PerCon: Supporting the Management and Reuse of Wearable Sensor Data

Frank Shipman, Su Inn Park, Ricardo Gutierrez-Osuna, Jongyoon Choi

Center for the Study of Digital Libraries and Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77845, USA

{shipman, rgutier}@cse.tamu.edu

Abstract. This paper describes the problem of managing wearable sensor data and the design of an environment to support this activity. We are engaged in studies collecting data from custom-made and off-the-shelf wearable and mobile sensors to explore the impact of context and physiological state on cognitive performance. The data being collected has the potential to assess many potential hypotheses beyond the original hypothesis being explored. In order to achieve this potential, infrastructure is required to manage, access, preserve, and share the various types of digital objects recorded from individual research participants. As with many studies of individuals over time, there are a number of data streams correlated via timestamps for each participant. To manage this data we are developing PerCon. PerCon processes and integrates the individual datasets so as to be loosely coupled with the analysis techniques used in our first stage of research. In addition, PerCon provides services such as searching, indexing, browsing and visualization. PerCon's APIs support access to and processing of the resources by user and researcher applications. A result of our activity is an initial representation of the relationships among the heterogeneous resources to facilitate their reuse. Longer term, this will be separated into domain-independent and domain-dependent ontologies of the data types and resources involved. The overall result is an architecture and initial instantiation for e-Science and e-Health digital libraries.

Keywords: Wearable sensors, e-Science, e-Health, cyberinfrastructure, digital libraries, data management, data reuse, data repository

1 Introduction

E-Science has the vision of sharing science data across research groups to support data reuse. Based on the technologies such as grid/cloud computing, database systems, and distributed collaboration support, e-Science emerged in the areas of physics, earth-science, and bio-informatics, where voluminous datasets are common and the need for infrastructure to manage and share data sets is more obvious. For example, data management practices and digital library systems for supporting data management are being explored for geographical sensor data [5], [30].

The application of e-Science to emerging fields relying on wearable sensors to study physiological and psychological phenomena in the world is challenging. Unlike geosciences or astronomy, there is no pre-defined spatial model about which to organize the data for sharing. Unlike physics, studies involving wearable sensors often lack the control over context that make interpretation of data easier across experimental setups. Indeed, the e-Science and e-Health applications of wearable sensors are valuable because they allow study participants to go (approximately) about their own life while data is collected.

There are numerous examples of researchers developing wearable sensors for studying contextual effects on individuals. With data reuse being a major reason for funding agencies beginning to require researchers to include data management plans, such studies pose many difficulties. First, custom-made sensors often lack the consistency of mass-produced devices. Each sensor built may perform slightly differently, requiring an understanding of each sensor in order to interpret the data collected.

Second, some wearable sensors generate grossly different data based on the individual and their overall physiology. For example, some electrical signals recorded vary based on the conductance of the participant's body which is, in turn, affected by their body fat. Third, even for the same individual, the placement of the wearable sensors affects the data collected. So, for a study where the participant removes the sensor each night to sleep, the data from each day has to be initially recalibrated. Overall, interpretation of the data requires knowledge of each sensor, each participant, and each individual data collection event.

The above difficulties led us to begin development of a digital library for managing wearable sensor and related data called PerCon (for **Per**sonal/**Con**textual data environment). PerCon is a digital library in that it supports data management, provides a method for searching for and retrieving data streams, and includes APIs for sharing data with other applications. In addition, it includes middleware capabilities supporting data stream analysis and end-user interfaces for visualization and interpretation. We anticipate two types of PerCon use: (1) researchers collecting study data to evaluate specific hypotheses about relationships between the data streams being collected, and (2) personal users who are collecting data about themselves for health monitoring.

We first describe the use of wearable sensors in science. Section 3 surveys related work in the fields of digital libraries and tools for wearable sensor data analysis and sharing. Section 4 presents the PerCon architecture and system. Section 6 discusses open issues. We conclude with lessons and future work.

2 Wearable Sensors in E-Science

Wearable sensors provide infrastructure for exploring the effects of variables related to physiological state, mental state, physical activity, and/or context on one another. To do so, participants in studies wear or carry sensors with them as they go about their life or perform some pre-scripted activity. The data streams generated are then analyzed for evidence of interactions between the variables being monitored.

The number and variety of wearable sensors being used in such research continues to grow. They include commercially-available sensors that measure physiological variables and are often intended to be used for monitoring data related to an individual's health. Examples include sensors for measuring blood pressure, cardiac activity (e.g., electrocardiography, phonocardiography, heart rate monitoring), respiratory activity (e.g., breathing rate, effort and volume), brain activity (e.g., electroencephalography, near-infrared spectroscopy), motor activity (e.g., electromyography, actigraphy), perspiration and electrodermal activity (e.g., galvanic skin conductance) [20]. Other sensors are designed and built specifically for a particular research project or to avoid limitations of commercial sensors. Examples include implantable sensors (e.g., blood glucose, blood pressure, therapeutic radiation) and smart textiles (e.g., cardiorespiratory activity, body temperature, skin conductance). A third class of "sensor" that is common today is the mobile phone. These devices often include multiple sensors (e.g. accelerometers, compass, GPS, microphones, cameras, etc.) that can be used to collect data about the individual and their environment [31].

