Computers & Geosciences 153 (2021) 104791

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Harmonizing heterogeneous multi-proxy data from lake systems

Gregor Pfalz [a,b,c,d,*], Bernhard Diekmann [a,b], Johann-Christoph Freytag [c,d], Boris K. Biskaborn [a,b,**]

[a] Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Research Unit Potsdam, Telegrafenberg A45, 14473, Potsdam, Germany
[b] University of Potsdam, Institute of Geosciences, Karl-Liebknecht-Str. 24-25, 14476, Potsdam-Golm, Germany
[c] Einstein Center Digital Future, Robert-Koch-Forum, Wilhelmstraße 67, 10117, Berlin, Germany
[d] Humboldt-Universität zu Berlin, Unter den Linden 6, 10099, Berlin, Germany

ARTICLE INFO

Keywords:
Data integration
Entity-relationship diagram
Sediment cores
Conceptual data framework
Arctic lakes

ABSTRACT

When performing spatial-temporal investigations of multiple lake systems, geoscientists face the challenge of dealing with complex and heterogeneous data of different types, structure, and format. To support comparability, it is necessary to transform such data into a uniform format that ensures syntactic and semantic comparability. This paper presents a data science approach for transforming research data from different lake sediment cores into a coherent framework. For this purpose, we collected published and unpublished data from paleolimnological investigations of Arctic lake systems. Our approach adapted methods from the database field, such as developing entity-relationship (ER) diagrams, to understand the conceptual structure of the data independently of the source. We demonstrated the feasibility of our approach by transforming our ER diagram into a database schema for PostgreSQL, a popular database management system (DBMS). We validated our approach by conducting a comparative analysis on a set of acquired data, hereby focusing on the comparison of total organic carbon and bromine content in eight selected sediment cores. Still, we encountered serious obstacles in the development of the ER model. Heterogeneous structures within collected data made an automatic data integration impossible. Additionally, we realized that missing error information hampers the development of a conceptual model. Despite the strong initial heterogeneity of the original data, our harmonized dataset leads to comparable datasets, enabling numerical inter-proxy and inter-lake comparison.

1. Introduction

On-going global warming impacts Arctic landscapes through the "Arctic amplification" effect, where temperatures in the Arctic exceed the average Northern Hemisphere surface air temperature change (Biskaborn et al., 2019b; IPCC, 2014; Miller et al., 2010). Lake systems are thereby among the most valuable, but at the same time also the most complex climatic archives of the earth as they enshrine various environmental information into their sediment (Bradley, 2015; Brauer, 2004; Cohen, 2003). The regional and global climate as well as non-climatic influencing factors both affect the sedimentation process of lake systems (Fritz, 2008; Wilke et al., 2016; Zolitschka et al., 2015). Understanding them helps to improve our perception of the earth system.

Analytical data derived from determining lake sediment properties, also known as proxy data, are essential for reconstructing lake histories, as they indicate change of environmental conditions (Bradley, 2015). While scientists continue to collect new data from lake systems each year, thorough data handling of already existing datasets might help to fill remaining knowledge gaps of past changes. The quality of these older datasets varies depending on different factors, such as date of creation, individual project goals, available laboratory resources, and personnel bias (Cai and Zhu, 2015; Heidorn, 2008; Wang et al., 2001). When integrating these existing datasets into a coherent framework and reporting standard, we can work with higher reliability and reproducibility thus enabling large-scale synthesis studies.

The number of repositories containing valuable data for paleolimnological studies has increased in recent years (Elger et al., 2016;

* Corresponding author. Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Research Unit Potsdam, Telegrafenberg A45, 14473, Potsdam, Germany.
** Corresponding author. Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Research Unit Potsdam, Telegrafenberg A45, 14473, Potsdam, Germany.
E-mail addresses: Gregor.Pfalz@awi.de (G. Pfalz), Boris.Biskaborn@awi.de (B.K. Biskaborn).

https://doi.org/10.1016/j.cageo.2021.104791
Received 26 June 2020; Received in revised form 1 April 2021; Accepted 16 April 2021
Available online 24 April 2021
0098-3004/

Latif et al., 2019; Muster, 2018). Various studies using these repositories have already shown the effectiveness of multi-proxy, multi-site investigations through the synthesis of data from various sources (e.g., Bouchard et al., 2016; Kaufman et al., 2020; PAGES 2k Consortium, 2017; Subetto et al., 2017). Khider et al. (2019) recently proposed a reporting standard for new and past ('legacy') paleoclimate datasets, which includes the reporting of metadata information and measured data from lake sediments. The positive reception of their proposed model shows the consensus within the paleoclimate research community in favor of uniform standards for reporting measurement data of various kinds.

Still, there are non-digitalized, unprocessed or unpublished data hidden on local storage devices, old field and lab books as well as in hand-written documents (Curry and Moosdorf, 2019; Heidorn, 2008). Some 'legacy' datasets might not meet the requirements of such a standard as proposed by Khider et al. (2019). This prevents older data from being included in any multi-site investigations. However, these datasets are potentially invaluable sources of almost forgotten knowledge (Muster, 2018).

In this paper, we present a *conceptual* integration approach to enable a comprehensive comparison of datasets of varying quality from laboratory analysis of lake sediment. The specific objectives of this paper are (I) to provide a conceptual entity-relationship (ER) model for merging heterogeneous multi-proxy data into a common framework from a database-centric perspective, and (II) to translate the conceptual model into a reference implementation using the PostgreSQL database management system (DBMS) to perform a comparative analysis on acquired and transformed data. Our approach will allow scientists to perform their data analysis on the integrated data with less effort compared to an analysis using the raw (original) data.

2. Methods

2.1. Data collection

For the cleansing and integration process presented in this paper, we used a collection of published and unpublished laboratory data and corresponding metadata from lake sediment cores. The majority of data came from online data repositories (e.g., Pangaea, GFZ Data Service) or institute-internal data sources of the Alfred Wegener Institute (e.g., expedition reports, personal communication). For the purpose of reproducibility and tracking, we collected and stored (meta) data about the sources of the collected data. This data is accessible in the repositories mentioned in the 'Code and data availability' section.

We manually curated the laboratory data by using different data validation approaches (Pannekoek et al., 2013; Sun et al., 2011). We assessed both laboratory data and metadata hereby on their completeness, consistency, accuracy, and precision (Batini et al., 2009; Batini and Scannapieca, 2006; Sebastian-Coleman, 2013). In a first step, we performed type checking, as laboratory data is known to be primarily numeric. We substituted unsuitable characters by numeric values using the Python package 'pandas' (Reback et al., 2020). A physical range check followed the previous check to ensure that values do not exceed physical ranges (Sun et al., 2011). If values exceed their logical physical range, we then removed them from the dataset. We standardized names of common proxies (e.g., 'Aluminum' to elemental symbol 'Al') or associated units (e.g., core lengths from centimeters to meters) to ensure consistent naming across all datasets. We logged all cleaning actions carried out during the entire validation process for provenance reasons. If possible, we examined original files from measuring instruments as well as original publications to avoid any conversion errors. We consulted the corresponding responsible scientist for any clarification when needed.

We selected the following information as minimum requirements for metadata information to be included in our study:

- unique core identifier ('*CoreID*'),
- geographical information (latitude, longitude),
- information about the field campaign (name, year),
- site name,
- lake type,
- water depth at coring location, and
- composite (i.e. cumulative) core length, derived from overlapping core segments.

The proposed preselection helped us to distinguish clearly unique entries of lake sediment cores from other entries. If the meta data was not included in the acquired dataset, then we searched in related literature to determine the uniqueness of the core. We generated unique identifiers for the cores using the pattern '*{FirstAuthorLastName} {LakeID} {ExpeditionYear} {CoreNumber}*', when no unique identifier was available. CoreIDs, names of field campaign, and sites are stored in English using the Latin alphabet. The information about latitude and longitude is recorded as decimal degrees (minimum precision: three decimal places), while water depth and core length are given in meters (precision: two decimal places). Besides using the geographical location to validate the uniqueness of a sediment core, we further used the coordinates in ArcGIS, Google Earth, and HydroLAKES database (Messager et al., 2016; Meyer et al., 2020) to check that the scientists correctly placed and assigned site names to cores.

In total, there were 70 metadata information entries for unique core sediments. Fig. 1 illustrates the geographical distribution of this metadata compilation. It shows that the availability of information with high latitude values (50° N to 90° N) dominates the spread. The dominance of the latitude values in this range is due to the clear research focus by the Alfred Wegener Institute and its partners on the Arctic.

2.2. Conceptual approach

During data collection, we encountered variety of structures and inconsistent data, which makes data of different sources incomparable. In order to facilitate a better integration of those heterogeneous datasets into a database, we decided to design a unified scheme based on the existing structures to store all base values. We defined as base values all values, which cannot be derived from any other data values in any of the data source. Under this assumption, the first step within the (conceptual) schema design approach was to develop generic concepts, which should resemble aspects of limnological studies. We followed the principles of database design for a redundant free data representation. As such, a (conceptual) data model describes all aspects of the real world by identifying relevant entities and the relationships among them in the domain of discourse (Codd, 1970; Elmasri and Navathe, 2009; Teorey et al., 2008).

Both, entities and relationships are associated with specific attributes, i.e. descriptive properties that are inherent for each of them in the real world. For instance, researchers conduct a core drilling at a specific geographical location (described by latitude and longitude) and water depth with a particular drilling device. The resulting core has a unique core identifier (e.g., '*International GeoSample Number*' (Conze et al., 2017), or simply '*CoreID*') and a composite core length. Therefore, we reduced this process to one core-specific entity *Drilling* with five attributes (see Fig. 2), while one attribute is a composite attribute ('*Geo-information*') and another is the key attribute ('*CoreID*') of this entity.

