D9.1: Data Management Plan.

Project acronym: AQuRA
Project full title: Advanced Quantum Clock for Real-World Applications
Grant agreement no.: 101080166

Due Date:	Month 6
Delivery:	Month 6
Lead Partner:	UvA
Editors:	UVA, MEN, NKT, IXB, CNRS, QQU, VEX, PTB, UMK
Reviewer:	PTB
Dissemination Level:	Public
Status:	Internal review
Approved:	No
Version:	V1.1

This project has received funding from the *European Union's Horizon Innovative Actions programme* under grant agreement No 101080166 (AQuRA project).
This deliverable reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

DOCUMENT INFO

Date and version number	Author	Comments
v0.1, 25/04/2023	Ineke Brouwer & Filipe Marques Teixeira	First draft
V1.0, 05/05/2023	Ineke Brouwer & Filipe Marques Teixeira & Florian Schreck	Draft circulated within AQuRA consortium
V1.1, 23/05/2023	Ineke Brouwer with input from Several beneficiaries	Draft for internal review
V1.2 23/05/2023	Ineke Brouwer	Final draft for submission

GLOSSARY

PM	Project Manager
DS	Data Steward

TABLE OF CONTENTS

Executive summary

This report documents how data management will be organised within the AQuRA consortium. It is based on the model data management plan as provided for Horizon Europe projects. This document describes the types of data collected within the consortium and the efforts towards making the data FAIR.

1 List of contributors

Name	Organisation	Role(s)
Ineke Brouwer*	UvA	PM / Author
Filipe Marques Teixera	UvA	DS / Author / Editor
Florian Schreck*	UvA	Author

* Personnel employed partly or fully under the AQuRA project.

2 Data summary

2.1 Data re-use

As the work in this project builds on the expertise and previous experiments of the project partners, existing data from previous work of the partners may be used. Data of other groups will not be used, except secondary data available in published articles.

2.2 Data types and formats

All data is recorded digitally, and all data is stored in the long run since it would require an excessive amount of work to filter data into useful and not useful compared to simply spending a tiny bit more effort to conserve everything. The data rate is so low that it's trivial to conserve everything.
Data that we created previously ourselves will be used, but no other previous data. Data is in the same format as the new data created during this project. No data is purchased.
Data will be shared upon reasonable request. If the data is protected by IP (such as our CAD files or descriptions of construction processes) access to these data can be denied.
Data packages belonging to the publications of academic AQuRA partners will be deposited in data repositories, see below.
Usually standard file formats are used, but exceptionally data may be recorded in proprietary file formats owing to the kind of measurement equipment or numerical tools used. Data we will record comprise:

Raw data:

- Fluorescence and absorption images of ultracold atom clouds (TGA) Quick preview as BMP.
- Metadata with each experimental run, containing parameters used (ASCII list)
- Sometimes additional data with each run, such as oven temperature, lock traces, photodiode voltages, etc. (ASCII list and tables).
- List of all runs with preliminary analysis (ASCII table)
- Continuously tracked environmental and health data of experiment, e.g. water flows, oven temperature, lab temperature, humidity.
- Data created by numerical simulations or fits (mathematica or matlab files, python or C++ program readable files, ASCII files, images as BMP, PNG, TGA, JPG)

In addition we store:

- Photos illustrating the construction process (JPG)
- Labbook (e.g. OneNote, classical paper)
- Program source code (C++, Python, Mathematica, Matlab) integrated into Github.
- CAD designs (e.g. Altium designer, Solid Works, Inventor)
- Files created for brainstorming, documentation, reports, minutes, articles (e.g. .xlsx, .docx, .txt, ASCII, ODT)
- Presentations (e.g. pptx)
- Manuscripts (LaTeX .tex or .docx) Integrated into overleaf, arXiv, SciPost, etc.
- Tabular data (sensors (experiments; e.g. from photodiodes, temperature, magnetic field or current sensors) (e.g. CSV, ASCII, .xlsx)
- Videos (sensors (experiments), videos / documentation) (.MP4)

File formats used:

- ASCII text (most well-known text format)
- .bmp and .tga for raw image data (BMP: simplest 8-bit grayscale format used by any image processing program. TGA: simplest 16-bit greyscale image format)
- .jpg for photos (e.g. showing parts of the apparati or construction steps. Well known compressed image format.)
- CAD files (e.g. SolidWorks, Inventor, Altium designer. These are the CAD programs licensed by our partners. There is not much choice of professional design programs and these programs are quite widely used.)
- .m, .mb (Matlab, mathematica), .org/.opju (origin) for data analysis and simulations (standard programs in research)
- .cpp .h (C++) for computer control, data capture and data analysis (standard for computer control in our research groups)
- Python for data analysis (standard programming language for this task)
- OneNote for electronic labbooks (standard for labbooks in our groups)
- .pdf for scans of paper labbooks (standard for scans)
- .xlsx sheets for various purposes (parts lists, order information, comparison of construction alternatives,...)
- .pptx: presentation of data (standard)
- .tex (LaTeX): manuscripts (standard in our research community)
- .docx: manuscripts
- Julia (.jl) for numerical simulations

