



This PDF is a simplified version of the original article published in Internet Archaeology. Enlarged images which support this publication can be found in the original version online. All links also go to the online version.

Please cite this as: Kandel, A.W., Haidle, M.N., Hochschild, V., Sommer, C. and Kanaeva, Z. 2023 The Aggregation of ROAD Data in the ARIADNE Pipeline: pitfalls and successes, Internet Archaeology 64. <https://doi.org/10.11141/ia.64.9>

The Aggregation of ROAD Data in the ARIADNE Pipeline: pitfalls and successes

Andrew W. Kandel, Miriam N. Haidle, Volker Hochschild, Christian Sommer and Zara Kanaeva

In this article we describe an online database about human evolution, called the ROCEEH Out of Africa Database (ROAD), and discuss our experience in aggregating Palaeolithic data from ROAD in the ARIADNE data processing pipeline. As of April 2023, ROAD contains more than 2400 localities in Africa and Eurasia dating between three million and 20,000 years ago. The database is transdisciplinary by nature and includes cultural artefacts, human and animal fossils, and plant remains. These finds are stored in a relational database, which is part of a structured, web-based, geographic information system. The process of preparing ROAD data for integration with ARIADNE taught us lessons about our own dataset, which we share here.

1. Introduction to ROCEEH and ROAD

The Role of Culture in Early Expansions of Humans ([ROCEEH](#)) is a research centre of the Heidelberg Academy of Sciences and Humanities funded for 20 years through the Academies' Programme of the Union of German Academies. One product of the research centre is the ROCEEH Out of Africa Database ([ROAD](#)) (Kandel et al. [2023](#)). ROAD contains information about archaeological, palaeoanthropological, palaeontological and palaeobotanical localities in Africa and Eurasia from three million to 20,000 years ago (Figure 1).

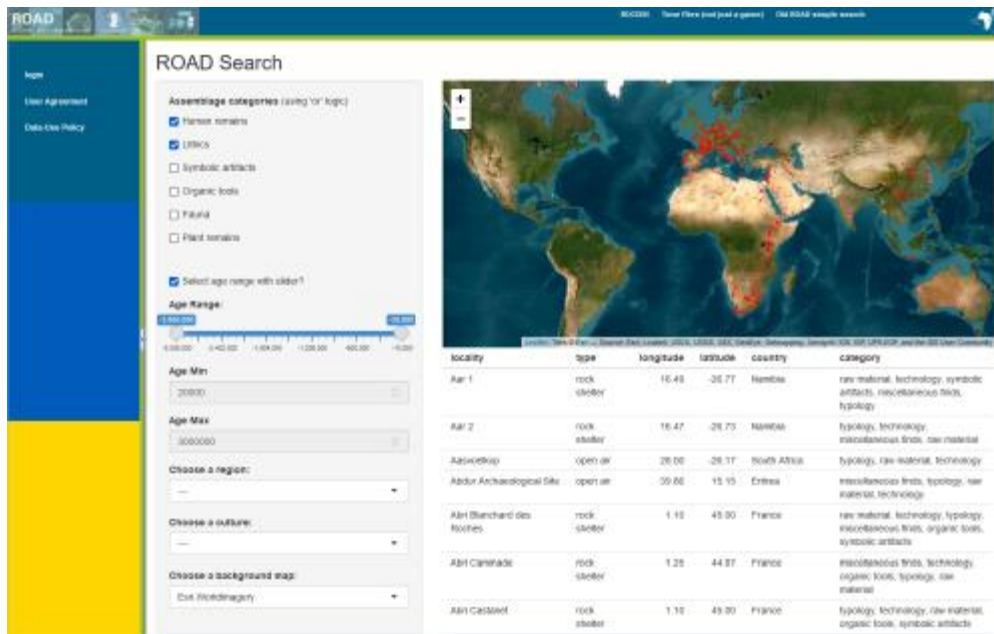


Figure 1: Browser view of [ROAD's entry page](#) showing the results of a simple query for localities with both human fossils and stone artefacts. By entering the name or clicking on a site, a user can download a ROAD Summary Data Sheet as a PDF. (Query executed on 11 April 2023 without login)

The ROCEEH team met frequently to discuss the concept of the database as the project began in 2008 and developed its structure over the next year. After the creation of a logical model by means of the Entity Relationship Diagram and its implementation, the research team began entering data in ROAD in late 2009. Since then, the multidisciplinary research team has integrated over 2400 localities containing more than 23,000 assemblages collected from over 5100 publications written in English, French, German, Italian, Spanish, Portuguese, Russian and Chinese, among other languages. Because ROAD contains vast amounts of previously inaccessible information that can be explored using innovative methods in data science, it serves as a valuable resource for those studying human evolution (Haidle *et al.* [2020](#)).

