**LIS 4220 Final Project: Profiles and Write Up**

Brigitte Kirchner

LIS 4220 – Data Curation

Dr. Matthew Mayernik

2/23/2021

**Data and Metadata Profile**

The dataset included here are from a series of six experiments done in a study by Caroline Curwen, a PhD candidate at the University of Sheffield, entitled “Synaesthesia for reading written musical keys”. In this study, from what I can gather from the abstract on Figshare, as well as an abstract of the conference proceedings these data/this study was presented at in 2019, Curwen did correlation and comparisons between how people with synesthesia (synesthetes) responded to the stimuli of seeing a written version of a key (in one of three randomized ways: written in treble clef, bass clef, and as a word) and a color, using both a group of synesthetes as well as a control group to test for reaction times, congruency/incongruency, and variance of responses when only shown the stimuli in certain intervals.

The data were uploaded to Figshare in September 2020, updated in November 2020, and have a total of 12 spreadsheets, including raw data collected from the subjects of the study, as well as ANOVA statistical analyses of these data. Though there is a pretty sizable amount of data and metadata provided with the dataset and on Figshare, there are some areas where the metadata could be enriched and improved upon. There doesn’t appear to be an explicit, published study as far as I’m aware, just references to a conference proceeding where the data may have been presented, and other publications by Curwen previously that mention music-color synesthesia.

**Data**

The data in this study are primarily quantitative, and measure things like reaction times in milliseconds to stimuli, calculations and other analyses of color perception, congruency and incongruency of matching the musical key to a particular color through various visual means, ANOVA correlation data and analyses between a control group and a group of synesthetes, and data involving repeating measures of music, as well as data involving very specific amounts of time (in milliseconds) in which the stimuli was shown in comparison to the number of errors made by the participants. Most of the data appears to have been collected and analyzed through IBM’s *SPSS Statistics V25* software.

There are various stakeholders for these data, including those involved in the study, (the 18[[1]](#footnote-1) participants, the researcher and conductor of the study, Caroline Curwen, the University of Sheffield), as well as those outside, including musicians with or without synesthesia, and anyone in the field of music psychology and that studies synesthesia of all kinds or other cognitive perceptions/phenomena.

As mentioned earlier, the 12 datasets, detailing 6[[2]](#footnote-2) different experiments, contain a variety of different types of quantitative data. Since the data were run and collected through a proprietary software, the file format that all of the data is collected in (.sav) is proprietary. Despite this, the data is published and presented through a Creative Commons 0 License (i.e. the Public Domain), so theoretically anyone can use the data if they can somehow access and open the files. There appears to be no restrictions on potential usage of the data. There are workarounds for the proprietary nature of the files if you don’t own the software, including open source software (like JASP Stats) and online conversion tools (like coN’verter, which is what I used to convert the .sav files to .xlsx, so that I could take a closer look at the data and try to actually determine what was being collected).

**Metadata**

The datasets themselves come with not much metadata—from the .xlsx files that I converted from the original .sav files, there are multiple pages in the same workbook; usually the first page is a “Rawdata” page, the second a “ListVariables” page, and the third a “Source” page.

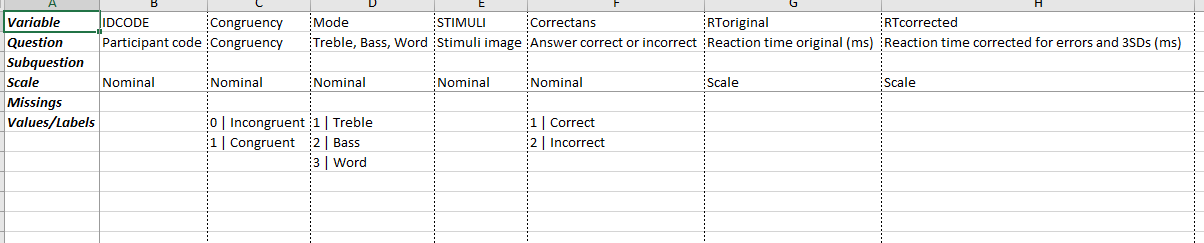
The “ListVariables” page sometimes has clarifications on what numbers in some columns mean.

