Physical Sciences Data Infrastructure (PSDI)

Resurrecting Second Harmonic Generation (SHG) Case Study
PSDI Report

Author: Matthew Partridge, Stephen Gow, Don Cruickshank, Jack Doyle, Samantha Pearman-Kanza, Jeremy Frey

Report Date: 23/04/2025

PSDI-Case-Study-Series:Report_001

Publishing Information

Resurrecting Second Harmonic Generation (SHG) Case Study
PSDI-Case-Study-Series:Report_001
Report Date: 23/04/2025
DOI: 10.5258/SOTON/PSDI0001
Published by University of Southampton

Funding Information

Physical Sciences Data Infrastructure
PSDI acknowledges the funding support by the EPSRC grants EP/X032701/1, EP/X032663/1 and EP/W032252/1

Title: *Physical Sciences Data Infrastructure Phase 1b EP/X032701/1*
Principal Investigator: *Professor Simon Coles*
Other Investigator: *Professor Jeremy Frey*
Co-Investigators: *Dr Nicola Knight & Dr Samantha Kanza*

Title: *PSDI Phase 1b EP/X032663/1*
Principal Investigator: *Dr Juan Bicarregui*
Other Investigator: *Dr Brian Mathews, Dr Vasily Bunakov, Dr Barbara Montanari*
Co-Investigators: *Dr Abraham Nieva de la Hidalga*

Project Details

Project Name	Resurrecting Second Harmonic Generation (SHG) Case Study
Project Dates	01/06/2023-01/10/2024
Website	PSDI Website

Project Team

Author Name	Affiliation	ORCID
Matthew Partridge	University of Southampton	0000-0001-5280-8309
Stephen Gow	University of Southampton	0000-0003-0121-1697
Don Cruickshank	University of Southampton	0000-0002-0777-0855
Jack Doyle	University of Southampton	LinkedIn Profile
Samantha Pearman-Kanza	University of Southampton	0000-0002-4831-9489
Jeremy Frey	University of Southampton	0000-0003-0842-4302

Project Description

This case study was undertaken as part of Pathfinder 3 to demonstrate how value can be elicited from legacy datasets and to capture the lessons that we can learn from these exercises, both in terms of re-using these techniques on future datasets, but also to help us consider how to store share and archive our current datasets to protect against these issues in the future.

Project Data & Materials

All of the relevant code and database scripts to create new versions of the resurrected databases, and to run the tableau visualisation files can be found here: 10.5258/SOTON/PSDI0002.

Contents

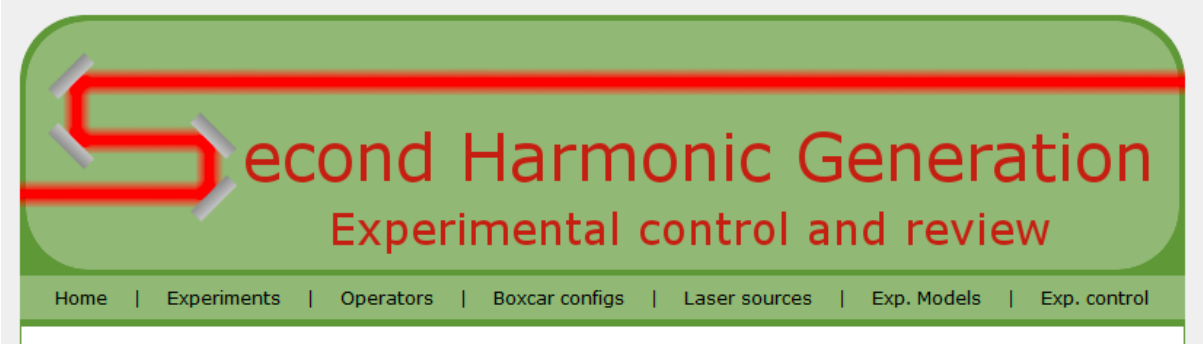

Figure 1: Original SHG Database header

1 Introduction

The preservation, accessibility, and reuse of scientific data are increasingly vital in contemporary research. Historical datasets, often termed legacy data, provide significant opportunities for new discoveries, validation of past results, and insights through modern analytical methods. Despite their potential value, legacy datasets frequently suffer from poor documentation, outdated formats, and technological obsolescence, complicating their reuse. This report addresses these challenges through the reconstruction of databases originating from Second Harmonic Generation (SHG) experiments, conducted as part of a doctoral research project completed in 2011 at the University of Southampton.