The variety of sensors and sensor systems result in data sets with very different data practices with respect to which sensors are used, how they are positioned or carried by participants, and what are the sensor settings/sampling rates. Even so, there is a fundamental similarity in that the studies generate a set of time-synchronized data streams. With metadata describing the details of the study, such data could be shared among related studies and could be used to explore hypotheses beyond those of the original study. The core features necessary for such a data management system are the recording of metadata indicating relationships among data streams (e.g. participant ID, date, time) and the details necessary to interpret the individual data stream (e.g. sensor ID, settings, electrode placement). In addition to metadata attached to individual data streams, the repository needs information about each sensor and each participant.

Our own research on wearable sensors for health and cognitive monitoring has become the main source of heterogeneous data for the repository. As part of this work, we have developed a wearable sensor platform that allows experimenters to collect physiological and contextual variables using a wireless network of small, light-weight and unobtrusive sensor units. To date, the system includes sensors for measuring heart rate, respiratory effort, muscle activity, skin conductance, body acceleration and GPS, uninterrupted for periods of up to 12 hours [8], [9]. The system has also been integrated with an Android-based mobile phone, which allows us to prompt the user to complete various cognitive tasks. The sensor system is currently being used in two multidisciplinary studies aimed at monitoring mental workload/stress and creative cognitive performance, both in ambulatory settings ("cognition in the wild").

The design and development of PerCon has been motivated by our goal of managing the wearable sensor data we collect for our own reuse across these varied interests. Initial suggestions for the use of OAI-ORE to model the relationships among data elements [21] rely on local repositories that store the necessary data and metadata. We see PerCon as such a local repository, defining middleware services that enable semantics-based querying of sensor data objects, coordinate the presentation of temporal data, and analyze the coherence, correlation, and patterns among the selected data streams.

In addition it includes middleware capabilities supporting data stream analysis and end-user interfaces for visualization and interpretation. We anticipate two types of PerCon use: (1) researchers collecting data from study participants to evaluate specific hypotheses about relationships between the data streams being collected, and (2) personal users who are collecting data about themselves for personal health monitoring.

Within the context of PerCon, this paper describes: (1) designing the repository to manage the datasets containing interrelated datastreams and other structured and unstructured data, (2) processing and integrating different forms of time-series data for synchronized presentation, (3) locating/selecting data streams for analysis via queries defining relations between streams, and (4) introducing an architecture and applications that can be extended beyond the initial domains of context-aware cognition and ambulatory health monitoring. Ultimately, through our digital library system, we intend to examine the potential for data reuse in the more general field of wearable sensors. Moreover, we expect this research will provide insight into requirements for cyberinfrastructure in the medical and social sciences.

3 Related Work

This project builds on a wealth of prior research in the areas of abstract models of digital libraries, digital libraries for science data, and tools for wearable sensor data.

3.1 Abstract Models of Digital Libraries

Fundamental abstractions and models/architectures of digital libraries have been defined and established in earlier efforts. McCray and Gallagher addressed underlying principles for digital libraries development [16]. Gonçalves et al. explored and defined fundamental concepts for digital libraries in their 5S (Streams, Structures, Spaces, Scenarios, and Societies) model [11]. The DELOS Network of Excellence on Digital Libraries [7] introduced a reference model for systematic approaches to digital libraries defining four perspectives (end-user, designer, system administrator,

application developer). In that model, conceptual frameworks are represented in six core domains – content, user, architecture, policy, quality, functionality . PerCon is grounded on the above principles and formal abstractions. For example, the concepts found in 5S can be identified in PerCon; streams are sequences of personal sensor data, structures are representational frameworks for metadata specification, spaces are the repository/database, services via requests/events implied by scenarios, and the scientists and personal users constitute societies. However, we are developing PerCon as a substantial instance of a digital library rather than as a conceptual framework.

Scientific literature digital libraries such as CiteSeerχ [15], NDLTD [10], and the SAO/NASA Astrophysics Data System [13] are built around service-oriented architectures that inspired the PerCon architecture. For example, Citeseer comprises three layers: storage, application, and user interface layer. Since each library's focus is unique (e.g. on semantic web services or searching engine via distributed servers/repositories), the models are specialized to functions matching that focus (e.g. crawling, storage). Opensource digital library systems, e.g. Fedora [26],21] and DSpace [are versatile environments for creating and managing collections. However, they were mainly designed for content management and acquisition. PerCon, with its goal of integrating data analysis, interpretation, and visualization, includes middleware and application layer services not found in the digital libraries mentioned above. Even though some of these have similar three-layer abstract models to that of PerCon, internal operations of layers are different due to different intent.

Models of data ingestion and workflow also inform the design of PerCon. Workflow models provide systematic procedures for data management and analysis. A digital object's status is recorded as metadata, enabling interfaces for monitoring the data ingestion and analysis process, such as found in Kepler [1], Taverna [14], and Trident [

3.2 Digital Libraries for Science Data

A number of projects have applied digital library technologies to collections of science data [16]. Brettlecker and colleagues integrated data stream management into the OSIRIS-SE digital library. The emphasis of their work is on enabling the reliable collection of data streams in case of communication failures [5]. They do not consider requirements for the sharing, analysis, and visualization of the data streams.

