PerCon: Support for Heterogeneous Data Management and Analysis via Mixed-initiative interaction

Su Inn Park
Texas A&M University
Department of Computer Science and Engineering
College Station, TX, 77843
joshsuihn@tamu.edu

Abstract

This research proposes a digital library system that allows users to manage a large number of heterogeneous datasets and to support a mixed-initiative interaction for data analysis among the related datasets. 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 datasets for analysis is more obvious. With the data explosion occurring in many areas, one crucial issue is the heterogeneity of data sources due to different data platforms and/or environments. Increasing interdisciplinary research and advances in devices, tools, and software for scientific data management generate more and more heterogeneous data. As the scientific community and industry face an increasing amount of diverse and interrelated data, data analysis is becoming more challenging to discover meaningful information and knowledge. Examples of scientific data management and analysis in various domains include digital libraries. However, the digital libraries to date have their uniqueness depending on specific research domains. Beyond a domain-specific data environment, this proposed digital library provides a general data infrastructure/platform applicable to various research fields and supports human-computer interaction-based data analysis. In order to achieve this, three important inquiries are made into the digital library: software requirements and capabilities for heterogeneous data management, a visual workspace environment for translating data into information and knowledge, and mixed-initiative framework for data analysis. A digital library system called PerCon is being developed as a substantial instance of a digital library rather than as a conceptual framework. PerCon is more than a typical digital library, as it integrates data management with data manipulation, presentation, and analysis capabilities. In the long term, the proposed digital library aims to explore the potential for data reuse in the more general field of heterogeneous data management and analysis.

1. INTRODUCTION

E-Science has the vision of sharing scientific data to support the reuse of data across different research groups. 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 datasets for analysis is more obvious. With the data explosion occurring in many areas, one of the crucial issues in e-Science is the heterogeneity of data sources due to different data platforms and/or environments. Increasing interdisciplinary research and advances in devices, tools, and software for scientific data management and the services generate more and more heterogeneous data.

As the volume and the need to share heterogeneous data have increased, long-term research on heterogeneous data management has been performed within the database community. As a traditional approach for data management, the data repository and (relational) database have offered well-defined structures and schema to store and access the original data objects as well as computed/filtered datasets including metadata [11]. In addition, many interdisciplinary areas have contributed to managing various types of data. For processing and integrating the heterogeneous data collections, algorithms based on various models, transformations, and filtering methods have been explored [8].

As the scientific community and industry face an increasing amount of diverse and interrelated data, data analysis is becoming more challenging. Since many researchers have collected heterogeneous data, exploratory visualization techniques have been developed to provide interaction with large and complex datasets and to give users broad bandwidth of understanding interpretation [10]. With the visualization strategies, data manipulations, location, and interaction within systems have been integrated for data analysis [15]. However, research on heterogeneous data management and analysis has been relatively less undertaken especially for a systematic and integrated environment/infrastructure.

The objective of this proposed research is to develop a digital library system that allows users to manage and analyze a large number of diverse but interrelated datasets. Ultimately, the proposed digital library system aims to support mixed-initiative interaction for data analysis among the related heterogeneous datasets and to explore the potential for data reuse in the more general field of data management and analysis. This proposal describes:

- designing the repository and the database to manage structured/unstructured datasets and encode multiple scientific standards in metadata,
- introducing a software architecture and application for services that can be extended beyond the initial domains of data management,
- processing and integrating different forms of data (e.g. time-series, geospatial, textual, multimedia data),
- providing a Visual Knowledge Builder (VKB) [12] workspace in which a user and the system can explore and interpret information visibly in multiple presentations (e.g.

temporal, thematic, and spatial composition), and discover knowledge from data, and

- analyzing data via mixed-initiative interaction by browsing and locating related data objects/sources beyond the query or menu selection.

The rest of this research proposal is organized as follows. Section 2 describes related work in the fields of data management, analysis, and interaction within digital libraries. Section 3 describes the proposed methodologies and system. The current status and plan of the proposed system are provided in Section 4 and 5, respectively. The evaluation plan is discussed in Section 6. Finally, Section 7 addresses the contributions.

