Manuscript for submission to the *Journal of Computer Applications in Archaeology*

XRONOS: an open data infrastructure for archaeological chronology

Joe Roe [*]
University of Bern
joseph.a.roe@unibe.ch

Clemens Schmid
Max Planck Institute for Geoanthropology

Setareh Ebrahimiabareghi
University of Bern

Caroline Heitz
University of Bern

Martin Hinz
University of Bern

ABSTRACT XRONOS (https://xronos.ch) is an open data infrastructure for the backbone of the archaeological record – chronology. It aims provides open access to published radiocarbon dates and other chronometric data from any period, anywhere in the world. Collating a number of recent regional and global compilations of dates, XRONOS offers the most comprehensive radiocarbon database yet published, with over 350,000 date and 75,000 site records. It also provides a foundation for expanding the systematic collection of chronometric information beyond radiocarbon, with support for typological and dendrochronological dates and a generalisable data model that can be adapted to other methods of absolute dating. Automated and semi-automated quality control processes ensure that data from diverse sources is continuously integrated and standardised, making it significantly easier to filter data of interest and reducing the need for manual data cleaning by end users. In this paper we describe the concept and implementation of XRONOS in relation to the state of the art in chronometric data-sharing, and evaluate its potential as a general-purpose open repository and curation platform for archaeological chronology.

KEYWORDS open data; chronology; chronometry; radiocarbon dating; dendrochronology; typological dating

```
Attaching package: 'gt'

The following object is masked from 'package:cowplot':

    as_gtable

Linking to GEOS 3.12.2, GDAL 3.9.0, PROJ 9.4.1; sf_use_s2() is TRUE
```

[*] Corresponding author.

1 Introduction

Chronology is the backbone of the archaeological record. It is needed for X, Y, and Z. Open access to chronological data is therefore a critical...

- Profusion of open C14 data in the last 20–30 years calls for global solutions
- Range of other chronological information used by archaeologists is largely untouched
- From uploading CSVs to (social) infrastructure

2 State of the art

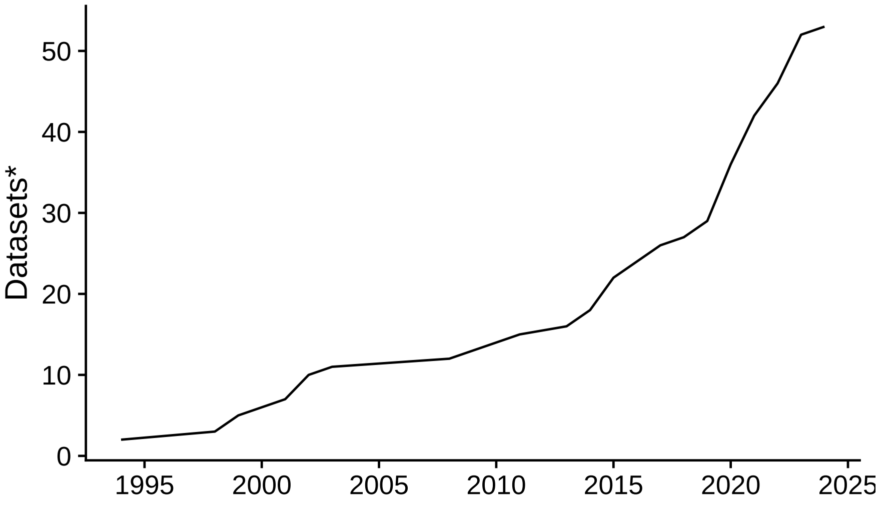
2.1 Compilations of radiocarbon dates

Though 'open data' in archaeology is a relatively recent phenomenon (**CITE**), the open publication of compiled radiocarbon dates has a substantial prehistory. Starting in the 19XXs, radiocarbon laboratories shared and compiled as 'date lists' published on a regular basis in the journal *Radiocarbon* (**CITE**). However, as the number of labs and volume of radiocarbon dates being produced grew, this paper-based format became impractical: the last 'date list' appeared in 19XX, and was not replaced by another form of systematic data-sharing or dissemination. Additionally, because the date lists were sourced from radiocarbon laboratories directly—not those who collected the sample—they typically included only very limited contextual information. On the eve of the AMS revolution there was an effort to create a computerised 'International Radiocarbon Database' (Kra 1988)—already by 1989 described as a "much needed, long overdue enterprise" (Kra 1989)—but it never came to fruition.

Thus, even though radiocarbon data comes from a relatively limited number of sources (**IntCalLabList**) and has relatively standardised reporting conventions (Millard 2014; Bayliss 2015), in practice the only way to produce aggregated datasets in recent decades has been to manually search through relevant literature for dates reported by the submitter of the sample. This already laborious process is further hampered by a significant inconsistency in how much authors adhere to measurement reporting conventions, a lack of conventions on the reporting of *contextual* information, and weak or nonexistent disciplinary norms regarding the responsibility to publish results openly in a timely fashion.

Despite these inefficiencies, there have been a profusion of published radiocarbon compilations since the decline of the date list. Our review of the literature identified 61 published since 1994. This is almost certainly an undercount, because our firsthand knowledge of regional literature was limited to Europe and West Asia and many resources only ever existed in 'grey' formats (e.g. websites that were not indexed and no longer exist). We also restricted ourselves to structured datasets disseminated primarily in a digital format; 'date lists' in printed periodicals and gazetteers were excluded. A full list of the datasets we identified is presented in appendix XXXX.