2.3 Purpose of data collection / generation

Raw data are collected as output of scientific experiments or numerical calculations. Secondary data is produced by data analysis. Raw data records the outcome of an experiment or a numerical simulation. Experimental data enables the understanding of an experimental apparatus, the physics it is designed to explore or exploit, or allows to use that physics for another purpose, such as measuring a frequency. Numerical data approximately describes the behaviour of a physical system and can be used to understand novel phenomena, guide experimentalists in their construction of their experiments, or helps to interpret experimental data. Analysing raw data leads to a dataset that distils the essence of the phenomena under investigation out of the raw data.
In addition to these types of data, we also collect secondary data, such as the code of all the self-written computer programs used for our research, labbooks, the JPG files of photos documenting the construction of our apparati, the LaTeX files of articles, CAD files, and so on, see list of data formats.

2.4 Relation to the objectives of the project

The collected data serve the purpose of gaining insights into the operation of optical atomic clocks and related cold-atom experiments. Most of this understanding will be derived from experimental data whose analysis will be presented in the associated publications. Analysed data (and in very rare cases possibly also raw data) will put other researchers into the positions to build on our work. The secondary data is required to understand the exact steps that have been taken to undertake an experiment or to analyse its data and publish it.

2.5 Expected data size

The AQuRA partners generate 0.1 to 1 TB/year of raw data in total. This is processed into at the most 1 GB/year of processed data. Secondary output is 1-10 GB/year.
All data will be stored on the storage systems provided by the partner institutions.

2.6 Origin / provenance of the data

Raw data is created by devices (e.g. optical clocks) at the partner locations and by numerical simulations. Analysed data and secondary data is distilled from raw data by the partners during their research.
The raw data from our clock might be marginally interesting for other people building Sr optical lattice clocks.
The error signal trace from a running clock can be used by specialists to detect signs of certain types of dark matter.

2.7 Data utility outside the project

The raw data is most useful to the group having generated it. For example, archived data is useful to debug the small details of a specific apparatus if something suddenly stops working, to reproduce the analysed data presented in a publication, or to fit an improved theoretical model to existing data. Experimental raw data is most of the time not useful to other experimental groups because it is unlikely that they have an apparatus that is similar enough to be directly comparable. Some theoretical raw data, such as computer codes or sample simulation results, can be useful for other groups to compare their theory models or build on previous work.
Analysed data, as published in articles or PhD theses, is useful for other groups since the analysis distils the essence out of the raw data, which can be transferred to other apparati. The steps taken from raw data to publication are useful to verify the validity of data analysis and its assumptions in light of new knowledge.
With each published article we will provide all datasets that contain data to reproduce the figures and tables. Implementations may vary between the different institutes. For example, UvA will, upon publication of an article, store a data package containing all raw data together with the corresponding lab book entries, self-written experiment control and data analysis programs, simulations/fits and simulation/fit results, together with a description of the steps undertaken during data analysis. This data package will allow verification of all claims made in the paper and additional data analysis (e.g. comparison of data to more advanced models). This data package is stored on figshare, which provides a DOI. PTB provides for all published articles datasets that contain data to reproduce the figures and tables. For derived results as Allan deviations, we will provide in addition the underlying data trace. Furthermore, we will include data on intermediate results of the data analysis, data reduction, or simulation that allow verifying the applied procedures. Data will be stored either on the respective journal's repository or the PTB Open Access Repository (oar.ptb.de).
All data except those involving IP of the partners will be made available in a timely manner upon reasonable request. Data involving IP of the partners, such as CAD files or construction procedures, can be distributed after successful negotiations for access to that data. There is no need to alter data before sharing, such as anonymizing it, since beyond IP issues there is no problem with sharing data.

3 FAIR data

3.1 Making data findable, including provisions for metadata

3.1.1 Identifiability of data and persistent identifier

We use DOIs as persistent identifier.
We use Dublin Core as a metadata standard and the minimum metadata provided for published datasets will cover amongst others title, type of data, creators, publication date and related publications. Each publication dataset will contain a README file describing the relationship of data from raw data over data analysis code to analysed data and publication figures and tables.