2. RELATED WORK

2.1 Digital Libraries for Data Management

There have been many proposed architectural approaches to digital libraries to manage data. As an amount of the data and the number of data types are increasing, several architecture approaches to (distributed) data store and processing have been researched. Digital libraries such as Massively Parallel Processing Databases [7], Hadoop-based digital libraries [16] are examples for big data management including data processing.

With regards to scientific data management, digital library instances in various research fields have been developed. Based on data integration, researchers on bioinformatics, as a representative data-intensive science, have developed databases and computational/statistical analysis tools to explore large-scale genome sequencing. Genbank [3] is an implementation of huge databases functioning as a type of fine-grained digital library system. Geography datasets have also been incorporated in domain-oriented data libraries. For instance, the Alexandria Digital Library [14] provides search services from collections of geographically referenced materials. This breadth has motivated the use of a combination of domain-independent ontologies, domain-dependent ontologies, and personal ontologies for representing the potential relationships within datasets.

2.2 Digital Analysis in Digital Libraries

For data analysis, computational, statistical, visualization-based, and human-computer interaction-based methods has been applied in digital libraries. Primarily, visualization in digital libraries helps to explore data/information and to lead to analytical reasoning for data interpretation and analysis in significantly increasing data size and complexity. In particular, associated with diverse data sources, various scientific data have been processed using computational and statistical methods and then visualized within digital libraries. For example, Bernard et al. [4] proposed metadata visualization with respect to content-based similarity to lead the user to find relationships between data items. As an effort toward visual representation of textual data, Abbasi and Chen [1] applied a linguistic feature-based visualization technique for analysis and categorization. Beyond type-dependent visualization methods, Curdt et al. [6] managed and visualized heterogeneous scientific data in multiplicity of interdisciplinary subprojects.

Compared to computational/statistical/visualization-based data analysis within digital libraries, data analysis through human-computer interaction is not still largely explored while interaction modeling/framework [5], or interaction-based applications for design, configuration, and task planning have been articulated to date.

2.3 Three Approaches to Interaction

Interactions between a user and the digital library system falls into three major approaches – interaction of system control, human control, and human-system interactive control. First, system-controlled interaction is employed in automated systems that take full control and guide users to perform a task. Systems, such as Google, are representative examples of automated digital libraries for information discovery/retrieval. The second approach for interaction is a direct manipulation, where users take initiative in terms of manipulating, displaying, and evaluating data. Direct manipulation is supported by well-designed interfaces, functions, and data visualization. For example, various systems researched by Shneiderman [13] approached direct user manipulation with user interfaces and visualization techniques. Finally, as a flexible and collaborative interaction, mixed-initiative interaction combines automated services with direct manipulation to provide various while a user performs a task. Conventionally, the system agent(s) computes related data and information according to the user's responses and infers user's interests. Scheduling system, Lookout [9], and AIDE for exploratory data analysis [2] are examples.

3. Approach

For heterogeneous data management and analysis, a digital library system called PerCon (the system has been initiated for **Per**sonal/**Con**textual data environment) is being developed as a substantial instance of a digital library rather than as a conceptual framework. PerCon is more than a typical digital library, as it integrates data management with data manipulation, presentation, and analysis capabilities. This section proposes architectures, interfaces, management support, and mixed-initiative interaction for data analysis in PerCon.

3.1 Design of Architecture and Capabilities

The underlying design of PerCon is inspired by a layered architecture for managing the interconnections and interoperations among the diverse software components and data resources collected from other institutions. As Figure 1 illustrates, the system architecture constitutes three layers: resource layer, middleware layer, and application layer. The figure also shows the core capabilities at each layer and the components within it.

The Resource Layer provides functionalities to store and preserve the original data objects, computed and filtered datasets, and metadata featuring a repository and a database. To focus on data-specific challenges and applications, the database includes descriptive metadata and thumbnails of the digital objects by supporting consistent and effective access. In addition, semantic data integration occurs. However, since any heterogeneous data rarely fits into PerCon database, the repository is designed to include loosely connected data sources for heterogeneous data environment. The repository stores and manages the raw or processed digital objects and indices of the data objects. In addition, in this layer, a web server accesses data resources in the repository and database to support web-based services, such as search. Finally, the resource layer also encompasses information about data provenance for data control access and domain-specific data processing.