The original PhD project [1] extensively explored the properties of molecular orientations at liquid-air interfaces through SHG experiments. These experiments were particularly valuable in providing insights into molecular adsorption dynamics and interfacial properties, areas critical to advancing physical and analytical chemistry. A comprehensive database system and an associated graphical user interface (GUI) were initially created to manage, visualise, and analyse the experimental data. At the time, these were regarded as robust tools for data exploration and model validation, complete with a documented data structure and online visualization capabilities.

However, by 2023, significant obstacles to data reuse had emerged. Many details about experimental protocols and data curation methodologies had been lost due to personnel changes and technological advances. The original data, stored on legacy computing systems, were believed to be inaccessible or corrupted. Furthermore, the original GUI, developed with technology now largely obsolete, had ceased to function effectively. Consequently, the comprehensive dataset was at risk of permanent loss, depriving the research community of valuable data and insights.

The current reconstruction project, therefore, had two fundamental objectives: firstly, to recover and reconstruct the original databases to ensure their integrity and completeness, and secondly, to redevelop an interactive visualization interface to facilitate exploration and analysis of the experimental results. This effort aligns closely with the broader goals of the Physical Sciences Data Infrastructure (PSDI) initiative, which aims to preserve valuable scientific data for long-term usability and promote best practices in data management and reuse.

This report details the comprehensive process undertaken to revive the SHG databases, including data recovery from legacy storage, format translation, and schema reconstruction. It also covers the redevelopment of visualization tools using contemporary software, identifying significant issues encountered during the project and the solutions implemented. Finally, it evaluates the overall success of the project in achieving its goals and outlines recommendations for future activities to ensure ongoing access and utility of these important experimental results.