Naturally, entities have an interdependency with other entities. Teorey et al. (2008) provides guidance for how to design entities and their relationships. As an example for paleolimnological studies: It is only possible to conduct a core drilling at one location at any given time; hence, there is a one-to-one relationship between the entities *Drilling* and *Lake*. Other binary data modeling cardinalities are one-to-many or many-to-many relationships (Garcia-Molina et al., 2002; Teorey et al., 2008). In general, entity-relationship diagrams, also known as '*ER diagrams*' (Figs. 2 and 3), or alternatively, diagrams using the Unified

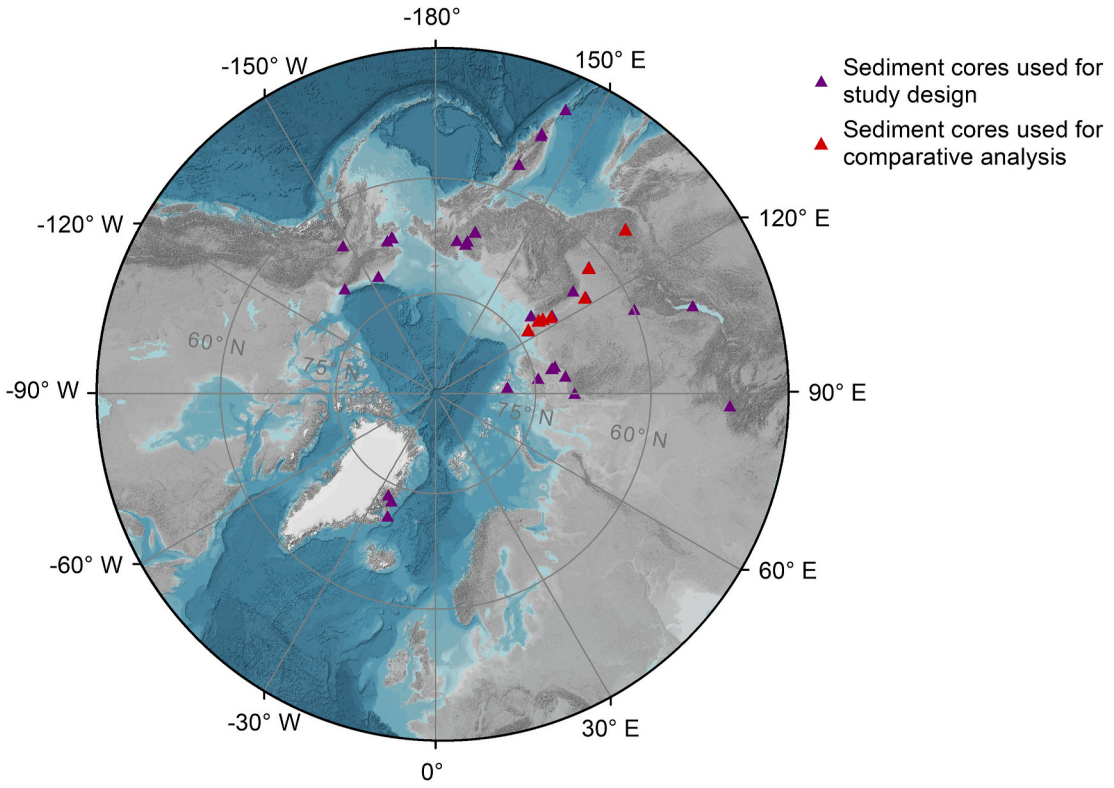

Fig. 1. Geographical distribution of lake sediment cores used for the study design (triangles, n = 70). Red triangles (n = 8) indicate lake sediment cores used for the comparative analysis of total organic carbon (TOC) and bromine (Br) content shown in this study. ArcGIS Basemap: GEBCO Grid 2014 modified by AWI. The outer ring in the graphic corresponds to 45° N. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

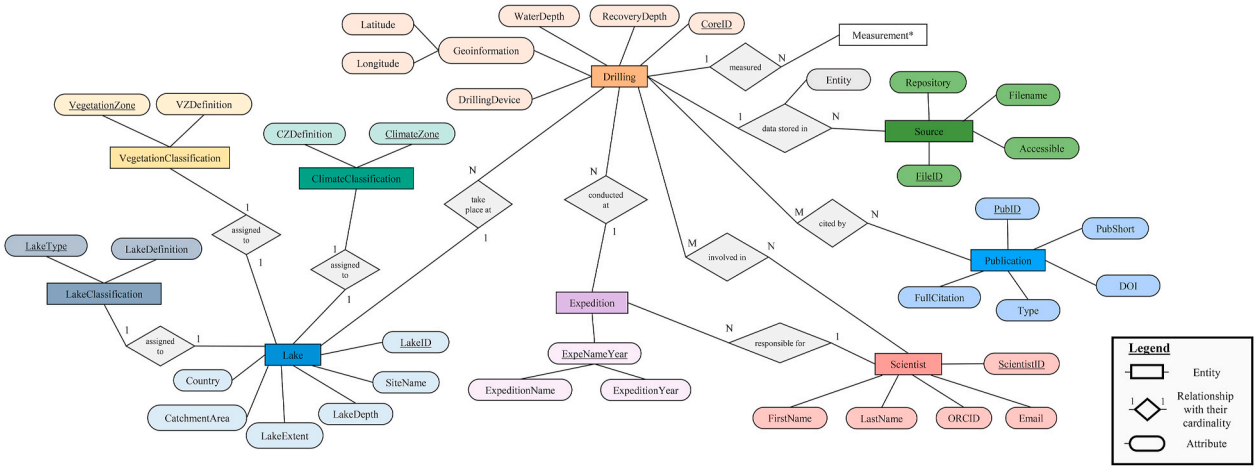

Fig. 2. Entity-relationship diagram of nine core-specific entities (rectangular boxes) with their attributes (boxes with rounded corners) connected through relationships (diamond-shaped objects). The core-specific entities refer to information about core retrieval and associated information. Entity 'Measurement' in the right upper corner establishes the connection between both entity groups.

Modeling Language (UML) (Fig. 4, Fig. 5), allow us to visualize entities together with their relationships (Chen, 1975; Garcia-Molina et al., 2002; Teorey et al., 2008). This visual representation accelerates the implementation of the developed concept in a database management system (DBMS).

In our case of paleolimnological studies, we identified two groups of entities: measurement-specific and core-specific entities. Measurement-specific entities represent individual laboratory measurements for proxy determination. Core-specific entities do not only characterize the core retrieval, but also operational metadata describing the drilling. The operational data includes data about the surveyed lake, field campaign/expedition, responsible scientist, and publications. This allocation further supports initiatives for optimal metadata management and investigating for possible systematic errors, i.e. data trustworthiness (Batini et al., 2009; Bertino and Lim, 2010).

Besides the superordinate entity *Drilling* in the group of core-specific entities (Fig. 2), we split the field campaign into the entities *Lake*, *Expedition,* and *Scientist*. *Lake* describes further details about the lake at which the drilling took place. The *Lake* entity also relates to the entities *ClimateClassification*, *VegetationClassification*, and *LakeClassfication* entity, which describe the climate, vegetation and lake origin at time of core retrieval to the lake, respectively. For operational data about the

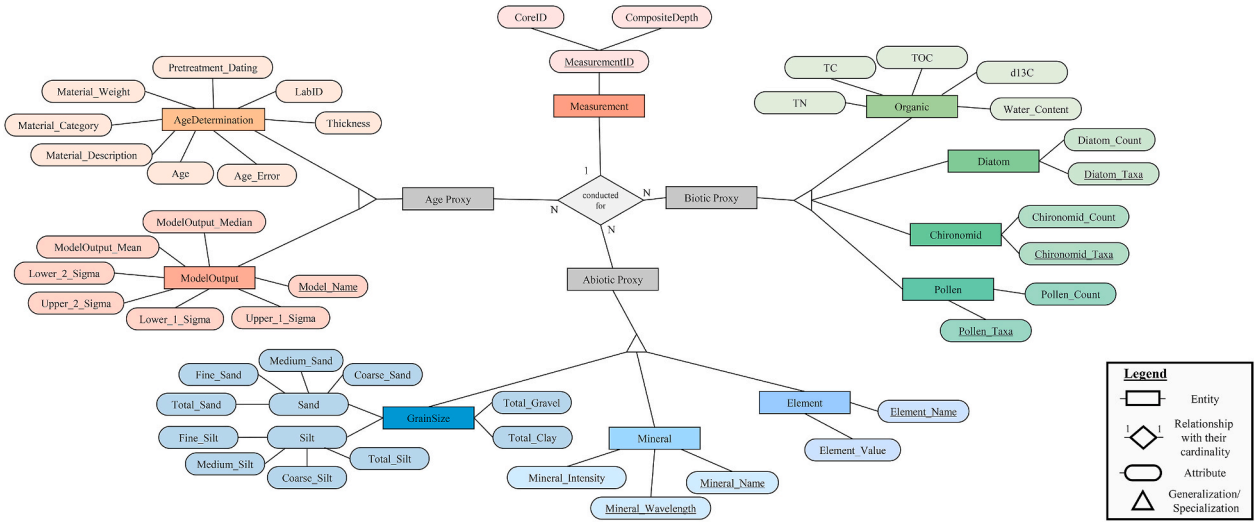

Fig. 3. Entity-relationship diagram of ten measurement-specific entities (rectangular boxes) with their attributes (boxes with rounded corners) connected through relationships (diamond-shaped object). Measurement-specific entities are consistent with the measured laboratory data of sediment cores.