The Middleware Layer of the architecture provides four major functionalities: data ingestion, data access, visualization, and data analysis. The first functionality, data ingestion, provides three services: data processing, data integration, and provenance

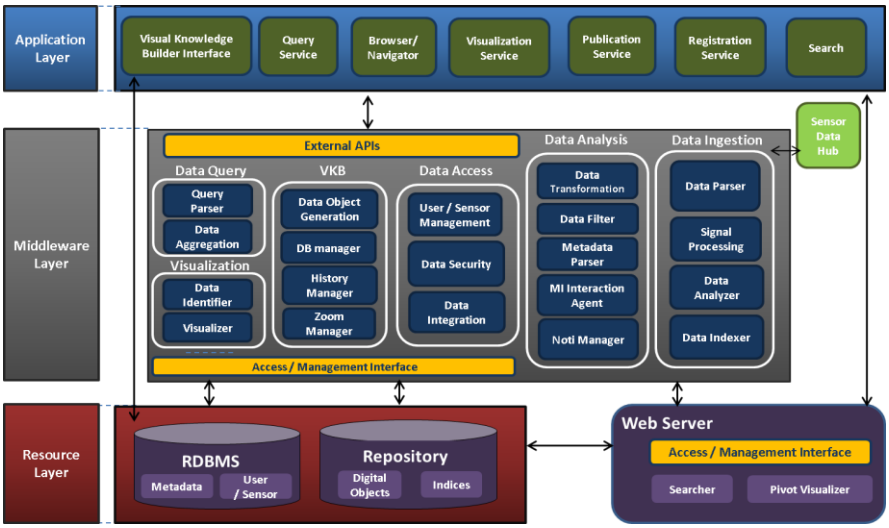

Figure 1: PerCon's architecture and software components

recording. The processing service includes a data reliability check to support the services in the application layer; this ensures that the data object content is well formed for the data object type. The processing service also provides support for processing raw data to generate computed data objects and tables/matrixes, encode metadata, and build the initial knowledge base accessed by the analysis framework. In particular, the computed data/knowledge tables include information about initial relationships between data objects. In turn, the integration service performs association of the independently captured and processed digital objects and metadata based on the knowledge base. The domain ontology contains relevant knowledge for sharing between heterogeneous domains and has a potential to be transferred to other integration processes for semantic integration. Finally, the provenance service extracts metadata regarding the data provenance and the description of the process of data generation, if it is not raw data. The second functionality of the Middleware Layer is to enable application layer components to access and interact with the resource layer's components through a set of external APIs. A query-processing module parses a query and determines the communication necessary with the resource layer. The third functionality of the Middleware Layer is data visualization, which instantiates data as a visual object form on VKB workspace depending on data types or queries. The visualization functionality includes various visual and spatial attributes to represent visual objects on the workspace. The individual object instances have their own threads for supporting data manipulation. The final functionality of the Middleware Layer is data analysis, which is driven by an analysis agent. This agent connects data objects and knowledge tables in the Resource Layer to states and variables related with user and workspace, ranging from the highly interactive to the highly automated. The agent for data analysis functionality supports mixed-initiative interaction between a user and PerCon for detecting a variety of dynamic relationships depending on queries, exploration, and events of users and analytics, rules, and knowledge base of the system.

Finally, the Application Layer enables end-users and external systems to access the content in the digital library. This layer implements user interfaces for browsing, searching, and visualizing the contents of the resource layer. It also incorporates a data registration interface that enables insertion, update, and deletion of datasets. Furthermore, the Application Layer implements a publication interface that enables remote access to certain contents of the digital library.

