DRUM: Data-repository for biomedical ultrasound metrology

Elly Martin and Bradley Treeby
Department of Medical Physics and Biomedical Engineering, University College London

Background and Summary

Experimental measurements are critical for the development of medical ultrasound software and devices, including for validation of modelling tools and for comparison of measurement equipment and protocols. Data sharing encourages reproducibility and consistency across labs, and provides access to other researchers who may not have the equipment or expertise to conduct their own measurements.

To promote and enable sharing of ultrasound metrology data across the international community, a new open access subject repository has been established under the University College London Research Data Repository. This will enable storage, curation and sharing of data in the long term (at least 10 years). Contributions from research groups across the community are welcomed with the prerequisite that the data supports a peer reviewed publication. All datasets will be released under suitable creative commons licences and assigned a digital object identifier which will form a permanent link to the data which can be used for retrieval and citation of datasets.

Methods

Scope

The data repository is intended for datasets of experimental measurements relating to biomedical ultrasound. Other areas of acoustics (e.g., seismology) or other areas of biomedicine (e.g., biology) are out of scope.

Data storage infrastructure

The dataset will be hosted in the open access UCL Research Rata Repository under the project name Ultrasound Metrology. The repository user interface is supplied by Figshare (https://knowledge.figshare.com/) and supported by UCL Library Services. The repository will allow indefinite storage of research data (guaranteed for at least 10 years). Datasets will be curated and converted to a common storage format (hierarchical data format) containing a common set of meta-data, assigned a unique DOI and published under a suitable creative commons licence.

The subject repository has an accessible webpage where datasets can be browsed or searched. This can be accessed via https://rdr.ucl.ac.uk/projects/DRUM_Data-repository_for_biomedical_ultrasound_metrology/73251. Datasets are deposited along with descriptive metadata (see below) and assigned a digital object identifier (DOI). Multiple datasets belonging to a single study should be published under a single record. One example of this is several sets of measurements made under different conditions and published in a single journal article.

Data file-format

All datasets will be stored in a standard file format which is intended to ensure that complete and consistent details are made available for all studies. This will maximise usage of the data, enabling for example, replication of the study either in experiment or simulation for validation purposes.

Datasets will be stored in Hierarchical Data Format (HDF5), a cross platform file format which can support large data volumes and heterogeneous data types. Broadly, HDF5 files are organised in a hierarchical structure divided into groups and datasets (equivalent to folders and files). The individual datasets contain a description of the study, details of the ultrasound source and driving conditions, the measurement equipment, the propagation medium, and the data itself, for example, measured waveforms. A detailed description of the standard file format is given in Appendix 1.

All datasets to be deposited in the repository must be converted to the standard HDF5 format. For the initial stages of the repository, we are happy to work with authors to convert their data to the correct format. In time, converters for the files generated by automated scanning equipment (e.g., those made by Precision Acoustics and Onda) will be made available. A standard set of tools for reading and displaying the data will also be made available for common programming languages (including MATLAB, Python, Julia, and C++).

To aid identification of files by study, and where a single record is composed of multiple data files, all files should be named using the project.studyName field described in Appendix 1. This gives a short name for each file in the format: institute acronym - unique short description - two-digit number.

Supplementary data

In addition to the data specified in the standard data format, further data may exist which would be useful to future users of the data. This supplementary data should be deposited along with the datasets to ensure it remains available in the long term. Examples of this may include photographs of the experimental set up, medical image data (e.g. CT scans of the propagation medium), transformation matrices that relate medical images to source coordinates, and additional details about the source or sensor (e.g. datasheets, element maps). There is no strict specification for the format of these supplementary data, however, they should be stored in common and easily readable file formats, the contents should be easy to interpret, and they should have descriptive names. All supplementary data files should be grouped within a single zip file.

Procedure for depositing data

Authors wishing to deposit data should contact Elly Martin (elly.martin@ucl.ac.uk) and Bradley Treeby (b.treeby@ucl.ac.uk). The following items are required for submission:

1. One or more datasets in the standard HDF5 format.
2. Completed pre-submission form (which captures the required metadata).
3. (Optionally) A single zip file containing supplementary data.

After screening, datasets will be deposited by UCL staff (Elly Martin and Bradley Treeby) on behalf of the authors. Each dataset will be assigned a digital object identifier and published under the chosen licence. The record will be accompanied by metadata which contains keywords to allow searching and retrieval of the data from the repository.

