parameters: an array of parameters used with specified units (if required)

Each record object (within the record array) contains:

- header: with georeferenced, site name, country etc
- series_list: an array of data series associated with this particular record
- refs: an array of references (details given in the bibliography)

Not all of these elements need be present. In the example given above for example the data for 10.1126/science.1189395 is just presented as two radiocarbon date-lists which are given as project level data series. A radiocarbon date-list data series (R_Datelist) is defined as a JSON object which contains:

- series: a name for the series
- parameter_list: a comma-separated list of parameters in addition to those in the standard R_Datelist
- refs: an array of references

- data: an object which contains arrays of:
    - labcode (the laboratory code for this radiocarbon date, or a double lab code if the pretreatment and measurement are from different labs: this should be a unique string)
    - site (the name of the site from which the samples come)
    - region (optional)
    - country
    - longitude
    - latitude
    - sample (the reference for this particular sample)
    - material (the material type of the sample itself such as bone, etc)
    - fraction (optional – but recommended when available: the fraction dated such as collagen, lipid, humin, etc)
    - species
    - qual (a specific flag for dates that cannot have their radiocarbon date specified in terms of a mean and standard uncertainty: it is normally blank, ">" for greater than dates where only r_date is used, or "m" for modern dates where r_date and r_date_sigma are not used)
    - r_date (the uncalibrated radiocarbon date, as specified by Stuiver and Polach 1977)
    - r_date_sigma (the standard uncertainty in r_date)
    - F14C (optional – but highly recommended where available: the F14C value)
    - F14C_sigma (optional: the standard uncertainty in F14C)
    - measurement_year (optional: the year of measurement)
    - d13C (the stable carbon isotope value if, and only if, this has been measured independently by IRMS)

Further parameters, for example relating to sample pre-treatment can be added if available. The IntChron site itself is set up to receive these files in JSON format and either pass them on to the requester if .json has been specified, or converts them to another format if required. The site also has links to open the data either in OxCal, or in the Integration tool (see below).

JSON objects should be flagged at the top level with information about the object type, the and the details of the compilation such as the person or organisation responsible (using the ORCID for individuals), the date of compilation and of retrieval. Details are given in the schema.

Integration tool

The data format used within IntChron has been developed with chronological analysis in mind. As part of the initiative there is also an IntChron Integration tool which makes use of the data format for display, import and export of the data and also for various forms of analysis. If you go to any dataset on the IntChron site you will see a link 'Open as: json' which will open the dataset within this tool. You can also open the tool directly from the toolbar button named [IntChron].

From within this tool you can look at a dataset or import data from IntChron using the [File > Import] menu item.

This tool is still under development but will integrate many of the other online tools already available including:

- OxCal: there are methods included to generate age-depth and other models from within the tool and to run simple calibrations of datasets in the background. In time this will be the primary way to organise data that you wish to use within OxCal models and to make these data available on publication.
- INTIMATE: the tools for applying time-transfer functions for comparing data on different timescales (Bronk Ramsey et al. 2014) is already incorporated into this tool, and IntChron has been set up to read data from the INTIMATE database to facilitate this. The methods here also allow for the use of age-depth models to generate time series for proxy data in environmental records.
- Tephra: work is underway to integrate the RESET tephra database (Bronk Ramsey et al. 2015) and associated tools into this same package. This will allow users to plot and compare tephra glass chemistry.

Overall this online application is intended to provide a whole range of different tools which can be applied to data indexed within IntChron or data which users have of their own and is as yet unpublished. The aim is to provide a comprehensive open data management system for chronological projects.

Laboratories and third-party databases

It is fairly simple for labs and other data holders to make data available through this new mechanism providing that they have an index of publications associated with their lab-codes. It may be (as with the ORAUD) that only a subset of dates is available in this way – but this will still make a valuable addition. The main other obstacle is the details of the bibliographic information; if there is a DOI available then IntChron has mechanisms to pull all other associated information from CrossRef so this should be no problem. Where publications don't have DOIs there is more work in ensuring that the bibliographic information is formatted as required.