Figure 1: Example of the "ListVariables" page from the Exp 2 - Synaesthete Data workbook

For example, the dataset “Exp 2 – Synaesthete Data” has 7 columns of data, and multiple columns assign numbers as a value that means true/false, whether the stimuli was presented in treble, bass, or word mode, and whether or not the answers are incorrect. The “Mode” column assigns that 1 = Treble, 2 = Bass, and 3 = Word. This is about the extent of the metadata that comes packaged with the datasets.

Luckily, there are much more metadata on the initial Figshare page for the dataset, including an abstract that details some of the study’s methodologies and findings, categories that this dataset was considered under (including: Musicology/Ethnomusicology, Psychology and Cognitive Sciences, Creative Arts, Media and Communication Curriculum and Pedagogy, Performing Arts, Sensory Processes, and Perception and Performance). These subject headings appear to come from the Fields of Research controlled vocabulary, according to the XML documentation. Specifications of this vocabulary can be found in the Resources section of this paper.

There are also what appears to be natural language/keyword tagging for these datasets as well on Figshare, which definitely have a narrower focus on music-color synesthesia and music psychology.

I was also able to find some metadata about the studies when searching for any publication of the study, by finding it on DataCite.org, which has the metadata in the DataCite Metadata Schema, and allows it to be easily converted to XML or JSON. Having the metadata in this schema helps with the discoverability, identification, and citation of the data.

**Enriching metadata for discovery and use**

Mostly, there’s not much that needs to be changed to the metadata that exists, just that it could potentially be centralized better. Having to go through various different sites and look through publications to try and find out more about the data, even when I had the datasets downloaded into a file format I could open, was cumbersome. Some of the subject heading choices seem a bit unrelated, but otherwise they appear to come from a particular standard. The data itself is very easy to find and cite, because it has a Digital Object Identifier (DOI), but there could be more ways to make the context better to understand. I think some better detailing of how the various datasets relate to each other within this study would be useful as well; one can infer how they relate, but it isn’t explicit, and would be very difficult for someone who has no familiarity with either the music or cognitive psychology fields. I have a music background, and there are still some aspects – those concerning perception and what some of the data are—that I’m still not completely sure about.

One thing I noticed is that, especially if someone is looking for these data and isn’t totally familiar with the field or the categories/subjects that the data is posted under, is that there might be some discoverability issues related to spelling. Because Curwen and the University of Sheffield are both in the UK, they use British English spellings as opposed to American English spellings (like synaesthesia vs synesthesia, colour vs color, etc). Having only one variation on the spelling might not necessarily bring up the study or the data if a searcher doesn’t know of these variations, especially because these spelling variations are about the main subjects.

The last thing I would add, in terms of helping with relationship data, is having some connections to previous studies that Curwen has done—from her website, she lists two publications (neither of which mention this study/this data), which could be used as good background information gathering for someone not necessarily familiar with this specific sub-field. On her website, there is also a section for “Conferences” which lists the title of what I can assume is a paper or presentation that directly relates to the study, though there only exists an abstract (which is different than the abstract/description of the study on the Figshare page) with some more details on the study the data were used in. Searching for any studies started with searching for doing a basic keyword search of “University of Sheffield Curwen” and finding her website as the second result (the first result being her ResearchGate page). Searching through both the University of Denver’s library discovery system and Google Scholar brought me to her published articles, of which there are two. Since the data was only uploaded to Figshare late last year, and because her presentation on it was in 2019, I imagine that either a paper is forthcoming/this is part of her dissertation, and any explicit study that outlines this study is not available yet.

**Repository Profile**

For the purposes of this project, determining a repository to store these datasets was challenging in a few respects. After searching through the Registry of Research data Repositories, I found a repository called PsychData, which feels like it would be a decent, if disciplinary, fit for Curwen’s datasets. PsychData, though it appears to be in the process of being integrated with a larger, not dataset/data specific repository called PsychArchives, has some clear benefits for researchers in psychology, and a mostly clear process for data submission and access after submission. It also appears to provide help with processing and making the data available to researchers.