The number of available compilations has steadily increased since around 1995 Figure 1. The first generation came around the turn of the century and consisted mostly of online databases with a web frontend. These included some databases operated by radiocarbon labs, for example the Oxford Radiocarbon Lab (ORAU) and the Belgian Royal Institute for Cultural Heritage (KIK-IRPA), and essentially represented a continuation of their date lists in a digital format. The majority, however, were compiled from the literature and individual researchers interested in a



* Excludes 8 datasets with an unknown year of publication

Figure 1: Cumulative number of radiocarbon compilations published since 1995

particular region. Notable early examples include ANDES 14C in 1994 (Central Andes, Michczyński et al. 1995), CARD (Canada, Gajewski et al. 2011) and RADON (Europe, Raetzel-Fabian 1999) in 1999, and CANEW in 2001 (Near East, Reingruber and Thissen 2005). From 2010, coinciding with broader shifts in scientific publishing (Tenopir et al. 2011), it became more common to publish standalone 'open data' products in the form of journal supplements, archives in repositories and/or data papers; the *Journal of Open Archaeology Data*, launched in 2012, has been a prominent venue for this latter category. Most recently there has been a trend towards providing version-controlled plain text data via platforms such as GitHub, reflecting the broader adoption of these tools amongst computational archaeologists over the last decade (Batist and Roe 2024). The shift from online databases towards more static but more preservable open data products is welcome, given how many databases from the first generation have subsequently ceased to be accessible. Version-controlled repositories are particular well-suited to data compilation projects because they allow for continued updates whilst still providing snapshot 'releases' that are citeable and can be archived in long-term repositories.

```
Warning: There were 3 warnings in `stopifnot()`.
The first warning was:
i In argument: `m49_macroregion = countrycode(ISO3_CODE, "iso3c",
  "un.region.name")`.
Caused by warning:
! Some values were not matched unambiguously: ATA, CPT, XA, XB, XC, XD, XE, XF, XG, XH, X
i Run `dplyr::last_dplyr_warnings()` to see the 2 remaining warnings.
```

```
Joining with `by = join_by(m49_region)`
```

Although this body of work has greatly improved the accessibility of radiocarbon dates and supported significant methodological advances (Crema 2022; Crema et al. 2024), some limitations are apparent.

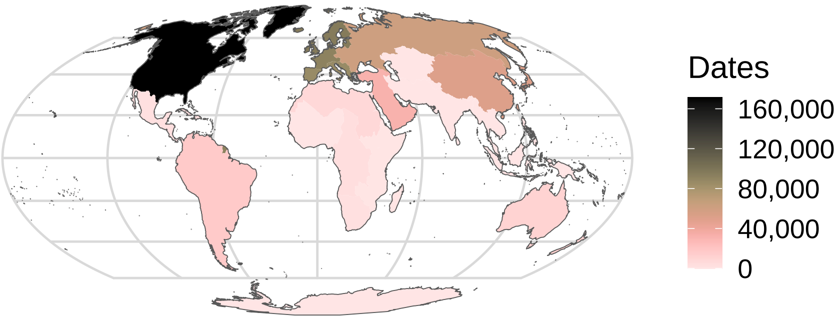


Figure 2: Geographic coverage of published regional radiocarbon compilations

The geographic coverage of regional radiocarbon compilations is markedly uneven (Figure 2). Europe and, especially, North America are very over-represented (as has already been observed by Chaput and Gajewski 2016, and others). South America, West Asia, and East Asia are reasonably well-covered, but there practically no systematically compiled dates from East or West Africa, Central or South Asia, or Mainland Southeast Asia. This is probably explained in part by a lower volume of archaeological research and access to radiocarbon dating in these regions, but a lack of attention in compilation work must also be a factor. For example, radiocarbon dating has been an established part of Indian archaeology since at least 1961 (Kusumgar, Lal, and Sarna 1963), but we have not able to locate a single systematic compilation of dates from the Indian subcontinent.

Datasets based on literature review also become out of date almost immediately upon publication, due the the constant stream of new dates. Unfortunately this applies to many databases that are in theory continuously updated, as it is common to see them become unmaintained and or unexpectedly become unavailable. Of the 61 published datasets we identified, 31 were intended to be continuously updated, but only 11 had received updates in the last two years. The average 'lifespan' of a dataset from its publication to its last update is 3.94 years. Most radiocarbon datasets we reviewed were compiled with a specific goal in mind (e.g. a particular analysis) and, even where there is the intention to keep them updated afterwards, the exigencies of scientific production combined with the labour-intensive nature of the process make that difficult to achieve in practice.

- The coverage of these databases is uneven, systematically biased, and duplicative of each other (**CITE**)?

The fragmentation of the radiocarbon record into regional datasets also hinders analysis at larger scales. Although the core elements of a radiocarbon date—laboratory identifier, radiocarbon age, measurement error—are more or less standardised, there is no such consistency in contextual information on the sample or site. Such contextual information is important not just for the interpretation of dates, but for 'chronometric hygiene' (filtering out unreliable dates based on sample information, see e.g. Pettitt et al. 2003) and for correcting for known systematic errors such as the marine reservoir effect (Alves et al. 2018). Most published datasets incorporate all or part of earlier compilations, duplicate records are also very common, but deduplicating them is not a trivial problem due to format variations (see Section 4.2). These issues are by no means impossible to overcome, but adds a significant amount of data-cleaning effort to a process that is otherwise very amenable to standardisation.

- Not all are open
- Not all are machine readable

2.2 Open data infrastructures

The profusion of radiocarbon compilations over the last decade has naturally prompted many to think globally:

- The first available synthetic radiocarbon database was **c14bazAAR** (Schmid, Seidensticker, and Hinz 2019), an R package that provides an index of openly published radiocarbon databases and a common interface for retrieving them and performing basic data cleaning. Because c14bazAAR downloads data from its original source repositories, rather than mirroring it, it only includes resources that have been published in a fully open and machine-accessible format. Despite this limitation, it has global coverage and is still the largest collection of radiocarbon dates available (**FIG<empty citation>**).