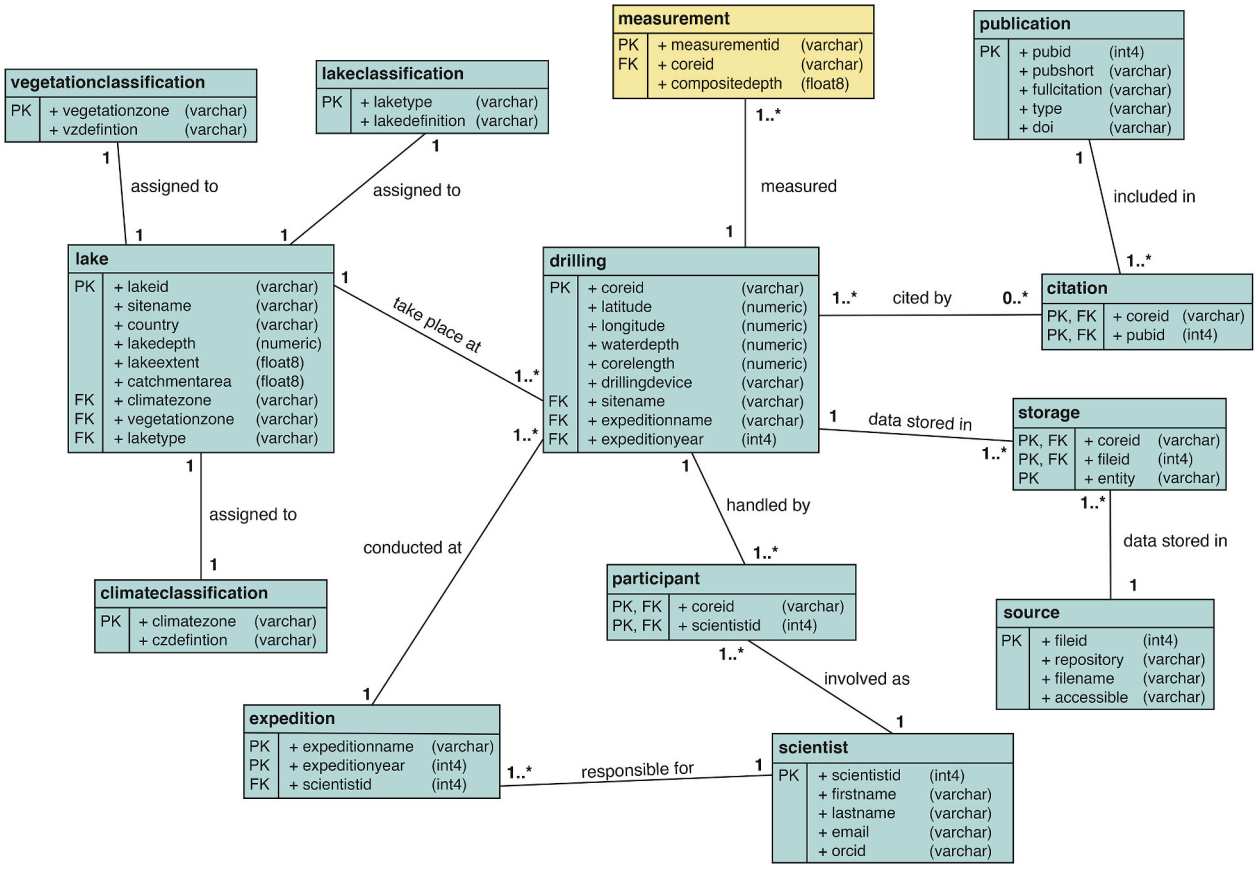

Fig. 4. Unified Modeling Language (UML) diagram of core-specific entities for the reference implementation. Entities (rectangular boxes) consist of their name, attributes and their PostgreSQL data type in tabular form, and an indication whether the attributes are primary keys (PK), foreign keys (FK), or both. The entities are connected by a relationship (solid line) to another entity. The numbers on the solid line indicate the cardinality of that relationship. This figure includes the entity 'measurement' to show connection to the other derived measurement-specific entity group. List S1 in the supplementary material further describes each entity.

field campaigns in relation to the drillings, we designed both *Expedition* and *Scientist* entities for temporal attribution and contact details. *Publication* consists of the important publications relating to each drilling, while *Source* gives information about the files used to produce the measurement-specific entities and therefore allows us to reproduce them with higher precision.

For the measurement-specific entities (Fig. 3), the attributes describe measured quantities of the individual laboratory measurements. It was therefore vital for us to gain an understanding of the different laboratory methods used to analyze proxies. Based on the available data, we determined eleven proxies that were analyzed frequently in the datasets (Table 1). We examined each related study closely on its applied

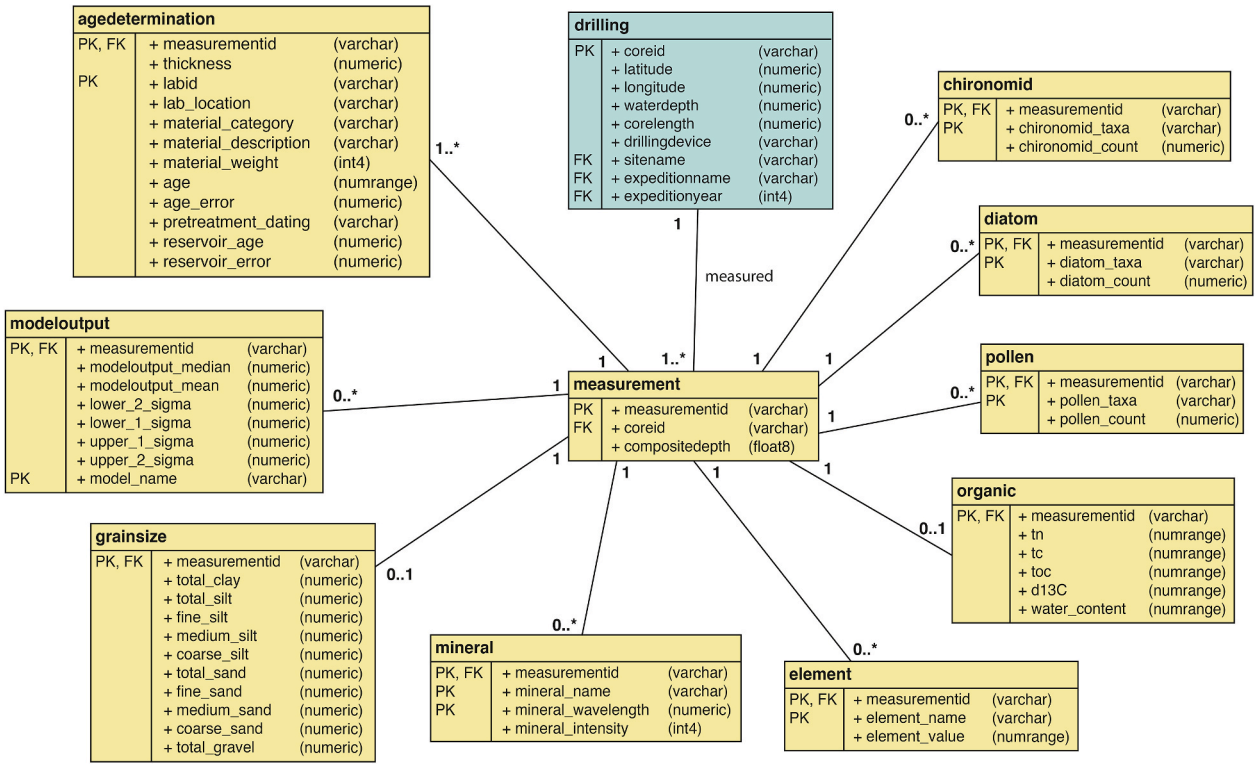

Fig. 5. Unified Modeling Language (UML) diagram of measurement-specific entities for the reference implementation. Entities (rectangular boxes) consist of their name, attributes and their PostgreSQL data type in tabular form, and an indication whether the attributes are primary keys (PK), foreign keys (FK), or both. The entities are connected by a relationship (solid line) to another entity. The numbers on the solid line indicate the cardinality of that relationship. This figure includes the entity 'drilling to show connection to the other derived core-specific entity group. List S1 in the supplementary material further describes each entity.

Table 1
Selection of proxies, which were frequently determined in the acquired laboratory datasets.

Abiotic proxies	Biotic proxies
Elements	Diatoms
Minerals	Chironomids
Grain size	Pollen
Water content	δ13C
	Total organic carbon (TOC)
	Total carbon (TC)
	Total nitrogen (TN)

methodology to determine said proxies. While most studies followed the same methodological approach, we found variations for three proxies, namely for elements, grain size, and Total Organic Carbon (TOC) (Fig. S1 in the supplementary material provides a pictorial representation). In our design, we decided to focus on one approach as a representative for both TOC and grain size. For the elemental proxy we kept the method selection optional.

Regarding the attributes for measurement-specific entities, we decided for the lowest common denominator to avoid artificial inflation. For instance, for a measurement independent abstraction of elemental data, the reduction leads to one entity *Element* with two attributes for (i) the name or symbol of the element, and (ii) the associated value at a specific depth inside the core. To place an emphasis on the applied method, we appended the unit of the measurement to the element symbol in the attribute '*Element_Name*', e.g., Aluminum measured in parts-per-million as '*Al_ppm*'. This strategy is only possible for elemental data, as the unit of measurement does not change and values from different methods for this proxy are not directly comparable.