**PsychData and PsychArchives**

PsychData, also called the Research Data Center, is a discipline specific data repository created and operated by the Leibniz Institute for Psychology (ZPID) in Germany. The Leibniz Institute functions as a research body for German speaking countries (“About”, n.d.), though there doesn’t appear to be any location restrictions for data submission. PsychArchives also calls itself an “international repository for digital research objects” (“PsychArchives Terms of Use”, 2019), and appears to have no restrictions on submission. Some of the sub-disciplines that PsychData has collected include:

1. Clinical Psychology
2. Developmental Psychology
3. Educational Psychology
4. Work, Organizational, and Business Psychology
5. Cognitive Psychology
6. Social Psychology
7. Personality Psychology, and
8. Other Areas. (“Datasets”, n.d.)

When looking into PsychData, I noticed that it appears to be closed to general submissions, and that if you want to submit through PsychData, any data will automatically be transferred and integrated to PsychArchives[[3]](#footnote-3). For the purposes of answering the rest of the questions about the procedures and processes, I will therefore defer to the documentation provided for submission by PsychArchives for submission particulars. It is worth noting that while PsychData and PsychArchives appear to be very similar in mission and purpose (as they are created and operated by the same institution) PsychArchives also is a repository for any and all parts of research and all of the associated objects – which they call “Data Research Objects” (DROs)— including articles, pre-prints, code, audio/visual data, as well as research data and datasets.

**Choosing PsychData/PsychArchives**

Choosing a data repository for data that belong to studies that are multi- or interdisciplinary can be a challenge, due to the fuzziness in scope. From what I’ve seen during my research of data repositories, if the data are interdisciplinary, they might go to a multidisciplinary repository, or they will go into a repository for a specific institution or university. While multidisciplinary repository does have space for these kinds of data, I based my decision for picking a disciplinary repository for these datasets because from what I could find, the University of Sheffield already has a multidisciplinary data repository, called ORDA – The University of Sheffield Research Data Catalogue and Repository. Since the purposes of this assignment would be to bring this data to a repository, I felt like it would make sense to choose a repository it isn’t already in, and one that isn’t tied to a specific institution.

Another challenge, since the data cross two different disciplines, was which one to defer to when choosing a disciplinary repository. I ended up going with a psychology repository like PsychData based on the fit of the data with other data in the respective repositories. First, the study and data measure the cognitive processing of notated music and written keys as words, rather than specifically composed music, made me lean more towards the psychology aspect. Though the data deals with music-color synesthesia, the types of data fit better in a scientific repository, rather than a Musicology or Music repository. After searching in the Registry of Research data Repositories, there is a category for “Musicology” and one for “Fine Arts, Music, Theatre and Media Studies”, the first category having 13 results and the second 66, some of which overlap. From exploring some of these repositories, I noticed that most of the data types were for notated music/sheet music collections in image files, audio or video files, and clips of music or art. Since the data in this study were measuring response times and included statistical analysis data/ANOVA analyses, this data didn’t seem to fit with what was already in the Musicology repositories.

**Data Submission**

As PsychArchives appears to be open for data and other research submissions in the psychology and adjacent disciplines, there is a plethora of documentation for how to go about actually submitting data. PscyhArchives has a general how to page that lists 17 different publication types (“How to Submit”, n.d.); for datasets and research data in particular, they also ask that you use their submission assistant to contribute files as well as the “Other” category content type form. In order to use the submission assistant, you have to either have an ORCID account, or a Leibniz Psychology account. As I didn’t feel comfortable making either of these accounts at this time, I didn’t look too closely at the submission assistant. This form downloads an .xlsx document, and has a list of metadata fields, of which only 7 are required:

1. Author/creator
2. Date of first publication
3. Title
4. Name of License for publishing
5. A way to agree to their terms of service (you would need to type AGREED)
6. Abstract/description
7. Language of content

Other fields include keywords, subjects, possible citation, publication status, and tags that are either visible or hidden from the public view of PsychArchives but are used for filtering/searching in the repository.

There are also quality and technical requirements for submission into PsychArchives (“Quality and selection as well as technical guidelines”, n.d.). First, since you will be using the submission assistant, one must have an ORCID account or a Leibniz Psychology account. Some of the submission guidelines are that objects need to belong in the psychology discipline or other disciplines that fit within the scope of PsychArchives. Research data and datasets must also be submitted with a codebook, which will contain all the relevant metadata to help another research interpret the data. There is a maximum file size of 5 GB, though larger submissions may be accepted upon request. Their specifications on metadata vary from type of object to type of object—from what I can gather from that “Other” form, it appears to be in the Dublin Core metadata schema. They also have preferred file formats; for tabular data – which is the kind that are in Curwen’s datasets – they prefer .csv files, though they will also accept .sav or .xlsx formats if necessary along with the .csv files (p. 1). All of this information seems like it would fit within the Submission Information Package (SIP), as defined by the OAIS reference model.