Criteria for accepting data for the repository

The following criteria must be fulfilled before the data can be deposited:

- The data must form part of a peer reviewed publication, or the value of the data to the research community must otherwise be demonstrated.
- The data must be stored in the standard HDF5 format.
- All compulsory fields in the HDF5 file format must be populated.
- The pre-submission form must be completed.
- Authors must confirm they own the copyright and have permission to share the data. It is the authors responsibility to check with funders, institutions, and publishers as to any sharing and copyright rules.
- Authors must confirm they are not uploading personal data that could identify a living individual, or confidential data that would contravene a third-party agreement.
- Authors must consider whether the data to be published may be licensed commercially before deciding to freely release it to the public.

Metadata

For each record, standard metadata fields are also completed at the repository level. This information helps to make the data discoverable within the repository. The metadata includes a brief explanation of the content of the dataset, for example, linking the datafiles to specific sections of a publication. These fields should be completed as part of the pre-submission form.

Publication and usage of data

By default, all data will be published (rather than remaining private). Those wishing to deposit should check guidance given by their funding agency to ensure they have permission to make the data available. When the data is deposited, an embargo can be applied if necessary. If this is required, the timeframe for this should be stated along with a reason for the embargo. A DOI can be generated when the record is generated, before the data is published.

Licences

The data are released under creative commons licences, of which there are several variations, the default is CC BY, but any other can be specified at submission. The most appropriate should be chosen at the time of submission, it will not be possible to change this once the data is published. The different types of licence are outlined below, further details can be found at https://creativecommons.org/licenses/. Users of the data must adhere to the terms of the licence under which the datasets are published.

Licence code	**Licence name**	**Terms**
CC0	no copyright	any re-use permitted
CC BY	Attribution	any re-use permitted, authors must be credited
CC BY-SA	Attribution - ShareAlike	any re-use permitted, authors must be credited, resulting work must be shared with the same licence
CC BY-ND	Attribution - NoDerivatives	any re-use permitted, authors must be credited, can't be shared in any other format
CC BY-NC	Attribution - NonCommercial	commercial use not permitted, authors must be credited
CC BY-NC-SA	Attribution - NonCommercial - ShareAlike	commercial use not permitted, authors must be credited, resulting work must be shared with the same licence
CC BY-NC-ND	Attribution - NonCommercial - NoDerivatives	commercial use not permitted, authors must be credited, can't be shared in any other format

Note, in some circumstances, other licence types (e.g., as required by different institutes) can also be used. Please discuss this with Elly Martin and Bradley Treeby if required.

Digital object identifiers

A digital object identifier (DOI) is a unique code which permanently identifies a document and provides a link to the document on the internet. Each dataset will be allocated a DOI when it is deposited in the repository. This can be used as a means of dissemination and accessing the data and should be used to reference the datasets in future work resulting from use of the data.

Appendix 1: Data file structure

STRUCTURE	REQ	DIMS	TYPE	UNITS	DESCRIPTION	EXAMPLE
project						
└─userName	Y	[1]	char	-	Name/s of person who acquired the data.	‘Elly Martin’
└─userAffiliation	Y	[1]	char	-	Name of institute or company.	‘University College London’
└─studyStartDate	Y	[1]	char	-	Start date of the study (i.e., the date of the first scan) in ISO 8601 format YYYY-MM-DDZ.	‘2018-02-03+00:00’
└─studyEndDate	Y	[1]	char	-	End date of the study (i.e., the date of the last scan) in ISO 8601 format YYYY-MM-DDZ.	‘2018-05-08+00:00’
└─studyName	Y	[1]	char	-	Short name of the study in the format: institute acronym, followed by a unique short description, ending with a two-digit number (to allow for multiple datasets within a study). This does not need to be defined by users, but is assigned on entry to the database.	‘UCL-repeatability-study-01’
└─studyFileNumber	Y	[1]	int	-	Numerical identifier for the file. Used when multiple files form a single study.	1
└─studyDescription	Y	[1]	char	-	Short description of the study, this could be, for example, the title of the related publication.	‘Investigation of the repeatability and reproducibility of hydrophone measurements of medical ultrasound fields.’
└─studyReference	N	[1]	char	-	DOI of accompanying published peer reviewed paper.	‘https://doi.org/10.1121/1.5093306’
source						
└─sourceProperties						
└─description	Y	[1]	char	-	Description of the source.	‘Sonic Concepts H-151 driven with a short burst’
└─positioningSystem	Y	[1]	char	-	Description of the positioning system.	‘Precision Acoustics UMS, automated tilt/rotate, fixed x/y/z’
└─sourcePosition	N	[3]	float	m	x/y/z position of the entire transducer (i.e., translation of the transducer coordinate system relative to the scan axes origin). Define if scannedComponent = ‘sensor’ or ‘none’.	[0, 0, 0]