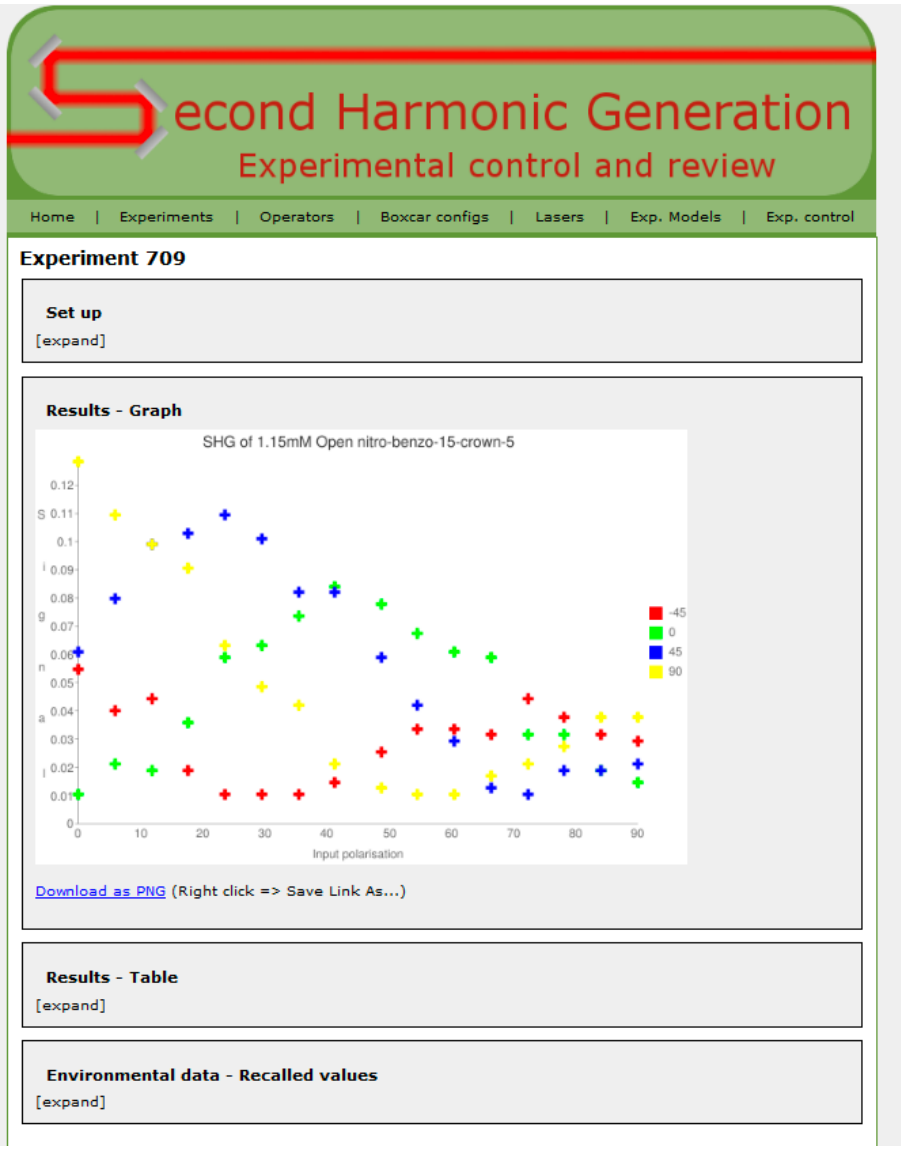

Figure 2: Original SHG polarisation plot visulisation

2 Methodology

2.1 Methodology intro

The original experimental data from the Second Harmonic Generation (SHG) experiments were stored on an array of RAID 5 disks, which formed the backup of a historic departmental computer. This data included comprehensive databases documenting experimental runs alongside environmental parameters such as temperature and humidity within the laboratory during the experiments. Files relating to data extraction, curation processes, and other miscellaneous information were also stored within this system. Later, additional critical data was retrieved from a DVD containing appendices from the original PhD thesis, primarily in Comma Separated Variable (CSV) and Microsoft Excel formats.

The RAID 5 data was initially extracted and provided as a zipped archive by the principal investigator involved in the original experiments. However, working with this archive posed immediate challenges due to inadequate file labelling and an unclear directory structure. Primarily, the data consisted of CSV dumps corresponding to each database table. These tables were originally stored using the now-deprecated MYISAM format, involving MYD files (data storage), MYI files (index files), and frm files (table properties) [2]. Due to format deprecation and compatibility issues with contemporary database management systems, the MYISAM tables could not be directly accessed or imported into modern MySQL environments [3].

2.2 Database Reconstruction and Curation

Given these incompatibilities, it was necessary to reconstruct the databases completely from the available CSV files. This process began with mining the original database schemas directly from the frm files using the Python-based utility mysqlfrm [4]. This tool efficiently recovered the original MySQL queries employed to create the database tables, simplifying reconstruction in most cases. However, a small number of tables required manual adjustments due to discrepancies and data-type incompatibilities with the current MySQL standards. Notably, timestamp entries previously represented as 00-00-00 00:00.00, which had been considered valid historically, were now unsupported and thus replaced by NULL values.

The schema of one critical database, named “shgdata” (figure 3), was initially documented in the original PhD thesis, though the actual database structure as extracted differed slightly from the documented version. The extracted version contained seven additional tables, while one table documented in the thesis schema was missing entirely. These discrepancies necessitated a careful comparison and verification process to ensure accurate reconstruction.

Validation of the reconstructed database involved meticulously plotting subsets of experimental data to generate graphs corresponding to original plots presented in the PhD thesis. The resulting plots matched precisely, thereby confirming the successful reconstruction of the database. Additional verification steps included cross-referencing data from the PhD thesis appendices, enabling the recovery and integration of previously missing data, particularly concerning sample concentrations and temperatures essential for meaningful analysis.

2.3 Visualization Interface Redevelopment

An important component of the original SHG data management infrastructure was an interactive graphical user interface (GUI), initially developed using web technologies including PHP and Google Charts. At the project’s inception, limited documentation and unavailable source code for this GUI significantly hindered the direct reproduction of the original interface.