ROAD is a relational database implemented and administered with a PostgreSQL database management system. It allows user interaction through its application called [ROADWeb](#), which is a web-based application written in .php, javascript and .html (Figure 1). ROAD and its applications are hosted on a dedicated server located at the University of Tübingen. The ROCEEH team purposely chose to use open access software to avoid common issues associated with proprietary software (e.g. cost and accessibility) in the hope of increasing its longevity.,

Although the FAIR Principles (Findable, Accessible, Interoperable, Reusable) for digital data did not exist at the inception of ROAD, the research centre has made strides over the past few years to make ROAD data more FAIR (Wilkinson *et al.* [2016](#)). The research team is currently working to incorporate its data into the Semantic Web and Linked Open Data and sees this solution as a viable option for prolonging the functionality of ROAD after the project ends in 2027. Almost all data in the Semantic Web are distributed using the Resource Description Framework (RDF),



a highly interoperable format developed by the World Wide Web Consortium (W3C) to describe data or metadata. In 2021, the ROCEEH team completed the development of an RDF data model (i.e. ontology) and the first RDF export of ROAD data, which we continue to update periodically. As a result of this effort, we are now able to use these data in other contexts, for example, as interoperable data on maps published by [ROAD in Wikipedia](#) pages about Stone Age cultures (or search Wikipedia for [Uluzzian](#) or [Gravettian](#)).

In accordance with major developments towards open science, ROCEEH registered ROAD with the repository re3data (www.re3data.org/), and published it under an open Creative Commons licence (CC BY-SA 4.0). Based on our experience with data models, thesauri and data synthesis, we have worked to promote the sustainability of the database by developing standardised practices. Our work was complemented by networks of collaboration with [ARIADNE](#), the [Coalition for Archaeological Synthesis](#), the [European Cooperation in Science and Technology](#) (COST) Action titled '[Integrating Neanderthal Legacy](#)' (iNEAL), and the [German National Research Data Infrastructure](#) (e.g. [NFDI4Objects](#)), among other associations.

2. Cooperation with ARIADNE

In January 2020, ROCEEH met the team from ARIADNE in Prato, Italy, to plan a timeline for the integration of data from ROAD into ARIADNE. As a result of this meeting, both teams agreed to a unified workflow. Subsequently, ROCEEH began using ARIADNE's data infrastructure to prepare ROAD data for integration into the ARIADNE GraphDB triple store. With the help of the [Getty Art & Architecture Thesaurus](#) (AAT), which provides a standardised vocabulary, and [PeriodO](#), which stores regionally defined chrono-cultural entities, ROCEEH successfully completed the first round of data integration in September 2021. Since then, users are able to search the ARIADNE portal to find the prehistoric data contained in ROAD, a function that enhances the use of both databases. The next update occurred in March 2022, followed by a third in March 2023.

The main focus of this article is to describe some of ROCEEH's experiences and highlight a few of the pitfalls and successes that our team encountered as we strove to make ROAD data available in the ARIADNE portal. We hope that others can learn from our experiences and are optimistic that such feedback will smooth out the journey for others.

3. Experience of Data Integration from ROAD into ARIADNE

Originally, we thought we would export all of the data stored in ROAD to ARIADNE, making it available to users as a complete data archive. But we soon realised that this action was not necessary. We recognised that ARIADNE was not intended to



serve as a long-term repository, but rather as a practical search engine to locate archaeological information.

In the end, we selected the most relevant information for each dataset in ROAD that we wanted to present in the ARIADNE portal, based on what we could integrate using ARIADNE's 3M mapping tool. The chosen values include site name (locality), a statement about why the site is important (summary), geographic coordinates (x, y) rounded off to protect site locations, archaeological culture (idarchstrat), age range of the layer (age_min, age_max), as well as the general classes (category) of finds such as cultural artefacts as well as human, animal and plant remains. Finally, we included a list of finds, which are represented in ARIADNE as subjects.