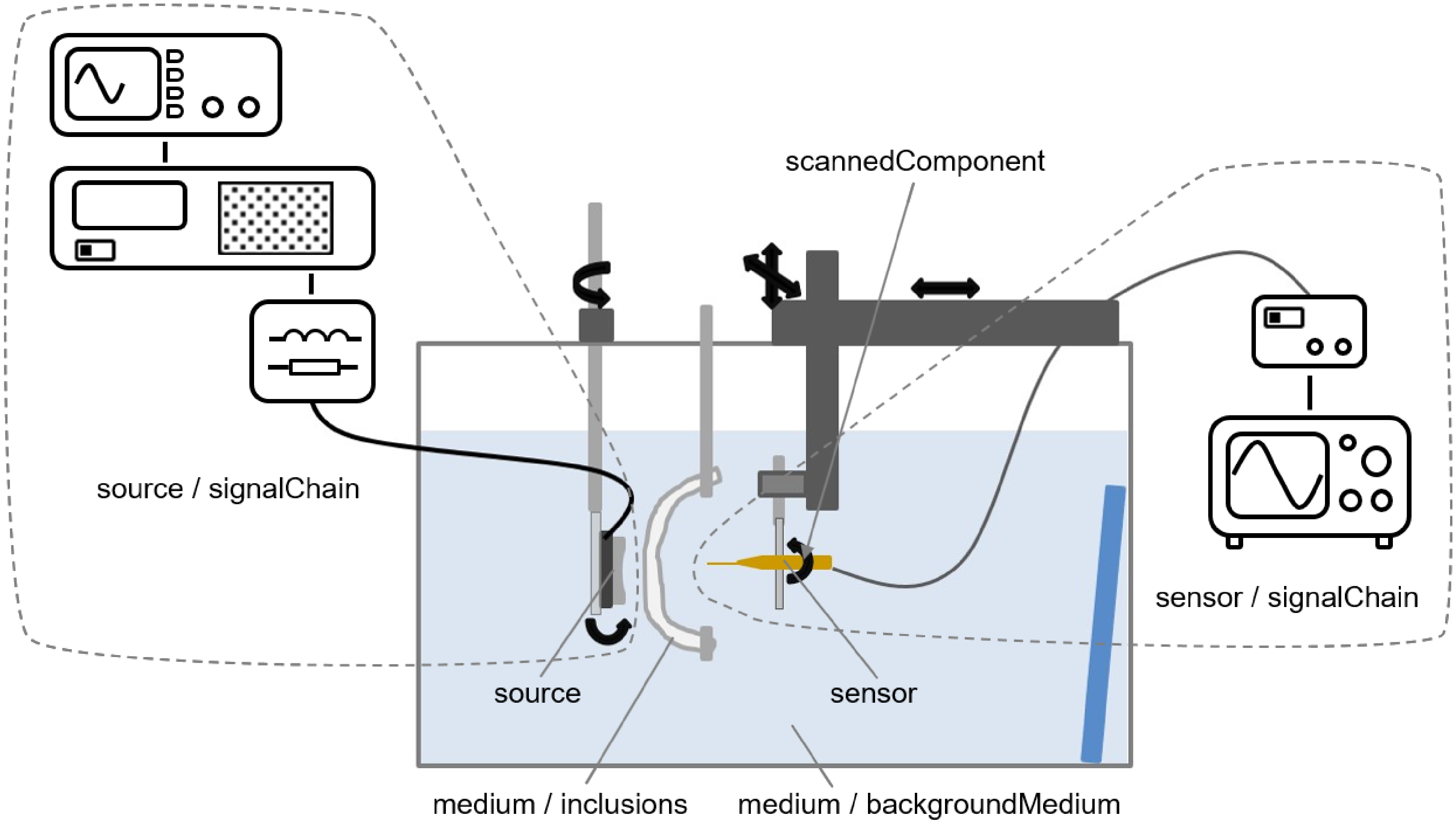
└─sourceOrientation	N	[3]	float	degrees	θ_1/θ_2/θ_3 rotation angles about x then y' then z'' (i.e., rotation of the transducer coordinate system relative to the scan axes). Define if scannedComponent = 'sensor' or 'none'.	[0, 0, 0]
└─centreFrequency	N	[1]	float	Hz	Centre frequency of the source.	1000000
└─numberElements	Y	[1]	int	-	Number of elements in the source.	2
└─elementSize	Y	[1]	char	-	Description of radius of curvature, aperture diameter, element dimensions, etc, with units.	'Aperture diameter 63.2 mm, radius of curvature 99.7 mm'
└─multiElementPositions	N	[numberElements, 3]	float	m	Define for multi-element transducers. x/y/z position of each element relative to the transducer origin.	[]
└─multiElementFocus	N	[numberElements, 3]	float	m	Define for multi-element transducers. Any point on the beam axis of the element. Defines the orientation of the element relative to the transducer coordinate system.	[]
└─**signalChain**	>=1				Group has an integer attribute called "number_components"	
└─**<component #>**						
└─modelName	Y	[1]	char	-	Model name of the component.	'33250B'
└─manufacturer	Y	[1]	char	-	Manufacturer of the component. Set to Custom if manufactured in-house or manufacturer is unknown.	'Keysight'
└─type	Y	[1]	char	-	Type of component, e.g. signal generator, oscilloscope.	'Arbitrary waveform generator'
└─serialNumber	N	[1]	char	-	Serial number if known.	'123456ABC'
└─settings	N	[1]	char	-	Description of settings of component.	'Sine wave 211 mV, 4 cycle burst, burst period 1 ms'
sensor						
└─**sensorProperties**						
└─description	Y	[1]	char	-	Description of the sensor.	'Precision Acoustics 0.2 mm membrane hydrophone'
└─positioningSystem	Y	[1]	char	-	Description of the positioning system.	'Precision Acoustics UMS, automated x/y/z, manual tilt/rotate.'
└─sensorPosition	N	[3]	float	m	x/y/z position of the sensor. Define if scannedComponent = 'source' or 'none'.	[0, 0, 0]