Borgman and colleagues have been studying and supporting sensor net data management practices by researchers in the physical sciences [5]. They found that "scientists often store data with minimal documentation and do little toward preservation …" Given the even higher standard of metadata requirements needed for interpreting wearable sensor data, this would seem disheartening for the practical potential of PerCon. One difference is that by integrating the initial data analysis tools with the data management capabilities, the metadata needed for the researcher's own interpretation will be necessarily provided to PerCon during data analysis rather than relying on data providers to attach the metadata when a data set is contributed to the repository.

As a result of earlier studies, Wallis et al. designed a Data Discovery Library where data sets are bundled into packages with attached metadata enabling potential reuse [30]. This is in contrast to the approach taken in PerCon, where the data objects in the library are individual data streams and the metadata encodes the relationships between data streams.

PerCon's domain-specific take on data digital libraries is far from unique. Bioinformatics as a representative data-intensive science has developed databases and computational/statistical analysis tools to explore large scale genome sequencing. Genbank [4] in the U.S, DDBJ [29] in Japan, and EMBL-Bank [2] in the U.K are huge databases functioning as a type of fine-grained digital library system. Similarly, healthcare systems, such as Google Health [12] and Microsoft HealthVault [18] , incorporate a domain-specific representation for managing and sharing personal health records. The domain of wearable sensors is broader than most of the above examples in that it can include many

different types of data for many different research purposes. The primary characteristic is that data sets are composed of interrelated data streams that are collected simultaneously from individual participants. This breadth motivates the use of a combination of domain-independent ontologies, domain-dependent ontologies, and personal ontologies for representing relationships within data sets. Suggestions for the use of OAI-ORE to model the relationships among data elements [19] rely on local repositories that store the necessary data and metadata. We see PerCon as such a local repository.

3.3 Tools for Wearable Sensor Data

Software tools for the analysis of wearable sensor data can be grouped into two broad categories: general purpose analysis tools and product-specific tools. General purpose tools include MATLAB and GNU-based implementations such as Octave and R. These computing environments are widely used in research circles and have a large number of user-contributed libraries, but are too sophisticated for non-programmer end users. Most wearable-sensor vendors provide tools to analyze sensor data. These can range from application suites with specific tools for each data type (e.g., EEG, EMG), such as those provided by Thought Technologies Ltd., (Montreal, Quebec, Canada) and g.tec (Graz, Austria), which are suited for researchers, clinicians and advanced users, to web applications such as those provided by FitBit (San Francisco, CA) and BodyMedia (Pittsburgh, PA), which are targeted for the health-minded consumer. These tools support analysis but lack facilities for sharing and reusing data.

There is a variety of research investigating techniques for capturing, analyzing, and locating data streams. In this context, efforts often emphasize how to efficiently work with data stream content in a database [11] and how to cope with real-time requests for data just being captured [10]. Such efforts can improve the components in PerCon but lack the rich metadata and provenance records required in our context.

More generic systems have been created for sharing data presentations and interpretations. For example, the Virtual Notebook System (VNS) was designed to support collaborative biomedical research by enabling researchers to combine metadata, interpretive text, and graphs of data [20]. Such tools support sharing research results but lack facilities for data analysis/sharing.

4 PerCon: A Repository for Interrelated Data Streams

PerCon is a combination of digital library and data analysis platform for wearable sensor data. Figure 1 shows the envisioned process for collecting, locating, and making use of data streams in PerCon. At the center is a database and data repository for the data. The lower left represents the collection and ingestion of data from wearable sensors and mobile devices, the lower right shows the location and visualization of data through applications, and the top shows the ontologies used to encode the characteristics of the data necessary for its analysis and use as metadata.

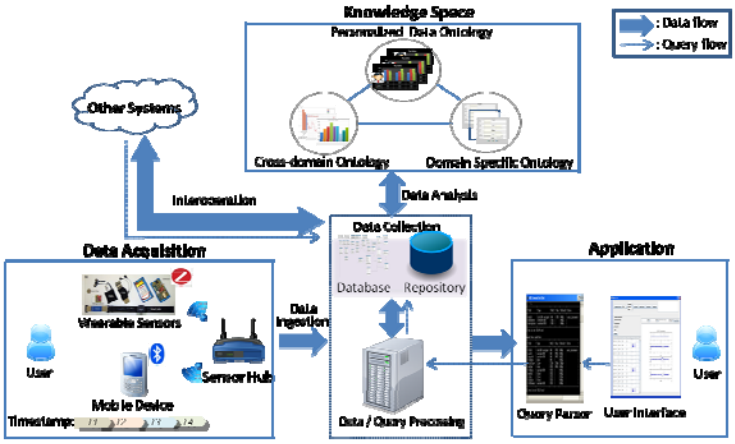

Fig. 1. Flow of information during data acquisition, location, and analysis in PerCon. Data collected is ingested into the repository (lower left). Ontologies enable recording of metadata necessary to enable reuse (top).Users access and interpret the data through applications (lower right).

During data collection, individuals (e.g. participants in studies) use a combination of wearable sensors and mobile devices to record information about their activities, physiology, and context. Periodically, this data is ingested into the PerCon repository. Data ingestion requires a combination of recording metadata about the data streams (e.g., subjectID, sensorIDs), and initial processing of the raw data streams to make them more amenable to analysis (e.g. normalization, time-windowing). Ingestion also generates feature-based indexes (e.g., signal statistics) into the data streams.

Data location occurs through a combination of searching and browsing. Queries are based on a combination of data-specific metadata (e.g. time, sensor type) and data features within the data streams (e.g. heart rate above a threshold value).