Ultimately, we derived nine core-specific (Fig. 2) and ten measurement-specific (Fig. 3) essential entities from the acquired data. The entities *Measurement* and *Drilling* establish the linkage between the two entity groups. When analyzing sediment cores, scientists extract multiple proxy measurements from a specific depth along the core's length (one-to-many relationship). Hence, unique identifiers in *Measurement* contain a composite attribute consisting of the composite depth and the corresponding core identifier. As the core identifier '*CoreID*' is also the key attribute of the *Drilling* entity, it enables us to extract additional operational information belonging to the measurement.

In the second step of our conceptual approach, we had to ensure that a comparison between different datasets is feasible. Harmonizing laboratory data from geographically dispersed cores necessitates finding a common anchor point between those datasets. The sampling scheme of the individual proxies strongly depends on the depth within the sediment core, which itself depends on the research questions posed. Laboratory measurements could therefore be taken very frequently (i.e. every one to 2 mm), less frequently (i.e. every five to 10 cm), or where distinctive changes within the sediment core are visible. The time axis is the only constantly running physical quantity and common denominator on which we can place all measurements. For this conversion, one might use existing age-depth modeling software.

While users can set defined depth resolutions for the age-depth relationship within the modeling software to match varying proxy resolutions, a complete conversion of the proxies from individual depth-dependent to joint age-dependent will leave blanks. Logical approximations or interpolations of the individual proxy have to fill those vacant time slices when seeking comparable multi-site investigations (Birks and Birks, 2006). There are various well-established techniques available, such as in-filling techniques, spline interpolation or machine-learning based interpolation (Birks, 2012).

At first, we define the proxy with the lowest age resolution as the *base proxy* after the conversion from depth-dependent to age-dependent. We

then match other proxies with a higher resolution (*desired proxies*) to the base proxy. That is to say, we select values of the desired proxy at the time slices equal to the time slices of the base proxy. If the value of the desired proxy is not available at the exact time slice of the base proxy, we perform an interpolation of the desired proxy. We hereby follow the advice by Blaauw (2012) to avoid the use of overfitted values which could potentially result in misinterpretations. If we were to excessively interpolate values of a proxy with a lower resolution in order to fit the curve of proxies with a higher resolution, we would increase the likelihood of a misinterpretation. Additionally, internal lake dynamics influence biological activity within lake systems, which means that a higher or lower abundance of biological proxies within a sediment core might be an indication for a specific event (Biskaborn et al., 2019a). It is therefore debatable, whether an interpolation of biological proxy is reasonable, or we should use Bayesian modeling approaches instead (Huntley, 2012). For simplicity, we assume in our study that interpolation is feasible for the proxies involved in this study.

To synchronize values between sediment cores, we use binning to create equally spaced bins of time. We calculate the optimal bin size using the mean of the maximum proxy age resolution for the base proxy across all sediment cores. We then select the interpolated central valu es for each proxy measurement at the same interval as the bin size. We use all remaining values within an age bin to calculate the minimum and maximum value range for each proxy.

2.3. Comparative analysis

To set up for our comparative analysis, we implemented the developed ER data model as a database schema for a PostgreSQL database system (Version 11.2; PostgreSQL Global Development Group, 2018). For the implementation, we used the open-source software tool DBeaver (Version 7.0; DBeaver Community, 2020). Figs. 4 and 5 provide a visual representation of the reference implementation as UML diagram. List S1 in the supplementary material shows the individual entities with their attributes, a short explanation, the PostgreSQL data type, and an example.

After implementing the schema in PostgreSQL, we had to transform the available (raw) data to fit the proposed schema. Therefore, we set up a spreadsheet template with the same schema as the database to support the integration process. We show an example spreadsheet template for our reference implementation in the repository mentioned in the 'Code and data availability' section of this paper. The standardized data was then inserted into the database using a Jupyter notebook (Kluyver et al., 2016) using the package 'SQLAlchemy' (Bayer, 2012).

We performed the comparative analysis in a separate Jupyter notebook – for more information on the code we refer to section 'Code and data availability'. We selected total organic carbon (TOC) content and X-ray fluorescence (XRF) measured bromine (Br) content to showcase the basic functionality of our approach. Previous studies conducted by Biskaborn et al. (2016) and Kalugin et al. (2007) showed that bromine is a good indicator for changes of organic content in lake sediment and should therefore agree well with the TOC content (Rothwell and Croudace, 2015). Both proxies were measured in eight sediment cores within our database. First, we converted the measurement depths to the median ages from the corresponding age-depth model. To allow a transparent and comprehensible (re-)modeling of already existing age-depth relationships, we gathered all information regarding laboratory age analysis with its associated uncertainty. For a better reproduction of the age-depth relationship, we stored further information regarding the age determination. This information included the description of the dated material, involved laboratory, pretreatment methods, and thickness of the dated sediment layer, if bulk sediment was dated.

For age-depth modeling, we used the existing open-source MATLAB software package 'Undatable' (Lougheed and Obrochta, 2019) with the improved IntCal20 calibration curve (Reimer et al., 2020). We implemented an additional script to access all eight sediment cores from the database and then computes age-depth models in bulk (Undatable settings: nsim = 10^5, bootpc = 30, xfactor = 0.1). This recalculation reduces potential biases introduced by the authors and possible differences between modeling software output (Trachsel and Telford, 2017; Wright et al., 2017). Age-depth relationships produced during our harmonization process are no replacement for the original relationships identified by the contributing authors (cf. McKay and Kaufman, 2014). We stored the resulting output created by the modeling software as '*ModelOutput*' in the database using an additional script.

We determined our base proxy for our comparative analysis using the mean proxy age resolution. We also considered the impact the higher resolution proxy had on the data, if we were to use the proxy with higher resolution as the base proxy. To enable a synchronized comparison of TOC and bromine, we used interpolation to replace missing values by approximated values. We applied a piecewise polynomial interpolation for existing gaps using the Python package 'SciPy' (Virtanen et al., 2020). We assessed all cores on their maximum proxy age resolution for the base proxy to determine the optimal bin size. Ultimately, we binned each measurement into its respective age bin and determined from all measurements the minimum and maximum value ranges within each bin.

3. Results and discussion

Using research data from external sources always contains the risk that undocumented transformations (knowingly or unknowingly) changed the data after a laboratory analysis. This might lead to erroneous and incomprehensible data. Therefore, we contacted the responsible scientists for further inquiries regarding the data handling to avoid propagating possible errors. If there were inconsistencies between different approaches, then we documented this circumstance for traceability. Due to the design of the data model, we provide data, which allows a scientist to retrieve the original data files and publication for each proxy dataset. Such reference supports the important concept of lineage thus providing an improved contextualizing of the data, which might be important for further use of the data. We claim that good data cleansing can foster an interoperability amongst geoscientist and the use of automated data integration tools. However, the biggest challenge during the harmonization process of our collected dataset was the handling of varying data qualities. The most noticeable inconsistency was the heterogeneous structure within the data. While almost all data from online repositories followed syntactic rules, data from other sources did not stay within a coherent framework. Therefore, over the course of our investigation, we had to exclude the possibility of automated data integration. If future datasets follow the FAIR principle (Findable, Accessible, Interoperable, Reusable), we are convinced that automated data integration becomes possible (Latif et al., 2019; Stall et al., 2018; Wilkinson et al., 2016).

The use of a database management system (DBMS) for the comparative analysis has clear advantages over loosely connected, personal spreadsheets without the ability for integration. Currently, measurement data exists in different labs on different computer without the ability for a common usage and understanding. Our approach presents first steps towards a data-driven integration of such data. There are multiple reasons for using database techniques in the context of data transformation and integration resulting in set of homogenized data for further analysis Once the data are in a standardized format, the database provides high availability, high flexibility, synchronization, error recovery, and great efficiency. Integrated datasets further support data integrity within the database. Despite new developments in database research, relational database management systems (RDBMSs) still provide the best fit for laboratory data from paleolimnological studies as most of data generated by measuring instruments can be stored and accessed in a tabular form. Other geoscientific databases, such as Neotoma, Pangaea, or GTN-P, proved the reliability of RDBMSs (Biskaborn et al., 2015; Diepenbroek et al., 2002; Williams et al., 2018).

Additionally, most database management systems provide interfaces (APIs) to a series of programming languages, which makes it easier to retrieve and to analyze the stored data efficiently and effectively (cf. Elmasri and Navathe, 2009; Teorey et al., 2008).

Fig. 6 (A) shows the untransformed output from the reference implementation, where TOC and bromine content are dependent on the depth within each sediment core. We then transformed all values from depth-dependent to age-dependent (Fig. 6 B). Fig. 6 (B) illustrates the TOC and bromine values against their corresponding median age derived from the age-depth modeling software 'Undatable' (Lougheed and Obrochta, 2019). What stands out in Fig. 6 is the variability of consecutive measurements along the X-axis. Still, we determined elemental and TOC measurements to have the highest and third highest age resolution in our reference implementation, respectively (see Table 2). The resolution depends highly on the level of automation and treatment processes for each proxy. We can measure elemental data from non-destructive X-ray fluorescence (XRF) core scanning without any pretreatment at a depth resolution of 2 to 5 mm. Other proxy groups such as diatoms, chironomids, or pollen require time-consuming pretreatment and microscope-based analyses performed by individual scientists. Scientists accommodate the additional preparation by commonly taking fewer samples.