Consequently, a decision was made to redevelop this interface using Tableau, a modern data visualization tool. Tableau was selected for its user-friendly capabilities and robust support for interactive visualizations, despite limitations in precisely replicating the original GUI functionality (figure 4). The redeveloped interface successfully restored the essential functionalities,

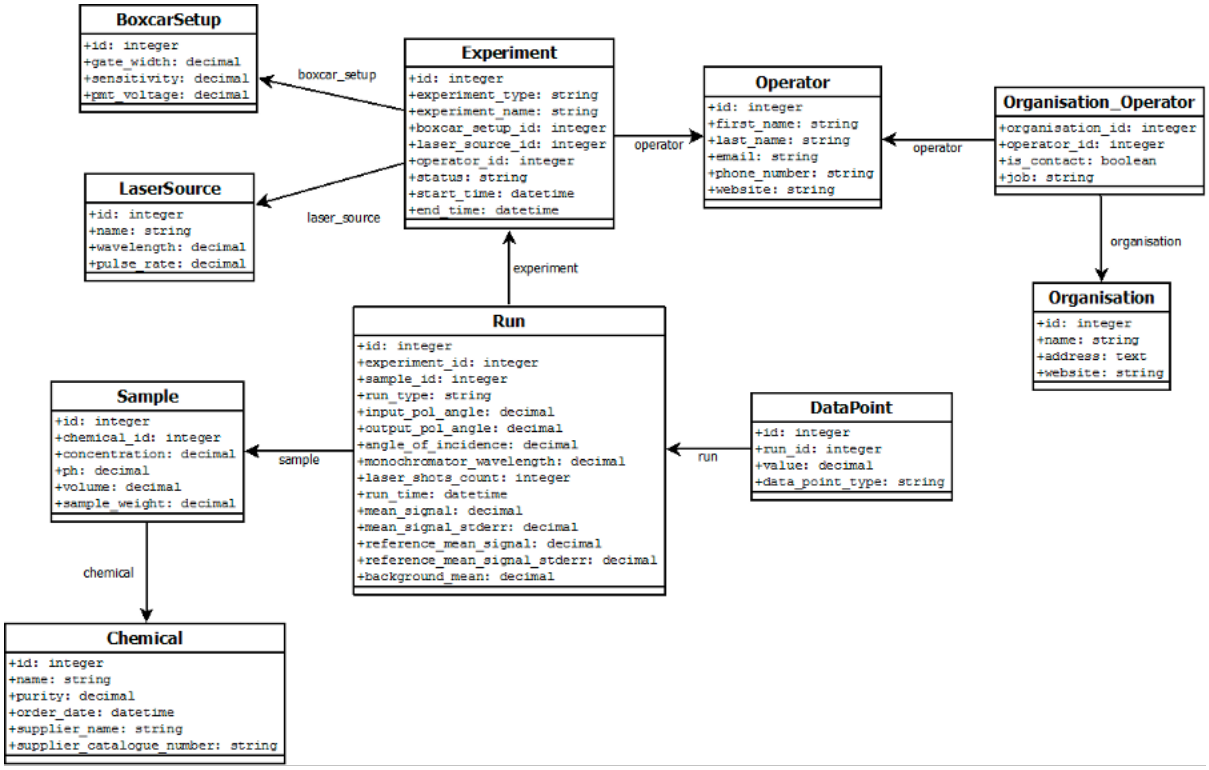

Figure 3: Schema for the "shgdata" database

including interactive timelines for experiment runs, detailed polarisation plots, and real-time environmental data tracking such as laboratory temperature and humidity. The new dashboard offered intuitive interactivity, allowing users to navigate seamlessly between different experimental datasets and visualizations, significantly enhancing the data exploration experience compared to the original system.

The process involved initially working with database extracts to manage performance issues arising from handling around 20 million individual data points. After successful prototyping and testing, the dashboard was extended to the complete dataset, ensuring both performance and comprehensive coverage of the experimental data.

3 Results

The reconstruction of the SHG databases was successfully completed, overcoming significant initial challenges related to format incompatibilities and incomplete documentation. The databases were rebuilt entirely from the original CSV dumps extracted from the RAID 5 storage array. Schemas recovered from the original .frm files via the Python-based mysqlfrm tool provided a reliable blueprint for this reconstruction, with minimal manual intervention needed for data-type adjustments. Particularly critical was addressing outdated timestamp formats, replacing invalid entries with NULL values for database integrity.

Verification of the database reconstruction process was accomplished by comparing newly generated plots to the original graphs presented in the PhD thesis. These comparisons yielded perfect matches, conclusively validating the successful recreation of the database. Notably, this verification step also identified gaps in the data, primarily regarding concentration and temperature information crucial for interpreting two specific experiment types: concentration sweeps and isotherms at multiple polarisations.