- Another indexical approach is the **IntChron** project (Bronk Ramsey et al. 2019), which exposes data from multiple sources and exposes them with a common JSON-based web interface. The IntChron specification is open, meaning that radiocarbon labs or compilation projects can implement it independently and thereby allow end users to access their data through a common itnerface (though to our knowledge it has so far only been adopted by databases associated with the Oxford Radiocarbon Lab). The JSON format also lends itself to the implementation of wrapper libraries, for example the 'rIntChron' package gives direct access to IntChron-indexed databases in R (**CITE**).

- **p3k14c** (Bird et al. 2022) instead compiles multiple source databases into a single flat file dataset, with a similar level of coverage to c14bazAAR. The major advantage of this approach is that the data is made internally consistent and has been manually cleaned to an extent, which makes it particularly well-suited to global analyses. The downside is that without the continuous link to the source databases present in the c14bazAAR and IntChron, it can only be kept up to date manually with periodic re-releases. An accompanying package (**CITE<empty citation>**) provides direct access to the p3k14c dataset in R.

2.3 Beyond radiocarbon

- Dendro

- Typology

- Other radiometric methods

- Application-specific

- XRONOS aims to bring all these together, starting with radiocarbon, dendro-, and typology

- There are lots of other types of chronometric data!

  - Dendro - in certain places (e.g. Switzerland!), way more important than radiocarbon
  - Other radiometric/absolute dating methods, OSL, etc.
  - Typology - still the backbone of many chronologies
  - Application-specific: sea level dating, rock patina, etc.

- Compilation of other types of dates has been much more limited.

  - Exceptions: …?

For European dendrochronological data, the situation is even less favourable for archaeological purposes:

- NOAA stores some global datasets, but has its main focus in the United States and, with its only 34 European prehistoric datasets, is not useful for the study of the European past.

- The ADS database is compiled and maintained by the UK-based Vernacular Architecture Group and is limited to data published annually in the journal Vernacular Architecture, containing only medieval or later dates for the UK.

- DendroDB has a European focus but only provides data for historical periods.

- The "Dendrochronological Database" of the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) has a worldwide focus on collected natural wood samples and has different goals than a database for archaeological wood material (furthermore, although announced a long time ago, this database is still not functional).

- Although "Digital Collaboratory for Cultural Dendrochronology (DCCD). An international digital data library for dendrochronology" is well connected in the context of archaeological data services (e.g. ARIADNE), it is focused on the Netherlands (more than 2/3 of the data, while 0.08% of the Quercus data refer to Switzerland) and contains very few prehistoric data (about 2.5%). Furthermore, the overall activity in this database has decreased from 3846 new project records in 2010-2014 to only 83 since then (until the end of 2019).

- Typology

  - perio.do
  - Often embedded in other types of database (e.g. gazetteers)
  - But nobody has treated it as a form of chronometric data in its own right?

What is completely missing worldwide so far, is a database that encompasses both, 14C and dendro data. However, to efficiently conduct data driven modelling of human-environmental dependencies and dynamics on the past, such a data base seems indispensable for future research.

3 Concept

XRONOS inherits its basic structure from RADON (Raetzel-Fabian 1999; Hinz et al. 2012; Kneisel, Hinz, and Rinne, RADON-B – Radiocarbon Dates Online (Version 2014). Database for European 14C Dates for the Bronze and Early Iron Age; Rinne et al., Rado.NB), with a database-backed web application and a data model that separates radiocarbon dates, contextual information, and sites. Our overall aims in developing XRONOS is to bring this model, which RADON has operated on for XX years, up to date, to generalise it to other types of chronometric information, and to transform it from an online database to a data infrastructure that supports the continuous ingestion, curation, and open dissemination of archaeological chronologies from diverse sources.

3.1 Design goals

XRONOS is our answer to Kintigh's call (Kintigh 2006) for digital infrastructures that don't just provide access to chronological data but enables researchers to "archive, access, integrate, and mine disparate data sets". It parallels and draws inspiration from several similar initiatives in archaeology, such as And outside of it:

The principal design goals of the software are to combine all available sources of radiocarbon and other chronometric data in single database; develop robust tools for the continuous ingestion and refinement of this data; and disseminate this data within an open and FAIR framework, embedding it in the wider world of Linked Open Data in archaeology and beyond.

Our concrete goals in the design and implementation of XRONOS are the following:

1. Provision of access to 14C and dendrochronological data with meta information on archaeological contexts via web-based Open Access frontend
2. Web frontend with processing functions (filters and queries, export to various data formats, 14C calibration & chronology formation) for in place analytical functions
3. machine readable interface — API — for direct access e.g. by open source statistical environments (eg. R) for analytical purposes or dendroanalytical software
4. licensing and versioning of individual data using eg. DOI
5. Storage in a repository with user and role-specific access rights as well as encryption for data privacy
6. Ensuring sustainable data security and usage through compliance with standards, appropriate backup solutions and high-availability technologies

- Link to limitations identified in the state of the art section

3.2 Data model

- FIGURE: conceptual scheme

At the base of the XRONOS data model are sets of spatiotemporal coordinates or, as we call them, *chrons*. In an archaeological context, we conceptualise a chron as an assertion linking human activity with a particular point in space and time. Our data model currently encompasses three types of chron: radiocarbon dates, typological dates (e.g. 'Early Neolithic') and dendrochronological dates. However we anticipate that the concept will accommodate other types of absolute and relative dating techniques, as the scope of the database expands.