3.2 System Interfaces

The proposed system interface is structured as the rationale of architecture design and capabilities. PerCon constitutes three straightforward interface components: a repository browser, a metadata viewer, and a metadata workspace to support low-level to high-level data management. The interface is the repository browsing interface, which organizes the digital objects as a hierarchy. Thus, raw and computed data objects are found together. The interface for browsing metadata examines the metadata associated with the individual data files being examined. The interface allows users to browse the metadatasets that match the current data type selection and to view relationships between the datasets. Finally, PerCon embeds VKB with its features including a history mechanism to facilitate the comprehension of visualized information and formalize interrelated knowledge. It supports a variety of visualizations for individual data types. As a user requests or queries data, the results appear as new visual objects in the workspace, and then the result properties and relationships are determined with regard to user-system interaction. This enables users not only to view and manipulate a variety of heterogeneous data objects simultaneously, but also to integrate and interpret formalized information among the visually represented data objects. The VKB workspace is a shared workspace for a user and system.

3.3 Heterogeneous Data Management

With low-level data management based on the architecture, the proposed system offers capabilities and functionalities as a high-level heterogeneous data management platform using the VKB workspace.

When a user accesses heterogeneous data, PerCon provides data management tools and applications on individual data objects. For example, editing/plotting tools, web browsing, multimedia data streaming application, database tables, and timeline are instantiated on the VKB workspace through query, data manipulation, and interaction. In particular, search engine based on IR mechanisms can be represented and the search results can be organized on the workspace. In many cases, metadata standards and structured languages are employed for encoding characteristics of the data to enable data comprehension and to reuse across projects. Hence, PerCon also features structured standard tools and applications such as an XML and JSON viewer.

By extending previous VKB data presentation capabilities, PerCon provides a model for visually representing data according to different requirements using appropriate individual applications on the VKB workspace. Diverse data visualizations across multiple dates, users, data types, etc. can be explored and browsed. As the amount of data objects increases in size and number, PerCon presents both overview and details in overview

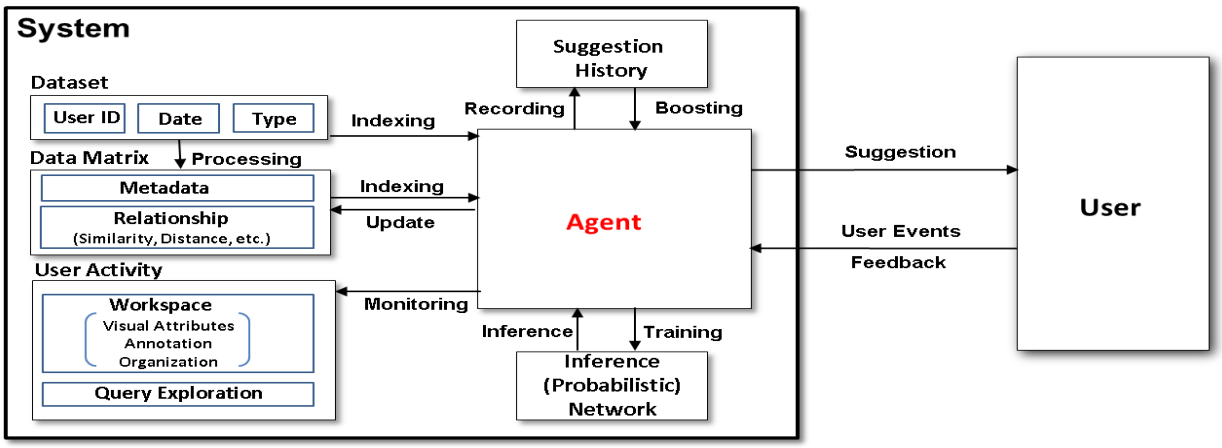

Figure 2: Design of mixed-initiative interaction framework in PerCon

so that users can explore data. For example, PerCon employs a coloring technique for representing a large quantity of data in an array. Also, a calendar type visualization that maintains the data provenance of creation and exploratory information is provided. The controls on the calendar type data object enable access to detailed visualizations, which contains hyperlinks to individual data, as well as dataset overview.

PerCon serves as a multi-parameter intelligent monitoring system in a synchronized manner. As the data often is recorded from different sources in parallel, an integrated visualization of the various signals is necessary to identify patterns/relations as well as form and assess new hypotheses. For example, PerCon augments the time-series plots with textual annotations of the data or textual data collected. Or, multiple time-series datasets are plotted in one synchronized application. The integrated visualization helps to find correlations between data streams. Moreover, correlations and other relationships that are not seen or detected visually in a single parameter view can be evaluated. This allows users to visually observe interrelated changes. Thus, PerCon supports more investigative analyses and various hypotheses testing.