One last important aspect of data submission is that there are two different but related services to help someone in depositing research and in getting the required documentation and help for the data submission. The first is the Submission Assistant tool that I mentioned previously. The Leibniz Psychology Institute also offers a tool called the “DataWiz” which appears to assist researchers throughout the entire research process (including the archiving of data). Besides these automated tools, contact information for various people within the institute are provided both on PsychData and on PsychArchives pages.

**Data Access and Storage/DIP**

When researching PsychData, it is worth noting that there doesn’t appear to be a way to deposit data directly to PsychData anymore, at least not on the English version of the website. The data access and use policies for PsychData and PsychArchives is also different. PsychData requires that if one is to use the data archived there, one needs to fill out a data use agreement form, which asks for information such as first and last name, the institution you study at or do research for, telephone number, email address, physical address, and the purpose of the data, which you fill out as a form on the site itself or download as a PDF (“Data use agreement,” n.d.). On the other hand, PsychArchives does have data freely accessible. In both instances, the metadata is accessible on the site without any sort of restriction.

The way one searches for the available datasets is also different from PsychData to PsychArchives. In PsychData, one can browse studies by category, but not search for data through a search box or with a query (as none exists). PsychArchives, on the other hand, has many different ways to access the data, including searching, browsing by category, browsing by subject, and can filter once you’ve searched for data. Once you’ve found a dataset, you can download the associated files directly. The metadata that is displayed on each platform is different as well—PsychArchives, probably due to the breadth of objects they store in the repository uses Dublin Core elements, while PsychData appears to use both Dublin Core and DDI standards. What appears to be included in the Dissemination Information Package (DIP) are the enhancement of the metadata and data deposited—by introducing and including controlled vocabularies, which isn’t really touched upon much in the submission part. The DIP would also, potentially assign a DOI to the research data, and provide linking to the study the data is supplemental to. All in all, if one is still able to submit data to PsychData, it would be a worthwhile repository for Curwen’s datasets on music-color synesthesia.

**Additional Information**

**Data citation**

As there is no study available or published by Curwen that details this data, having a consistent citation would be extremely important for anyone who might need to find this dataset. As this dataset was given a DOI through DataCite, there is an autogenerated citation in a variety of different citation styles, including APA, Harvard, MLA, etc. Given that this data would a) be potentially housed in a psychology-specific repository, APA would be the most appropriate citation style. Thus, the preferred citation would look like this:

Curwen, C. (2020). *Synaesthesia for reading written musical keys* [Data set]. The University of Sheffield. <https://doi.org/10.15131/SHEF.DATA.13140086.V2>

This citation includes the title of the dataset, the researcher, the publisher, and a DOI link.

**Long term preservation**

For the most part, long-term preservation of these data will mostly come down to making sure that all the files and possible metadata are converted to open file formats and that there are measures in place to make sure the files remain accessible. One way that I solved this problem was to convert the original data files from the proprietary .sav format to .csv, though it wasn’t necessarily a lossless or clean conversion process. CSV files can be opened with any kind of text editing or word processing programs, some of which come installed or open source software can be downloaded, and thus don’t rely on proprietary software like .sav or even .xlsx files do. Though I’ve included all three file types, one would need specific tools/work arounds to access .sav or .xlsx files.

One other aspect that helps with the long-term preservation is that by submitting this dataset to another repository, this can be used as a backup or as another institution that will have the time and resources to provide preservation services. PyschArchives, as they are funded and created by a European research institute, has the ability to make long-term preservation a core part of their data repository.

**Copyright License**

This dataset, as released on Figshare, already has a CC0 license, meaning that it is in the public domain. A CC0 license allows any who wish it to use, reuse or share this data without any restrictions. This is done for most research outputs that are deposited to Figshare, to help with searching of the data, citation of the data, and sharing of it (“CC0”, Examples section, para. 2); thus, having this license still makes sense if depositing the datasets elsewhere.