Based on these results, we selected TOC as our base proxy for the comparison of TOC and bromine. Fig. 7 compares the results from using (A) TOC and (B) bromine as base proxy for the matching process. Panel A is following our approach of choosing the proxy with the lower resolution as the appropriate base proxy and generating corresponding interpolated bromine values for each TOC value, if needed. In panel B we show that using the higher resolution proxy as base proxy instead, leads to an overestimation of specific events due to necessity of excessive interpolation. With the results from panel A we started to calculate optimal bin size and minimum and maximum value ranges for each bin. We determined 700-year bins to be the optimal bin size for our comparative analysis. Fig. 8 illustrates bromine and TOC values for all sediment cores binned into 700-year bins with their minimum and maximum value ranges. Through this approach we are able to transform

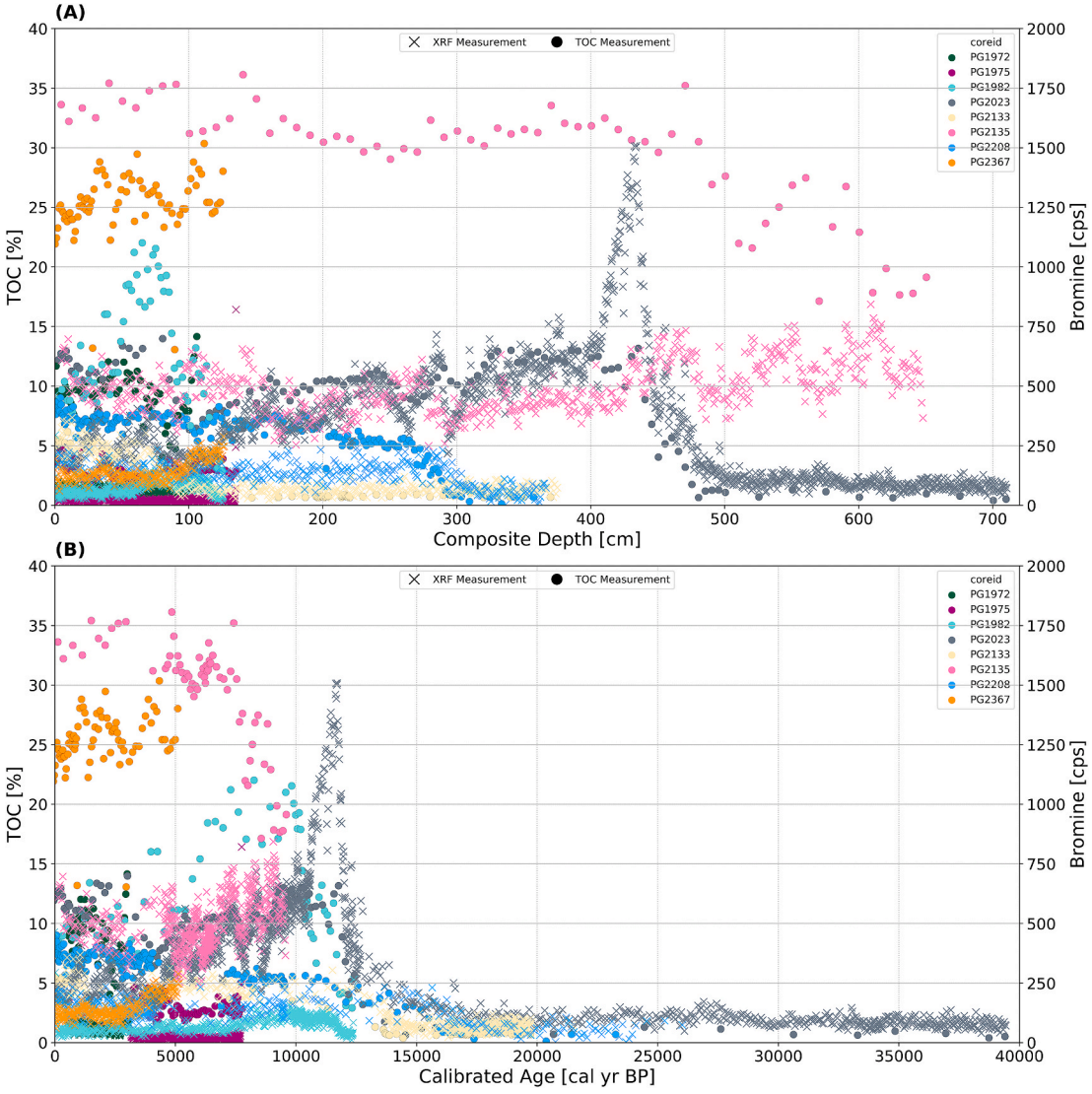

Fig. 6. Comparison of total organic carbon (TOC) and bromine (Br) content for eight selected sediment cores within the reference implementation and after being age-transformed. Panel A shows TOC and bromine measurements against the individual composite depth in centimeter of each sediment core, as existent in the reference implementation. In panel B, same measurements are transformed from composite depth to the corresponding calibrated median ages. The X-axis in panel B is based on calibrated median ages (calibrated years Before Present, cal yr BP) derived from age-depth modeling software Undatable (Lougheed and Obrochta, 2019) using the IntCal20 calibration curve (Reimer et al., 2020). Color codes are consistent over all measurements and plots for each sediment core. Circles markers represent TOC measurements in percent and cross markers indicate bromine measurements in counts per seconds (cps) using X-ray fluorescence (XRF). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Statistics of proxy sampling and proxy age resolution for each proxy in reference implementation. Q_{25}, Q_{50}, and Q_{75} correspond to the 25% quantile, median, and 75% quantile within each proxy resolution, respectively. Note: Total organic carbon (TOC) and Total nitrogen (TN) were measured together, hence the resolution and number of data points are the same.

Proxy	Number of data points	Proxy age resolution [yr]				Proxy sampling resolution [cm]			
		Mean	Q_{25}	Q_{50}	Q_{75}	Mean	Q_{25}	Q_{50}	Q_{75}
Element	8388	33.38	9.00	13.71	30.00	3.97	0.50	1.00	1.02
TOC/TN	2130	187.65	22.67	72.00	158.00	5.99	1.95	4.75	9.65
TC	1169	181.92	37.75	73.75	125.15	3.88	1.95	2.02	5.29
δ13C	1166	163.24	16.67	55.00	152.00	6.86	1.78	6.24	9.94
Pollen	760	217.07	51.00	126.00	185.00	8.84	3.28	8.78	10.93
Grain Size	462	312.61	76.50	134.00	355.00	15.51	2.06	7.93	18.99
Diatom	437	372.99	82.88	173.75	356.75	9.88	4.63	9.67	12.74
Mineral	418	295.37	77.00	119.50	232.50	12.22	4.77	5.87	9.48
Chironomid	152	472.87	126.87	236.50	566.75	14.08	12.41	15.01	16.68

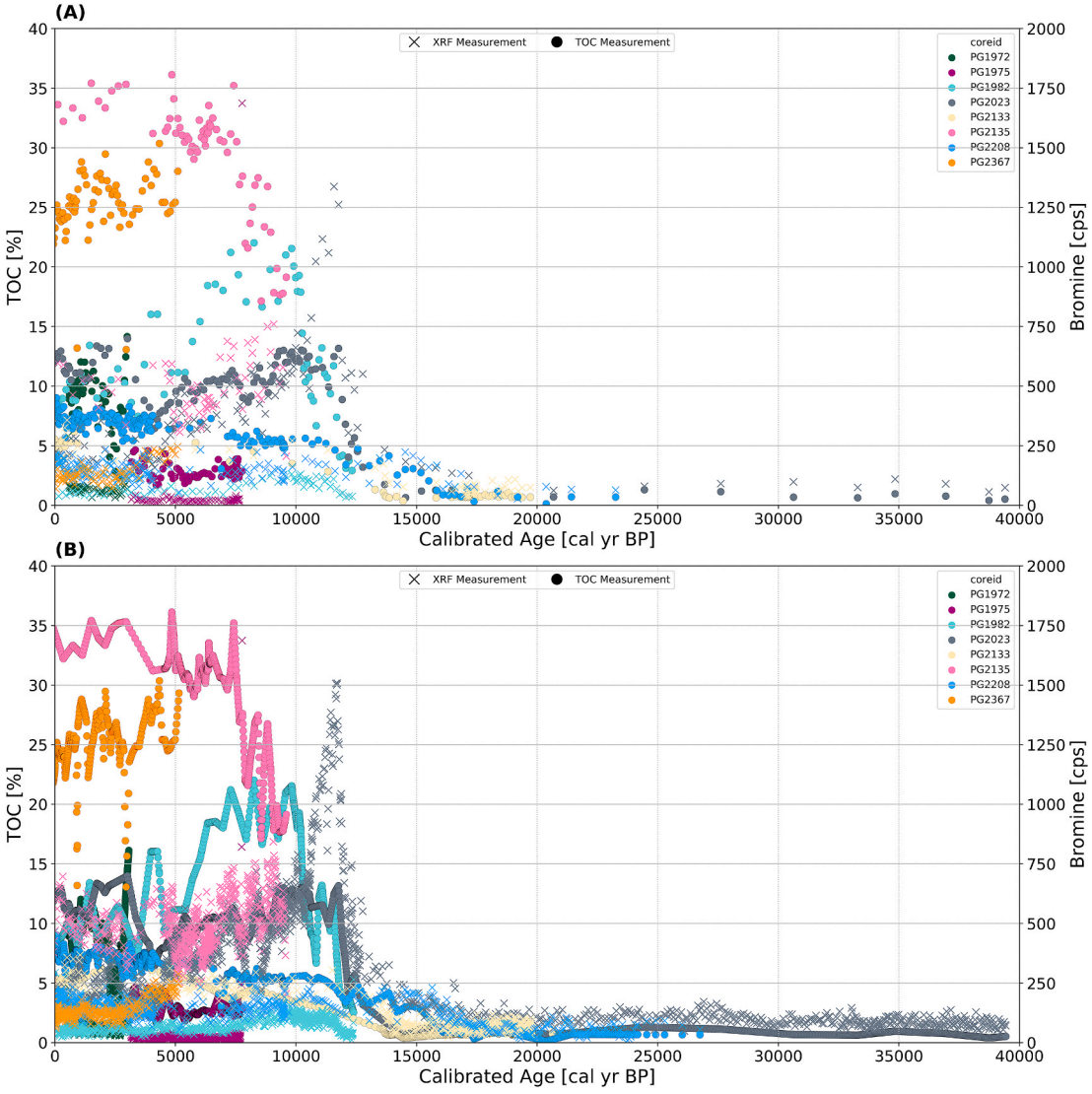

Fig. 7. Difference in interpolation approaches to synchronize measurements from total organic carbon (TOC) and bromine (Br). Panel A shows bromine values being synchronized to measurements of base proxy TOC. Panel B displays TOC values being resampled to match measurements in higher resolution of the base proxy bromine. In the case that the exact value of the desired proxy was not present at the specific age of the base proxy, we applied a piecewise polynomial interpolation to the desired proxy. Panel B therefore demonstrates a case were harmonization would result in a strong bias of resulting interpolated values. Color codes are consistent over all measurements and plots for each sediment core. Circles markers represent TOC measurements in percent and cross markers indicate bromine measurements in counts per seconds (cps) using X-ray fluorescence (XRF). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