To extend data management capabilities as a more general infrastructure for heterogeneous data management, PerCon provides the external APIs to create different applications on the workspace. Other types of data with required management tools/applications can be integrated and accommodated in the workspace with less effort. This design rationale enables fine-grained data manipulations, adaptation of new types of data to visual workspace, and maximizing bandwidth utilization.

Finally, PerCon provides methods and platforms for presenting various data and formalizing information for data management at the human level. When applications according to the data types and queries are instantiated and structured on VKB workspace, PerCon enables complete translation of data into information presentation as well as integration of potential knowledge in multi-disciplinary datasets. Consequently, PerCon allows users to translate heterogeneous data into information for possible knowledge discovery.

3.4 Mixed-initiative Interaction Data Analysis

A key challenge to increasing data analysis ability via interaction with PerCon depends on knowing relations among large amount of heterogeneous datasets and recognizing what user's activities and interests are focused on in the given tasks or hypotheses. This is caused by the fact that user actions and preferences play an important role in representing user's information needs. Thus, PerCon adopts highly collaborative interaction, or mixed-initiative interaction for data analysis. An analysis agent in PerCon integrates ontologies including encoded metadata and data data/feature tables (e.g., similarity, correlation, and clusters in various domains). Then, the agent associates the knowledge bases with user events and feedbacks to guide the search for relevant patterns and similarities between datasets in (semi)automated methods. The agent monitors user's behaviors with data objects including activities of information analysis in the VKB workspace, and then matches the behaviors on pre-defined user event types. Simultaneously, the agent continues to check user feedback, whether suggestions are accepted or rejected. The agent uses the user's activities and feedback to model constraints/variables/states which are features for employed machine learning algorithms. Along with the machine learning techniques, the agent detects patterns of data and suggests those interrelated/correlated data sources most related to user activities.

The mixed-initiative interaction agent employs dynamic and adaptive methods for inferring user's task-related intentions and interests which are evolving. PerCon continually attempts to improve its ability to provide valuable contributions by performing background learning so that it tracks the user's interests or work practices as they change over time. The VKB attributes (e.g., object annotation, spatial arrangement, and visual attributes) and user's events are applied on probability networks/models and correlation rules. In addition to user activities/events on data objects and the attributes, suggestions are made to the users based on the data/feature tables which are another key variables. As new data are ingested, the agent updates and monitors state-of-the-data matrix/tables and connects user events to data objects. Based on these internal learning and computational processes, when the agent infers a relationship between the observed variables and data formalization, it helps a user to test hypotheses as confirmatory data analysis. In particular, to manage the shifts and make the relationships among the user's event sequences lucid, the agent (1) estimates different levels of user goals with event sequences and (2) measures confidence level of inference to determine whether or not suggestion/recommendation is made for data analysis.