└sensorOrientation	N	[3]	float	degrees	$\theta_1/\theta_2/\theta_3$ rotation angles about x then y' then z''. Define if scannedComponent = 'source' or 'none' or if the sensor is placed at an angle.	[0, 0, 0]
└frequencyResponse	Y	[5, N]	float	[Hz, V/Pa, %, radians, %]	Frequency response of the sensor given as [frequency, amplitude, amplitude uncertainty, phase, phase uncertainty].	$\begin{bmatrix} 100e3 & 1.22e-7 & 9 & -0.0011 & 9 \\ 150e3 & 1.27e-7 & 9 & -0.0017 & 9 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$
└elementSize	Y	[1]	char	-	Description of the size of the sensing element, e.g., diameter, radius of curvature.	'0.2 mm diameter'
└**signalChain**					Group has an integer attribute called "number_components".	
└**<component #>**						
└modelName	Y	[1]	char	-	Model name of the component.	'D1602'
└manufacturer	Y	[1]	char	-	Manufacturer of the component. Set to Custom if manufactured in-house or manufacturer is unknown.	'Precision Acoustics'
└type	Y	[1]	char	-	Type of component, e.g. preamplifier, hydrophone, oscilloscope.	'Differential membrane hydrophone'
└serialNumber	N	[1]	char	-	Serial number if known.	'D1602-95'
└settings	N	[1]	char	-	Description of settings of component.	[]
medium						
└backgroundMedium	Y	[1]	char	-	Description of the measurement medium.	'Water'
└conductivity	N	[1]	float	S/m	Conductivity of the background medium.	5.5e-6
└dissolvedGasContent	N	[1]	char	-	Description of dissolved gas content including units.	'5 ppm'
└tankSize	Y	[1]	char	-	Description of the measurement vessel.	'0.6 by 0.6 by 1 m rectangular tank'
└inclusions	N	[1]	char	-	Description of any inclusions (e.g., heterogeneities, scatterers, etc) in the field.	[]
data					Group has an integer attribute called "number_datasets"	
└**<dataset #>**						
└**drivingConditions**	Y					
└drivingRegime	Y	[1]	char	-	Description of driving regime. Valid settings are 'burst' (N-cycle burst, with number of	'burst'