Chrons are conceptually useful because they emphasise that different types of archaeological 'dates', drawn from different sources, have essentially the same information content: the location of an event in space and time. We thereby avoid privileging certain sources of chronological data over (as might be the case if, for example, we treated 'period' as a fixed attribute of a site) and can accommodate contradictory (e.g. differences of opinion on typological classification). This is important given that XRONOS aspires to be an authoratative 'backbone' with a global scope, so we cannot realistically impose a single chronological scheme or resolve conflicting information provided by specialists. They are useful practically because they expose a common interface for attributes that all types of chronological information share, such as a *terminus post quem* (TPQ), *terminus ante quem* (TAQ), and midpoint estimate. This allows applications that use XRONOS' data model (including XRONOS itself) to collate chronological data from multiple sources, without necessarily having to be aware of the pecularities of each type of dating.

In order to unify chronological information in the form of a chron, we need a common chronological 'coordinate system'. The natural choice is a *calendar probability distribution*, which expresses the probability that an event occurred as a function of time on a calendric scale. Most archaeologists are familiar with working with this kind of representation in the form of calibrated radiocarbon dates, but it can be extended and generalised to essentially any kind of chronological information (**CITE**). For example, in aoristic analysis (**CITE**), a periodic time estimate (e.g. the event occurred in the Neolithic) is conceptualised as a uniform probability distribution over the timespan between the known start and end dates of that period. A similar model is used in OxCal (**OxCal**) to integrate prior chronological information from diverse sources. In practical terms, this model means that the canonical representation the time component of any chron in XRONOS', regardless of source, is a probably distribution over the set of calendar years (arbitrarily measured in years Before Present) in which it could have plausibly occurred. Further statistics, e.g. a midpoint estimate or TPQ/TAQ range, can be derived from this distribution using well-known methods. In this way, we can support many different types of date and much of the implementation of XRONOS can be agnostic to the source of chronological information.

Chrons are located in space through association to a *sample* – the physical object from which a chronological determination was made. The location of samples is represented with geographical coordinates and an associated coordinate reference system (CRS), though since in practice the precise location of single samples is rarely available, this property is usually inherited from the site. We also record relevant metadata on the nature of the sample. For radiocarbon dates, for example, we follow established conventions (Millard 2014) in recording the type (e.g. charcoal, charred seed) and, where applicable, taxonomic designation (e.g. TODO, TODO) of the organic material used for dating. For typological dates, an ideal scenario would be for the sample to represent the particular object from which an inference was made (e.g. 'Natufian' might be inferred from 'lunate-type microlith'). In practice, the best we can glean from most published datasets is the type of material used (e.g. 'pottery', 'lithics'). The same sample can be associated with multiple chrons, including different types of chron. This is useful, for example, for representing replicate radiocarbon dates on the same sample, or radiocarbon dates and dendrochronological made on the same section of wood for wiggle-matching.

Further contextual information is associated with *contexts* and *sites*. The site is the primary geographic container for chronological information. As already mentioned, we typically record the spatial location of chrons using this entity, though it is possible to modify this by providing specific coordinates at the sample level. Sites also have attributes describing their conventional name or names in different languages and are associated with a flexible 'site type' typology that combines information on their form and function.

A context represents the specific find-context of a sample, e.g. an architectural feature, stratigraphic unit, or phase. Since the units and conventions for recording such information vary greatly between different regions and archaeological traditions—and XRONOS is designed with global data in mind—we leave the question of what a context precisely represents open, and only record an unstandardised, free text label for it. Crucially, however, contexts can have a self-referential association to other contexts belonging to the same site. This allows it to encode arbitrary relational structures between contexts, whether they be hierarchical (e.g. phases and sub-phases) or graphical (e.g. stratigraphic). In this way, it can serve as a foundation for chronological modelling.

The series of relations `[chron] > sample > context > site` links the chronological and contextual sides of the XRONOS data model. Each step is a many-to-many association, meaning for example that it is possible to attach multiple chrons to the same sample (e.g. replicated radiocarbon dates on the same material), multiple *types* of chrons to the same sample (e.g. radiocarbon dates on tree-rings for wiggle-matching). Since this kind of information is rarely systematically recorded in our source databases, there are currently few actual records that make use of this feature of the data model. However, we hope it will provide a foundation for … ?

Metadata is incorporated into XRONOS' data model at the level of the individual records (e.g. all records store their data of creation and last modification) and through two additional nodes: bibliographic references and versions. Bibliographic references store information on the source of a record in the standard BibTeX (**CITE<empty citation>**) format and can be linked through many-to-many associations to sites or chrons. Versions are a special type of record that are associated with all other records (including bibliographic references) and store the previous versions of those records as a series of changesets. In this way all changes to data are recorded and can be reconstructed (or reversed) precisely. This 'paper trail' also stores contextual metadata, e.g. who made the change and why. It also means that records that are deleted can be reviewed or restored from their stored version history, which is never discarded. Together these two systems provide a transparent record where the data in XRONOS comes from and how it has been altered, which we view as essential in a scientific data infrastructure.