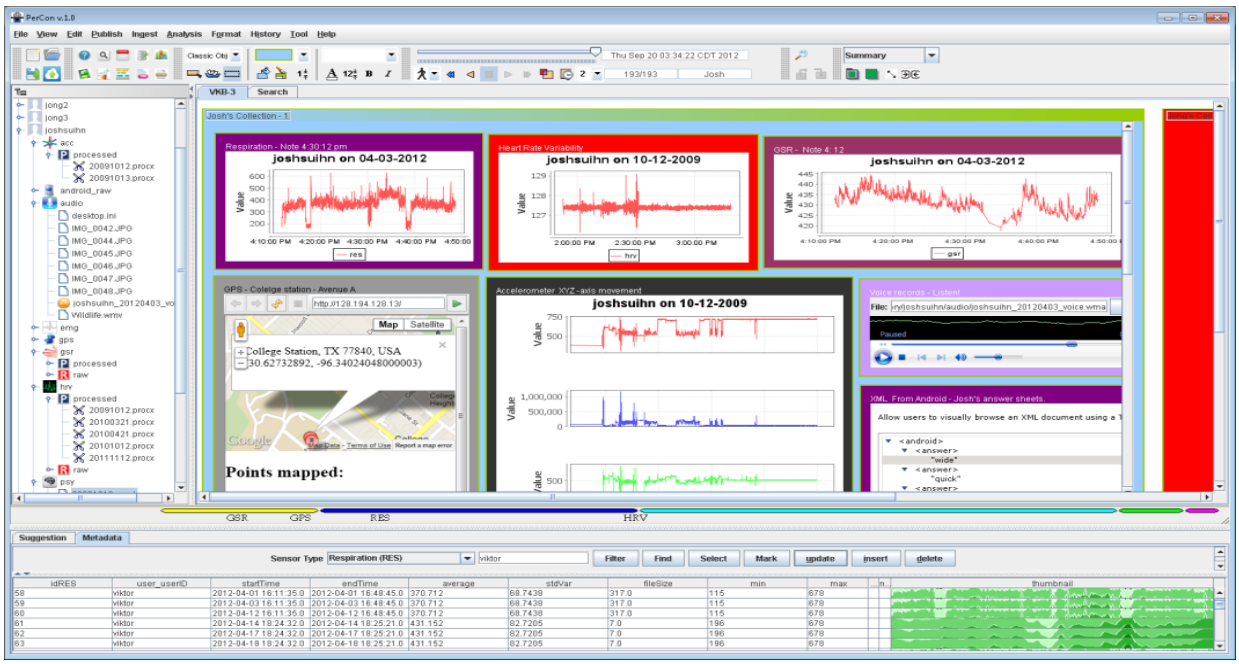

Figure 3: Example of PerCon instantiation with diverse wearable sensors and mobile device data

Consequently, via the mixed-initiative interaction in PerCon, the data analysis component allows the users to deeply explore data, refine data formalization and the analysis results, and incrementally formalize user-driven information space in an (partly) autonomous way. The user-generated information space is not a static collection of information, but rather a dynamic resource that may increase its capacity of knowledge. The interactive information formalization allows the information space to be converted into knowledge space, which gives other users potential answers as well as new perspectives/insights into the structured information. Furthermore, the agent helps to formulate and assess hypotheses and to enhance system usability.

4. CURRENT STATUS

PerCon is being instantiated for this research by developing a desktop application that includes storage, middleware, and application layer services using Java, C#, Javascript, and Matlab technology. In addition, several services in PerCon are empowered with web-based accessibility. Since the architecture of PerCon is designed, frameworks and components in each layer for data ingestion, diverse services and applications on VKB and interfaces have been instantiated, as Figure 3 shows.

5. FUTURE WORK

PerCon's mixed-initiative interaction framework is being instantiated based on user's intentions/activities/events, VKB attributes, and knowledge bases. The significant variables and states attainable with user activities are being designed and developed. In addition, various data/knowledge tables for connecting the variables and the states to relevant data objects are being newly created based on similarity, time translation, transformation, and clusters. Thus, to maintain working memory of recent interactions with users, the internal database in the framework for recording the variables is being largely extended. The so-called agent in the framework will include networks and mechanisms that let users guide the system and allow them to be guided for sharing understanding. The agent will primarily harness probabilistic networks to reason/infer user's attention, intention and task under uncertainty. PerCon's agent for mixed-initiative interaction ties inference networks to the mechanisms where the system can search data or relationships associated with the user activities. To make the agent/framework adaptive to user's response, continuous learning by observing user's activities and feedback will allow the system to present better responses to users.

6. EVALUATION PLAN

Considering the overall effect of various components and design, data management capabilities will be evaluated through user studies. Referring to the system workflow, the user studies will constitute a series of tasks: ingesting data, locating and searching indexed data objects and metadata to explore its contents via user interface, instantiating data objects in visual representation, identifying salient features in visualized data, assessing data correlation issues, formalizing information to explain and interpret data, and reporting results and findings.