					cycles defined below) or ‘pulse’ (driven by a pulser).	
└─drivingNumberCycles	N	[1]	int	-	Number of cycles. Define if drivingRegime is set to ‘Burst’.	4
└─drivingFrequency	N	[1]	float	Hz	Driving frequency. Define if drivingRegime is set to ‘Burst’.	1100000
└─pulseRepetitionFrequency	Y	[1]	float	Hz	Repetition frequency of the drive signal.	1000
└─drivingAmplitude	Y	[numberElements]	float	-	Numerical drive amplitude, e.g., output power, voltage, or setting.	73.4
└─drivingAmplitudeDescription	Y	[1]	char	-	Description of the parameter defined under drivingAmplitude.	‘Peak to peak driving voltage measured at matching network in V’
└─drivingPhase	N	[numberElements]	float		Define relative phases for multi-element sources if known.	[]
└─notes	N	[1]	char	-	Field to place additional notes, for example, pulser settings.	[]
└─scanData						
└─scanDescription	Y	[1]	char	-	Description of the scan.	‘Lateral scan through the focus, -20 mm to 20 mm.’
└─scannedComponent	Y	[1]	char	-	Description of which component is scanned. Allowable options are ‘source’, ‘sensor’ or ‘none’.	‘sensor’
└─scanAxesOrigin	Y	[1]	char	-	Description of where [0, 0, 0] is in the coordinate system.	‘Centre of transducer element.’
└─scanAxesLabels	Y	[1]	char	-	Labels for the scan axes, e.g., ‘x’, ‘xy’, ‘xyz’.	‘x’
└─scanDataUnits	Y	[1]	char	-	Units for the acquired data.	‘V’
└─scanNumDim	Y	[1]	int	-	Number of scan dimensions, e.g. 1 for line scan, 2 for planar scan.	1
└─scanSize	Y	[scanNumDim]	int	-	Number of scan points in each dimension.	201
└─scanTotalPoints	Y	[1]	int		Total number of scan points.	201
└─scanPointSpacing	Y	[scanNumDim]	float	m	Spacing of scan points.	0.2e-3
└─scanCoordinates	Y	[scanTotalPoints, 3]	float	m	xyz positions of each scan point.	$\begin{bmatrix} -0.0200 & 0 & 0.0980 \\ -0.0198 & 0 & 0.0980 \\ \vdots & \vdots & \vdots \end{bmatrix}$
└─scanAngles	N	[scanTotalPoints, m]	float	degrees	Rotation angles of each scan point. A description of angular coordinate system (i.e., the points of rotation) must be given in the scanAnglesDescription.	[]

└─scanAnglesDescription	N	[1]	char	-	Description of the angular coordinate system (points of rotation).	[]
└─scanTemperature	N	[1] or [scanTotalPoints]	float	°C	Temperature acquired with each waveform, or mean temperature during the scan.	20.1
└─scanStartTime	N	[1]	char		Date and time stamp of file creation in ISO 8601 format, e.g., YYYY-MM-DDThh:mmZ.	'2018-05-04T16:21:46+00:00'
└─scanEndTime	N	[1]	char		Date and time stamp of file creation in ISO 8601 format, e.g., YYYY-MM-DDThh:mmZ.	'2018-05-04T16:29:05+00:00'
└─scanRate	N	[1]	char		Rate of scan, define if scan start and end times are not defined.	[]
└─waveformData	Y*	[scanTotalPoints, waveformRecordLength]	float	scanDataUnits	Measurement data, e.g. voltage waveforms. * either waveform or complexData must be provided.	$\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}$
└─waveformTriggerDelay	Y*	[scanTotalPoints]	float	s	Trigger delay time for each waveform (start of acquisition window), must be defined when waveform data is given.	$\begin{bmatrix} \\ \\ \vdots \end{bmatrix}$
└─waveformRecordLength	Y*	[1]	int	-	Number of samples in each waveform, must be defined when waveform data is given.	2000
└─waveformSamplePeriod	Y*	[1]	float	s	Waveform sample period, must be defined when waveform data is given.	10e-9
└─waveformAverages	Y	[1]	int	-	Number of averages acquired for each waveform or complex pressure measurement.	32
└─complexData	Y*	[scanTotalPoints]	float	-	If only the amplitude and phase is stored (e.g. for quasi steady state field scans), the complex pressure should be given here. * either waveform or complexData must be provided.	[]
other						
└─<custom name>	N	-	-	-	Additional information can be stored under the other heading. Use descriptive titles and camelCase.	-

Example experimental set up with reference to data fields:


scannedComponent
source / signalChain
sensor / signalChain
source
sensor
medium / inclusions
medium / backgroundMedium