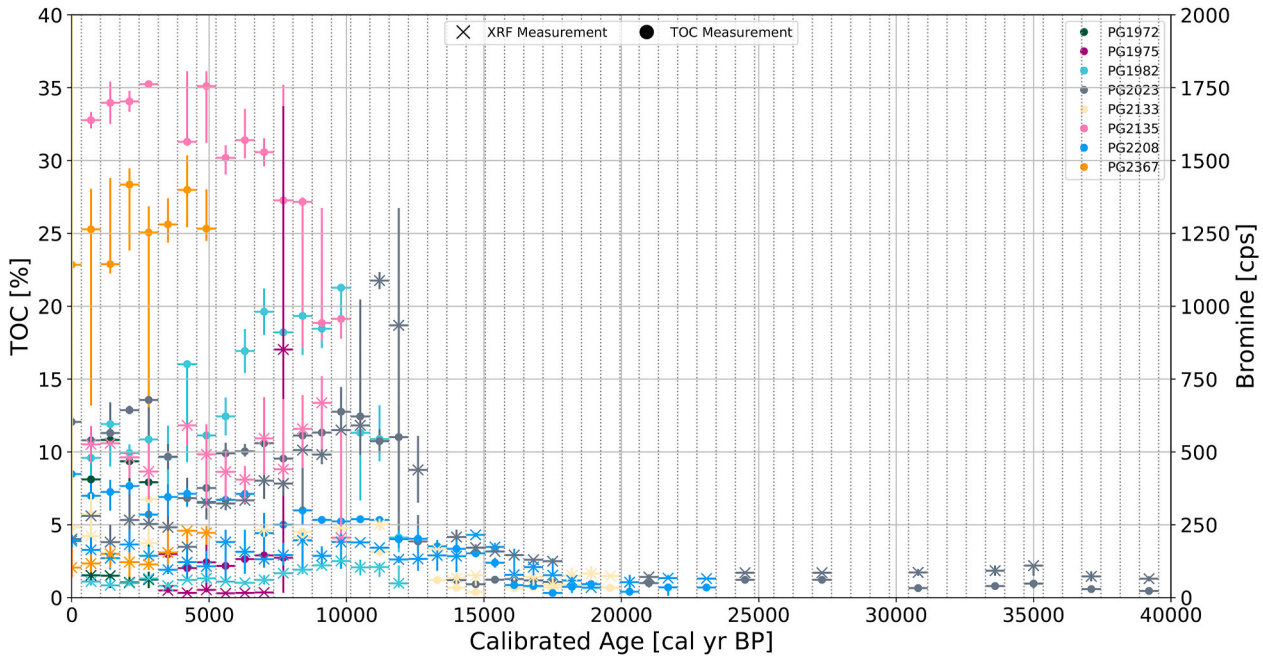

Fig. 8. Synchronized values for both total organic carbon (TOC) and X-ray fluorescence (XRF) derived bromine (Br) measurements using 700-year binning. We calculated the bins using the mean of the maximum resolution for the base proxy of each involved sediment core. Markers represent the interpolated value at the center of the bin at a continuous 700-year interval. We included all measurements within each 700-year bin to determine minimum and maximum values within each bin, represented by connected vertical lines. Circles markers represent TOC measurements in percent and cross markers indicate bromine measurements in counts per seconds (cps) using X-ray fluorescence (XRF). Dashed vertical lines display boundaries between 700-year bins. Horizontal lines display the age range for which measurement are applicable. Color codes are consistent over all measurements for each sediment core. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

both high and low resolution data into one single matrix containing a minimum amount of null values. By defining value ranges for each bin, we quantify the uncertainty of interpolated values at the center of each bin. We claim that through our approach, sediment cores are now comparable on both a temporal scale and an inter-proxy level.

However, natural systems hold some degree of uncertainty, as the analysis of proxies itself inhibits inevitable uncertainties (Amrhein, 2019; Goswami, 2014; Reschke et al., 2019). Therefore, it is the responsibility of scientists to handle and to report the uncertainties with their data measurements. By default, manufacturers of measuring devices usually report error intervals for their devices. The laboratory staff on-site refine the accuracy of these devices through the implementation of improved calibration methods. In many cases, however, the inaccuracy/deviation of results is not reported, visible in publications, or stated in any supplementary material. This situation is a serious obstacle in a multi-site investigation, especially when minor alterations in the data can determine distinct points of change. As a result, we omit static uncertainties and error information from our conceptual model and reference implementation in favor of dynamical error adjustment in the comparative analysis and appraisal of the comparable multi-proxy results.

While we designed our approach with the clear research focus on Arctic lake systems, we believe that our conceptual approach could be applicable to other research areas, such as long-term based in-situ data record (Su et al., 2018; Zeng et al., 2019). The practical implementation depends on the deliberate selection of the reference frame, i.e. universe of discourse (Elmasri and Navathe, 2009), and choosing the appropriate entities and relationships for the abstracted aspects of the real world. Further on, it is crucial to consider the specific domain knowledge and long-term scientific goals of the harmonization, when converting our conceptual approach into another domain.

4. Conclusions

The goal of this study was to provide paleolimnologists with a conceptual framework to integrated heterogeneous multi-proxy data from lake systems. The conceptual data model allows scientists to integrate heterogeneous data into a common database for further comparative analyses. We presented additional steps to prepare datasets for multi-site statistical investigation. We found that heterogeneous structures within the data, differing methods for determining proxy values, and missing error information still pose major challenges in developing a comprehensive data model. However, we concluded that despite strong initial heterogeneity our harmonized dataset still leads to comparable values, enabling numerical inter-proxy and inter-lake comparison.

Code and data availability

This study used multiple Python scripts for visualization, database connection, calculation and interpolation. The codes are available at GitHub (https://github.com/GPawi/MAYHEM). A SQL script to create a blank database following the introduced conceptual data model is provided in the same repository. Likewise, further files containing accessible links to the used datasets and contact details for unpublished data can be found there. Contact details comprise name of research group and personal communication address of working group leader.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the support of the Helmholtz Einstein

Berlin International Berlin Research School in Data Science (HEIBRiDS), the Alfred Wegener Institute - Helmholtz Center for Polar and Marine Research and its Open Access Publication Funds, the Potsdam Graduate School, the Einstein Center Digital Future and the Humboldt University of Berlin. The BMBF project PALMOD (grant no. 01LP1510D) supported our research as well. We thank Levent Vorpahl and Daria Kapustina for their support in the collection of metadata. We thank the four anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2021.104791.

Author contribution

GP wrote the manuscript with inputs from all the co-authors. BKB, BD and JCF supervised the works of GP. GP developed the conceptual framework and conducted the data analysis. JCF further supported the development of the entity-relationship diagram. BKB and BD provided unpublished data and metadata to the analysis.

