iqClock

A Data Management Plan created using DMPonline

Creators: Cynthia Crul, Markus Gellesch, Florian Schreck

Affiliation: University of Amsterdam (Universiteit van Amsterdam)

Funder: European Commission (Horizon 2020)

Template: Horizon 2020 DMP

ORCID iD: 0000-0001-8225-8803

Grant number: 820404

Project abstract:

Optical clocks are amazingly stable frequency standards, which would be off by only one second over the age of the universe. Bringing those clocks from the laboratory into a robust and compact form will have a large impact on telecommunication (e.g. network synchronization, traffic bandwidth, GPS free navigation), geology (e.g. underground exploration, monitoring of water tables or ice sheets), astronomy (e.g. low-frequency gravitational wave detection, radio telescope synchronization), and other fields. Likewise, techniques developed for robust clocks will improve laboratory clocks, potentially leading to physics beyond the standard model. To make this a reality, we have founded the iqClock consortium, assembling leading experts from academia, strong industry partners, and relevant end users. We will seize on recent developments in clock concepts and technology to start-up a clock development pipeline along the TRL scale. Our consortium represents a nucleus for a European optical clock ecosystem, which will continuously deliver competitive products and foster the development of clock applications. Our first product prototype will be a field-ready strontium optical clock, which we will benchmark in real use cases, such as network synchronization (TRL 6). This clock will be based on a modular concept, already with the next-generation clocks in mind, which our academic partners will realize (TRL 3-4). By their operation principle, these optical clocks are more robust than the current ones and have come into reach by recent breakthroughs, some of which achieved by our partners. We will leverage the foundational work by the consortia QuantERA Q-Clocks and JRP f17 USOQS, which have joined partners with us, and translate their work into a higher TRL. To increase our impact and to broaden our industry base, we will reach out to all stakeholders, train the next generation of quantum engineers, educate and listen to end users, and enrich the exchange of scientific ideas.

Last modified: 23-04-2020

iqClock - Final review DMP

1. Data summary

State the purpose of the 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. Analyzing raw data leads to a data set that distills 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, the OneNote files of 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.

Explain the 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. Analyzed 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 analyze its data and publish it.

Specify the types and formats of data generated/collected

All data is recorded digitally.

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 its 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 negotiated.

Data packages belonging to the publications of academic iqClock 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)
- program source code (C++, Python, Mathematica, Matlab) integrated into Github.
- CAD designs (e.g. Altium designer, Solid Works, Inventor)
- files created for brainstorming, documentation, houshold keeping,...(e.g. .xlsx, .docx, .txt, ASCII)
- presentations (e.g. pptx)
- manuscripts (LaTeX .tex or .docx) Integrated into overleaf, arXiv, SciPost, etc.

File formats used:

- ASCII text (most well-known text format)
- .bmp and .tga for raw image data (BMP: simplest 8-bit greyscale 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 (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

Specify if existing data is being re-used (if any)

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.

Specify the origin of the data

Raw data is created by ultracold atom experiments at the partner locations and numerical simulations. Analyzed data and secondary data is distilled from raw data by the partners during their research.

State the expected size of the data (if known)

Each experiment generates up to a few TB/year of raw data. This is processed into a few GB/year of processed data. Secondary output is up to a few GB/year.

Theory partners generate a few GB/year of numerical data, processed data and secondary output.

All data will be stored on the storage systems provided by the partner institutions.

Since the amount of data that we expect to be exchanged between partners or with external parties is relatively small we do not expect challenges to share the data between sites. We expect that most requests for data will be for a single sample of raw data, analyzed data or secondary data, which can easily be transferred over internet, e.g. by providing it upon reasonable request on our webservers or distributing it by KPN secure file transfer. In the highly unlikely case that transfer over internet is not possible we can ship hard disks by post.

Outline the data utility: to whom will it be useful

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 analyzed 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.

Analyzed data, as published in articles or PhD theses, is useful for other groups since the analysis distills 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.

UoB and UvA: With each published article we will store a data package containing all raw data together with the corresponding labbook 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.

