



Commentary

The Problem with Dates: Applying ISO 8601 to Research Data Management

Kristin Briney

University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Abstract

Dates appear regularly in research data and metadata but are a problematic data type to normalize due to a variety of potential formats. This suggests an opportunity for data librarians to assist with formatting dates, yet there are frequent examples of data librarians using diverse strategies for this purpose. Instead, data librarians should adopt the international date standard ISO 8601. This standard provides needed consistency in date formatting, allows for inclusion of several types of date-time information, and can sort dates chronologically. As regular advocates for standardization in research data, data librarians must adopt ISO 8601 and push for its use as a data management best practice.

Correspondence: Kristin Briney: briney@uwm.edu

Keywords: standards, data information literacy, date-time formatting, data standards

Rights and Permissions: Copyright Briney © 2018



All content in Journal of eScience Librarianship, unless otherwise noted, is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Dates are a common element of managing research data. Researchers regularly record dates as data points, write dates in research notebooks, label observations by date, and communicate dates to collaborators. Dates also represent a significant hurdle in data cleaning due to inconsistent and culturally specific formatting. For example, depending on where you are in the world, “9/1/91” can represent either September 1, 1991 or January 9, 1991. The same date may also be written “Sept 1, 1991,” “01-09-1991,” “1.Sep.1991,” etc. Normalizing dates is an annoyance, yet not an uncommon issue when working with research data.

Data librarians use a variety of strategies for managing and normalizing dates. This represents a huge gap in our data management toolkit, given the prevalence of date data and our expertise with standardization. Date-time formatting should be considered within the suite of regular research data management advice that data librarians dispense. This commentary asserts that data librarians should adopt the international date standard, ISO 8601 (International Organization for Standardization 2004), to format dates and liberally advise researchers to do the same.

As librarians, we are familiar with standards and it should come as no surprise that a standard exists for formatting dates. ISO 8601 was first developed in 1988, bringing together several existing ISO standards for date and time. It is currently in its third edition, dating from 2004, with updates expected in the near future. Other ISO 8601-based date and time standards exist, such as the W3 Note on Date and Time Formats (Wolf and Wicksteed 2018) and RF3339 (Internet Engineering Task Force 2002), with more non-ISO 8601 standards within specific cultures and software tools.

There are many benefits to using a consistent date format and ISO 8601 in particular. Consistent dates are easier to process and easier to reformat, if necessary, and can reduce ambiguity regarding the exact date to which a value refers. ISO 8601 is an internationally recognized standard that can be used to create that consistency. The standard comes with added benefits that the format is extensible, allows for sorting, and enables mathematical comparison between dates. For extensibility, the standard actually consists of several different variants under one umbrella standard, allowing researchers to also include extra information like time (more on this below). With respect to sorting, ISO 8601 formatted dates sort chronologically as information is ordered from largest unit of time to smallest; this gives the standard an edge in usability. Finally, ISO 8601 expresses all date information numerically which facilitates easier calculation and comparison when using dates as data. Given the prevalence of dates in research data, ISO 8601 is a natural standard to adopt.

As mentioned above, ISO 8601 is a standard comprised of several variants. The most readily adoptable are the date formats YYYY-MM-DD or YYYYMMDD. So September 1, 1991 would be written as either 1991-09-01 or 19910901. Both are acceptable under the ISO 8601 standard, though the version with dashes is more human readable. Adoption of one or the other may also depend on software requirements or character limitations.

A few other useful formats under the ISO 8601 umbrella include:

- Year and month: YYYY-MM (e.g. 1991-09)
- Year: YYYY (e.g. 1991)

- Date and time: YYYY-MM-DDTHH:MM:SS (e.g. 1991-09-01T11:00:00)
- Year and week: YYYY-Www (e.g. 1991-W35)
- Year, week, and day: YYYY-Www-D (e.g. 1991-W35-7)
- Year and ordinal day: YYYY-DDD (e.g. 1991-244)

Note that the week starts on Monday and time uses a 24-hour clock. It is too much to cover every ISO 8601 variation in this short commentary—see the standard itself for more specifics—but this list of most useful variations highlights the standard’s breadth. While YYYY-MM-DD is probably the most commonly used format, specific research needs will dictate the use of other variants.