3.1.2 Discoverability of data and metadata provision

At the point of publishing (e.g., in reports or peer-reviewed journal articles) metadata for all research data throughout the project will be recorded at the facilities of the partner where the data was generated originally. This metadata will adhere to Dublin Core. UvA uses figshare, which is searchable on the figshare website. PTB will publish metadata along with the data discussed in Sec. 2.7. When relevant and applicable, research data will be shared through facilities at the project partner where the data was initially recorded.
We use Dublin Core as a metadata standard and the minimum metadata provided for published datasets will cover amongst others title, type of data, creators, publication date and related publications.
To be allowed for exchange and re-use between researchers, institutions, organisations, countries, etc., we adhere to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins.

3.1.3 Naming conventions used

AQuRA consortium level data: The naming convention for files is defined in the project handbook.
Experimental partners: The raw data is automatically recorded after each run of an experiment. The data is uniquely identified by storing it in folders and files with unique names. The file names of all analysis data, code used for analysis, created secondary data, etc. are given in the "readme" file included in each publication data package, together with the description of the flow of data from raw data to published figure.

3.1.4 Approach for clear versioning

Experimental partners: the raw data is automatically recorded after each run of an experiment. The data is uniquely identified by storing it in folders and files with unique names. The file names of all analysis data, code used for analysis, created secondary data, etc. are given in the "readme" file included in each publication data package, together with the description of the flow of data from raw data to published figure.
Similarly we make sure that all other files are also uniquely identifiable, usually by adding the creation date in the filename or the permanent storage folders name.

3.1.5 Metadata standards

We use Dublin Core and embed metadata in readme files contained in the data packages.

3.1.6 Approach towards search keywords

figshare publication data package will have keywords to optimise the possibility for discovery and then potential re-use and a DOI, which, as far as possible, will be referred to in the associated publication.

PTB Open Access Repository (oar.ptb.de) lists keywords and provides search functionality. Entries are uniquely identified by a doi.
The data package links can also be obtained in the traditional way, i.e. by writing the authors. Publications can be found with standard search tools, such as Google or the search function of the arXiv. arXiv allows searches for author names or keywords in titles, whereas Google also allows keyword search in the fulltext.

3.1.7 Metadata harvesting and indexing

figshare publication packages follow standard metadata formats, and therefore allow for harvesting and indexing using standard procedures.

3.2 Making data accessible

3.2.1 How the data will be made accessible

Data that allow the derivation of meaningful results to the field will be shared in the form of scientific publications and reports. As conventional in the field, data will be available from the authors of the respective publications (or from the group in case the author is no longer available) upon request. A data package as described above is available with each publication.
Where relevant and applicable, research data will be shared in a searchable way in accessible repositories, as outlined in the previous section.
Data that touches upon intellectual property (IP), such as CAD files or construction procedures, can be made available after successful negotiations with the IP holders. IP cannot be made available without prior negotiation since this would hurt the rights of the IP holders.
Data accessibility implementations may vary between different beneficiaries. For example, the UvA uses figshare to make data accessible upon publication and PTB uses PTB Open Access Repository (oar.ptb.de) or the respective journal's repository.

3.2.2 Deposition of data and associated metadata, documentation and code

Data storage systems vary between beneficiaries. For example, for the UvA, all data is stored on UvA servers. Published data will be available through the figshare publication package, where figshare ensures the allocation of a digital object identifier. PTB will follow the documentation code set PTB library for the PTB-OAR. Data is stored on PTB servers. Experimental and theoretical data generated by Menlo will be stored on Menlo servers.

3.2.3 Which data is made available and why

If no IP is involved, data will be made openly available upon reasonable request. (Reasonable is used here to avoid being spammed by requests.)
Publication data packages are at least available upon reasonable request. Several partners use open file sharing services for the distribution of publication data packages: The UvA uses figshare, while PTB uses PTB OAR or journal repository.
Final analysed data is always openly available in publications.
Data that touches upon IP, such as CAD files or construction procedures, can only be made available after successful negotiations with the IP holders, since doing otherwise would hurt their rights. All partners reserve the right to deny access to data when the disclosement of said data is in conflict with the economical interest of a given partner.

3.2.4 Access to restricted data

Unless stated otherwise below access to all data will be provided by the research group upon reasonable request. Access to data that touches upon IP will be restricted and shared only after

successful negotiations with IP holders. During this process, the identity of the person accessing the data will be ascertained.
Publication data will be made openly available. For the UvA, this will be achieved through the publication data package on figshare. Other beneficiaries may use other platforms for sharing publication data.
There is no specific need for a data access committee, but the UvA's Data Steward and Data Protection Officer are available upon request in case external guidance on data access requests is needed.

3.2.5 Access to metadata

For the UvA, metadata of publication data is included in the figshare publication package. PTB includes metadata in the available datasets. Licences are regulated by the respective provider.
For all partners, publicly available data is openly available under CC BY-NC-SA 4.0 licence with no embargo. All publicly available data and associated metadata is guaranteed to remain available and findable for 10 years.
We will use standard data formats, e.g. CSV, JSON, BMP, etc., such that no special software is needed to read the data.