The next section describes the layered architecture being developed to isolate the PerCon software components. A presentation of PerCon as it currently exists follows.

4.1 PerCon's Layered Architecture

The goal of combining digital library and data analysis components led us to develop a layered architecture for managing the interconnections among the many and varied necessary software components. As illustrated in Figure 2, the architecture constitutes three layers: a Resource Layer, a Middleware Layer, and an Application Layer; the figure also describes some of the core capabilities at each layer. As is typical for layered software development, software components in a layer are accessed only by software components in the next higher layer.

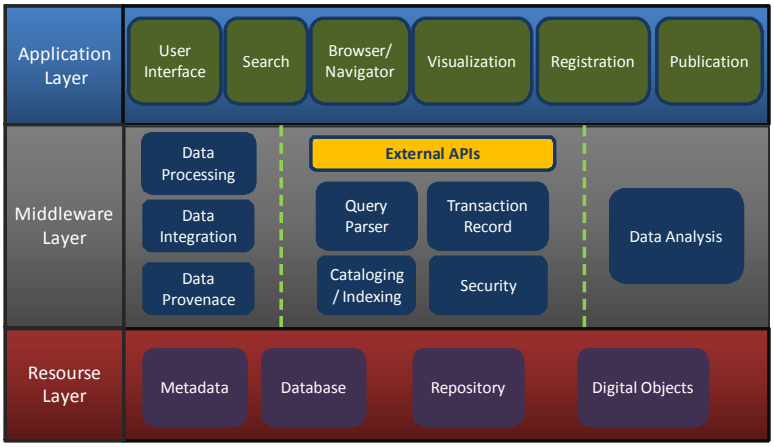

Fig. 2. PerCon's architecture manages dependencies among software components. The resource layer stores the digital objects and their metadata; the middleware layer has facilities for ingestion, analysis and search; and the application layer includes interfaces for locating, manipulating and visualizing the data.

The Resource Layer provides capabilities related to storing and preserving the original data objects (e.g. data streams), computed and filtered datasets, and metadata. A repository is used to store and manage the large data files and some basic (e.g. Dublin Core) metadata. A database is used to store more complete metadata for the digital object and indexes into the data objects in the repository. Metadata standards like the OpenGIS Sensor Model Language (SensorML) provide methods for encoding many characteristics about sensors. Unfortunately, it appears that the domain of wearable sensors (where people wear sensors collecting data about themselves and their context) is different enough from geographic information systems (where sensors measure features in a geographic space) that such a model would have to be greatly extended/modified for our efforts.

The Middleware Layer of the architecture can be divided into three classes of functionality: data ingestion, data access, and data analysis.

- ***Data ingestion*** involves a combination of data processing, data integration, and provenance recording. To enable the services for the application layer, the data processing framework includes a data reliability check to ensure that the data-object content is well-formed for the data-object type. Data objects of unknown types can be stored with relationships to other data objects encoded as metadata but without indexing and analysis capabilities. The data processing framework also provides modules for preprocessing raw data for feature (metadata) extraction and to generate computed data sets. The data integration framework synchronizes the digital objects with timestamps based on an understanding of the experimental setup and/or communication between devices/sensors collecting data. Finally, data ingestion records metadata concerning the provenance of the data: project ID, researcher ID, data type, sensor ID, creation date, participant ID, and a description of the process of data generation if it is not raw data (e.g. which raw data stream(s) it comes from and how it was generated).
- ***Data access*** includes providing application layer tools to access resources stored/managed by the resource layer through a set of external APIs. A query-processing module parses a query and

determines the communication necessary with the resource layer (e.g. database and repository). This includes determining if a query can be satisfied with existing index structures or if data objects need to be processed. For example, a query asking for skin conductance data at times when the participant's heart rate exceeded some threshold would be sped up if the database contains a precomputed index of heart rate maxima per time window enabling fast location of the times desired. Additionally, the data access layer is responsible for ensuring the requester has access to the data being requested.

- ***Data analysis*** capabilities included are initially driven by the domain, data types, and goals of the research. Some analysis capabilities process data streams for visualization (e.g. computing window averages, standard deviations, principal components scatter plots) while others are meant to locate patterns in the data. Thus, the analysis capabilities support a variety of human-driven hypothesis testing, ranging from the highly interactive to the highly automated.

Finally, the Application Layer includes end-user and external systems accessing the content in the digital library. User interfaces for searching and browsing the contents of the storage layer are provided. When a user retrieves data stream information, visualization tools help the user understand and explore the data recorded and expectations based on hypothesized domain models. A data registration interface enables data to be added, modified, or deleted. Finally, a publication interface is enables remote access to the contents of the library.

4.2 PerCon Instantiation and Vision

The vision of PerCon presented in section 4.1 is being instantiated for use in our own research. As such, the current PerCon system components are highly tailored to a very specific set of sensors, data types, and data analyses. Here we describe the current state of the database and repository, the data ingestion and analysis components, the interfaces and services provided, and the API for developing additional services and providing external access.

4.2.1 Database and Repository in Resource Layer

The resource layer of PerCon is instantiated through the combination of a data file repository and a relational database to maintain metadata and indexes into the data files. The repository includes both raw and computed data files, which can greatly vary in size depending on what type of signal is being recorded and its sampling rate. As is common in such repositories, the relational database maintains metadata describing the data objects in the repository and their relationships. In addition, the relational database maintains indexes into the data files for particular events.