In practice, applying the ISO 8601 standard leads to it being used in both research data and metadata. ISO 8601 has the benefit that all date information is expressed numerically, allowing for easier calculation, smoother analysis, and comparison between date values. Additionally, some software packages expect dates in the ISO 8601 format, such as in the “lubridate” library in R (Grolemund and Wickham 2011). The one analysis tool that will likely be the most challenging when working with ISO 8601 is Excel. Excel has a long history of mangling dates (Bahlai 2014; Woo 2014; Kosmala 2016; Broman and Woo 2018) and even interpreting non-date data as dates (Ziemann, Eren, and El-Osta 2016), so it should not be surprising that its date problems extend to ISO 8601. There are a few strategies for working with ISO 8601-formatted dates in Excel. First, the cells can be reformatted into ISO 8601, though this is a cosmetic change and can easily be reverted (this is because reformatting only alters the display and not the underlying configuration in which Excel stores date information). Second, dates can be represented as YYYYMMDD and interpreted by Excel as an 8-digit number. Finally, date parts can be stored in separate columns, one each for year, month, and day. The latter represents the best option as it is least likely to be mangled yet the information remains readily computable (Bahlai and Pawlik 2016). Always refer to the specifics of your preferred analysis tool for how it does or does not support ISO 8601-style dates.

Dates in metadata are another important use of ISO 8601. In many cases these dates act like dates appearing in a dataset, as discussed in the previous paragraph, but there is a special case worth further consideration: dates in file names. ISO 8601 and file names are a match made in heaven. The reason for this is that 8601-formatted dates sort chronologically. In combination with a consistent file naming scheme, this makes for wonderfully organized files. One useful example is in the file names of meeting notes, such as “Meeting_2018-10-31.docx.” Given a whole group of such files, it is simple to sort and scan through documents to find what one needs.

While ISO 8601 has many uses within research data management, it isn’t perfect. One problem is that ISO 8601 is based on the western, Gregorian calendar, which is not used in all countries. Additionally, while the standard can theoretically handle BCE (Before the Common Era) dates, it is not an ideal format for this information. Moreover, few people are familiar with ISO 8601, which may lead to date confusion. This is compounded by the fact that some 8601-formatted dates are less human readable.

As data librarians advise researchers to adopt more standardized workflows, we should not forget to apply date standards to this work. ISO 8601 is a natural partner for research data management, yet there are many examples of data librarians not utilizing this standard. I have adopted ISO 8601 liberally in my own data, my own file names, and in committees to which I belong, and will never go back—the benefits I reap from readily scanned file names and easily analyzed dates are simply too great. I therefore urge my peers to learn the benefits of this standard themselves and, in turn, advocate for its adoption with the researchers they advise.

Acknowledgments

The author thanks Yasmeen Shorish for her valuable feedback on a draft of this commentary.

Disclosure

The author reports no conflict of interest.

References

- Bahlai, Christie. 2014. "Dealing with Dates as Data in Excel." *Practical Data Management for Bug Counters*. <https://practicaldatamanagement.wordpress.com/2014/07/02/dealing-with-dates-as-data-in-excel>
- Bahlai, Christie, and Aleksandra Pawlik. 2016. "Data Organization in Spreadsheets: Dates as Data." *Data Carpentry*. <https://datacarpentry.org/spreadsheet-ecology-lesson/03-dates-as-data>
- Broman, Karl W., and Kara H. Woo. 2018. "Data Organization in Spreadsheets." *The American Statistician* 72(1): 2-10. <https://doi.org/10.1080/00031305.2017.1375989>
- Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with Lubridate." *Journal of Statistical Software* 40(3): 1-25. <http://dx.doi.org/10.18637/jss.v040.i03>
- International Organization for Standardization. 2004. "ISO 8601:2004." <https://www.iso.org/standard/40874.html>
- Internet Engineering Task Force. 2002. "RF3339." <https://www.ietf.org/rfc/rfc3339.txt>
- Kosmala, Margaret. 2016. "Beware This Scary Thing Excel Can Do to Your Data!" *ECOLOGICAL BITS*. <http://ecologybits.com/index.php/2016/07/06/beware-this-scary-thing-excel-can-do-to-your-data>
- Wolf, Misha, and Charles Wicksteed. 2018. "Date and Time Formats." *W3C*. <https://www.w3.org/TR/NOTE-datetime>
- Woo, Kara. 2014. "Abandon All Hope, Ye Who Enter Dates in Excel – UC3 :: California Digital Library." *UC3 Blog*. <https://uc3.cdlib.org/2014/04/09/abandon-all-hope-ye-who-enter-dates-in-excel>
- Ziemann, Mark, Yotam Eren, and Assam El-Osta. 2016. "Gene Name Errors Are Widespread in the Scientific Literature." *Genome Biology* 17(1): 177. <https://doi.org/10.1186/s13059-016-1044-7>