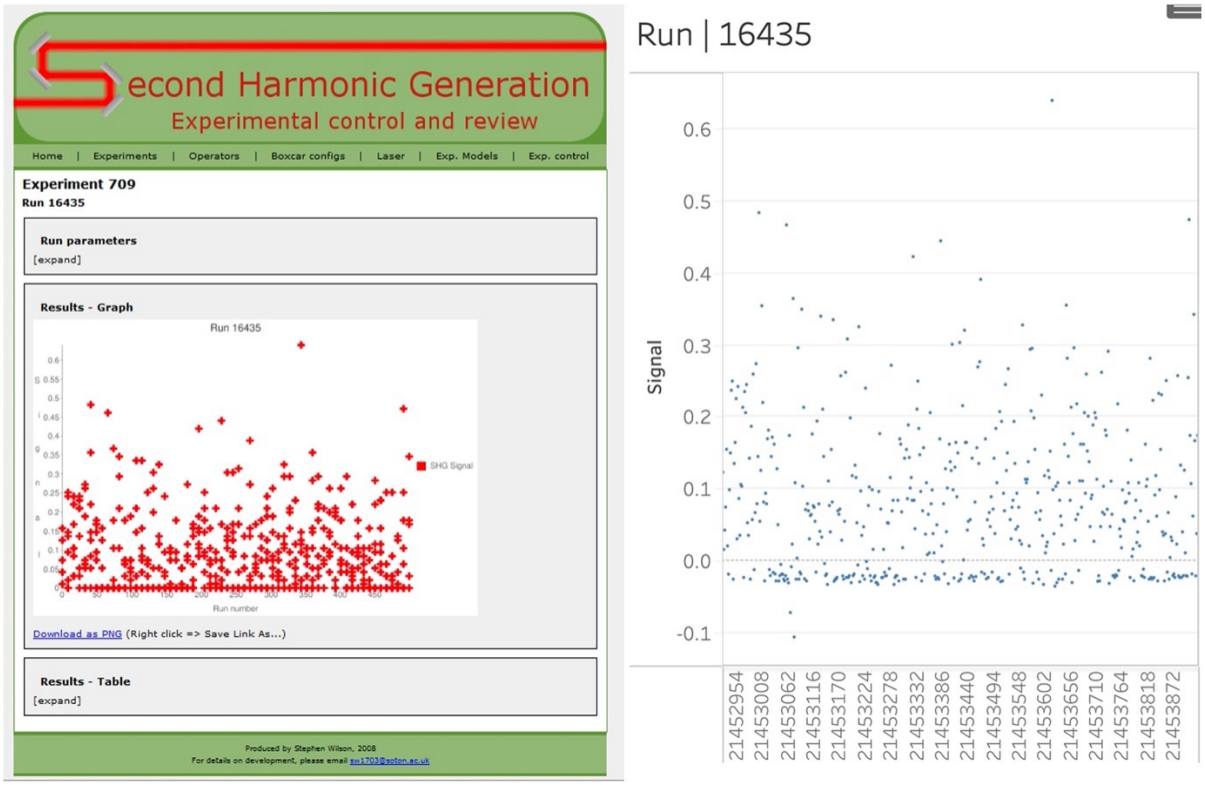

Figure 4: Results output in original GUI (left) and Tableau (right)

3.1 Identification and Integration of Missing Data

Upon further examination, significant gaps were confirmed concerning the sample concentration and temperature data, elements essential for the comprehensive interpretation and analysis of the SHG experimental data. Initially, it was unclear whether these gaps were due to improper data extraction or whether the original dataset was incomplete. Cross-referencing the reconstructed databases against supplementary data available from the PhD thesis appendices clarified this issue, indicating that the concentration data had never been fully integrated into the original databases.

A portion of this previously missing data was successfully recovered from the appendices, accounting for roughly 40% of the necessary concentration information. This data, although incomplete, was crucial in providing a partial restoration of the experimental dataset. During this recovery process, it was discovered that the database's intended structure—specifically regarding unique identifiers (sample IDs)—was not consistently followed in the original data management, complicating the integration effort. To address this, a revised database schema was employed to accurately map concentration values to experimental runs.

