# The Kymata Atlas: Data Preservation Policy

Version 1.00 (Author: AT, 17/08/16) Initial drafting of document.
Version 2.00 (Authors: AT, SL, MT 22/11/16) Edits following amendments from Somaya Langley and Marta Teperek.
Version 2.01 (Authors: AT, SL, MT 31/01/17) Addition of data breach procedure, following discussions with Kieren Lovell (IM Implementation Manager, IS).

## 1. Aims

The aim of this document is to set out how the developers of Kymata intend the data associated with the atlas to be preserved in perpetuity.

## 2. Standards

We aim to a minimum of Level 2 of the National Digital Stewardship Alliance's *Levels of Preservation* Framework (Phillips et al., 2013), which includes guidelines for data storage, file fixity, metadata and file format control. In addition, we aim to:

- retain files are in their 'original' file format *with as much* metadata (such as technical, structural metadata etc.) preserved. For example, retaining the

**lastModified** date for each file as well as the original filename and (relative) file path (which might be needed to reconstruct the data).

- retain files at the highest quality possible, where preservation format standards are not achievable.
- never 'up-sample' files (e.g. never turn a compressed file into a lossless file).
- If a 'normalised' file format is created – e.g. a WAV file is turned into an MP3, then both files are retained.

## 3. Content coverage

Within the context of the Kymata research, three types of content are intended to be preserved: **raw measurement data** (that is to say, the electro-magnetoencephalographic recordings of individuals brains and associated data), **hypothesis data** and **result data**. There are five types of raw measurement data that are associated with Kymata, and one form of result data.

Raw measurement data:

- **Experimental stimuli** The stimuli experienced by the participants being recorded. (.wav, .tiff, etc.)
- **Raw EMEG measurements of individuals (plus individual structural meshes)** The electro- and magnetoencephalography recordings of the individual participants experiencing the stimuli, together with the structualT1 scans of their brains. (.fif format).
- **Average EMEG measurements (uses average structurals)** The above recordings, averaged together into a single, average recording. (.fif format)
- **Average source current estimates (uses average structurals)** Each individual raw electro- and magnetoencephalography recording can be used to estimate individual source current estimates. When these source current estimates are averaged, these result in the average source current estimates. (.stc format)

Hypothesis data:

- **Hypothesized functions**. The mathematical transforms that the brain may (or may not) be engaged in, in code format. (.py, .mat, .gcc, .java, etc)

Result data:

- **Processing pathway map** The processing map generated when Kymata is fed the average source current estimates, the experimental stimuli and hypothesized functions. The format of this map is a directed, acyclic graph.

## 4. Overview of preservation strategy

In general, Kymata will aim to preserve data in formats that are open and free, although in some cases this will not be possible because the underlying data is saved in a proprietary format. Where a proprietary format is in use, any migration or normalisation to another formats will be informed by best practice within the international digital preservation community, if this exists. If migration or normalisation occurs, the original proprietary file format will also be retained.

Personally identifiable data (eg. **Raw brain-activity measurements of individuals**) is not shared outside the Kymata Development Group, and is always kept securely (see section 6). All data types that are made publically accessible (for instance on kymata.org) contain no personally identifiable data (either because the data is anonymized or averaged), and as such are not bound by the Data Protection Act 1998. For more information about data sharing, see the Kymata Data Sharing and Data Access Policy.

The following is a list of current files formats:

| Data type | File formats |
|---|---|
| **Experimental stimuli** | .wav, .tiff (open, free) |
| **Raw brain-activity measurements of individuals (plus individual structural meshes)** | .fif format (proprietary Elekta, free to open) |
| **Average EMEG measurements (uses average structurals)** | .fif format (proprietary Elekta, free to open) |
| **Average source current estimates (uses average structurals)** | .stc (open, free) |
| **Hypothesized functions** | .py (open, free)<br>.mat (proprietary Mathworks)<br>.c (open, free),<br>.c++ (open, free),<br>.java (open, free)<br>etc.<br><br>*This list is not exhaustive, functions can be written in any programming language.* |
| **Processing pathway data** | .JSON (open, free) |

All data is generated in-house (and not by third parties) so it should not require specific ingress procedures.