As an example, consider when a sensor captures skin conductance over a period of time. The raw data file, as captured and stored in real time by the sensor, will be saved in one file. A computed file will normalize the data based on which sensor was used, the participant's baseline measurement, etc. Additional files might be present that record computed measurements over the normalized data (e.g. mean and standard deviation for a time window). The database includes metadata that links these files and identifies the relationships among them. The database might also include one or more indexes into the data, such as the mean, the highest value, and the lowest value for each five minutes of normalized data. This index enables rapid responses to queries for events, such as locating the time periods where a particular value is above or below a threshold and for the window-averaged normalized data during those periods in order to present high-level graphs of the data. Particular portions of repository files are then accessed to localize events and to present detailed graphs.

Metadata in the database is organized based on a representation of the domain of the research. In Figure 1, this domain representation is found in the Knowledge Space, consisting of a domain-independent (or cross-domain) ontology, a domain-specific ontology, and a personal ontology. Currently, PerCon has a single ontology that is used to represent domain-independent, domain-dependent, and personal ontologies. This is similar to the single schema used in [13] except our

schema is aimed at supporting project-specific sensors as well as commercial sensors. An example of content in the ontology is the hierarchy of sensor types used to collect data. Sensors are divided into those that capture physiological data, those that capture activity data, and those that capture contextual data. Heart rate monitors and skin conductance sensors are in the first category. Due to some of these sensors being made within our research group, those categories are further subdivided into particular models/designs and then individual instances of sensors. In our current study, an activity sensor is a smartphone application that prompts the participant to answer a series of questions at various points during their day. Separate smartphone applications capture contextual data, e.g. sound pressure levels, throughout the study.

The ontology also represents the relations between data files. Thus, the forms of computation over data files used to generate other data files are represented as a hierarchy. Classes of functions used to generate computed data files include normalization functions, filtering functions, and statistical functions. The preservation of the metadata about data collection (name, date, capture device, digital object types, etc.) and metadata describing the relationships between the digital objects combined with the knowledge base enables understanding of the raw and computed data files.

There is valuable information that is not possible to formally represent within an ontology or similar knowledge representation [24]. PerCon includes metadata fields and data object types for recording such descriptive content. In our context, a descriptive data object is the training protocol/document for teaching participants how to use the smartphone application. Descriptive metadata might mention that the participant indicated that the skin conductance sensor had to be disconnected and reconnected at some point during the day.

4.2.2 Processing and Analysis in Middleware Layer

The middleware layer of PerCon currently supports ingestion of data into the repository and simple data analysis and query facilities. Data ingestion begins with a preprocessing step to verify whether the raw data has been recorded correctly. In particular, the first check is to make sure the data file format is as expected (i.e. the data can be parsed if it is of a type meant to be manipulated). This step consists of testing an XML data file against the appropriate Document Type Definition (DTD) or, for raw data files, ensuring that the custom parser can read the data file without error. In addition, the ingestion may invoke a sensor-specific "sanity check" program to determine if the data has the characteristics expected. For example, such a check for a breathing sensor would make sure the respiratory signal was not clipped or that the respiration rates were within a reasonable range. The user ingesting the data file is notified of any issues.

Ingestion also generates computed data files and indexes into the data files. In our research, each measurement consists of a timestamp (in milliseconds) and the signal value. Other sensors might record data with implicit timestamps based on a consistent data rate or record event-based data, such as the times when users complete tasks. Thus, pre-processing of the raw data transforms the timestamp (which is based on the processor's clock) into a user-readable time to the second. It also adds the results of a window average mean and variance to the computed data file. The ingestion process, including all processing steps and generation of processed data files, is dated and recorded for later ability for provenance analysis.

Data analysis capabilities of the middleware layer currently include a set of time-windowing and normalization algorithms. Planned data analysis capabilities will enable frequency-domain analyses, such as Fast Fourier Transforms.

Currently, PerCon includes very simple query capabilities. Users can request data by specifying a participantID, a data type (can be raw, processed and stored in a permanent file, or computed based on an analysis capability and stored in a temporary file), and a time window. As it is currently being used as a research group digital library/repository, PerCon's instantiation does not yet include an access control module. The design of appropriate access control capabilities is interrelated with the data privacy issues discussed later.

4.2.3 Services and User Interface

Access to the data in PerCon is through application-layer services and tools. Currently, a single desktop application provides support for data ingestion as well as for locating and visualizing data in the PerCon repository. We are developing access capabilities via a Web-based interface for members of the research team that are not located in our building (and do not share the same file server) and new VKB Data Objects in the Visual Knowledge Builder (VKB) [23] to support data analysis in spatial hypertext.

The PerCon desktop application is shown in Figure 3. At the top of the application is the data ingestion interface where research team members add data to the repository. Dialog boxes lead the user through the process of attaching metadata and provide feedback on whether there are problems or anomalies with the data file.