**Human Subject Considerations**

One final aspect of note is that of the possible human subject considerations. This study appears to have used human subjects in their data. Luckily, from what metadata there is on the study, it looks like any private information has already been anonymized. Participants in the study were split into two groups—one with nine synesthetes, and one with nine ‘controls’ (presumably, people without music-color synesthesia)—but they were labeled by participant number, and it is unclear if that is a persistent and unchanging number across the six experiments. Thus, any human identifiable data has been anonymized to the furthest extent possible. Because of the nature of this data—where the researcher explicitly seeks out and measures cognitive responses of people who have synesthesia—this appears to be the only identifying aspect to the data.

References

*CC0.* (n.d.). Creativecommons.org. Retrieved February 23, 2021 from <https://creativecommons.org/share-your-work/public-domain/cc0/>

*Datasets.* (n.d.) Psychdata.de. Retrieved February 10, 2021 from <https://www.psychdata.de/index.php?main=search&sub=browse&lang=eng>

*Data use agreement.* (n.d.). Psychdata.de. Retrieved February 10, 2021 from <https://www.psychdata.de/index.php?main=take&sub=vertrag&lang=eng>

*How to submit*. (n.d.). Psycharchives.org. Retrieved February 10, 2021 from <https://www.psycharchives.org/contribute>

Leibniz Institute for Psychology. (n.d.). *About*. <https://leibniz-psychology.org/en/institute/about/>

*PsychArchives terms of use*. (2019, March 20). Psycharchives.org. Retrieved February 10, 2021 from <https://www.psycharchives.org/static/about/terms_of_use.pdf>

*Quality and selection as well as technical guidelines.* (n.d.). Psycharchives.org. Retrieved February 10, 2021 from <https://www.psycharchives.org/static/about/technical_guidelines.pdf>

Other Resources

Curwen’s website: <https://carolinecurwen.com/>

Conferences page with link to downloadable word document of her abstract for the “Synaesthesia for reading and playing musical keys” paper: <https://carolinecurwen.com/conferences/>

Datasets and metadata on Figshare: <https://figshare.com/articles/dataset/Synaesthesia_for_written_musical_keys/13140086>

Dataset on DataCite:

<https://search.datacite.org/works/10.15131/shef.data.13140086.v2>

Fields of Research schema/vocabulary: <https://www.abs.gov.au/ausstats/abs@.nsf/0/6BB427AB9696C225CA2574180004463E>

ORDA data repository from the University of Sheffield: <https://orda.shef.ac.uk/>

SPSS .sav file to Excel .xlsx file online converter: <https://secure.ncounter.de/SpssConverter/Convert/SpssToExcel>

Information on JASP Stats – open source statistical tool that can read .sav files: <https://jasp-stats.org/>

Other publications by Caroline Curwen

Curwen, C. (2018). Music-colour synaesthesia: Concept, context and qualia. *Consciousness and Cognitions, 61,*94-106. <https://doi.org/10.1016/j.concog.2018.04.005>

Curwen, C. (2020). Music-colour synaesthesia: A sensorimotor account. *Musicae Scientiae*, 1-20. <https://doi.org/10.1177/1029864920956295>

1. Though the abstract discusses 18 participants, some of the data in the datasets has the participants having IDs that aren’t just 1-18, though all of those numbers are included in one data set or another; I’m not sure if participants got assigned other numbers during the study, or if there were more than 18 participants involved. [↑](#footnote-ref-1)
2. This is a bit confusing to me; from the abstract on Curwen’s website, they appear to have conducted 3 experiments. From the data and the abstract on Figshare, however, it looks like each file is saved as “Exp #” which goes up to 6, and the abstract states that they conducted 6 experiments. Though some of these might not be experiments, as they do contain ANOVA correlation results, so as far as my understanding goes, I believe there where 6 experiments. [↑](#footnote-ref-2)
3. I found this through a couple of sources, though it doesn’t say it outright on the English version of the PsychData site. For whatever reason, any of the actual data submission forms are only available on the German version of the site itself [on this page](https://www.psychdata.de/index.php?main=give&sub=uebergabe&lang=ger) (which does not appear to have an English equivalent; I relied on Google Translate and a friend who speaks German to help me translate this) and [from the Institute’s services page](https://leibniz-psychology.org/en/services/archiving/), they mention that The Research Data Center at ZPID (which is PsychData) will be integrated with PyschArchives. [↑](#footnote-ref-3)