Guidelines for functional preservation:

- We will try to ensure continued readability and accessibility of all data held in Kymata
- Preserve a minimum set of metadata associated with each file, including:
    - **original filename**
    - **original file path** (relative, rather than absolute)
    - **lastModified date**
    - **SHA-1 checksum hash**
    - metadata about the software dependencies for each file including:
        - **tool** - name of the tool/software used to create the file
        - **tool version** - the version of the software used

    - **additional dependencies** - any additional libraries/plugins used
- Items will be migrated to new file formats where necessary, but the original will always be retained.
- Where possible, software emulations will be provided to access un-migrated formats.

# 5. Methods / levels of preservation

Guidelines for data retention:

- Items will be retained in perpetuity.
- We employ best practice data management procedures to ensure preservation including keeping three backup copies of preservation material.
- For results data, preservation of permalinks is of utmost importance as these are used in scientific publications.

# 6. Implementing the strategy (operational details)

## 6.1 Location of storage media

- **Raw brain-activity measurements of individuals.** This data is stored on the MRC Cognition and Brain Sciences Unit (CBU) servers, and does not leave these servers except under specific circumstances (see the Kymata Data Sharing and Data Access Policy).

- **Experimental stimuli, average EMEG measurements, average source current estimates** This is stored on the Psychology Storage Server, located in the Psychology Department, University of Cambridge.

- **Hypothesized functions.** The central repository is stored on Github.com (https://github.com/kymata-atlas/), with a remote copies stored on the Kymata Server, located in the Psychology Department, University of Cambridge.

- **Processing pathway data.** This is stored on the Kymata Server, located in the Psychology Department, University of Cambridge.

## 6.2 Security of storage media

- **Raw brain-activity measurements of individuals.** The MRC Cognition and Brain Sciences Unit is committed to taking all necessary precautions to ensure the physical safety and security of all data collections that it preserves. The server rooms are equipped with multiple key entries linked to an on-site alarm system. Only CBU members have (password protected) virtual access to the data. Please contact the MRC-CBU IT dept. for more information.

- **Experimental stimuli, Average EMEG measurements, Average source current estimates**. Only the project lead and Kymata developers have (password protected) access to the data.

- **Hypothesized functions** See https://help.github.com/articles/github-security/ Only the project lead and Kymata developers have (password protected) admin access to the data and only these people have the ability to modify this data. The data is stored and made available as open source, so anyone can copy and reuse it.

- **Processing pathway data** The University is committed to taking all necessary precautions to ensure the physical safety and security of all data collections that it preserves. The server room where the Kymata sever and backup rooms are stored are equipped with multiple key entries. Only the psychology IT department have physical access to the server room. Only the Kymata lead and developers have (password protected) virtual access to the data.

## 6.3 Data breach procedure

- **Raw brain-activity measurements of individuals.** The MRC Cognition and Brain Sciences Unit IT department will follow their own internal procedures (contact: MRC CBU IT Head).

- **Experimental stimuli, Average EMEG measurements, Average source current estimates, Hypothesized functions and Processing pathway data** Follow UIS Information Management *Information Security Breach* Form (contact: UIS IM Implementation Manager).

## 6.4 Data back-ups

- **Raw brain-activity measurements of individuals.** The CBU server data is routinely backed up, with at least three copies stored online, nearline and offsite.

- **Experimental stimuli, Average EMEG measurements, Average source current estimates** This routinely backed up on the server and off-site (to the plant sciences department).

- **Hypothesized functions** The central repository is stored on Github.com (https://github.com/kymata-atlas/). Github back-up these repositories off-site (https://help.github.com/articles/github-security/). We also store remote copies on the Kymata Server, located in the Psychology Department, University of Cambridge.

- **Processing pathway data** This routinely backed up on the server and off-site (to the plant sciences department).