3.3 Linked data

The XRONOS data model presents several opportunities to link to other resources as linked open data (LOD) (**CITE**). We use controlled vocabularies in the form of the GBIF Backbone Taxonomy (for taxonomic descriptions of samples) via its API, and in future plan to extend this to for example site types (using the Getty and/or DAI). Standardising field using these vocabularies to automatically then link the two entities. Beyond this, there is also a concordance between XRONOS entities and other specialised data infrastructures in archaeology, such as Perio.do (**CITE<empty citation>**, mapping to typological chrons) and gazetters like Pleiades or Vici (**CITE<empty citation>**, mapping to site names). Moving beyond archaeology, Wikidata (**CITE<empty citation>**) already includes many of the concepts represented in the XRONOS model (e.g. archaeological sites, Q). Linking to Wikidata is especially useful because it dissimenates—and thereby preserves—the data compiled in XRONOS beyond the narrow field of archaeology. It also allows us to enrich the database with contextual information that otherwise be beyond our scope and resources, for example embedding multilingual descriptions of sites from Wikipedia articles on site records where the site has been linked to a Wikidata item.

Conversely, we encourage others to use XRONOS as a linked open data resource by providing stable URLs and a machine-readable interface for every record. Such usages could look like, for example, using a XRONOS URL as a canonical representation of a single radiocarbon date (e.g.). We also plan to implement an IntChron ((**CITE**)) interface to XRONOS, to allow it to be indexed through that existing standard.

4 Implementation

Following a short pilot project in 2019, the first phase of development of XRONOS was completed in 2021–2024, supported by a project grant from the Swiss National Science Foundation. The web interface (https://xronos.ch) has been publicly accessible since July 2021. Though we envisage XRONOS as a continuously-developing open source project and 'living database', the following offers a snapshot of progress at the end of our first grant-funded implementation phase. We do not aim to be comprehensive, but rather to describe some key elements of XRONOS' current implementation that illustrate how our concept has been realised in practice.

4.1 Software architecture

The XRONOS data model is implemented as a relational database using the free and open source database management system PostgreSQL. However, apart from backups and other routine maintenance procedures, all interaction with the database is via a web application, which thus forms the core of XRONOS' architecture. The XRONOS web application is written in Ruby and uses CRUD (Create, Read, Update, Delete) and MVC (Model, View, Controller) patterns as implemented in the Ruby on Rails framework. The choice of this somewhat traditional or 'boring' architecture—as opposed to the non-relational/semantic technologies that are more *à la mode* in the digital humanities (Fan 2018; Hyvönen 2020; Schloen and Prosser 2023)—was purposeful, motivated by a sense that it offers a better chance of longevity and maintanability. Our aim is for the software architecture to be as boring as possible so that the scientific contents can be as interesting as possible.

The XRONOS web application exposes two distinct user interfaces: a graphical user interface accessed through a web browser; and an application programming interface (API). Both interfaces follow a REST (Representational State Transfer) pattern (Verborgh et al. 2015), where each resource (e.g. a single radiocarbon date, a single bibliographic reference, or a single user) is statelessly mapped to a single address. Users can then interact with resources at these addresses using a preditable and uniform interface based on HTTP verbs. For example, the radiocarbon date RTD-8904 is represented by the address https://xronos.ch/c14s/156205. Users can view information on this resource by sending a GET request to that address, regardless of which interface they are using, and authorised users can modify it using POST, PATCH, or DELETE requests. The bibliographic reference associated with this date (Richter et al. 2017) is similarly represented at the address https://xronos.ch/references/17778, and can be accessed at that address using the same interface as the radiocarbon date. These uniform REST interfaces are another example of a boring architectural choice that make it easier for us to enrich the scientific contents of XRONOS, by adding new types of modular resource that represent new scientific entities.

This basic REST pattern is augmented by seven 'actions' (following the standard pattern in Rails application) that express different ways of interacting with a resource: index, show, destroy, new, create, edit, and update. The 'show' action represents interaction with a single resource, as described above. The 'index' action, which lists resources of a given type (e.g. https://xronos.ch/c14s for radiocarbon dates), is worth special mention because it is through this that the filtering logic at the core of XRONOS' two interfaces is implemented. By passing a query as HTTP GET parameters to the index action of a resource, the list returned the user is modified to only include records that match that query. For example, https://xronos.ch/sites?site%5Bcountry_code%5D=CH (the part of the URL after the `?` character encodes the SQL WHERE clause `country = 'CH'` as a GET parameter) lists sites in Switzerland . More complex queries can be executed using nested parameters. For example, https://xronos.ch/c14s?c14%5Bsample%5D%5Bmaterial%5D%5Bname%5D=charcoal (encoding that the `c14` table should be joined to the `material` table via `sample`, followed by the WHERE clause `material.name = 'charcoal'`) lists radiocarbon dates obtained from charcoal samples . Uniquely, index actions can also respond with the result in a tabular data format (i.e. `.csv`).

4.2 Data ingestion and curation

The chronological data in XRONOS comes from a variety of sources, including published structured datasets in repositories and journal supplements, other online databases, literature review, and direct input from collaborators. Our aim is not just to 'mirror' these sources as they are, but integrate them into a single curated and continuously updated database. For the purposes of ingestion, we classify data resources into three categories: static resources, such as supplementary data in published papers, which are imported once; versioned resources, updated on a periodic basis, which we import after each new version; and live resources, which are continuously updated and therefore continuously imported. Records of each type are imported into XRONOS in as close to their original state as possible, i.e. without any corrections or standardisation applied. This ensures that any subsequent changes (even immediate, automatic ones) are entered into the record's version history, so that the source of any deviations or potential errors can always be reconstructed. The version history also records the direct source of the data for attribution purposes. In addition, a bibliographic reference to the original resource is attached to ensure that the source is clearly attributed even if the record is merged with another one.