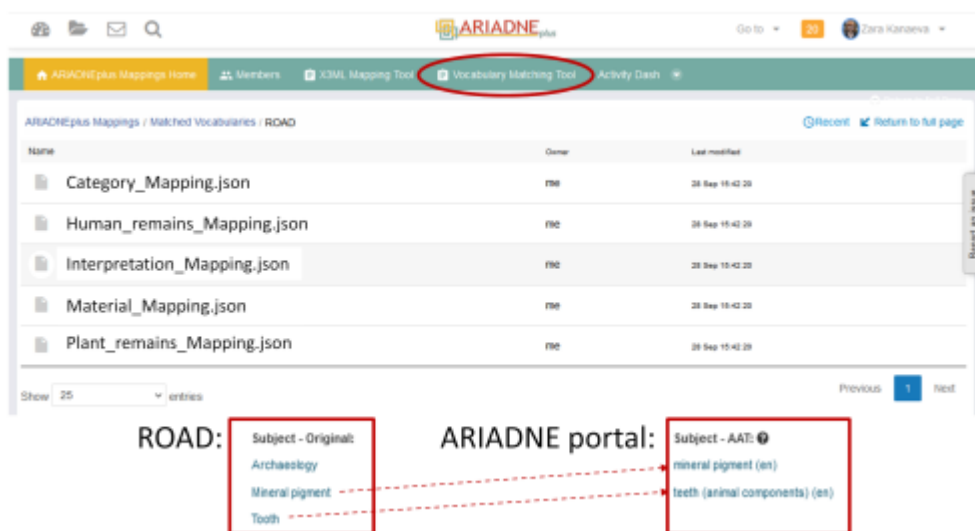


Figure 2: Using the vocabulary matching tool in ARIADNE to match vocabulary in ROAD

We also embedded a link that a user can click to generate a ROAD Summary Data Sheet (Bulus *et al.* 2020). These dynamic PDFs are an important feature of ROAD. They provide an overview of each site entered in ROAD and include more information than that described above; for example, the PDFs include the specific contents of the assemblages and a list of references used. The PDFs help users learn more about a given site. They also serve as a means of data control to ensure the data are correctly entered in ROAD. Furthermore, if a user desires details beyond the scope of the PDFs, they can access, query and download data directly from ROAD, since ROCEEH freely provides all of its data to registered users.

All data from external providers must first be converted into an RDF triple store format, as required for the ARIADNE GraphDB. This transformation is made with the framework called the ARIADNE aggregation pipeline. This pipeline uses three primary mappings: 1) export structure mapping onto the ARIADNE ontology (AO-Cat), the so-called 3M mapping, 2) vocabulary mapping to the Getty AAT (Figure 2), and 3) chronology mapping to PeriodO. When we began, the main difficulty that we encountered was that we did not know how the results of these three mappings would be displayed in the ARIADNE portal.



Since ROAD data do not include the Getty AAT vocabulary, we first needed to define such mappings. For the Getty AAT mappings, we used three categories of matches: exact, close and broad. Many of the Getty AAT subject mappings were intuitive, so that 'mineral pigment' in ROAD matched exactly with the same Getty attribute, while 'tooth' in ROAD could easily be paired with 'teeth (animal components)' in Getty. However, in some cases we experienced a loss of granularity because the Getty AAT contains limited coverage of some palaeoanthropological and Palaeolithic subjects.

While ROAD contains detailed information about the evolution of hominins, not all of these genera are listed in Getty AAT. For lack of a better choice, we decided to broadly match the genus *Ardipithecus* with 'hominini' and 'hominidae' in Getty AAT. As a result, in the ARIADNE portal, the user sees the content of the field 'Subject - Original' as '*Ardipithecus*', which stems from its ROAD entry. The content of the field 'Subject - Getty' is a broader match with two of the Getty AAT entries, 'hominini' and 'hominidae'.

The mapping onto the chrono-cultural database PeriodO offered us more challenges than we expected. PeriodO contains data about cultures that can be ascribed to different periods in various regions by different users. We also noticed that some users in PeriodO lumped Palaeolithic periods together with more recent periods, which expanded their time frames into unrealistic historical epochs. To streamline the process, we decided to use the same chrono-cultural entities as defined in ROAD. In fact, the ROAD attributes matched well with those of PeriodO (Figure 3). The only catch we had was in dealing with similar cultural entities from different regions.