TUW: For first author articles, a data package will be provided, which will contain source files for the paper (both for text and figures), scripts (in Wolfram Mathematica, julia, python etc), and ASCII files with calculated data readible by respective scripts. In some cases, when the volume of the calculated data is too high, TUW plans to upload only the scripts essential to reproduce these data.

UIBK: some additional data, e.g. molecular potential curves, are published as supplementary material to articles, inheriting the DOI of that article.

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

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

At the point of publishing (e.g., in reports or peer-reviewed journal articles) relevant metadata for all research data throughout the project will be recorded at the relevant facilities of the partner where the data was generated originally. For instance, at UoB, metadata will be documented in the UoB's current research information system PURE. These records will be searchable within the University's Research Portal, FindtIt@Bham (the library discovery tool) and also via Google. UvA uses fighshare, which is searchable on the figshare website. The University of UCPH uses university servers (UCPH Electronic Research Data Archive (UCPH ERDA)) and provides access to data on reasonable request using a DOI.

When relevant and applicable, research data will be shared through facilities at the project partner where the data was initially recorded. In the case of the UoB, this facility would be the eData repository (https://edata.bham.ac.uk/) and which makes the datasets discoverable through search engines like Google.

We use 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.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

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.

Outline naming conventions used

iqClock 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.

Outline the approach towards search keyword

UoB's eData repository (https://edata.bham.ac.uk/) makes the datasets discoverable through search engines like Google.

UvA's figshare publication data package will have a DOI, which, as far as possible, will be referred to in the associated publication.

UCPH's Electronic Research Data Archive (UCPH ERDA) is accessible using DOIs.

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.

Outline the 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.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

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

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

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. The following partners use open file sharing services for the distribution of publication data packages:

UoB: eData

UvA: figshare

TUW, UIBK: zenodo

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.

Specify how the data will be made available

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.

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

Most data can be accessed without special software. Self-written source code used to analyze the data (e.g. C++, Python, matlab, mathematica) is contained in the publication data packages, include a description of the data analysis flow from raw data through analysis software to final publication figures.

Specify where the data and associated metadata, documentation and code are deposited

UoB: eData

UvA: UvA servers for everything, figshare for publication data packages

UCPH: Data, metadata and documentation are deposited on university servers (UCPH Electronic Research Data Archive (UCPH ERDA))

TUW, UIBK: zenodo

other partners: servers of the respective institution

Specify how access will be provided in case there are any restrictions

Unless stated otherwise below access to data will be provided by the research group upon reasonable request.

UvA: access to data packages on figshare is open.

UoB: access to data packages on eData is open.

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

As much as possible we'll 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.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Standard vocabulary of the ultracold atom field will always be used.

The data description included in all publication data packages is sufficient to allow a researcher who is knowledgeable in our field and has access to all commercial software licenses (e.g. Mathematica, Matlab, Origin) required to redo the data analysis and reproduce the figures.

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

Except data protected by IP of the partners (CAD files, etc.):

CC BY-NC-SA 4.0, no embargo

Specifically this license applies to all raw data recorded by our experiments.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

UvA, UoB and TUW: 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.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

Analyzed data, as published in articles or PhD theses, is useful for other groups since the analysis distills 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.

Describe 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 randomized fashion where possible, in order to avoid biases. Typically raw data is raw data is recorded automatically after each run of an experiment. Data formats, naming and organization: see above.
All partners: The data analysis leading to published data is verified by all authors of a publication.

Specify the length of time for which the data will remain re-usable

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

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

Data entry costs:
Covered by Other Direct Costs from the iqClock budget
Data storage costs:
- Data storage systems provided through partner institutions. Costs are covered by institutions.

Clearly identify responsibilities for data management in your project

Roles:
- data capture, documenting and archiving: PhD students, postdocs or other personnel working on the project
- oversight: project leader of each partner

Describe costs and potential 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.

4. Data security

Address data recovery as well as 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 University on its campus to allow maximum data integrity and security. Frequent back-ups at regular intervals are designed to minimize the risk of data loss.
Moreover, all project partners use the filesharing facilities provided by the university which come with the just previously outlined benefits.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

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

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

-