Table 1: Automatically-recognised data quality issues currently implemented in XRONOS

Issue	N	Description
Sites		
MISSING_COORDINATES	4452	Missing geographic coordinates

INVALID_COORDINATES	0	Geographic coordinates fall outside the earth's ellipsoid
MISSING_COUNTRY_CODE	1221	Missing data on what country the site is in
Samples		
MISSING_MATERIAL	138177	Missing data on the sample material
MISSING_TAXON	138182	Missing data on the sample taxon
MISSING_CRS	0	Sample has coordinates but the coordinate reference system is not given
Taxons		
UNKNOWN_TAXON	9260	Sample taxon has not been matched to the GBIF Backbone Taxonomy
LONG_TAXON	509	Description of the sample taxon is implausibly long
Radiocarbon dates		
MISSING_C14_AGE	447	Missing radiocarbon age
VERY_OLD_C14	76	Radiocarbon age older than the effective range of the method (50 ka)
MISSING_C14_ERROR	585	Missing measurement error
MISSING_D14C	238875	Missing δ13C measurement
MISSING_D14C_ERROR	238875	Missing δ13C measurement error
MISSING_C14_METHOD	242233	Missing data on radiocarbon dating method (conventional, AMS, etc.)
MISSING_C14_LAB_ID	1233	Missing laboratory identifier
INVALID_LAB_ID	16053	Laboratory identifier does not match the standard format (e.g. 'Abc-1234')
MISSING_C14_LAB	350190	Missing data on the radiocarbon laboratory
Bibliographic references		

MIXED_REFERENCE	21933	Bibliographic reference appears to combine multiple publications
MISSING_BIBTEX	42109	Bibliographic reference without structured data in BibTeX format

Once ingested, we apply a number of automated and semi-automated quality control processes to integrate new data into the existing database. Controlled vocabularies are used in a number of places in the data model (**FIG<empty citation>**?), and we use thesauruses to automatically standardise these fields as much as possible. For example, the taxonomic description of samples is controlled using GBIF's backbone taxonomy (**CITE<empty citation>**), and we also use a thesaurus service provided by GBIF to automatically change variant or obsolete taxonomic names to the canonical version. If the system is not able to standardise a field using the available thesaurus, it is flagged for manual correction. A wide variety of other potential data quality issues (e.g. missing data on what country a site is in) are also flagged for human review by this system Table 1, which can often be semi-automated (e.g. suggesting close matches in the thesaurus or the country indicated by the record's coordinates).

A final critical component of XRONOS' data curation system is duplicate handling. We import data from many overlapping resources (many of which incorporate each other either in whole or in part), so duplicate records are common. The end result of standardising and correcting a record is also often to create a duplicate: e.g. the same sample imported from one source as 'oak' but another as '*Quercus* sp.' will become a duplicate pair as '*Quercus*', and thus be recognised as a single sample. Such exact duplicates can be merged automatically, with the oldest record becoming the authoratitive version, but detecting fuzzier duplicated information (e.g. differences in the spelling of site names) has proved a more difficult problem. As of writing there are therefore still many duplicate records in XRONOS that need to be manually resolved, but we hope to automate much more of this work in the future.

4.3 User interfaces

The graphical user interface (GUI) to XRONOS, accessed through a web browser (e.g. at https://xronos.ch), uses REST resources and actions as the building blocks for various interfaces through which users can browse, search, retrieve, and analyse chronometric data. Each action on each resource is represented by a page, though not all of these are publicly accessible. Pages representing REST resources directly are supplemented by a number of synthetic interfaces, for example the 'data browser' (https://xronos.ch/data), which facilitates more complex filtering, or the search interface (https://xronos.ch/search). The GUI also includes several resources which are not part of XRONOS' scientific data model, for example documentation pages, user profiles and news articles; these are currently not exposed in the API.

Access to various 'backstage' interfaces for creating, editing, and deleting data, and monitoring data quality is managed using a user permissions system. Currently only authorised users affiliated with the XRONOS project can access these, but in the future we intend to support open registration and expose editing interfaces to all authenticated users. For this reason, there is no sharp division between a 'public' and 'private' areas – viewing/querying data and editing/curating data share the same architecture and interface patterns.

The XRONOS API uses the same addresses as the web-based GUI (with the exception of some of the synthetic interfaces mentioned above) but responds with machine-readable data in JSON format, rather than a HTML page. This response can be triggered by appending `.json` to the address or by including a HTTP `content-format` header in the request. Though users can make such requests manually and parse the data with one of several off-the-shelf tools, the primary intended uses of this interface is to provide access for 1) programmatic clients to XRONOS and 2) other web services. The XRONOS R package (**CITE**) is an example of a programmatic client; it uses the API to facilitate direct querying and retrieval of data from XRONOS in the R statistical programming language (**CITE**), which is widely used for computational applications in archaeology (**CITE**). Similar libraries could be developed in other programming environments used for scientific computing, such as Python or Haskell. The API also provides the foundation for other web services to access XRONOS directly, to embed chronological information in other contexts or otherwise make use of its data resources.

An overarching principle of this software architecture is that all interaction with XRONOS' data store, and as much of the data processing and 'business logic' of responding to REST requests as possible, is directed through the same server-side routines. First and foremost, this allows us to provide multiple interfaces (i.e. the GUI and API, perhaps more in the future) without duplicating these elements of our codebase. It also improves accessibility for users accessing XRONOS through devices with limited processing capability or through text-only browsers. More broadly, avoiding reliance on client-side processing, e.g. with Javascript or Web Assembly (WASM)—which would be the other option—allows us keep our client interfaces simple (in most cases plain, semantic HTML pages and self-contained stylesheets) and therefore, we hope, sustainable in the face of constantly-evolving client-side technologies and standards. It does have the weakness that, in practical terms, XRONOS is difficult to run and relies on the continued existence of a maintained external server. We have however tried to mitigate this by providing clearly-documented source code and regular data dumps so that, if our instance of XRONOS disappears, or if one simply does not want to use it, it is possible for others to host a XRONOS server of their own. As the sustainability of scientific software and data infrastructures is a pressing problem (a point we will return to in the conclusion), in the future it may be desirable to support further decentralisation through, for example, a federated server-server model (**CITE**).