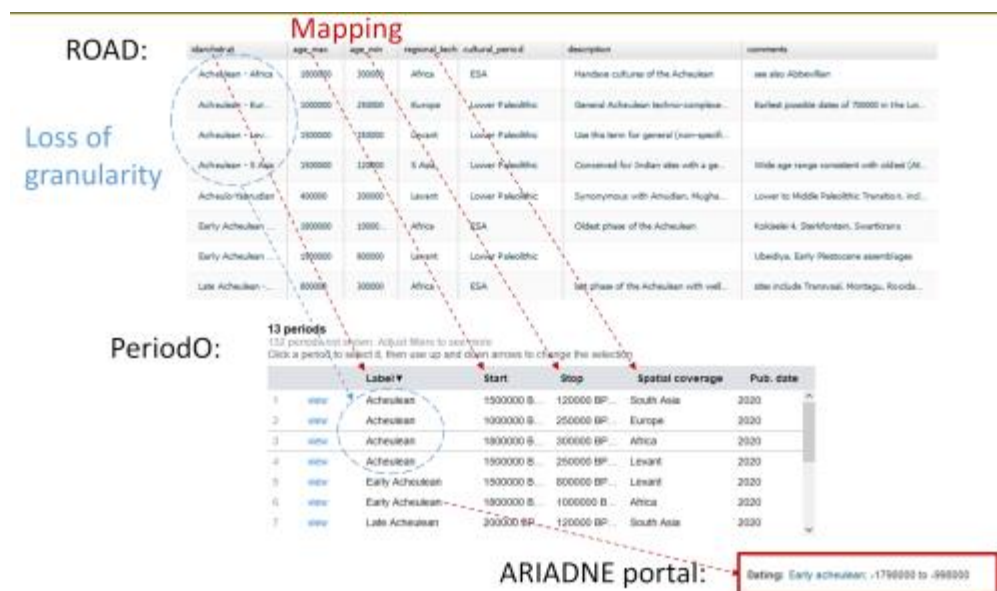


Figure 3: Mapping the chrono-cultural entities in ROAD to PeriodO, and the result in ARIADNE

In ROAD we incorporate the culture and region within a single designation. For example, we define 'Late Acheulean - Africa', since this period has a range in time and space that differs from 'Late Acheulean - Europe'. However, if the cultures do



not match exactly, ARIADNE does not indicate the age range of that culture. The solution to this dilemma was to define the culture as a generalised 'Late Acheulean', but then to specify the region as Africa or Europe. In addition, we needed to create an alternate label called 'Late Acheulean - Africa' or 'Late Acheulean - Europe'. In this way, the entries in ROAD could still be paired with the alternate labels in PeriodO and then recognised within the ARIADNE pipeline so that age ranges could be ascribed to the cultural assemblages. Finally, we encountered difficulties updating our own entries in PeriodO, so owe our thanks to Adam Rabinowitz and Ryan Shaw who patiently added our new data using patches they generated from an Excel list that we periodically supplied.

A further issue was that ARIADNE signifies dates before the Christian era (BCE) in negative years and that the ARIADNE pipeline automatically subtracts 2000 years from data in PeriodO. Thus a date of 500,000 years before present appears as -498,000 years in ARIADNE, which gives a false impression of its precision. We discussed this issue with the ARIADNE team, and contemplated disregarding the 2000 years and presenting the data as -500,000 years. While this might not be correct, it would better convey the imprecise nature of such old dates and simultaneously compensate for the large standard errors, often several thousands to tens of thousands of years, associated with many radiometric techniques applied to date Palaeolithic contexts. Only with ages younger than about 50,000 years would this 'rounding' start to become a significant factor, but even then would represent a manageable error within the time frame at the end of the Palaeolithic.

Another setback occurred when we tried to map ROAD attributes to those of ARIADNE using the 3M mapping tool. In ROAD, we can date sites using radiometric results, in addition to the chrono-cultural stratigraphy applied in ARIADNE. While we also use chrono-cultural dating, we could not bring the radiometric ages of finds in ROAD into ARIADNE's GraphDB. The issue was that the model that describes AO-Cat offered no appropriate resource class for establishing the radiometric ages of the finds, even though this feature is present in ROAD. This represents another area where granularity was lost.