Assuming the publication information is available the lab then needs to provide a url which allows the data to be accessed. As an example we will use the SADB discussed above. The following URL directly accesses this database:

https://c14.arch.ox.ac.uk/sadb/public.php?ref=

If called on its own this will retrieve the full bibliography for that database. From within this then if you choose a publication, for example with a DOI of 10.1080/0067270x.2018.1436740, the data for this (in JSON) can be retrieved using:

https://c14.arch.ox.ac.uk/sadb/public.php?ref=doi:10.1080/0067270x.2018.1436740

Because of the indexing within IntChron the same information can also be viewed in different formats through IntChron at:

https://intchron.org/doi/10.1080/0067270x.2018.1436740

but what is happening when you do that is IntChron is retrieving the data from the SADB and then reformatting it for the user as required. This means that a laboratory participating only needs to provide the data in the one (JSON) format.
In the case of labs, it will also be useful to users if you can provide radiocarbon dates generated for them in the IntChron format. This is because users can then use this data (without sharing it with others) with the analytical tools designed to work with the IntChron format. We already do this for users of the Oxford Radiocarbon Accelerator Unit.

Users of chronological data

The integration tool described above has two related functions. One is to enable users to work with data retrieved from IntChron. The other is to enable them to format their own data in the same way, both for their own use and for sharing.
The way this is intended to work is that users will work on a project, usually building on a mixture of published data (already available on IntChron) and new data. When they come to publish the project they can make their project file available (either on the IntChron site associated with their ORCID, or on some other data archive) and indexed through IntChron so that others can make use of the data for other research. In many cases it may be that the users are using unindexed data which is already published, in this case their static project file(s) can be an alternative to the construction of a fully-fledged database with all of the associated issues of long-term maintenance.

Tool developers

The data format is intended to be a fully open format and easy for tool developers to understand and use. The hope is that the IntChron initiative can help developers of new tools to have easy access to chronological data through the RESTful interface. We also intend, where possible to embed further tools such as Bacon (Blaauw and Christen 2011) into the online Integration tool.

CONCLUSIONS

The IntChron initiative aims to address major bottlenecks in the easy availability of chronological data. It provides a very simple generic data format, specifically designed to assist the development of chronological analysis tools, and an indexing mechanism which allows data to be distributed across a whole range of different databases, data archives, and websites (see Figure 1). The system is inherently extensible in that different data sources can add their own additional parameter definitions but the key parameters required for chronological analysis (such as the uncalibrated radiocarbon dates) share a common format. Participation of laboratories and other major data providers will help to make this useful to a broader range of users and ensure that data is as widely disseminated and cited as possible.
Any laboratories wishing to participate in data sharing or tool developers wishing to use the file format should contact the corresponding author.

REFERENCES

Blaauw, M., & Christen, J. A. (2005). Radiocarbon peat chronologies and environmental change. Journal of the Royal Statistical Society Series C-Applied Statistics, 54(4), 805–816.

Blaauw, M., & Christen, J. A. (2011). Flexible paleoclimate age-depth models using an autoregressive gamma process. Bayesian Analysis, 6(3), 457–474.
Bronk Ramsey, C. (2008). Deposition models for chronological records. Quaternary Science Reviews, 27(1-2), 42–60.
Bronk Ramsey, C. (2009). Bayesian analysis of radiocarbon dates. Radiocarbon, 51(1), 337–360.
Bronk Ramsey, C., Albert, P., Blockley, S., Hardiman, M., Lane, C., Macleod, A., Matthews, I. P., Muscheler, R., Palmer, A., & Staff, R. A. (2014). Integrating timescales with time-transfer functions: a practical approach for an INTIMATE database. Quaternary Science Reviews, 106, 67–80.
Bronk Ramsey, C., & Lee, S. (2013). Recent and planned developments of the program OxCal. Radiocarbon, 55(2-3), 720–730.
Bronk Ramsey, C., Housley, R. A., Lane, C. S., Smith, V. C., & Pollard, M. A. (2015). The RESET tephra database and associated analytical tools. Quaternary Science Reviews, 118, 33–47.
Lanos, P. and Philippe, A. (2017) Hierarchical Bayesian modeling for combining dates in archaeological context. Journal de la Soeciete Francaise de Statistique158 (2), pp.72-88.
Loftus, E., Mitchell , P. and Bronk Ramsey, C. (submitted) An archaeological radiocarbon database for southern Africa.
Millard, A.R. (2014) Conventions for reporting radiocarbon determinations. Radiocarbon, 56(02), pp.555-559.
Rowland, J.M. and Bronk Ramsey, C. (2011) Online C14 database for Egypt. Egyptian Archaeology, 38, pp.33-34.
Reimer, P.J., Brown, T.A. and Reimer, R.W. (2004) Discussion: reporting and calibration of post-bomb C-14 data. Radiocarbon, 46(3), pp.1299-1304.
Stuiver, M. and Polach, H.A. (1977) Discussion: reporting of ^{14}C data. Radiocarbon, 19(3), pp.355-363.