References

Amrhein, D.E., 2019. How large are temporal representativeness errors in paleoclimatology? Clim. Past Discuss 1–26. https://doi.org/10.5194/cp-2019-10.
Batini, C., Scannapieca, M., 2006. Data Quality, Data-Centric Systems and Applications. Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-33173-5.
Batini, C., Cappiello, C., Francalanci, C., Maurino, A., 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41 https://doi.org/10.1145/1541880.1541883.
Bayer, M., 2012. SQLAlchemy. In: Brown, A., Wilson, G. (Eds.), The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks. aosabook.Org.
Bertino, E., Lim, H.-S., 2010. Assuring Data Trustworthiness - Concepts and Research Challenges, pp. 1–12. https://doi.org/10.1007/978-3-642-15546-8_1.
Birks, H.J.B., 2012. Overview of numerical methods in palaeolimnology. In: Tracking environmental change using lake sediments. Springer, Netherlands, pp. 19–92.
Birks, H.H., Birks, H.J.B., 2006. Multi-proxy studies in palaeolimnology. Veg. Hist. Archaeobotany 15, 235–251. https://doi.org/10.1007/s00334-006-0066-6.
Biskaborn, B.K., Lanckman, J.P., Lantuit, H., Elger, K., Streletskiy, D.A., Cable, W.L., Romanovsky, V.E., 2015. The new database of the global terrestrial network for permafrost (GTN-P). Earth Syst. Sci. Data 7, 745–759. https://doi.org/10.5194/essd-7-245-2015.
Biskaborn, B.K., Subetto, D.A., Savelieva, L.A., Vakhrameeva, P.S., Hansche, A., Herzschuh, U., Klemm, J., Heinecke, L., Pestryakova, L.A., Meyer, H., Kuhn, G., Diekmann, B., 2016. Late Quaternary vegetation and lake system dynamics in north-eastern Siberia: implications for seasonal climate variability. Quat. Sci. Rev. 147, 406–421. https://doi.org/10.1016/j.quascirev.2015.08.014.
Biskaborn, B.K., Nazarova, L., Pestryakova, L.A., Syrykh, L., Funck, K., Meyer, H., Chapligin, B., Vyse, S., Gorodnichev, R., Zakharov, E., Wang, R., Schwamborn, G., Bailey, H.L., Diekmann, B., 2019a. Spatial distribution of environmental indicators in surface sediments of Lake Bolshoe Toko, Yakutia, Russia. Biogeosciences 16, 4023–4049. https://doi.org/10.5194/bg-16-4023-2019.
Biskaborn, B.K., Smith, S.L., Noetzli, J., Matthes, H., Vieira, G., Streletskiy, D.A., Schoeneich, P., Romanovsky, V.E., Lewkowicz, A.G., Abramov, A., Allard, M., Boike, J., Cable, W.L., Christiansen, H.H., Delaloye, R., Diekmann, B., Drozdov, D., Etzelmüller, B., Grosse, G., Guglielmin, M., Ingeman-Nielsen, T., Isaksen, K., Ishikawa, M., Johansson, M., Johannsson, H., Joo, A., Kaverin, D., Kholodov, A., Konstantinov, P., Kröger, T., Lambiel, C., Lanckman, J.-P., Luo, D., Malkova, G., Meiklejohn, I., Moskalenko, N., Oliva, M., Phillips, M., Ramos, M., Sannel, A.B.K., Sergeev, D., Seybold, C., Skryabin, P., Vasiliev, A., Wu, Q., Yoshikawa, K., Zheleznyak, M., Lantuit, H., 2019b. Permafrost is warming at a global scale. Nat. Commun. 10, 264. https://doi.org/10.1038/s41467-018-08240-4.
Blaauw, M., 2012. Out of tune: the dangers of aligning proxy archives. Quat. Sci. Rev. 36, 38–49. https://doi.org/10.1016/j.quascirev.2010.11.012.
Bouchard, F., Macdonald, L.A., Turner, K.W., Thienpont, J.R., Medeiros, A.S., Biskaborn, B.K., Korosi, J., Hall, R.I., Pienitz, R., Wolfe, B.B., 2016. Paleolimnology of Thermokarst Lakes : a Window into Permafrost Landscape Evolution 1. https://doi.org/10.1139/AS-2016-0022.
Bradley, R.S., 2015. Paleoclimatology: Reconstructing Climates of the Quaternary. In: Paleoclimatology : Reconstructing Climates of the Quaternary Second Edition, third ed. Elsevier, Oxford. https://doi.org/10.1029/eo081i050p00613-01.
Brauer, A., 2004. Annually Laminated Lake Sediments and Their Palaeoclimatic Relevance, pp. 109–127. https://doi.org/10.1007/978-3-662-10313-5_7.
Cai, L., Zhu, Y., 2015. The challenges of data quality and data quality assessment in the big data era. Data Sci. J. 14, 1–10. https://doi.org/10.5334/dsj-2015-002.
Chen, P.P.S., 1975. The entity-relationship model: toward a unified view of data. ACM SIGIR Forum 10, 9. https://doi.org/10.1145/1095277.1095279.
Codd, E.F., 1970. A relational model of data for large shared data banks. Commun. ACM 13, 377–387. https://doi.org/10.1145/357980.358007.
Cohen, A.S., 2003. Paleolimnology: the History and Evolution of Lake Systems. Oxford University Press, New York.
Conze, R., Lorenz, H., Ulbricht, D., Elger, K., Gorgas, T., 2017. Utilizing the international geo sample number concept in continental scientific drilling during ICDP expedition COSC-1. Data Sci. J. 16, 1–8. https://doi.org/10.5334/dsj-2017-002.
Curry, P.A., Moosdorf, N., 2019. An open source web application for distributed geospatial data exploration. Sci. Data 6, 1–7. https://doi.org/10.1038/sdata.2019.14.
DBeaver Community, 2020. DBeaver [WWW Document]. URL. https://dbeaver.io/. accessed 3.11.20.
Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., Wefer, G., 2002. Pangaea - an information system for environmental sciences. Comput. Geosci. 28, 1201–1210. https://doi.org/10.1016/S0098-3004(02)00039-0.
Elger, K., Biskaborn, B.K., Pampel, H., Lantuit, H., 2016. Open research data, data portals and data publication - an introduction to the data curation landscape. Polarforschung 85, 119–133. https://doi.org/10.2312/polfor.2016.009.
Elmasri, R., Navathe, S.B., 2009. Grundlagen von Datenbanksystemen (Fundamentals of Database Systems). Pearson.
Fritz, S.C., 2008. Deciphering climatic history from lake sediments. J. Paleolimnol. 39, 5–16. https://doi.org/10.1007/s10933-007-9134-x.
Garcia-Molina, H., Ullman, J.D., Widom, J., 2002. Database Systems: the Complete Book, an Alan R. Apt Book. Prentice Hall.
Goswami, B., 2014. Uncertainties in Climate Data Analysis. University of Potsdam.
Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science. Libr. Trends 57, 280–299. https://doi.org/10.1353/lib.0.0036.
Huntley, B., 2012. Reconstructing palaeoclimates from biological proxies: some often overlooked sources of uncertainty. Quat. Sci. Rev. 31, 1–16. https://doi.org/10.1016/j.quascirev.2011.11.006.
IPCC, 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland.
Kalugin, I., Daryin, A., Smolyaninova, L., Andreev, A., Diekmann, B., Khlystov, O., 2007. 800-yr-long records of annual air temperature and precipitation over southern Siberia inferred from Teletskoye Lake sediments. Quat. Res. 67, 400–410. https://doi.org/10.1016/j.yqres.2007.01.007.
Kaufman, D.S., McKay, N.P., Routson, C.C., Erb, M., Davis, B.A.S., Heiri, O., Jaccard, S., Tierney, J.E., Dätwyler, C., Axford, Y., Brussel, T., Cartapanis, O., Chase, B.M., Dawson, A., de Vernal, A., Engels, S., Jonkers, L., Marsicek, J., Moffa-Sánchez, P., Morrill, C., Orsi, A., Rehfeld, K., Saunders, K., Sommer, P.S., Thomas, E., Tonello, M., Tóth, M., Vachula, R., Andreev, A., Bertrand, S., Biskaborn, B.K., Bringué, M., Brooks, S., Caniupán, M., Chevalier, M., Cwynar, L., Emile-Geay, J., Fegyveresi, J., Feurdean, A., Finsinger, W., Fortin, M.-C., Foster, L., Fox, M., Gajewski, K., Grosjean, M., Hausmann, S., Heinrichs, M., Holmes, N., Ilyashuk, B., Ilyashuk, E., Juggins, S., Khider, D., Koinig, K., Langdon, P., Larocque-Tobler, I., Li, J., Lotter, A., Luoto, T., Mackay, A., Magyari, E., Malevich, S., Mark, B., Massaferro, J., Montade, V., Nazarova, L., Novenko, E., Pařil, P., Pearson, E., Peros, M., Pienitz, R., Płóciennik, M., Porinchu, D., Potito, A., Rees, A., Reinemann, S., Roberts, S., Rolland, N., Salonen, S., Self, A., Seppä, H., Shala, S., St-Jacques, J.-M., Stenni, B., Syrykh, L., Tarrats, P., Taylor, K., van den Bos, V., Velle, G., Wahl, E., Walker, I., Wilmshurst, J., Zhang, E., Zhilich, S., 2020. A global database of Holocene paleotemperature records. Sci. Data 7, 1–34. https://doi.org/10.1038/s41597-020-0445-3.
Khider, D., Emile-Geay, J., McKay, N.P.P., Gil, Y., Garijo, D., Ratnakar, V., Alonso-Garcia, M., Bertrand, S., Bothe, O., Brewer, P., Bunn, A., Chevalier, M., Comas-Bru, L., Csank, A., Dassié, E., DeLong, K., Felis, T., Francus, P., Frappier, A., Gray, W., Goring, S., Jonkers, L., Kahle, M., Kaufman, D., Kehrwald, N.M.M., Martrat, B., McGregor, H., Richey, J., Schmittner, A., Scroxton, N., Sutherland, E., Thirumalai, K., Allen, K., Arnaud, F., Axford, Y., Barrows, T., Bazin, L., Pilaar Birch, S.E.E., Bradley, E., Bregy, J., Capron, E., Cartapanis, O., Chiang, W, H., Cobb, K.M., Debret, M., Dommain, R., Du, J., Dyez, K., Emerick, S., Erb, M.P.P., Falster, G., Finsinger, W., Fortier, D., Gauthier, N., George, S., Grimm, E., Hertzberg, J., Hibbert, F., Hillman, A., Hobbs, W., Huber, M., Hughes, A.L.C., Jaccard, S., Ruan, J., Kienast, M., Konecky, B., Le Roux, G., Lyubchich, V., Novello, V.F.F., Olaka, L., Partin, J.W.W., Pearce, C., Phipps, S.J.J., Pignol, C., Piotrowska, N., Poli, M.-S., Prokopenko, A., Schwanck, F., Stepanek, C., Swann, G.E. A., Telford, R., Thomas, E., Thomas, Z., Truebe, S., Gunten, L., Waite, A., Weitzel, N., Wilhelm, B., Williams, J., Williams, J.J.J., Winstrup, M., Zhao, N., Zhou, Y., 2019. PaCTS 1.0: a crowdsourced reporting standard for paleoclimate data. Paleoceanogr. Paleoclimatol. 34, 1570–1596. https://doi.org/10.1029/2019PA003632.
Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Loizides, F., Schmidt, B. (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas, pp. 87–90.
Latif, A., Limani, F., Tochtermann, K., 2019. A generic research data infrastructure for long tail research data management. Data Sci. J. 18, 1–11. https://doi.org/10.5334/dsj-2019-017.
Lougheed, B.C., Obrochta, S.P., 2019. A rapid, deterministic age-depth modeling routine for geological sequences with inherent depth uncertainty. Paleoceanogr. Paleoclimatol. 34, 122–133. https://doi.org/10.1029/2018PA003457.
McKay, N.P., Kaufman, D.S., 2014. An extended Arctic proxy temperature database for the past 2,000 years. Sci. Data 1, 1–10. https://doi.org/10.1038/sdata.2014.26.