To partly compensate for this, we decided to incorporate the age determinations in ROAD (age_min, age_max) into the 'Original ID' designated in ARIADNE. For example, Original ID 'Grotta Grande of Scario, (59, 60), age: 80000-110000' signifies that assemblages 59 and 60 from the site Grotta Grande of Scario date between 80,000 and 110,000 years. This is useful information since this faunal assemblage is not associated with cultural finds and would therefore lack an indication of age in ARIADNE. In addition to such cases with faunal remains, the same issue applies for human and plant remains in ROAD, since these types of finds can come from contexts that are not associated with a specific culture, and would therefore lack age information in ARIADNE.

There was another positive outcome of adding the age estimates from ROAD to the Original ID in ARIADNE. In cases where the cultural designation is general, for example Upper Palaeolithic (defined in ROAD as from 22,000-43,000 years), we could use the results of absolute dating to specify a more precise age for the assemblage (e.g. 25,000-25,700 years), as shown in Figure 4. In most cases, by using the absolute dating, ROAD provides more precise dating information than



would a general cultural designation such as Oldowan (c. 1.5-2.6 million years), Acheulean (c. 0.3-1.8 million years) or Middle Palaeolithic (45,000-350,000 years). Furthermore, certain methods of relative dating, for example, Marine Isotope Stages, magnetostratigraphy and biostratigraphy, can also be integrated to help refine the age estimates in ROAD.

ARIADNE PORTAL

ROAD - locality: Aghitu-3 Cave, finds: archaeology, culture: Upper Paleolithic - Eurasia

Description

Aghitu-3 Cave is located at the base of a basalt outcrop along the Mousian River valley. This UP site yields the earliest evidence for modern human behavior in Armenia dated between 38-24 ka cal. Notable finds include shell beads, an eyed bone needle and other bone tools, plus many laminar tools made from obsidian sourced up to 260 km away.

Metadata

Original ID: Aghitu-3 Cave, 2, age: 25860-25760

Landing page: [http://www.roadh.net/ariadneplus/ariadneplus/PDF.php?locality/Armenia/Aghitu-3 Cave](http://www.roadh.net/ariadneplus/ariadneplus/PDF.php?locality/Armenia/Aghitu-3%20Cave)

Language: English

Resource type: Site/Inscription

Subject: Digheath
Archaeology
Obsidian

Dating: Upper paleolithic: ~18000 to ~26000

Place: Armenia

Type: Dataset (Provisional record)

Publisher: Heidelberg Academy of Sciences and Humanities

Issued: 2021-04-23

Last updated: 2021-05-06

Responsible person and organisations

Creator: Project "The Role of Culture in Early Expansions of Humans" (ROCEEDH)

Contributor: Project "The Role of Culture in Early Expansions of Humans" (ROCEEDH)

Owner: Heidelberg Academy of Sciences and Humanities

Responsible: ROCEEDH team

License information

Access Rights: <https://creativecommons.org/licenses/by/4.0/>

Resource links

View resource at provider

Join | Print | PDF | Cite | Report an issue

Resource is part of

ROCEEDH Out of Africa Database

Thematically similar

Thematically similar resources based on terms in common of

Subject & Time period:

Cave of Pige de Diable

Cave of Alcobertas

ROAD - locality: Jhinal Cave, finds: archaeology, culture: Upper Paleolithic - E Asia

ROAD - locality: Jhinal Cave, finds: archaeology, culture: Upper Paleolithic - E Asia

ROAD - locality: Jhinal Cave, finds: archaeology, culture: Upper Paleolithic - E Asia

ROAD - locality: Jhinal Cave, finds: archaeology, culture: Upper Paleolithic - E Asia

ROAD - locality: Jhinal Cave, finds: archaeology, culture: Upper Paleolithic - E Asia

ROAD - locality: Jebel Faya FWY 9E1, finds: archaeology, culture: Upper Paleolithic - Eurasia

Tags

Upper paleolithic

Armenia

Figure 4: Screenshot taken from the ARIADNEplus website showing the results of a search for the Upper Palaeolithic site of Aghitu-3 Cave in Armenia. By clicking on 'Landing page' URL, a user can download a ROAD Summary Data Sheet (Bulus *et al.* 2020). This PDF summarises the results of Aghitu-3 Cave and is generated directly from ROAD without the need to log in

Once we could view our data in the ARIADNE staging portal, we realised that many assemblages had the same appearance, which was confusing to an uninitiated user. We therefore decided to consolidate similar assemblages into a single record and denoted this consolidation through the Original ID. For example, Original ID: 'Krapina, (159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171), age: 120000-140000' shows that several assemblages share the same age and the same information contained in the export. This does not mean that the assemblages are



identical, because in ROAD they contain additional information that helps to distinguish the finds from one another.