3.2 Making data interoperable

3.3.1 Interoperability of the data

As much as possible we will use standard file formats such as ASCII, BMP, TGA, and the formats of standard software packages such as Latex, Mathematica, Matlab, Python etc.
Each publication data package includes metadata using Dublin Core, included in readme files.

3.3.2 Standard vocabulary for the data

Standard vocabulary of the ultracold atom field will always be used. Detailed information on accepted wording in the context of metrology is given in "International vocabulary of metrology" (https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf/f0e1ad45-d337-bbeb-53a6-15fe649d0ff1)
The data description included in all publication data packages is sufficient to allow a researcher in our field to redo the data analysis and reproduce the figures.

3.4 Increase data re-use

3.4.1 Licensing for data re-use

For the publication data, documentation needed to validate data analysis and facilitate data-reuse are included in the metadata of the publication data package. This will be made available under CC BY-NC-SA 4.0 licence with no embargo. For the UvA, this is included in the standardised figshare package, other beneficiaries may employ different platforms.

For the restricted data that is protected by the IP of the partners, this documentation will be made available only after successful IP negations.

3.4.2 Data available for reuse

Publication data packages will be made available for reuse immediately after peer-reviewed publication.
All other data that is not touching on IP is made available upon reasonable request within at the most a few weeks.
Data touching on IP is made available within a few weeks after successful negotiations.

3.4.3 Length of time for which the data will remain re-usable

Data will remain on the servers of partner institutions for up to 10 years after collection, with the intention to store it for the foreseeable future.

3.4.4 Data re-use by third parties

Analysed data, as published in articles or PhD theses, is useful for other groups since the analysis distils the essence out of the raw data, which can be transferred to other apparati. The steps taken from raw data to publication are useful to verify the validity of data analysis and its assumptions in light of new knowledge. Other raw data is most of the time not useful to other experimental groups because it is unlikely that they have an apparatus that is similar enough to be directly comparable.
Data that touches upon intellectual property (IP), such as CAD files or construction procedures, can be made available after successful negotiations with the IP holders. IP cannot be made available without prior negotiation since this would hurt the rights of the IP holders.

3.4.5 Data quality assurance processes

Experimental groups: Data will be collected by digital interfacing of the experiments with oscilloscopes, cameras, ADCs, etc. Data-collection concerning the mapping of individual parameter spaces will be performed in a randomised fashion where possible, in order to avoid biases. Typically raw data are recorded automatically after each run of an experiment. Data formats, naming and organisation: see above.
All partners: The data analysis leading to published data is verified by all authors of a publication.

4 Other research outputs

If IP allows, other outputs (software, workflows, protocols, models, etc.) will be included in the repositories next to the research data.
The AQuRA project does not produce physical outputs (e.g. new materials, antibodies, reagents, samples, etc.).

5 Allocation of resources

5.1 Costs for making the data FAIR and how these costs are covered

The costs associated with making data and other research outputs FAIR consist of costs for data entry and costs for data storage. The costs for data entry are covered as Other Direct Costs from the AQuRA budget. The costs for data storage are covered by the individual institutions at which the data is stored at institutional storage systems.

5.2 Responsibility for data management

Roles:

- data capture, documenting and archiving: PhD students, postdocs or other personnel working on the project
- oversight: project manager, project coordinator and project leader at each partner
- External advise and data management plan: UvA Data Steward and Data Protection Officer

5.3 Costs and value of long term preservation

Costs are unknown to the partner as they are not explicit but covered by the institution. The potential value of long-term preservation is highly dependent on the data type, and will generally be small when analysis results have been published.

6. Data security

6.1 Data recovery and secure storage and transfer of sensitive data

Our research data do not contain personal or ethical sensitive data. Nevertheless, data in this project are to be stored in a secure manner by the project partner. All partners ensure that access to data is only possible with a personal password and optional additional security measures. Moreover, each project partner has to ensure that data is back-up at regular intervals.
Experimental partners: research data relevant to the project are stored at secure servers owned and operated by the institute on its campus to allow maximum data integrity and security. Frequent back-ups at regular intervals are designed to minimise the risk of data loss.
Moreover, all project partners use the file sharing facilities provided by the university which come with the just previously outlined benefits.

7. Ethics

7.1 Ethics or legal issues impacting data sharing

Research data from AQuRA do not contain personal data or ethically sensitive data and are not the result of any experiments which fall subject to an ethical review.

8. Other issues

8.1 Other national/funder/sectorial/departmental procedures for data management

UvA will make use of departmental rules, which are a subset of the rules described above.