To evaluate mixed-initiative interaction geared towards data analysis, a simple conjecture will be made: PerCon with an agent for data analysis is more effective and useful than without the agent. To prove this conjecture, first, in the aspect of human problem-solving processes, a sequence of data access, manipulation, exploration, and practices on the workspace without agent will be examined under given tasks. An individual user may have different exploration patterns and habits when performing a given task. Categorizations of the patterns or rules will be presumed and verified as a preliminary work for inferring user's attention/interests/tasks. Then, when a subject works on given tasks, how subjects interact with the system agent will be observed. The frequency of collaborative interactions time duration for completing tasks, and frequency of user's

acceptances/rejections will become direct metrics for evaluating the effectiveness and usefulness of the agent.

7. CONTRIBUTIONS

This research opens the door to more naturalistic (in situ) studies of integrated data management and interactive data analysis environment focusing on heterogeneity of datasets. The datasets can be quite large and have the potential to be valuable for many different research goals. An effort to mixed-initiative interaction capability will help to discover knowledge and to find rules that determine interrelationships between heterogeneous datasets. This study is also exploring the data/work practices surrounding data processing/manipulation, information organization, human-computer interaction. Furthermore, by adopting available metadata standards and repository communication, PerCon will contribute to scalability and interoperability with other scientific digital libraries. PerCon design and capabilities will provide insight into requirements for cyberinfrastructure in other fields such as medical and social sciences.

8. REFERENCES

[1] Abbasi, A., Chen H., 2007, Categorization and analysis of text in computer mediated communication archives using visualization, Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 11-18

[2] Amant, R., Cohen, P.,Interaction with a mixed-initiative system for exploratory data analysis, 1997, Proceedings of the 2nd international conference on Intelligent user interfaces, pp. 15 – 22

[3] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W., 2008. "GenBank", Nucleic Acids Research. Vol. 37. Iss suppl. 1. pp. D26-D31

[4] Bernard J., Ruppert T., Scherer M., Kohlhammer J., Schreck T., 2012, Content-based layouts for exploratory metadata search in scientific research data, Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, pp. 139-148

[5] Bryan-Kinns, K., Blandford, A., 2000, Harold Thimbleby Interaction Modelling for Digital Libraries, Procceedings of Workshop on Evaluation of Information Management Systems,

[6] Curdt, C., Hoffmeister, D., Jekel, C., Brocks, S., Waldhoff, G., Bareth, G., 2011, TR32DB - Management and visualization of heterogeneous scientific data, IEEE 19th International Conference on Geoinformatics, pp. 1-6

[7] Davis, E., 1983, Application of the massively parallel processor to database management systems, Proceedings of national computer conference, pp. 299-307

[8] Domenig, R., Dittrichm, K., 2000, A query based approach for integrating heterogeneous data sources, Proceedings of the ninth international conference on Information and knowledge management, pp. 453-460

[9] Horvitz, E., 1999, Principles of mixed-initiative user interfaces, CHI '99 Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 159-166

[10] Luo, W., MacEachren, A., Yin, P., Hardisty, F., 2011, Spatial-social network visualization for exploratory data analysis, Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, pp. 65-68

[11] Pirahesh, H., Mohan, C., Cheng, J., Liu, T., Selinger, P., 1990, Parallelism in relational data base systems: architectural issues and design approaches, Proceedings of the second international symposium on Databases in parallel and distributed systems, pp. 4 - 29

[12] Shipman, F., Hsieh, H., Maloor, P., Moore, J. M., 2001, The visual knowledge builder: a second generation spatial hypertext

[13] Shneiderman, B., Feldman, D., Rose, A., Grau X.G., 2000, Visualizing digital library search results with categorical and hierarchical axes, Proceedings of the fifth ACM conference on Digital libraries, pp. 57 – 66

[14] Smith, T. R. , Frew, J., 1995, "Alexandria Digital Library". In Megazine, Communications of the ACM. Vol. 38 Iss. 4

[15] Weaver, C., 2010, Cross-Filtered Views for Multidimensional Visual Analysis, IEEE Transactions on Visualization and Computer Graphics, Vol. 16 , Iss. 2, pp. 192 – 204

[16] Zhang, Y., Zhang, R., Chen, Q., Gao, X., Hu R., Zhang, Y., Liu, G., 2012, A Hadoop-based Massive Molecular Data Storage Solution for Virtual Screening, Seventh ChinaGrid Annual Conference, pp. 142 – 147