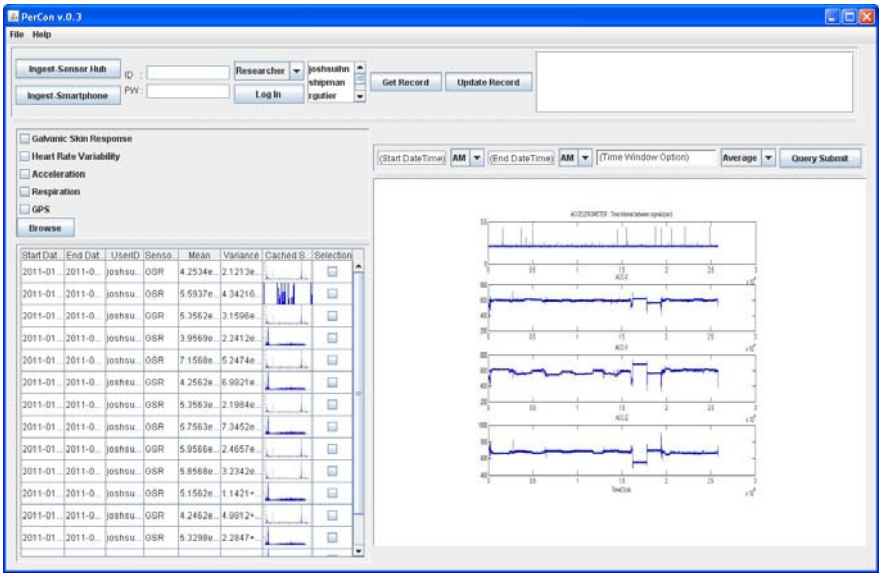

Fig. 3. Prototype interface for locating and viewing data in PerCon repository. Researchers use top region to login and ingest data. Sensor types of interest are selected and the resulting data files are browsed on the lower left. Data analysis and visualization is in the lower right.

Below this is the data location and visualization interface. On the left the user selects what types of sensor data are of interest. Currently, the data file types correspond to galvanic skin response, heart rate variability, acceleration, respiration rate, and GPS (global positioning system) data. Note that each of these data types may correspond to more than one type or model of sensor. The user may select more than one sensor type at a time when wishing to visualize potential correlations.

Below the data type selection (on the bottom left) is a browsing region for selecting among the data files that match the current selection. If there is no selection, all the data files are shown. The list includes the start and end time of the data stream, the userID, the sensorID, the mean and standard deviation, and a thumbnail graph of the data.

On the right, below the ingestion interface is where the user identifies the time window for the data and an operator such as average (mean) or variance for visualization. If two data streams are selected, the user can also select to view the correlation between the streams.

To display more than one data stream simultaneously, PerCon presents more than one plot in the visualization canvas (bottom right). The graph shown in the figure corresponds to accelerometer data. Here the time between readings is shown in the top graph (to provide a sense of how consistently the data was recorded) and the acceleration in each of the three dimensions is shown in the lower three graphs. The spikes in the top graph show that the accelerometer, while generally recording data at the expected rate, sometimes dropped a reading. The variance in the time between readings can create noise that the researcher must take into account.

This interface currently includes analysis and visualization capabilities designed to locate problems with data collection (e.g. issues with sensors, data recording, and data transfer). Additional capabilities will be added when the cross-subject analysis begins.

4.2.4 Interoperability & Compatibility

The vision of data reuse in e-Science implies that the original researcher cannot anticipate all uses of the data they are collecting. While such a situation poses numerous difficulties for the recording of sufficient metadata, at a first pass systems need to be extensible and interoperable to support alternative forms of analysis. We aim to support such interoperability through two capabilities. First, the PerCon middleware services provide an API for researchers to develop new applications (or tools that collect information from PerCon for analysis in existing applications). In the future, we plan to provide OAI-PMH and OAI-ORE interfaces to the PerCon repository when it is allowed and appropriate.

5 Planned Use Scenario

The following scenario describes our intended use of PerCon and thus is based on the actual domain and data types being collected and managed. As the project is currently at the stage of testing the sensors and smartphone applications prior to the actual data collection, experience to date is with data from pilot studies used to ensure the sensors, applications, and study protocol work as intended.

In this scenario, a multidisciplinary research team is exploring the effects of physiology and context on people's creativity in a natural setting. As such, physiological data is collected via wearable sensors communicating with a wearable sensor hub, contextual data is collected via sensors onboard a smartphone, and creative performance data is collected via a smartphone application. The smartphone and the sensor hub timestamp and store the data independently, but also communicate with each other via Bluetooth to periodically record the relationship between the two devices' clocks to enable synchronization between the data stored on the two devices.

As data is downloaded from the sensor hubs and smartphones upon completion of a participant's involvement, it is ingested into the data repository by members of the research team. As part of this process, the researcher must use the provided tools to ensure the quality of the data, generate the computed data objects from the raw data, and attach metadata to the raw and computed data objects. A number of research team members are likely to be involved in data collection and ingestion. Similarly, different research team members will be analyzing the data, potentially with different goals.

The data collected has value for assessing many independent hypotheses in addition to the original goal of assessing the effect of physiological state on creativity. Additional hypotheses include questions about the design of the wearable sensors and potential effects of context on physiological state.

The members of the research team designing the wearable sensors will perform detailed analyses of the data coming from the sensors to determine if the sensors are recording data as expected and to look for potential problems in their design. For example, comparing the data from early in a recording session to data recorded late in a session could identify trends due to changes in battery power or electrode contacts.