Messager, M.L., Lehner, B., Grill, G., Nedeva, I., Schmitt, O., 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nat. Commun. 7, 1–11. https://doi.org/10.1038/ncomms13603.

Meyer, M.F., Labou, S.G., Cramer, A.N., Brousil, M.R., Luff, B.T., 2020. The global lake area, climate, and population dataset. Sci. Data 7, 1–12. https://doi.org/10.1038/s41597-020-0517-4.

Miller, G.H., Alley, R.B., Brigham-Grette, J., Fitzpatrick, J.J., Polyak, L., Serreze, M.C., White, J.W.C., 2010. Arctic amplification: can the past constrain the future? Quat. Sci. Rev. 29, 1779–1790. https://doi.org/10.1016/j.quascirev.2010.02.008.

Muster, S., 2018. Arctic Freshwater – A Commons Requires Open Science. Springer International Publishing, Cham, pp. 107–120. https://doi.org/10.1007/978-3-319-66459-0_9. Arctic Summer College Yearbook.

Wilkinson, M.D., Dumontier, M., Aalbersberg, IjJ., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. Comment: the FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 1–9. https://doi.org/10.1038/sdata.2016.18.

PAGES 2k Consortium, 2017. A global multiproxy database for temperature reconstructions of the Common Era. Sci. Data 4, 1–33 https://doi.org/DOI: 10.1038/sdata.2017.88.

Pannekoek, J., Scholtus, S., Van der Loo, M., 2013. Automated and manual data editing: a view on process design and methodology. J. Off. Stat. 29, 511–537. https://doi.org/10.2478/jos-2013-0038.

PostgreSQL Global Development Group, 2018. PostgreSQL 11 [WWW Document]. URL. https://www.postgresql.org/. accessed 3.11.20.

Reback, J., McKinney, W., jbrockmendel, Bossche, J. Van den, Augspurger, T., Cloud, P., gfyoung, Sinhrks, Klein, A., Roeschke, M., Hawkins, S., Tratner, J., She, C., Ayd, W., Petersen, T., Garcia, M., Schendel, J., Hayden, A., MomIsBestFriend, Jancauskas, V., Battiston, P., Seabold, S., chris-b1, h-vetinari, Hoyer, S., Overmeire, W., alimcmaster1, Dong, K., Whelan, C., Mehyar, M., 2020. pandas-dev/pandas: Pandas 1.0.3. https://doi.org/10.5281/ZENODO.3715232.

Reimer, P.J., Austin, W.E.N., Bard, E., Bayliss, A., Blackwell, P.G., Bronk Ramsey, C., Butzin, M., Cheng, H., Edwards, R.L., Friedrich, M., Grootes, P.M., Guilderson, T.P., Hajdas, I., Heaton, T.J., Hogg, A.G., Hughen, K.A., Kromer, B., Manning, S.W., Muscheler, R., Palmer, J.G., Pearson, C., van der Plicht, J., Reimer, R.W., Richards, D.A., Scott, E.M., Southon, J.R., Turney, C.S.M., Wacker, L., Adolphi, F., Büntgen, U., Capano, M., Fahrni, S.M., Fogtmann-Schulz, A., Friedrich, R., Köhler, P., Kudsk, S., Miyake, F., Olsen, J., Reinig, F., Sakamoto, M., Sookdeo, A., Talamo, S., 2020. THE INTCAL20 NORTHERN HEMISPHERE RADIOCARBON AGE CALIBRATION CURVE (0–55 CAL kBP). Radiocarbon 1–33. https://doi.org/10.1017/RDC.2020.41.

Reschke, M., Rehfeld, K., Laepple, T., 2019. Empirical estimate of the signal content of Holocene temperature proxy records. Clim. Past 15, 521–537. https://doi.org/10.5194/cp-15-521-2019.

Rothwell, R.G., Croudace, I.W., 2015. Micro-XRF Studies of Sediment Cores, Developments in Paleoenvironmental Research. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-017-9849-5.

Sebastian-Coleman, L., 2013. Measuring Data Quality for Ongoing Improvement, Measuring Data Quality for Ongoing Improvement. https://doi.org/10.1016/C2011-0-07321-0.

Stall, S., Yarmey, L., Boehm, R., Cousijn, H., Cruse, P., Cutcher-Gershenfeld, J., Dasler, R., de Waard, A., Duerr, R., Elger, K., Fenner, M., Glaves, H., Hanson, B., Hausman, J., Heber, J., Hills, D., Hoebelheinrich, N., Hou, S., Kinkade, D., Koskela, R., Martin, R., Lehnert, K., Murphy, F., Nosek, B., Parsons, M., Petters, J., Plante, R., Robinson, E., Samors, R., Servilla, M., Ulrich, R., Witt, M., Wyborn, L., 2018. Advancing FAIR data in earth, space, and environmental science. Eos 99, 1–9. https://doi.org/10.1029/2018EO109301.

Su, Z., Timmermans, W., Zeng, Y., Schulz, J., John, V.O., Roebeling, R.A., Poli, P., Tan, D., Kaspar, F., Kaiser-Weiss, A.K., Swinnen, E., Toté, C., Gregow, H., Manninen, T., Riihelä, A., Calvet, J.C., Ma, Y., Wen, J., 2018. An overview of european efforts in generating climate data records. Bull. Am. Meteorol. Soc. 99, 349–359. https://doi.org/10.1175/BAMS-D-16-0074.1.

Subetto, D.A., Nazarova, L.B., Pestryakova, L.A., Syrykh, L.S., Andronikov, A.V., Biskaborn, B., Diekmann, B., Kuznetsov, D.D., Sapelko, T.V., Grekov, I.M., 2017. Paleolimnological studies in Russian northern Eurasia: a review. Contemp. Probl. Ecol. 10, 327–335. https://doi.org/10.1134/S1995425517040102.

Sun, S., Bertrand-Krajwewski, J.-L., Lynggaard-Jensen, A., Broeke, J. van den, Edthofer, F., Almeida, M. do C., Ribeiro, A.S., Menaia, J., 2011. Literature review of data validation methods. Sci. Technol. 47, 95–102.

Teorey, T.J., Buxton, S., Fryman, L., Güting, R.H., Halpin, T., Harrington, J.L., Inmon, W. H., Lightstone, S.S., Melton, J., Morgan, T., others, 2008. Database Design: Know it All. Morgan Kaufmann.

Trachsel, M., Telford, R.J., 2017. All age–depth models are wrong, but are getting better. Holocene 27, 860–869. https://doi.org/10.1177/0959683616675939.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A., Pietro, Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D.V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods. https://doi.org/10.1038/s41592-019-0686-2.

Wang, R.Y., Ziad, M., Lee, Y.W., Data Quality, 2001. Advances in Database Systems. Kluwer, Boston. https://doi.org/10.1007/b116303.

Wilke, T., Wagner, B., Van Bocxlaer, B., Albrecht, C., Ariztegui, D., Delicado, D., Francke, A., Harzhauser, M., Hauffe, T., Holtvoeth, J., Just, J., Leng, M.J., Levkov, Z., Penkman, K., Sadori, L., Skinner, A., Stelbrink, B., Vogel, H., Wesselingh, F., Wonik, T., 2016. Scientific drilling projects in ancient lakes: integrating geological and biological histories. Global Planet. Change 143, 118–151. https://doi.org/10.1016/j.gloplacha.2016.05.005.

Williams, J.W., Grimm, E.C., Blois, J.L., Charles, D.F., Davis, E.B., Goring, S.J., Graham, R.W., Smith, A.J., Anderson, M., Arroyo-Cabrales, J., Ashworth, A.C., Betancourt, J.L., Bills, B.W., Booth, R.K., Buckland, P.I., Curry, B.B., Giesecke, T., Jackson, S.T., Latorre, C., Nichols, J., Purdum, T., Roth, R.E., Stryker, M., Takahara, H., 2018. The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. Quat. Res. (U. S. A.) 89, 156–177. https://doi.org/10.1017/qua.2017.105.

Wright, A.J., Edwards, R.J., van de Plassche, O., Blaauw, M., Parnell, A.C., van der Borg, K., de Jong, A.F.M., Roe, H.M., Selby, K., Black, S., 2017. Reconstructing the accumulation history of a saltmarsh sediment core: which age-depth model is best? Quat. Geochronol. 39, 35–67. https://doi.org/10.1016/j.quageo.2017.02.004.

Zeng, Y., Su, Z., Barmpadimos, I., Perrels, A., Poli, P., Boersma, K.F., Frey, A., Ma, X., de Bruin, K., Goosen, H., John, V.O., Roebeling, R., Schulz, J., Timmermans, W., 2019. Towards a traceable climate service: assessment of quality and usability of essential climate variables. Rem. Sens. 11 https://doi.org/10.3390/rs11101186.

Zolitschka, B., Francus, P., Ojala, A.E.K., Schimmelmann, A., 2015. Varves in lake sediments - a review. Quat. Sci. Rev. 117, 1–41. https://doi.org/10.1016/j.quascirev.2015.03.019.