3.2 Redevelopment and Enhancement of Visualization Interface

The visualization interface was successfully redeveloped using Tableau software, effectively replicating and enhancing key functionalities from the original GUI. Four distinct visualization types from the original GUI were recreated:

1. Experiment Timelines: The timeline visualisation, although slightly altered from its original horizontal-bar format due to software limitations, effectively conveys the chronological progression of experiments using discrete points for start and end times (Figure 5).

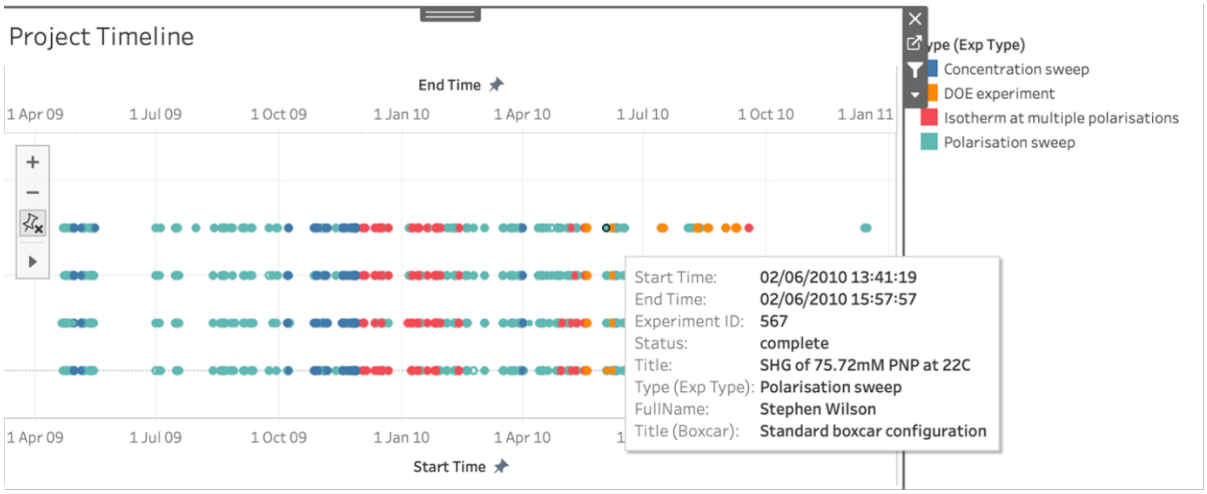

Figure 5: Project timelines in Tableau

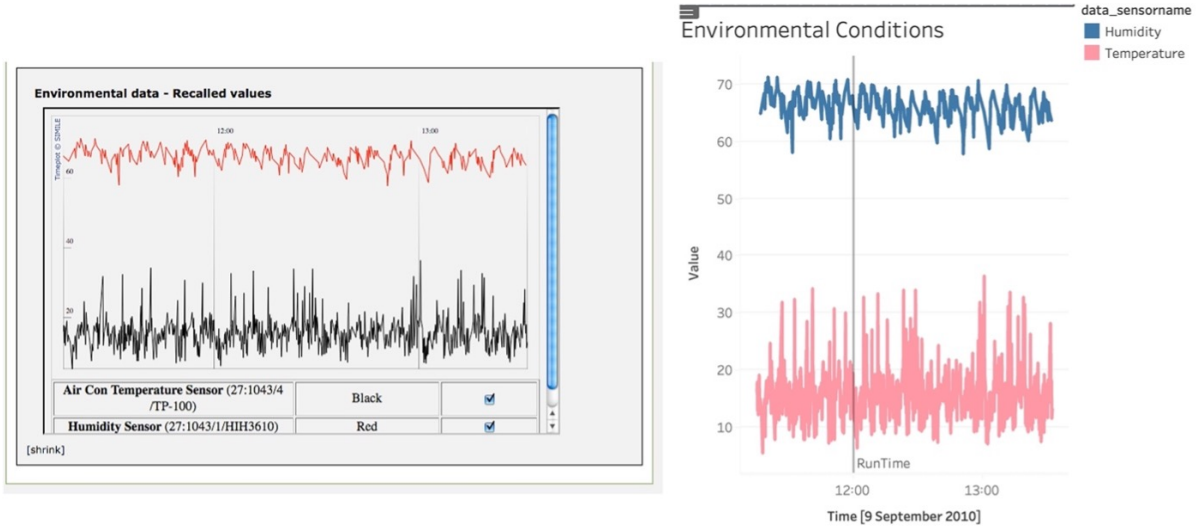

Figure 6: Temperature and humidity data in original GUI (left) and in Tableau (right)

2. Polarisation Plots: These were replicated nearly identically to the originals, featuring interactive capabilities allowing exploration of detailed experimental data points and metadata. The visualization preserved the original polarisation angle conventions used in the previous GUI.