The team members examining the results of the creativity tasks would first compare the data from the smartphone applications to data from similar tasks in their prior studies to determine if the data has similar distributions. Once the quality of the creativity data is determined, they will begin to examine correlations between physiological and contextual data with the creativity data. In particular, this project is examining the hypothesis that stress affects creativity and that this effect may not be monotonic – that people perform best with some, but not too much, stress. Another hypothesized outcome is that background noise affects performance on creativity tasks.

This project is just one of a number projects on which the faculty members in the research team are collaborating. Another project explores systems using wearable sensors to provide insight into one's own health. In this context, user interfaces supporting the analysis and presentation of the data from the wearable sensors would be provided to the participants in the study to determine if they learn about how their activities and context affect their physiology.

Once the results of the initial studies are known, we can envision additional uses for the data in the repository. For example, the study may show that there are subpopulations of participants with different correlations between variables, where different effects are found. In such a case, the data could be used to generate and test statistical or probabilistic models for classifying members of the population into different groups. Such models would be important for making systems that proactively help users learn about their own performance or physiology.

This scenario of data use shows the potential for an environment such as PerCon to support the reuse of wearable sensor data. Underlying this scenario, and the design of PerCon, are a number of assumptions that pose challenges for this vision to become a reality. These are discussed next.

6 Open Issues

There are a number of open issues related to digital libraries of wearable sensor data. Of these, here we focus on two practical issues, metadata consistency and sensor evolution, and one social issue, the privacy of wearable sensor data.

Metadata consistency is an issue for all libraries and repositories. In general, metadata inconsistency can lead to difficulties in locating content due to the resulting inconsistencies for users when browsing or searching the library. These problems are exacerbated when metadata is used to determine what computation is applied to the content of a library. For example, if the sensorID is incorrect then the results of any analysis are also likely to be incorrect but may not be obviously so to

the user. Thus, metadata errors in this context are more likely to lead to "garden path" problems [22], which are often not identified by the system or user until much later.

Another issue comes from the mutable nature of self-made wearable sensors. Each time a component is replaced or the sensor is otherwise changed, there is the potential that its data recordings will be altered. The procedures for converting from the raw data to the computed data are meant to take care of many of these issues (through calibration, normalization, etc.) but some changes can qualitatively affect the sensor. In the ideal, such a change is noticed and a new SensorID is attached to the data collected with the revised sensor. In practice, such changes are not always clear cut.

Perhaps the most challenging issue for the vision of e-Science with wearable sensors is the issue of participant privacy. Institutional Review Boards are justly conservative about the use and sharing of data from wearable sensors. Wearable sensors are often recording physiological data that could theoretically be used to evaluate a participant's health. In addition, the activity and context data from carried sensors (e.g. smartphones) may include information that could later be used to identify anonymized participants, similar to the previous use of IMDB and Amazon movie reviews to identify the identities of individuals in the Netflix dataset [17].

7 Conclusions and Future Work

Research projects using wearable sensors open the doors to more naturalistic (in situ) studies of physiological and other phenomena. These studies generate time streamed data sets that can be quite large and have the potential to be valuable for many different research goals. To enable reuse, we are designing and developing PerCon, an environment for managing and analyzing data sets including wearable sensor data. PerCon is designed to include storage, middleware, and application layer services.

The initial use of PerCon is as a research group library for physiological, psychological, and contextual data. Thus, the PerCon prototype includes data location, analysis, and visualization capabilities appropriate for the particular types of data being collected. These capabilities are built around an ontology representing the sensor types and data relationships found in this domain. Our current use of PerCon is aimed at identifying issues with our sensors and data capture process. Through continued use of PerCon in the context of health and cognitive monitoring, we will explore open issues in the design of repositories of wearable sensor data.

To extend the capabilities of PerCon and develop a full-fledged digital library system, we are exploring the workflow surrounding data collection, ingestion, and analysis. By adopting available metadata and repository communication standards, we will work on scalability and interoperability with other scientific digital libraries. There are number of open issues in the design of repositories of wearable sensor data. Due to these issues, we currently view/use PerCon as a research group-level library with an eye towards how limitations on data sharing might enable reuse while protecting participant privacy.

Acknowledgements

This work was supported in part by grant 10-49217 from the National Science Foundation.