5 Evaluation

This paper outlines the conceptual and technical infrastructure developed to realise these goals in XRONOS' initial phases of development (2019 and 2021–2024), including a generalised data model for site and radiocarbon information, extendable to other chronometric data; an R- and Ruby-based pipeline for continuous ingestion of data from a variety of sources; continuous, semi-automated data cleaning protocols; a Ruby-on-Rails application providing a web-based frontend to the data and a REST API for programmatic access; and an R package for interfacing with the API. We believe the XRONOS framework provides more open, more reliable, and more comprehensive access to chronometric data than previously available, as well as a foundation for its continuous expansion and refinement.

- XRONOS is useable, use XRONOS!
- Immediate development goals
- The sustainability challenge

XRONOS blends aspects of all three approaches to achieve the same aim of providing access to the global radiocarbon data through a common interface. Like c14bazAAR and IntChron, it is a ‘metadatabase’ that draws from existing data resources and maintains an explicit link to them. But like p3k14c, it integrates these into a single database and applies data curation processes to harmonise them and improve the quality of the information. It has a wider scope than c14bazAAR or IntChron, as it mirrors rather than directly retrieves the source data (allowing us to use resources that aren’t openly published), and does not rely on the authors of these sources to implement a common specification. It also goes beyond the functionality of p3k14c by providing systems for the contunious ingestion and curation of new data as it is published. However in general we see the approaches as complementary. A c14bazAAR parser for XRONOS is in development (https://github.com/ropensci/c14bazAAR/pull/150), and we also aim to provide an IntChron interface to XRONOS’ data in the near future. New data and corrections from p3k14c are incorporated into XRONOS as they are released.

6 Acknowledgements

The XRONOS project was carried under the direction of Albert Hafner. Individual contributors to the XRONOS database to date include Chiara Huwiler, Rivana Moser, Tomasz Chmielewski, and Stephanie Döppler. A complete and up-to-date list of acknowledgements for the project can be found at https://xronos.ch/about/acknowledgements.

7 Funding statement

This work was funded by the Swiss National Science Foundation (SNSF Project #198152) and the University of Bern (Initiator Grant, Caroline Heitz, Project ‘Time and Temporality’).

References

Alves, Eduardo Q., Kita Macario, Philippa Ascough, and Christopher Bronk Ramsey. 2018. “The Worldwide Marine Radiocarbon Reservoir Effect: Definitions, Mechanisms, and Prospects.” *Reviews of Geophysics* 56 (1): 278–305. https://doi.org/10.1002/2017RG000588.

Batist, Zachary, and Joe Roe. 2024. “Open Archaeology, Open Source? Collaborative Practices in an Emerging Community of Archaeological Software Engineers.” *Internet Archaeology,* no. 67 (July 18, 2024). https://doi.org/10.11141/ia.67.13.

Bayliss, Alex. 2015. “Quality in Bayesian Chronological Models in Archaeology.” *World Archaeol.* 47, no. 4 (August 8, 2015): 677–700. https://doi.org/10.1080/00438243.2015.1067640.

Bird, Darcy, Lux Miranda, Marc Vander Linden, Erick Robinson, R. Kyle Bocinsky, Chris Nicholson, José M. Capriles, et al. 2022. “p3k14c, a synthetic global database of archaeological radiocarbon dates.” *Scientific Data* 9, no. 1 (January 27, 2022): 27. https://doi.org/10.1038/s41597-022-01118-7. https://www.nature.com/articles/s41597-022-01118-7.

Bronk Ramsey, Christopher, Maarten Blaauw, Rebecca Kearney, and Richard A Staff Staff. 2019. "The Importance of Open Access to Chronological Information: The IntChron Initiative." *Radiocarbon* 61 (5): 1–11. Accessed April 20, 2019. https://doi.org/10.1017/RDC.2019.21. https://www.cambridge.org/core/journals/radiocarbon/article/importance-of-open-access-to-chronological-information-the-intchron-initiative/5D76092C90DED5500B2512E0AC287398.

Chaput, Michelle A, and Konrad Gajewski. 2016. "Radiocarbon Dates as Estimates of Ancient Human Population Size." *Anthropocene* 15 (September 1, 2016): 3–12. https://doi.org/10.1016/j.ancene.2015.10.002.

Crema, E. R. 2022. "Statistical Inference of Prehistoric Demography from Frequency Distributions of Radiocarbon Dates: A Review and a Guide for the Perplexed." *Journal of Archaeological Method and Theory* 29, no. 4 (December 1, 2022): 1387–1418. https://doi.org/10.1007/s10816-022-09559-5.

Crema, E. R., A. Bloxam, C. J. Stevens, and M. Vander Linden. 2024. "Modelling Diffusion of Innovation Curves Using Radiocarbon Data." *Journal of Archaeological Science* 165 (May 1, 2024): 105962. https://doi.org/10.1016/j.jas.2024.105962.

Fan, Lai-Tze. 2018. "On the Value of Narratives in a Reflexive Digital Humanities." *Digital Studies / Le champ numérique* 8, no. 1 (1 2018). https://doi.org/10.16995/dscn.285.