In the end, despite these minor setbacks we succeeded in integrating ROAD data into the ARIADNE portal and learned from this experience. Since the first export and upload of ROAD data to the ARIADNE portal, we have added periodic updates without major issues, most recently in March 2023.

4. Conclusion

The cooperation and networking efforts gained by working with ARIADNE were instrumental to the growth of our own project. We were motivated to contemplate our own perspectives for the sustainability of our data and to search for solutions. We licensed the database and registered it with several repositories. While the means for long-term archiving of ROAD data are not yet certain, we are reviewing various options to secure our data by the end of the project in 2027 and hopefully to keep a version of the database active.

Most importantly, the cooperation served to stimulate us to begin work on the RDF export of ROAD data in order to incorporate its data into the Semantic Web and Linked Open Data. In this case, we are able to define our own ontology for ROAD, so we do not need to limit the completeness of the data export, as we had to do for the ARIADNE framework. Furthermore, the cooperation helped us to improve our efforts to make ROAD data FAIR. This philosophy has become increasingly important in securing the future of Big Data in science (Kandel *et al.* [2022](#)). This topic also dovetails nicely with another success of our cooperation with ARIADNE, namely making ROAD data findable through the ARIADNE portal.

Finally, to explore the full potential of both ROAD and ARIADNE, we encourage you to visit our respective websites (www.roceeh.uni-tuebingen.de/roadweb/ and portal.ariadne-infrastructure.eu/) to discover what else these databases have to offer. Should you wish to explore ROAD further, ROCEEH provides expanded access for anyone interested. Simply complete the [User Agreement](#) to register for a user id and password.

Acknowledgements

The ROCEEH research centre of the [Heidelberg Academy of Sciences and Humanities](#)) was promoted by the Joint Science Conference of the Federal Government and the state governments of the Federal Republic of Germany in the Academies '[Programme of the Union of the German Academies](#)'. Since 2008, ROCEEH and the research and development of ROAD have been generously supported by the Federal Government of Germany (Federal Ministry of Education, Science and Research) as well as the states of Baden-Württemberg (Ministry of Science, Research and the Arts) and Hesse (Ministry of Science and the Arts).



Bibliography

Bolus, M., Bruch, A.A., Haidle, M.N., Hertler, C., Heß, J., Kanaeva, Z., Kandel, A.W., Malina, C. and Sommer, C. 2020 'Explore the history of humanity with the new ROAD summary data sheet/Durch die Menschheitsgeschichte mit dem neuen ROAD Summary Data Sheet', *Mitteilungen der Gesellschaft für Urgeschichte* **29**, 145-47. <https://doi.org/10.51315/mgfu.2020.29008>

Haidle, M.N., Bolus, M., Bruch, A.A., Hertler, C., Hochschild, V., Kanaeva, Z., Sommer, C. and Kandel, A.W. 2020 'Human origins – digital future, an international conference about the future of archaeological and paleoanthropological databases', *Evolutionary Anthropology* **29**, 289-92. <https://doi.org/10.1002/evan.21870>

Kandel, A.W., Haidle, M.H. and Sommer, C. (eds) 2022 *Human Origins – Digital Future: An International Conference about the Future of Archaeological and Paleoanthropological Databases*, Heidelberg: Propylaeum. <https://doi.org/10.11588/propylaeum.882>

Kandel, A.W., Sommer, S., Kanaeva, Z., Bolus, M., Bruch, A.A., Groth, C., Haidle, M.N., Hertler, C., Heß, J., Malina M., Märker, M., Hochschild, V., Mosbrugger, V., Schrenk, F. and Conard, N.J. 2023 'The ROCEEH Out of Africa Database (ROAD): A large-scale research database serves as an indispensable tool for human evolutionary studies', *PLoS ONE* **18**(8): e0289513. <https://doi.org/10.1371/journal.pone.0289513>

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva, Bonino, Santos, L., 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., 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., Wittenberg, P., Wolstencroft, K., Zhao, J. and Mons, B. 2016 'The FAIR guiding principles for scientific data management and stewardship', *Scientific Data* **3**, 160018. <https://doi.org/10.1038/sdata.2016.18>