## 6.5 Version control

- **Raw brain-activity measurements of individuals.** Each dataset has explicit version numbers which are reflected in dataset names (Dataset 1.00, 1.01, 2.00 etc.).

- **Experimental stimuli, Average EMEG measurements, Average source current estimates** Each dataset has explicit version numbers which are reflected in dataset names. No errata and corrigenda are included with the original record. The most recent version is clearly identified.

- **Hypothesized functions** Versions are controlled using git (https://git-scm.com), with the central repository stored on Github.com (https://github.com/kymata-atlas/).

- **Processing pathway data** The version control for processing pathways data is complicated. A set of pathways is dependent on a dataset, each of which has an explicit version number which is reflected in dataset names (the same as the raw and averaged measurements). Each dataset supercedes the previous dataset, but the old set of pathways will remain. The atlas' main URL will always link to the latest version, but if users are forward to the old version (through permalinks), they can continue to view this old data, but an alert will tell them that there is more recent data available. However, within the latest dataset, the evidence for pathway is constantly fluctuating while more data is added in. No version control takes place during this time. This only effects the latest dataset, so the previous dataset is the *stable* version.

## 6.6 Data removal and/or deprecation

- **Raw brain-activity measurements of individuals.** Items are never removed.

- **Experimental stimuli, average EMEG measurements, average source current estimates.** Items are never removed.

- **Hypothesized functions.** No hypothesized functions data will be deprecated. (The data is open source (version controlled via git), with the central repository stored on Github).

- **Processing pathway data.** Items are not normally removed, however, functions may be deprecated. When this happens, a replacement function will be provided, together with a link to this replacement version, where available, together with a note explaining the reasons for withdrawal. The metadata of withdrawn functions will only be searchable using certain filters.

  *Not sure what a function is? See https://kymata.org/documentation for more information.*

### 6.7 Permalinks persistence

- **Processing pathway data.** The processing pathway data is linked to from journal publications, thus all permalinks given out must remain persistent. (The persistent link format is https://kid.kymata.org/<kid>, where <kid> is a function's unique Kymata ID)

## 7. Sustainability, closure and succession plans

At present, the data held in the Kymata Atlas service is aimed to be available in perpetuity. While the costs of running the Kymata Atlas service are currently maintainable, if this changes in the future, closure plans will need to be implemented. In planning for this risk, there are several different closure plans proposed, depending on the type of data:

- Costs for storing the **raw brain-activity measurements of individuals** data is born by the MRC, and this data should exist in perpetuity. See the MRC CBU data preservation plan for more information (https://mrc-cbu.cam.ac.uk). Users wishing to access this data would have to contact the director of the MRC.

- Costs for storing **experimental stimuli, average EMEG measurements and source current estimates** data are born by the University of Cambridge, and this data should exist in perpetuity. This data would remain available to users online.

- There are no costs associated with the **hypothesized functions**: Github is free and open source.

- Costs for keeping the **processing pathway data** available are low**.** The data is stored on the Kymata Server, located in the Psychology Department, University of Cambridge. As this server has been paid for, there are no running costs, except the maintenance of the server, currently undertaken by the Kymata dev team. Keeping this data online is critical – unlike the other data, this database is linked to from journal papers. Due to this, if the project lead or developers are unable to carry on with the project, authority over the server must be given to the head of the University of Cambridge Psychology Department.

### *Succession plans*

If something happens to the project lead or the project lead leaves the Kymata project, authority over the atlas will be given over to the developers, with final say over the direction of the atlas given by the head of the Psychology Department. This would be done with written signed agreements made with each of the parties, detailing their ongoing/future responsibilities for the service.

## References

The Data Protection Act (1998) The Data Protection Act 1998, Chapter 29, UK Parliament, http://www.legislation.gov.uk/

Phillips, M., Bailey, J., Goethals, A., Owens, T. (2013) The NDSA Levels of Digital Preservation: An Explanation and Uses Proceedings of the Archiving (IS&T) Conference, Washington, DC, http://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf