A Platform for the Analysis of Qualitative and Quantitative Data about the Built Environment and its Users

Mike Simpson
Digital Institute
Newcastle University
Newcastle-upon-Tyne, UK
mike.simpson@ncl.ac.uk

Simon Woodman
Digital Institute
Newcastle University
Newcastle-upon-Tyne, UK
simon.woodman@ncl.ac.uk

Hugo Hiden
Digital Institute
Newcastle University
Newcastle-upon-Tyne, UK
hugo.hiden@ncl.ac.uk

Sebastian Stein
School of Computing Science
University of Glasgow
Glasgow, UK
sebastian.stein@glasgow.ac.uk

Stephen Dowsland
Digital Institute
Newcastle University
Newcastle-upon-Tyne, UK
stephen.dowsland@ncl.ac.uk

Mark Turner
Digital Institute
Newcastle University
Newcastle-upon-Tyne, UK
mark.turner@ncl.ac.uk

Vicki L. Hanson
Rochester Institute of Technology
Rochester, NY, USA
vlhics@rit.edu

Paul Watson
Digital Institute
Newcastle University
Newcastle-upon-Tyne, UK
paul.watson@ncl.ac.uk

***Abstract* - There are many scenarios in which it is necessary to collect data from multiple sources in order to evaluate a system, including the collection of both quantitative data - from sensors and smart devices - and qualitative data - such as observations and interview results. However, there are currently very few systems that enable both of these data types to be combined in such a way that they can be analysed side-by-side.**

This paper describes an end-to-end system for the collection, analysis, storage and visualisation of qualitative and quantitative data, developed using the e-Science Central cloud analytics platform. We describe the experience of developing the system, based on a case study that involved collecting data about the built environment and its users. In this case study, data is collected from older adults living in residential care. Sensors were placed throughout the care home and smart devices were issued to the residents. This sensor data is uploaded to the analytics platform and the processed results are stored in a data warehouse, where it is integrated with qualitative data collected by healthcare and architecture researchers. Visualisations are also presented which were intended to allow the data to be explored and for potential correlations between the quantitative and qualitative data to be investigated.

Keywords - built environment, data analytics, cloud computing, data visualisation, qualitative data, quantitative data

I. INTRODUCTION

Data collection is an important part of many projects, whether it be in research or in the management and evaluation of systems in a range of industries and disciplines. These projects often involve the collection of quantitative sensor data as well as the collection of qualitative data (such as observations about the system made by specialists). Currently, there are very few systems that are capable of simultaneously processing these disparate types of data, or which are capable of visualising/reporting on these data types alongside one another. Doing so would allow for correlations to be investigated and would provide greater insight into any observed patterns or behaviours in the data than would be possible using either data type in isolation.

One such problem area is the evaluation of the built environment - and how the design and management of the built environment affect the performance of the building and/or the well-being of its users. There are many reasons to study the built environment. For example, building managers may be interested in how staff respond during a fire drill, healthcare professionals may wish to monitor the welfare of patients in hospitals, and managers may be interested in studying the behaviour of users in schools, shopping centres, airports etc., in order to optimise the well-being of the users and/or the efficiency of the building.

In order to evaluate a built environment and its users, it is necessary to overcome two problems: data heterogeneity and data volume. The data required is necessarily heterogeneous as it includes both quantitative and qualitative data. Quantitative data includes data about the user's location within the environment, as well as data such as their activity levels. It may also include environmental factors (such as temperature, light levels, ambient noise etc.). Qualitative data may take the form of observations made by experts, such as built environment or healthcare professionals, and is less likely to be structured or machine-readable, but may provide additional context or understanding to the quantitative data.

In order to differentiate actual trends from 'noise' in the data, a significant amount of data needs to be collected. It is necessary to collect data from many users over many days (or potentially years, in order to adjust for seasonal averages etc.). Some resources, such as the UK Biobank [1], are based on the collection of data from thousands, or even hundreds of thousands, of users over several years. Additionally, many algorithms, particularly those used in physical activity analysis, require high-frequency data (100Hz+) [2].

It is often impractical to perform the storage and processing of data from studies that include this kind of physical activity analysis of multiple participants over long periods of time on a single machine, but a cloud-based platform may provide a solution to this problem.

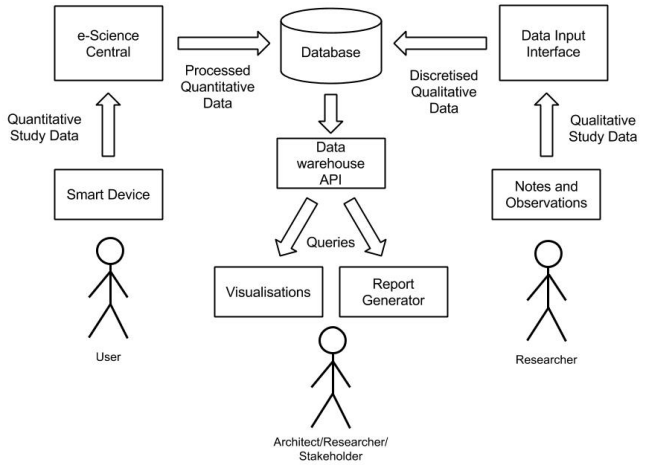

Fig. 1. BESiDE Data Flow Overview

This paper presents a system for the storage, processing and analysis of qualitative and quantitative data. This system solves the first of the problems described above through the automated collection and processing of sensor data, and the ability to discretise qualitative data and add it to the same database.

The second problem is addressed by using cloud-based data analytics and by unifying blob stores for large volumes of sensor data with non-relational databases to hold derived and processed data for further analysis and visualisation. As the platform utilises cloud computing, it is able to scale as required in order to meet demand. The system uses e-Science Central (e-SC) [3], a scalable, open-source, cloud-based platform for data analytics, which has been shown to work well for chemical modelling, stroke rehabilitation and health informatics (as discussed in Section II). The work presented in this paper has extended e-SC with connectors for accessing the data warehouse and with dashboard applications to visualise and derive value from the study data.

As a case study, we demonstrate how the system was used in the processing of data collected as part of the BESiDE [4] project. In this study, beacons and smart devices were used to monitor the movements and well-being of older adults living in residential care. Quantitative data was recorded using an Android application, which uploads the data to the analytics platform, where it triggers a series of analysis workflows. Additionally, qualitative data - in the form of interviews and observations made in the care home by architecture and healthcare researchers - was discretised and uploaded to the data warehouse. A series of visualisation tools are then able to query the database and produce a range of different views on the data. These tools allow researchers to explore the data and to investigate potential correlations between data the sensor data and the qualitative observations.

The focus of this paper is on the development of technologies that allow researchers in fields such as architecture to answer questions using a combination of qualitative and quantitative data. The findings of the BESiDE project from a built environment perspective are being prepared for submission to the relevant built environment publications.

The main contributions of this paper are:

- An extensible cloud-based platform for the automated collection, analysis and storage of quantitative sensor data, which enables the integration of qualitative data.
- Multiple visualisations tools designed to explore the data, including visualising the qualitative and quantitative data side-by-side.
- A discussion of the experience of developing the tool and recommendations for future research and development.

II. Related Work

A range of commercial products exist for fitness tracking and indoor location tracking, including GENEActiv [5], but the data from these products in isolation is insufficient for this kind of project as, for example, they do not allow us to investigate the relationship between physical activity levels and particular locations or features of the building. Commercial interfaces are also often designed with a narrow focus and many APIs are either closed or restricted, which hinders the automated collection and analysis of data from multiple devices simultaneously. An Android Wear smartphone app was therefore developed, as discussed in Section IV.

Tweet My Street [6] was a multi-disciplinary research project in which we used qualitative and quantitative data analysis to understand how people use social media. [7] Combining qualitative and quantitative data can be challenging, but several projects have collected both data types as part of a 'mixed methodology' study. Processing the qualitative data requires a process of discretisation (or 'quantising') of the data in order for it to be processed and stored by a machine. Many qualitative researchers will state that this process results in a loss of depth and flexibility, but it can make the process of analysing the data, particularly from a statistical standpoint, much easier [8]. Although different groups of BESiDE researchers were collecting both qualitative and quantitative data, as computer scientists designing the data analytics platform we decided to use a Priority-Sequence model, selecting the quantitative approach as our principal method, with the qualitative approach as a complementary method [9]. We did not initially 'interfere' with the collection of the qualitative data, but collaborated with the researchers after it had been collected, as discussed in Sections IV.C.5) and VI.

The e-Science Central platform has been successfully used in a number of projects ([10], [11], [12]), including Limbs Alive [13], which was a video game designed by clinicians to help with patient rehabilitation following a stroke. The game gathers therapeutic data from patients playing the game and uses e-SC to analyse the data. These results are then presented to patients and clinicians via web applications.

In other fields, projects such as DAME [14] gather data from aircraft engines via remote links and present this data to technicians via a Signal Data Explorer [15], enabling pattern recognition tools to be used to diagnose unusual engine operation. Both of these are examples of projects that involve automated remote data collection and cloud-based data processing, coupled with centralised expert analysis. This suggests that the e-SC platform is ideal for the automated collection of the quantitative sensor data that was collected during the BESiDE Project. The e-SC platform was also chosen because it is easily extensible, supports rapid setup and reconfiguration of the analysis pipeline through a graphical user interface, as well as for its scalability and ease of use for non-technical domain experts - such as care home stakeholders and researchers in fields other than computing. The use of e-SC minimises the amount of custom code necessary to implement the pipeline, while allowing custom code written in a variety of programming languages to be integrated into the workflows where necessary. e-SC is capable of being installed on a number of cloud services, including Azure and Amazon, and multiple machines can process the data in parallel in order to speed up the process. Although e-SC has not previously been used to process qualitative data, a number of extensions have been made to the system that allow qualitative data to be processed and stored in the same database as the quantitative data.

III. System Requirements

The design of the system presented in this paper was based on analysing three key requirements: the structure and semantics of the data to be stored; the processing requirements, including timeliness, granularity and scale; and the requirements of the visualisations - interactivity, context and accessibility to non-domain experts. Whilst the system was required to be able to process both quantitative and qualitative data, these data types share some characteristics when concerned with the built environment; the data is often concerned with the intersection of the contexts People, Time and Place, as illustrated in Fig. 2. That is, when the disparate streams of quantitative data are combined, they are able to locate a person at a point in time (or for a period of time) in a specific location.

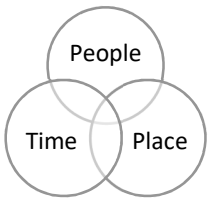

Fig. 2. Visualisation of the built environment data contexts; People, Place and Time.

For example, Paul went from the corridor to the common room at 10:32 am. Not all items of data will be at the intersection of the three contexts, some will only intersect two or three. This is especially true of the qualitative data where observations are made which only intersect two contexts, for example; there is a bingo competition in the common room at 10:30 am. Paul's presence in the common room at a similar time does not itself indicate participation in the bingo competition (but another qualitative observation might). One of the goals for the qualitative data analytics pipeline is to allow the data point to intersect with as many of these contexts as possible, usually through the integration of the location data with the metric being measured. Essentially, the environment should be monitored, the people taking part should be monitored and insights can be derived from analysing the intersection of those observations.

Quantitative data can be collected from sensors placed throughout the environment and from smart devices issued to the users. The system should be capable of collecting this data automatically and processing it in a timely manner. There is not a requirement to be able to do this in real time - but it should be as automated as possible and resilient to network partitions (i.e. the data may need to be stored locally on a gateway that is not connected to the Internet for security, ethical or performance reasons). In order to deal with large studies, the processing requirements are that the computational capacity can scale, in order to process the data within a reasonable amount of time. An equal requirement is flexibility - different researchers have experience in a variety of technologies from MatLab/R to Excel/Tableau. The flexibility is not limited to the analytics technology of choice - in order to maximise the value from studies which use such a platform, it should be possible to integrate new analytics and automatically apply them to historical data. This allows new ideas to be developed, tested and evaluated throughout a project rather than simply prior to the data being collected.

Qualitative data may be used to add missing context to the data, i.e. an observer may note that Paul is taking part in the bingo competition. However, qualitative data is typically considered to be more difficult to store and process in an automated fashion than quantitative data. Qualitative data is less likely to be structured and, in some cases, may not be machine-readable. Images and drawings may illustrate observations, particularly around the 'Place' context, that text cannot. Further, the qualitative data is less likely to be limited to a precise point in time, whereas much of the quantitative data is. For example, an observation may well be valid even if it has an approximate timestamp; it may explicitly note recurring events; it may observe the lack of an event.

Outputs from the processing components will include derived data, typically numerical values and aggregation reports. However, the final output and consumer of these derived values will be visualisations. Such visualisations are used to illustrate variations in the data, which may be hard to see numerically. Many of these will be straightforward charts that are common to any discipline, but a number will be required to be bespoke and interactive. Such 'active' visualisations allow the end user to explore the data in ways that are harder with 'static' reports. That is not to say that the latter does not have a part to play - particularly in studies with a medical context, reports are necessary in order to maintain archives of patient/participant data. However, active charts enable researchers to explore multidimensional data in an interactive manner - for example, by turning a data series on/off, exploring different chart types, removing outlying data or limiting the time frame displayed. For users with a high level of knowledge of the domain, this can be extremely powerful, and can enable research questions to be answered in a 'self-serve' manner without the platform developers supplying every feasible chart.

IV. Case Study - BESiDE

The goal of the BESiDE [4] project is the understanding of how the built environment can facilitate physical and emotional well-being in residential care homes. Currently, there is a lack of objective evidence about how existing care homes are utilised and how building design impacts well-being. Stakeholders, including architects and care home managers, are interested ways to evaluate care home facilities. It is hoped that the data collected by BESiDE will provide architects and stakeholders with information that can be used to improve the design and management of existing and future care homes [16].

The previous section outlined the requirements for a general-purpose system that could store, process and visualise qualitative and quantitative data about the built environment. Within the BESiDE project, some pragmatic decisions were made based on past experience, specific project requirements and time/budget constraints.

The granularity of location within a residential care home was defined to be a 'Zone' and identifiers were assigned for each Zone. In some cases, there is a 1:1 mapping between room and Zone, whereas in other cases there is a 1:N mapping with a larger room being broken up into multiple Zones. In some cases, this is because different areas of the room have different purposes and in others, this is because of a physical change in the environment, such as a change in flooring type. All data was associated with at least one Zone ID. For quantitative data, this was based on the algorithm described in subsection D, whereas for qualitative data this was assigned by the person making the observation. The unification of the qualitative and quantitative data in this manner allows comparisons to be made and enables the visualisations described in subsection E.

Not only are consistent zone identifiers necessary for aligning disparate data streams in this context, but so is a consistent representation of time. Owing to the fact that some of the observations (qualitative data) are being made manually, it was decided to standardise on the local time of observation rather than UTC. This minimises the likelihood of recording mistakes and removes the need to correct for daylight savings time during the data analysis. Given that the subjects of the study are people, time that the residents perceive it to be it is likely to be of more interest than an absolute value. Speaking of which, it is also necessary to identify the participating residents in a consistent way, in order to enable the union of all contexts.

In order to store both the quantitative and qualitative data, and to limit the amount of management and configuration needed to use the database, a combination of Amazon S3 [17] and MongoDB [18] was chosen. These technologies were already supported by the e-SC platform, which allowed us to develop the pipeline without requiring any additional implementation. The raw sensor values are stored in Amazon S3 and then processed into derived values which are stored in MongoDB. MongoDB is a document-oriented, non-relational database, which stores the transactional, non-blob data. This architecture makes it straightforward to reprocess the raw sensor data into new MongoDB collections if new analytical techniques are developed.

A number of devices were evaluated in order to collect the necessary quantitative data. Commercial activity trackers were considered unsuitable due to their closed-source algorithms, potential inaccuracy and closed ecosystems (as described in Section II). Because of their 'programmability' and form factor, Android smartwatches were selected for these studies, which enabled all of the quantitative data streams to be collected using a single device and then uploaded to e-Science Central for automated processing. Subsection D describes the e-Science Central workflows, which were developed to process the quantitative data - these can be automatically triggered by the arrival of new data via the e-Science Central REST API.

Both active web-based visualisations and static reports were developed for the BESiDE project, the former of which allow domain specialists to explore the data, and to view the qualitative and quantitative data side-by-side where possible (these are discussed in Subsection E).

A. *Configuring the Sensors/Smart Devices*

This Section describes how the sensors and smart devices were configured in order to instrument the care home, to detect interactions with staff and visitors, and to track resident activity levels.

1) Care Home Infrastructure

In order to track a resident's position, it was necessary to place a number of Bluetooth Low Energy (BLE) transmitter beacons in the public areas of the care home. These devices also include temperature sensors, which allows us to monitor the temperature within the instrumented areas.

As mentioned in Section I, it is not possible to include the floor plans of any of the care homes that participated in the study due to ethical and commercial concerns. However, Fig. 3 shows an approximation of an area within one of the care homes.

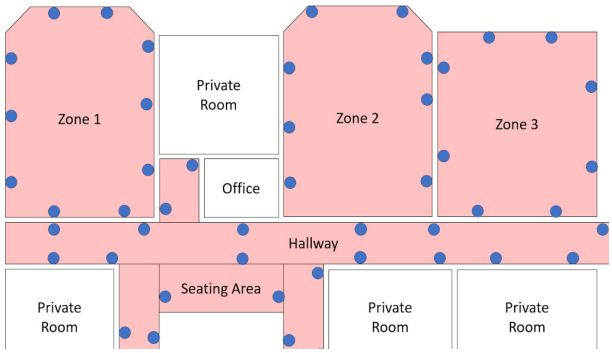

Fig. 3. Floor Plan showing beacon (blue dots) placement in the public areas (pink) of the care home

The instrumented (public) areas are shown in pink and the blue dots represent the approximate placement of the BLE beacons. Private rooms and off limits areas (white), such as the office, were excluded from the study.

The BLE beacons periodically broadcast data packets containing information about each beacon's identity, along with some auxiliary data. The power present in the signal at the receiver, called the Receiver Signal Strength Indication (RSSI), indicates the distance between the receiver and a particular beacon, and can be used to estimate the location of the receiver within an instrumented zone, provided that the receiver can detect signals from at least three beacons.

For this project, Estimote [19] Beacons were chosen due to their low cost, physical robustness and extensive configurability. The beacons were placed along the walls in public areas at a height of approximately 2.2m and spaced between 2-3m apart. The localisation algorithm requires a calibration (or "fingerprinting") phase to generate a map of the environment. Calibration data is collected at locations approximately 1m apart on in a regular grid in all instrumented spaces. At each location, BLE packets were recorded using a smartwatch worn by a researcher, who faced four perpendicular directions at each location for a period of 30 seconds. Modelling these variations requires additional work but provides a more accurate starting point for measurement than "on the fly" calibration methods and increases the precision of the localisation algorithm considerably.

2) Zone Definitions

The zones within the care home space are defined in JSON files, which stored in e-Science Central. As discussed earlier in this section, Zones could represent an entire room within the home, or a larger number of smaller spaces, including spaces that were identified as being 'of interest' by the researchers. For example, a large common room (such as Zone 1 in Fig. 3) could be broken down into multiple zones (i.e. 'TV area', 'socialising space' etc.).

3) Staff and Visitors

In order to track social interactions between the residents and care home staff/visitors, it is necessary to be able to identify these individuals and to detect when they come into proximity with the residents. A system was devised that used Estimote 'Nearables' (BLE beacons with a very small and flat form factor), which were made available to care home staff and visitors. The RSSI of packets from the Nearables are used to detect proximity between the person carrying the device and the participating residents, and extended periods of close proximity are considered to be a proxy of social interaction.

In order to protect the anonymity of staff and visitors, they select Nearables from a basket, rather than a specific Nearable device being assigned to a specific individual, and the unique identifiers of the Nearable devices are simply mapped to the roles 'Staff' or 'Visitor' when they are stored in the database.

4) Residents

In order to track residents' movements and activity levels, it was necessary to issue them with a smart device. These devices have Bluetooth receivers to detect the signals from the BLE beacons, accelerometers for activity tracking and microphones to measure the ambient noise levels (in decibels).

LG G smartwatches were used to capture the quantitative data streams for the BESiDE project, as this model had (at the time of procurement) the largest battery capacity, comparatively low weight, and sufficient disk capacity to store the study data until it could be uploaded to e-Science Central. The watches were modified such that they could be clipped onto the residents' clothing on the upper body, in order to improve the accuracy of the indoor location tracking. Residents were then asked to wear a smartwatch during waking hours for the duration of the study.

5) Obtaining Consent

In consultation with care home managers, a subset of residents was identified that were likely to be willing and able to take part in the study. These residents were initially approached by a researcher to informally establish a trust relationship and to assess their ability to give informed consent. Subsequently, the residents were invited to take part in the study, or their legal guardian was contacted to discuss giving consent on their behalf. This work received Ethical Approval from Dundee University.

B. Sensor Data Recording and Transmission

A mobile application based on Android Wear was developed for the project. The app supports recording of BLE packets, data from accelerometers, gyroscopes, magnetometers and the onboard microphone. It also supports manual event timestamping, which is useful for recording fingerprinting data (as discussed in Section A.1)). The application's user interface was also designed to allow the researcher to set study metadata (start time, end time, home ID, resident ID) before the device is issued to the resident.

Developing software for devices with limited processing and power resources requires careful consideration of threading and disk access. In order to ensure that the user interface remains responsive while sensor data is being recorded, all parts of the software that process sensor data are executed in background threads; one background thread for each sensor to avoid deadlocks. The data received from each sensor is buffered in memory and only sporadically written to disk, which also helps to avoid deadlocks and is more power efficient.

The following sensor data is captured:

- Accelerometer Data - (timestamp, a_x, a_y, a_z, sensor accuracy)
- Ambient Noise - (timestamp, loudness)
- Battery Data - (timestamp, battery level, temperature, battery health)
- BLE Data - (timestamp, MAC address, RSSI, device_name, TX_power, device_id)
- Event Logger Data - (timestamp, event ID)
- Magnetometer Data - (timestamp, m_x, m_y, m_z, sensor accuracy)
- Study Metadata - (timestamp, logging_data (true/false), resident_ID, home_ID)
- Temperature Data (from beacons) - (timestamp, MAC address, temperature, battery_voltage)

The data is stored as CSV files, which enables easy processing by e-Science Central workflow blocks and other tools. Files are stored in one folder per day for easy retrieval. When the researcher requests to transmit data, all data folders are compressed into ZIP archives, which are subsequently compressed and stored in a single 'master' ZIP archive, in order to minimise the amount of data that is transmitted and to minimise the risk of transmitting partial data.

An entire day's worth of data for a single user consists of:

- >100mb of Compressed (Zip) Data
- >600mb of Uncompressed CSV data

While the app was designed to be power efficient, only ~4 hours of data could be recorded on a single charge, requiring the devices to be swapped and recharged twice daily.

C. Qualitative Data

BESiDE researchers from other fields (such as architecture and healthcare) collected data in the care homes. The data was collected based on the AEIOU Framework [20], which focusses on Activities, Environment, Interactions, Objects and Users. This framework uses ethnographic methods (notes, photos, videos, interviews, etc.) as a 'lens to observe the environment.' In the case of BESiDE, this mainly included interviews with the staff/residents and observations about each home's layout and the behaviour of the residents. This data mostly took the form of handwritten notes, though some of the interview answers and observations were collected digitally.

After the initial studies were complete, we collaborated with the researchers to discuss the data that had been collected, and how it might be discretised in order to allow it to be stored in the database and be evaluated alongside the quantitative sensor data. Below are three examples of different types of qualitative data that were identified and added to the system.

1) Zone Observation Data

The researchers made a number of observations during their visits to the care homes, including observations about the residents, such as "residents were observed to complain about the temperature in <Zone 2>", as well as notes about the building architecture, such as "the window in the south of <hallway> has a view of the nearby countryside and residents were observed stopping there to admire the view". Many of these observations were associated with a specific zone within the care home and/or with a specific time (or time period), and it is these observations that were selected as being the best suited for discretisation (as they have data fields that correlate with the quantitative data in some way). Samples of these observations were discretised and stored in the database to enrich the behaviours observed in the quantitative data.

Zone observation data is stored in the following format:

- Time, Home ID, Zone ID, Observation.

2) Event Data

Events refer to organised activities that take place in the care home. For example, the home might host a bingo competition in a particular room at a particular time, which may be a one-off or recurring event. An 'Event' can also be used to refer to activities such as building/renovation work, visiting hours and open days. Including this data provides possible explanations for behaviours observed in the sensor data, such as why a particular room might be being highly utilised at a particular time.

A sentence describing the event was stored, along with the associated zone ID, the date and time of the event, and whether it was a recurring or one-off event. 'Event' data is stored in with the following fields:

- Date, Start Time, End Time, Home ID, Zone ID, Description, Recurring Event (true/false).

3) Interview Feedback

Another example of the qualitative data is feedback from interviews with the staff and residents.

We included mainly questions with discrete answers and these were broken into individual questions and their responses. These included questions such as "how would you rate the temperature/air quality/lighting in <this zone> (out of five)". Questions were chosen that related to specific zones or times/time periods, in order to allow the data to be unified with the sensor data.

In this case, the average score is stored (after the results had been manually processed to calculate the average scores), along with the question and the zone ID (and timestamp, where applicable).

The following fields are stored in the database:

- Home ID, Zone ID, Interview Question, Result.

4) Storing the Qualitative Data

For the initial implementation of the system (and initial development of the visualisation tools), samples of the data were processed manually and discretised so that they could be added to the database. The data was input by selecting samples of the researcher's' notes that could be easily discretised and storing them in CSV files. A script was then written to inserted the data from the CSV files into the same database as the quantitative data.

5) Improving the Collection of Qualitative Data

Once the data types had been identified, an attempt was made to improve the collection of such data in the future. A series of web pages, along with a mobile app designed to run on a tablet PC, were also designed to enable faster and more automated collection and processing of the data, as well as ensuring that the necessary ancillary data (i.e. Zone IDs) is collected (see 'Future Work').

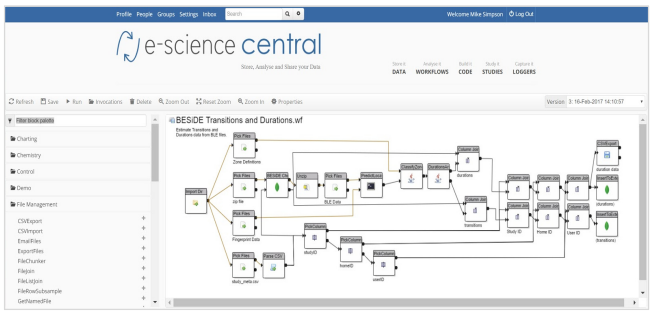

Fig. 4. The Location Data workflow and the e-Science Central interface

D. Data Analysis and Storage

The data is processed and analysed using the e-Science Central platform. The data from the smart devices is uploaded via an API, which allows the automatic triggering of the data processing workflows. Users can design Workflows to analyse the data by adding and connecting Workflow Blocks via a graphical user interface (as shown in Fig. 4). The tools included

with e-SC include workflow blocks for data import/export from CSV, data manipulation and transformation, machine learning and database management.

1) Pipeline Overview

In order to enable e-Science Central to achieve parallelism, the processing is subdivided into separate workflows. Although these subdivisions can take arbitrary structures, the most common form is that of a top-level workflow for coordination and sub-workflows to process independent data streams. It is this structure that was used during the BESiDE project to process the data from the devices and insert the results into the MongoDB database. The uploaded ZIP archives are extracted and then processed by a series of sub-workflows, each designed to process a separate data file within the ZIP archive, as shown in Fig. 5.

Some workflows perform relatively simple tasks, and mostly use the existing workflow blocks provided by e-SC: the Study Metadata workflow manages the metadata of each study (start time, end time, number and IDs of participants etc.); the Temperature workflow uses the data from the beacons to calculate the average temperature for a series of fixed time intervals for each zone; and the Ambient Noise workflow calculates the average noise level for fixed time intervals for each zone, based on aggregated data from the smart devices.

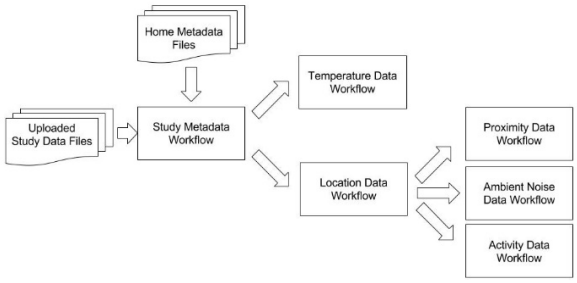

Fig. 5. BESiDE Pipeline Overview showing the various sub-workflows for processing the sensor data.

Other workflows include custom blocks that process the sensor data to infer resident location, social interactions, physical activity levels, etc. These workflows are discussed in more detail below.

2) Location Workflow

A key element of the pipeline is the Location Data Workflow. Initially, the system estimates the indoor location of each resident in consecutive, one-second intervals. Each estimate is represented as a 2D coordinate (x, y) corresponding to a position on the calibrated floor plan. The estimate is calculated using the fingerprint calibration data, which is stored in e-Science Central, and the recorded RSSI data from the beacons.

As described in Section A.1), a fingerprinting technique was implemented that models the distribution of RSSI values from the BLE beacons at locations on a regular grid throughout the public spaces of each home. This model represents the fingerprint of each grid location as an isotropic Gaussian distribution over vectors r : (r1, r2, ..., rn), where ri represents the RSSI of packets from the *i*th beacon, with n beacons in total. The mean and covariance of this distribution are estimated once from the recorded fingerprinting data. During inference, vectors r are constructed from one-second intervals of participant data and the (x, y) location is determined using maximum-likelihood estimation.

In preliminary experiments, this method was compared to trilateration, where each RSSI is directly mapped to a distance $\check{D}i$ between beacon i and the receiver. Locations (x, y) are estimated by minimising the mean squared error between the distances from (x, y) to beacon locations (x_i, y_i) and the estimated distances $\check{D}i$. This method requires only all beacon locations to be known, avoiding the need to record calibration data. However, the researchers observed a considerably higher localisation precision of, on average, 1.05m with our fingerprinting technique compared to 2.2m with trilateration, and it was decided that this level of accuracy justified the additional time required to record calibration data in each home.

Because of the large volume of data that was collected, it was decided the full sequences of estimated resident locations (i.e. one x,y location per second, per resident) would not be stored in the database. There are two reasons for this: firstly, the activity patterns of the participating residents are relatively sedentary; and secondly, the average error of 1.05m in each reading may imply movement when none occurred. The latter is particularly relevant given the likelihood of interference on the BLE RSSI caused by movement around the participant rather than the participant themselves.

Instead, the estimated location data is used to estimate when each resident moved from one zone to another ("transition data"), along with the amount of time that the resident spent in each zone between transitions ("duration data"). This approach significantly reduces the volume of data stored in the database, while still providing useful insights into zone utilisation and resident movements over the course of the study. To improve the quality of the data and account for the noise present in the BLE signal, participants are required to have multiple consecutive readings before they are considered to have transitioned from one zone to another. This results in higher quality but lower resolution data - the researchers can have a high level of confidence that a participant is in one zone as opposed to another, but they cannot identify where the participant is within that zone.

The durations and transitions data is used by the subsequent sub-workflows to allow social interactions, ambient noise and physical activity levels to be associated with a Zone ID; allowing the data to be associated with the resident's location at the time the measurements were made.

3) Proximity Workflow

The Proximity workflow estimates periods of social interactions from recorded Nearable broadcast packets. For each Nearable, intervals of continuous proximity are estimated by comparing a moving average of inter-packet durations to a predefined threshold. This threshold is chosen such that 50% of uniformly random packets are allowed to get

lost during continuous periods of close proximity. The colocation of a large number of Nearables was used to filter spurious detections of social interactions with unused Nearables in the baskets.

4) Activity Workflow

Measuring physical activity levels can be achieved using the simple summary statistic of Activity Counts [21] and this technique is used in this study. Sequences of accelerometer data are transformed into activity counts for epochs of one minute in length. Activity counts are estimated from sequences of acceleration samples a : (a_x, a_y, a_z) using Eq. 1, where g = 9.81, the magnitude of standard gravity.

$$ActivityCount = \sum | (a_x^2; a_y^2; a_z^2) - g | \qquad (1)$$

Each activity count estimate is associated with the corresponding zone ID, resident ID and timestamp, and is then stored in the database.

E. Visualisations and Reports

In this section, we discuss examples of the visualisations that were created to explore the data.

Recognising that there is no 'one-size-fits-all' visualisation solution that is capable of allowing stakeholders to properly explore the collected data, multiple visualisations have been produced to show different aspects and granularities of the data. This includes attempts to visualise the data from multiple sources in such a way as to enable correlations between the quantitative data from different sources to be studied (i.e. between physical activity and zones within the care home), as well as correlations between quantitative and qualitative data. All of the visualisations are customisable, with a range of options and filters that allow the user to explore the data.

In order to allow the visualisations to be accessed from anywhere, without requiring the user to install any additional software, the visualisation tools were developed as web applications, which run in a web browser and access the data via an API. In order to keep data about the residents secure, users are required to log into e-Science Central in order to access the visualisations.

1) Report Generator

The configurable Report Generator accesses the database and produces PDF reports from the study data. Reports display summary data about the study, including the participants and zones, displaying the data mainly in the form of tables and graphs. These include reports for each resident, showing details about their location, activity and interactions over the course of the study, as well as reports for each zone, showing details about the utilisation, resident activity counts, temperature and ambient noise levels. Reports also include summaries of qualitative data, as well as specific qualitative data relating to each zone, where applicable.

2) Data Explorer

The Data Explorer displays summary data from the study using interactive charts and graphs, For example, Fig. 6 shows the room utilisation charts (generated using synthetic data). Pie charts show data such as the total time spent per room, while Gantt-style charts show room occupancy and transitions for each resident over time. These visualisations allow the user to study resident movements and zone utilisation at a fairly high level.

The Gantt-style chart is a good indicator of whether residents move around the home frequently for short bouts or move less frequently for longer bouts. This insight is difficult to obtain from numerical data alone but very easy to see on the correct visualisation.

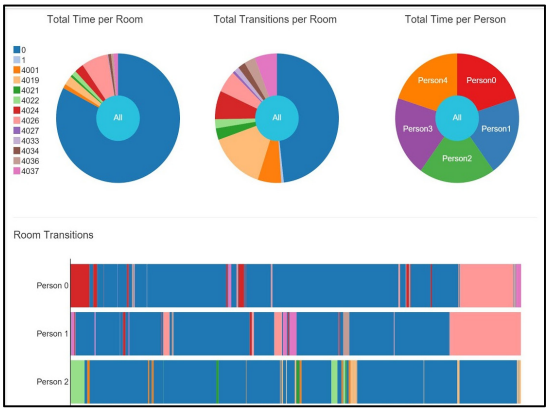

Fig. 6. BESiDE Data Explorer (synthetic data)

All of the charts are interactive, powered by the Dimensional Charting Javascript Library [22]. As the user makes selections on any of the charts, it triggers a filter over the data that updates all of the other charts accordingly. This provides an intuitive way to explore summary data from the study, treating the whole page as a single dashboard.

3) Floorplan Visualisations

The floor plan visualisation represents an attempt to visualise the collected data in the context of the built environment. The visualisation shows an abstracted floor plan of the instrumented rooms/zones of the care home. Residents' rooms and other private areas of the care home are excluded. Fig. 7 shows a section of the floor plan, showing the location of each of the participating residents at the specified time, using colour-coded markers and participant IDs.

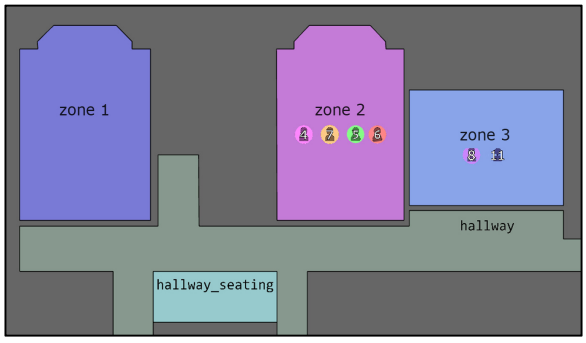

Fig. 7. The BESiDE Floorplan Visualisation showing resident locations.

The visualisations have been developed using WebGL [23], a Graphics API based on OpenGL. It executes using the HTML5 Canvas element, and so the visualisation can be used in most modern web browsers without requiring the user to install any additional software.

The data can also be replayed as an animation so that the user can observe the residents moving from zone-to-zone and look for any behavioural patterns that may occur. Filters can be applied to show data for individual residents, or for all residents, at a specific time or over a specific time range.

Alternatively, heat maps of the data show the time spent in different zones between a specified start and end time within the study, such as in the example shown in Fig. 8. This allows room utilisation over time to be visualised intuitively by showing the data on the map of the building. Other types of data can be visualised using heat maps analogously.

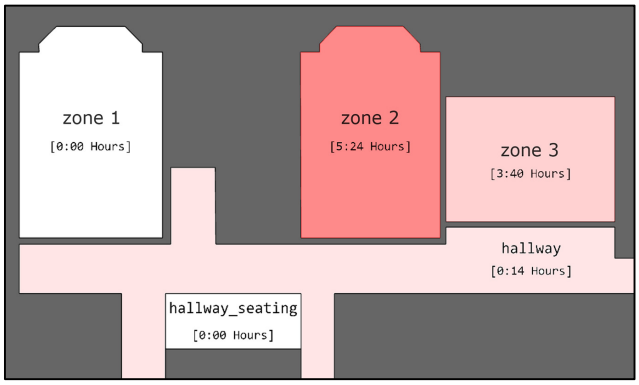

Fig. 8. The BESiDE floor plan visualisation showing room utilisation over time

Other data is visualised in a similar way. Heat maps show activity counts in shades of green to indicate in which zone the most activity took place. Zone temperature is visualised slightly differently; rooms are colour-coded blue if the temperature is below 18°C (and therefore considered to be cold) or red if the temperature exceeds 24°C (and is considered hot).

4) Visualising Qualitative Data

When viewing the quantitative data, a feed displaying the relative qualitative data can be enabled, as illustrated in Fig. 9. Any observations that are relevant to the Zone, participants currently in that Zone, or the currently selected time period are displayed in the panel. Users can also manually select a Zone, time period or participant to see any relevant observations that have been made.

1) Privacy Concerns

Because the collected data represents potentially vulnerable service users, we have to consider issues of privacy and security. Data has been anonymised - userIDs are stored, but no identifiable data about the user (name, age etc.) is stored - and abstracted floor plans are used in the visualisations so that no personally identifiable data about the care home or the participating residents, visitors or staff is stored or visualised. Additionally, stakeholders are required to log in to the BESiDE Web Visualisation application to view the data.

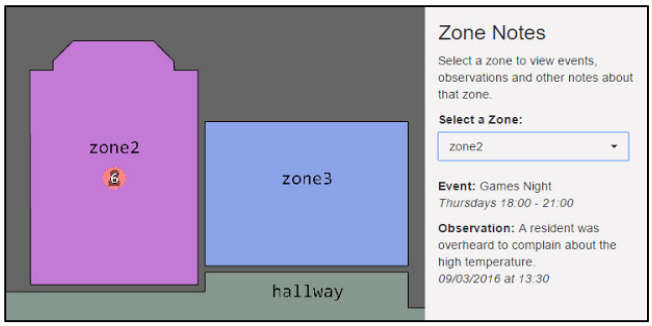

Fig. 9. The BESiDE floor plan visualisation showing the zone notes panel, which displays qualitative data

V. Conclusions

Our research has shown that it is possible to collect qualitative and quantitative data about a built environment system, and to process and visualise these disparate data types side-by-side, in order to enable correlations between the results to be explored, and enabling greater insight into the results that would be possible using either data type in isolation.

As a case study, we discussed the successful collection of data from a study of building utilisation by older adults living in residential care, and have highlighted the potential of the tool for use by architects, interior designers and care providers to improve the design and management of current and future care homes.

Using the quantitative data, the platform visualises physical activity and social interaction alongside location data and environmental measurements, and that by doing so it is possible to provide researchers and stakeholders with a way to gain insight into resident well-being, in addition to studying room utilisation within the care home. The system would also be suitable for medical research experts to study factors influencing physical activity. The addition of qualitative data to the system adds additional context for behaviours/patterns observed in the quantitative data. The suite of reports and visualisations described in Section VI demonstrates that a set of tools for domain-experts such as architects, designers and care staff can be developed. This allows them to intuitively access, explore and learn from data about the environment and human behaviour that has previously not been available.

Through our work on the BESiDE project, we have been able to develop and test a data processing system for evaluating the built environment. There was some difficulty in recruiting sufficient numbers of participants for the study, and so the amount of available data was limited, but the system, developed using e-Science Central, has been successfully used to perform automated analysis of quantitative data from sensors and smart devices. Qualitative data has also been added to the system and has been successfully used to produce a range of reports and visualisations, adding potentially valuable missing context to the quantitative data. We also now have a better idea of the sort of qualitative data that it is useful to collect about the built environment and, although we were unable to fully implement and test these new features within the available timescale, extensions to the platform have also been designed that will

enable easier processing of qualitative data in the future. We also have several ideas about how the system could be generified and extended so that it could be used in the analysis of other built environment scenarios.

VI. Future Work

Due to the timescales involved in the BESiDE project, the qualitative data collection was carried out by hand, largely using pen and paper. This was later transcribed and inserted into the data warehouse. In the future, we plan to develop an application that can be used by researchers to collect the qualitative data and automatically upload it to the platform. This will reduce transcription errors and force certain fields to be recorded consistently. Further, were not able to deal with image based observations, but a dedicated application could be developed that could upload these to Amazon S3 in a similar manner to the raw sensor data, which would allow them to be displayed alongside the text-based observations.

Acknowledgments

The authors would like to thank all study participants and care home staff for their cooperation and valuable feedback. The wider BESiDE team also deserves considerable acknowledgement - without their hard work and collaboration, this paper would not have been possible. This work was supported by the RCUK Lifelong Health and Well-being Programme grant number EP/K037293/1 - BESiDE: The Built Environment for Social Inclusion in the Digital Economy.

References

[1] U. Biobank. [Online]. Available: http://www.ukbiobank.ac.uk/. [Accessed 30 May 2017].

[2] S. Zhang, A. V. Rowlands, P. Murray and T. L. Hurst, "Physical activity classification using the GENEA wrist-worn accelerometer," *Med Sci Sport Exerc.,* vol. 44, no. 4, p. 742–748, 2012.

[3] "e-Science Central," [Online]. Available: http://www.esciencecentral.co.uk/. [Accessed 30 May 2017].

[4] "BESiDE - the built environment for social inclusion in the digital economy," [Online]. Available: http://www.beside.ac.uk/. [Accessed 16 May 2017].

[5] "GENEActiv," [Online]. Available: http://www.geneactiv.org/. [Accessed 30 May 2017].

[6] "Tweet My Street," [Online]. Available: http://www.tweetmystreet.info/. [Accessed 30 May 2017].

[7] G. Mearns, R. Simmonds, R. Richardson, M. Turner, P. Watson and P. Missier, "Tweet My Street: A Cross-Disciplinary Collaboration for the Analysis of Local Twitter Data".

[8] D. L. Driscol, A. Appiah-Yeboah, P. Salib and D. J. Rupert, "Merging Qualitative and Quantitative Data in Mixed Methods Research: How To and Why Not," *Ecological and Environmental Anthropology,* vol. 3, no. 1, 2007.

[9] D. L. Morgan, "Practical Strategies for Combining Qualitative and Quantitative Methods: Applications to Health Research," *Qualitative Health Research,* vol. 8, no. 3, pp. 362 - 376, 1998.

[10] P. Watson, Hiden H. and S. Woodman, "e-Science Central for CARMEN: science as a service," *Concurrency and Computation: Practice and Experience,* vol. 22, no. 17, p. 2369–2380, 2010.

[11] J. Cala, H. Hiden, S. Woodman and P. Watson, "Cloud computing for fast prediction of chemical activity," *Future Generation Computer Systems,* vol. 29, no. 7, p. 1860–1869, 2013.

[12] H. Hiden, S. Woodman, P. Watson, M. Catt, M. Trenell and S. Zhang, "Improving the scalability of movement monitoring workflows: An architecture for the integration of the hadoop file system into e-science central," *1st workshop on Digital Research.*

[13] S. Woodman, H. Hiden, M. Turner, S. Dowsland and P. Watson, "Monitoring of upper limb rehabilitation and recovery after stroke: an architecture for a cloud-based therapy platform," in *IEEE s-Science*, Munich, Germany, 2015.

[14] J. Austin, R. Davis, M. Fletcher, T. Jackson, M. Jessop, B. Liang and A. Pasley, "Dame: Searching large data sets within a grid-enabled engineering application," *Proceedings of the IEEE,* vol. 93, no. 3, p. 496–509, 2005.

[15] M. Fletcher, T. Jackson, M. Jessop, B. Liang and J. Austin, "The signal data explorer: a high performance grid based signal search tool for use in distributed diagnostic applications," in *Sixth International Symposium on Cluster Computing and the Grid*, 2006.

[16] V. L. Hanson and L. J. McIntyre, "BESiDE: The built environment for social inclusion through the digital economy," *The Application of Digital Innovation, Fifth Annual Digital Economy All Hands Meeting (DE 2014),* December 2014.

[17] "Amazon S3," [Online]. Available: https://aws.amazon.com/s3/. [Accessed 30 May 2016].

[18] "MongoDB," [Online]. Available: https://www.mongodb.com/. [Accessed 30 May 2016].

[19] "Estimote," [Online]. Available: http://www.estimote.com/. [Accessed 30 May 2016].

[20] "AEIOU Framework," [Online]. Available: https://help.ethnohub.com/guide/aeiou-framework.

[21] K. T. Khaw, N. Wareham, S. Bingham, A. Welch, R. Luben and N. Day, "Combined impact of health behaviours and mortality in men and women: the epic-norfolk prospective population study," *PLoS Med,,* vol. 5, no. 1, 2008.

[22] "Dimensional Charting Javascript Library," [Online]. Available: http://dc-js.github.io/dc.js/. [Accessed 30 May 2017].

[23] "WebGL," [Online]. Available: https://www.khronos.org/webgl/. [Accessed 30 May 2017].