References

1. Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludäscher, B., Mock, S. Kepler: An Extensible System for Design and Execution of Scientific Workflows, In: 16th Int. Conf. on Scientific and Statistical Database Management, (2004)
2. Baker, W., Broek, A. can den, Camon, E., Hingamp, P., Sterk, P., Stoesser, G., Tuli, M.A.: The EMBL Nucleotide Sequence Database, Nucleic Acids Research. 28 (1), 19-23 (2000)
3. Barga, R., Jackson, J., Araujo, N., Guo, D., Gautam, N., Simmhan, Y.: The Trident Scientific Workflow Workbench, In: IEEE Int. Conf. on eScience '08, (2008)
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W.: GenBank. Nucleic Acids Research. 37 (1), D26-D31 (2008)
5. Borgman, C.L., Wallis, J.C., Mayernik, M.S., Pepe, A.: Drowning in Data: Digital Library architecture to Support Scientific Use of Embedded Sensor Networks, In: 7th ACM/IEEE-CS Joint Conference on Digital Libraries (2007)
6. Brettlecker, G., Schuldt, H., Fischer, P., Schek, H-J.: Integration of Reliable Sensor Data Stream Management into Digital Libraries, In: 1st Int. Conf. on Digital Libraries: research and development.DELOS'07. pp. 66-76, (2007)
7. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Agosti, D.. SoergelM., Dobreva, M., Katifori, V., Schuldt, H.. 2007. "The DELOS Digital Library Reference Model Foundations for Digital Libraries. Version 0.98", The DELOS Network of Excellence on Digital Libraries. (December 2007).
8. Choi, J., Ahmed, B., Gutierrez-Osuna, R.. "Ambulatory Stress Monitoring with Minimally-Invasive Wearable Sensors". Technical Report (Nov. 1, 2010). Department of Computer Science and Engineering, Texas A&M University. (2010)
9. Choi, J., Gutierrez-Osuna, R.. 2010. "Estimating Mental Stress Using a Wearable Cardio-Respiratory Sensor", Proceedings of IEEE Sensors. (Nov. 1-4, 2010).
10. Fox, E. A., Hall, R., Kipp, NNDLTD: Preparing the next generation of scholars for the information age", New Review of Information Networking, Vol. 3, Iss. 1, pp. 59-76.
11. Gonçalves, M. A., Fox, E. A., Watson, L. T. , Kipp, N. A. . 2004. "Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries", ACM Transactions on Information Systems. Vol. 22 Iss. 2, (April 2004).
12. Google Health. http://www.google.com/health
13. Henneken, E. A., Accomazzi, A., Grant, C. S., Kurtz, M. J., Thompson, D., Bohlen, E., Murray, S. S.. 2009. "The SAO/NASA Astrophysics Data System: A Gateway to the Planetary Sciences Literature", 40th Lunar and Planetary Science Conference. (March 23-27, 2009).
14. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., Oinn, T.. 2006. "Taverna: a tool for building and running workflows of services", Nucleic Acids Research. (Jul. 1, 2006). vol. 34. No. Web Server issue, pp. W729-732
15. Li, H., Councill, I. G., Bolelli, L., Zhou. D., Song, Y., Lee, W-C., Sivasubramaniam, A., Giles, C. L.. 2006. "CiteSeerχ: a scalable autonomous scientific digital library", InfoScale '06 Proceedings of the 1st international conference on Scalable information systems.
16. McCray, A. T., Gallagher, M. E.. 2001. "Principles for digital library development", In Communications of the ACM. Vol. 44 Iss. 5 (May 2001).
17. McGrath, R. E., Futrelle, J., Plante, R., Guillaume, D.. 1999. "Digital library technology for locating and accessing scientific data", Proceeding of DL '99 Proceedings of the fourth ACM conference on Digital libraries.
18. Microsoft HealthVault. http://www.HealthVault.com/Microsoft/
19. Narayanan, A., Shmatikov, V.. 2008. "Robust De-anonymization of Large Sparse Datasets", Security and Privacy, 2008. SP 2008. IEEE Symposium on. (May 18-22, 2008).

20. Pantelopoulos, A., Bourbakis, N. G.. 2010. "A Survey on Wearable Sensor-Based Systems for Health Monitoring and Prognosis", Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. Vol. 40. Iss. 1, pp. 1 – 12. (Jan. 2010).

21. Pepe, A., Mayernik, M., Borgman, C.L., Van de Sompel, H. "From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web", Journal of the American Society for Information Science and Technology, Vol. 61 No. 3 (March 2010), pp. 567-582.

22. Shipman, F.M., Chaney, R.J., and Gorry, G.A. "Distributed Hypertext for Collaborative Research: The Virtual Notebook System", Proceedings of the ACM Conference on Hypertext, 1989, pp. 129-135.

23. Shipman, F.M., Hsieh, H., Airhart, R., Maloor, P., and Moore, J.M. "The Visual Knowledge Builder: A Second Generation Spatial Hypertext", Proceedings of the ACM Conference on Hypertext, 2001, pp. 113-122.

24. Shipman, F.M., Marshall, C.C. "Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems", Computer-Supported Cooperative Work, Vol. 8, No. 4 (Fall, 1999), pp. 333-352.

25. Smith, T. R. , Frew, J.. 1995. "Alexandria Digital Library". Communications of the ACM. Vol. 38 Iss. 4 (Apr 1995).

26. Staples, T., Wayland, R., S. Payette. "The fedora project: An opensource digital object repository system". D-Lib Magazine, vol. 9, April 2003.

27. Suchman, L. A.. 1987. "Plans and situated actions: the problem of human-machine communication". Cambridge University Press.

28. Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., Smith, M.. 2003. "The DSpace institutional digital repository system: current functionality", JCDL '03 Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. (May 27-31, 2003).

29. Tateno, Y., Fukami-Kobayashi, K., Miyazaki, S., Sugawara, H., Gojobori, T.. 1998. "DNA Data Bank of Japan at work on genome sequence data". In Nucleic Acids Research, 1998, Vol. 26, No. 1. Pp. 16-20.

30. Wallis, J. C., Mayernik, M. S., Borgman, C. L., Pepe, A.. 2010. "Digital Libraries for Scientific Data discovery and Reuse: From Vision to Practical Reality". In Proceedings of the 10th annual joint conference on Digital libraries (June 18-23).

31. Want, Roy. 2008. "You Are Your Cell Phone". In Pervasive Computing, IEEE. Vol. 7, Issue 2, pp. 2-4. (April-June 2008