3. Individual Experimental Runs: This feature displayed fine-grained data points within individual experimental runs. Notably, the new visualisation preserved negative signal values previously truncated to zero in the original GUI, maintaining greater transparency and accuracy of the raw data.

4. Temperature and Humidity Data: Environmental conditions throughout experiments were visualized similarly to the original interface, with enhancements such as interactive linkage to experimental run timestamps, thereby improving contextual understanding (Figure 6).

Interactivity among these visualisations significantly enhanced usability, allowing intuitive navigation between experiments, runs, and associated environmental data. An initial limitation—difficulty in handling the extensive dataset—was resolved by initially developing the Tableau dashboard using smaller database extracts before scaling to encompass the complete

dataset of approximately 20 million individual data points.

3.3 Original Interface Recovery and Legacy Issues

Unexpectedly, an intact copy of the original website interface was subsequently recovered from decommissioned virtual machine backups. Although this provided valuable reference points, substantial challenges arose when attempting to restore and share the original PHP-based interface due to incompatibilities with modern computing environments and discontinued external services such as Google Charts. Efforts to resolve these issues included recreating historical computing environments via virtual machines, highlighting broader difficulties inherent in digital preservation of legacy software systems.

Despite successful internal restoration, cross-platform compatibility issues hindered sharing this resource effectively across different operating systems, reinforcing the necessity and value of the new Tableau-based interface developed during this project.

4 Conclusions & Future Work

This project successfully reconstructed and modernised the legacy database from Second Harmonic Generation (SHG) experiments initially carried out in 2011, overcoming significant data compatibility challenges and filling critical gaps in metadata. By leveraging contemporary database management tools and Tableau-based interactive visualisations, the resulting infrastructure offers vastly improved accessibility, usability, and analytical capability compared to the original system.

The newly reconstructed database not only facilitates straightforward replication of previous analyses but also significantly enhances opportunities for deeper exploration of SHG data. Specifically, the inclusion and clearer presentation of experimental run timestamps, concentrations, and temperature conditions enable sophisticated time-based analyses previously unattainable. This advancement lays a critical foundation for future incorporation of expanded datasets, particularly time-domain SHG measurements, which require robust and precise temporal metadata integration for meaningful interpretation.

Moreover, the modern infrastructure implemented here provides an adaptable framework suitable for ongoing dataset expansion. Future research can easily incorporate new SHG datasets, benefiting from a structured and scalable database environment. It also allows researchers to apply advanced data analysis techniques, such as temporal correlation analyses and machine learning algorithms, thereby unlocking new scientific insights into molecular dynamics and surface interactions.

Moving forward, further efforts should include integrating additional recovered and future-generated SHG datasets, particularly focusing on comprehensive documentation and metadata standardisation. The methodologies and insights gained from this project offer a robust template for similar efforts to revitalise historical scientific datasets, significantly enhancing their analytical and research value.

5 Outputs, Data & Software Links

The following files have been made available via our SHG Dataset Record.

- new_shgdatabase.sql - This script will create the required databases
- shg_dashboard.twb - This is the tableau file to visualise the data in the SHG databases created by the script above

Please note that due to the size of the database scripts, these are stored on the University of Southampton file servers and can be requested via this form.

References

[1] Wilson SM. The SmartLab: experimental and environmental control and monitoring of the chemistry laboratory - ePrints Soton. Sotonacuk. 2011 06. Available from: https://eprints.soton.ac.uk/192833/.

[2] Agrawal V. MyISAM vs InnoDB: 7 Critical Differences. Hevo Data; 2021. Available from: https://hevodata.com/learn/myisam-vs-innodb/.

[3] MySQL Utilities. MySql; 2018. Available from: https://downloads.mysql.com/docs/mysql-utilities-1.5-en.pdf.

[4] MyISAM Overview; 2024. Available from: https://mariadb.com/kb/en/myisam-overview/.