Gajewski, K., S. Munoz, M. Peros, A. Viau, R. Morlan, and M. Betts. 2011. "The Canadian Archaeological Radiocarbon Database (Card): Archaeological 14C Dates in North America and Their Paleoenvironmental Context." *Radiocarbon* 53, no. 2 (January): 371–394. https://doi.org/10.1017/S0033822200056630.

Hinz, Martin, Martin Furholt, Johannes Müller, Dirk Raetzel-Fabian, Christophe Rinne, Karl-Göran Sjögren, and Hans-Peter Wotzka. 2012. "RADON - Radiocarbon Dates Online 2012. Central European Database of 14C Dates for the Neolithic and Early Bronze Age." *Jungsteinsite* 14:1–4. https://www.jna.uni-kiel.de/index.php/jna/article/view/65/116.

Hyvönen, Eero. 2020. "Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery." *Semantic Web* 11, no. 1 (January 1, 2020): 187–193. https://doi.org/10.3233/SW-190386.

Kintigh, Keith. 2006. "The Promise and Challenge of Archaeological Data Integration." *Am. Antiq.* 71, no. 3 (July): 567–578. Accessed August 15, 2019. https://doi.org/10.1017/S0002731600039810. https://www.cambridge.org/core/journals/american-antiquity/article/promise-and-challenge-of-archaeological-data-integration/037763619B121990D59B585523826A03.

Kneisel, Jutta, Martin Hinz, and Christophe Rinne. 2014. (RADON-B – Radiocarbon Dates Online (Version 2014). Database for European 14C Dates for the Bronze and Early Iron Age). https://radon-b.ufg.uni-kiel.de.

Kra, Renee. 1988. "Updating the Past: The Establishment of the International Radiocarbon Data Base." *American Antiquity* 53, no. 1 (January): 118–125. https://doi.org/10.2307/281158.

———. 1989. "The International Radiocarbon Data Base: A Progress Report." *Radiocarbon* 31 (3): 1067–1075. https://doi.org/10.1017/S003382220001273X.

Kusumgar, S., D. Lal, and R. P. Sarna. 1963. "Tata Institute Radiocarbon Date List I." *Radiocarbon* 5 (January): 273–282. https://doi.org/10.1017/S0033822200036894.

Michczyński, Adam, Andrzej Krzanowski, Mieczysław F. Pazdur, and Mariusz S. Ziołkowski. 1995. "A Computer-Based Database for Radiocarbon Dates of Central Andean Archaeology." *Radiocarbon* 37, no. 2 (January): 337–343. https://doi.org/10.1017/S0033822200030812.

Millard, Andrew R. 2014. "Conventions for Reporting Radiocarbon Determinations." *Radiocarbon* 56 (2): 555–559. Accessed February 2, 2018. https://doi.org/10.2458/56.17455. https://www.cambridge.org/core/journals/radiocarbon/article/conventions-for-reporting-radiocarbon-determinations/E4077EC6F1EE1C90C5170309E2C0CF9B.

Pettitt, P B, W Davies, C S Gamble, and M B Richards. 2003. "Palaeolithic Radiocarbon Chronology: Quantifying Our Confidence beyond Two Half-Lives." *J. Archaeol. Sci.* 30, no. 12 (December 1, 2003): 1685–1693. https://doi.org/10.1016/S0305-4403(03)00070-0.

Raetzel-Fabian, Dirk. 1999. "Editorial." *Jungsteinsite* (November 14, 1999).

Reingruber, Agathe, and L Thissen. 2005. "14C Database for the Aegean Catchment (Eastern Greece, Southern Balkans and Western Turkey) 10,000–5500 Cal BC." In *How Did Farming Reach Europe?,* edited by C. Lichter, 295–327. Byzas 2. Istanbul: Ege Yayinlar.

Richter, Tobias, Amaia Arranz-Otaegui, Lisa Yeomans, and Elisabetta Boaretto. 2017. "High Resolution AMS Dates from Shubayqa 1, Northeast Jordan Reveal Complex Origins of Late Epipalaeolithic Natufian in the Levant." *Sci. Rep.* 7, no. 1 (December 5, 2017): 17025. https://doi.org/10.1038/s41598-017-17096-5.

Rinne, Christoph, Jutta Kneisel, Martin Hinz, Martin Furholt, Nina Krischke, Johannes Müller, Dirk Raetzel-Fabian, et al. 2024. (Rado.NB). https://radonb.ufg.uni-kiel.de.

Schloen, Sandra R., and Miller C. Prosser. 2023. "The Case for a Database Approach." In *Database Computing for Scholarly Research: Case Studies Using the Online Cultural and Historical Research Environment,* edited by Sandra R. Schloen and Miller C. Prosser, 25–73. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-46696-0_2.

Schmid, Clemens, Dirk Seidensticker, and Martin Hinz. 2019. “c14bazAAR: An R package for downloading and preparing C14 dates from different source databases.” *Journal of Open Source Software* 4, no. 43 (November 25, 2019): 1914. https://doi.org/10.21105/joss.01914. https://joss.theoj.org/papers/10.21105/joss.01914.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. “Data Sharing by Scientists: Practices and Perceptions.” *PLOS ONE* 6, no. 6 (June 29, 2011): e21101. https://doi.org/10.1371/journal.pone.0021101.

Verborgh, Ruben, Seth van Hooland, Aaron Straup Cope, Sebastian Chan, Erik Mannens, and Rik Van de Walle. 2015. “The Fallacy of the Multi-API Culture: Conceptual and Practical Benefits of Representational State Transfer (REST).” *Journal of Documentation* 71, no. 2 (January 1, 2015): 233–252. https://doi.org/10.1108/JD-07-2013-0098.