My Desktop
Prepare & Submit Proposals
Proposal Status
Proposal Functions
Awards & Reporting
Notifications & Requests
Project Reports
Submit Images/Videos
Award Functions
Manage Financials
Program Income Reporting
Grantee Cash Management Section Contacts
Administration
Lookup NSF ID

Preview of Award 1640864 - Annual Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1640864
Project Title:	CIF21 DIBBs: EI: Vizier, Streamlined Data Curation
PD/PI Name:	Oliver A Kennedy, Principal Investigator Juliana Freire, Co-Principal Investigator Boris Glavic, Co-Principal Investigator
Recipient Organization:	SUNY at Buffalo
Project/Grant Period:	01/01/2017 - 12/31/2019
Reporting Period:	01/01/2017 - 12/31/2017
Submitting Official (if other than PD\PI):	Oliver A Kennedy Principal Investigator
Submission Date:	12/28/2017
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	Oliver A Kennedy

Accomplishments

*** What are the major goals of the project?**

The high-level goal of the project is to develop and integrate an end-to-end software stack, called Vizier, for data exploration, iterative curation, debugging, and data re-use. The Vizier stack links GProM (a fine-grained data

provenance system), Mimir (a system for uncertainty-provenance), and VisTrails (a workflow provenance system), and a front-end interface that combines the best features of spreadsheet and notebook data management systems.

The proposed task-specific deliverables are:

1. Production quality releases of GProM and Mimir
2. A provenance and ambiguity-aware data processing system supporting queries and updates
3. A notebook-style UI for VisTrails
4. A language specification for Vizier's DSL and compliant parser
5. A prototype front-end for Vizier
6. A back-end with support for the Vizier DSL
7. A beta version of Vizier
8. Connectors between Vizier and external data sources
9. Extended library of automated data curation operations and quality evaluation modules
10. Automatic provenance-based recommendation module
11. A full release of Vizier
12. A comprehensive set of version tree manipulations required by exploratory curation

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities:

Specific Objectives:

After a year of effort, we are converging on a prototype implementation of Vizier. We are ahead of schedule in some respects (Tasks 8, 9), and have pivoted on several points (Tasks 2, 4, 8).

Tasks 1, 6: We have implemented a more extensive testing infrastructure around both Mimir and GProM. Significant effort was taken to document both projects, and in particular to provide developer guidance documentation. We have developed an experimental branch of Mimir linking it with GProM, and replacing signifincant functionality with equivalent features in GProM. The resulting infrastructure is described in the products section of this report.

Task 2: As originally described, Task 2 consists of two subtasks: 1) Linking GProM and Mimir, and 2) enabling support for Re-Enactment queries within Mimir. The first subtask has been completed as originally planned. At this stage of the project we are working exclusively with static data. Hence, hence re-enactment based on database audit logs, as originally implemented in GProM, is hard to test. However, we do allow updates to the data through the Vizier UI. To retain provenance information, these updates are only applied virtually through what are effectively re-enactment queries. In summary, we have not added full-stack support for re-enactment yet, but have re-implemented it at the UI/Mimir level.

Tasks 3, 5: A notebook/spreadsheet-style UI for VisTrails is in progress. We have iterated through several conceptual prototype interfaces and settled on a tentative interaction model. We are in the process of implementing this user interface and a server component linking it to VisTrails. This interface is described in the products section of this report.

Task 4: Rather than implementing a complete language (language model, parser, etc...), we decided to use the VisTrails workflow model as an interface between the user-interface layer and the uncertainty-aware curation backend.

We have implemented a Mimir plugin for VisTrails, as described in the products section of this report, producing a full software stack consisting of VisTrails, Mimir, and GProM.

Task 8: We have started porting Mimir/GProM to Spark. This link will allow us to leverage Spark's existing data source connectors (HDFS, Relational DBs, Local Data, etc...), as well as capture a broader range of Vizier workflow stages.

Task 9: Since the grant was awarded, we have added lenses for missing key detection, annotation, explicit user-provided options, and key repair. We have also added an infrastructure called Adaptive Schemas for managing heuristic data cleaning with non-deterministic schemas.

Tasks 7, 10-12: These tasks are scheduled for Years 2-3

Significant Results: Published papers as listed in the products section of this report.

Key outcomes or Other achievements:

*** What opportunities for training and professional development has the project provided?**

This project has directly supported the training and professional development of 7 PhD students and 1 MS student, as well as 2 undergraduate students through our REU supplement. The project has indirectly contributed to the training and professional development of 5 MS students through UB's CSE-662 "Languages and Runtimes for Big Data", a project-based class that featured 2 Mimir-related projects and through independent study projects.

*** How have the results been disseminated to communities of interest?**

Research-related efforts have been disseminated through top research conferences, as described in the products section of this report.

Code artifacts are available through our GitHub site: https://github.com/VizierDB

*** What do you plan to do during the next reporting period to accomplish the goals?**

Our primary aim for the next year is to get a prototype in front of our collaborators, in particular analysts at the Center for Urban Science and Progress. We will continue bulletproofing efforts on the full software stack, and establish a public-facing analytics platform.

Products

Books

Book Chapters

Inventions

Journals or Juried Conference Papers
View all journal publications currently available in the NSF Public Access Repository for this award.

The results in the NSF Public Access Repository will include a comprehensive listing of all journal publications

recorded to date that are associated with this award.

Pimentel, João Felipe and Murta, Leonardo and Braganholo, Vanessa and Freire, Juliana. (2017). noWorkflow: a tool for collecting, analyzing, and managing provenance from python scripts. *Proceedings of the VLDB Endowment*. 10 (12) 1841 to 1844. Status = Deposited in NSF-PAR doi:10.14778/3137765.3137789 ; Federal Government's License = Acknowledged. (Completed by Freire, null on 12/20/2017) Full text Citation details

Zacharatou, Eleni Tzirita and Doraiswamy, Harish and Ailamaki, Anastasia and Silva, Cláudio T. and Freiref, Juliana. (2017). GPU rasterization for real-time spatial aggregation over arbitrary polygons. *Proceedings of the VLDB Endowment*. 11 (3) 352 to 365. Status = Deposited in NSF-PAR doi:10.14778/3157794.3157803 ; Federal Government's License = Acknowledged. (Completed by Freire, null on 12/20/2017) Full text Citation details

Spoth, W. and Arab, B. S. and Chan, E. S. and Gawlick, D. and Ghoneimy, A. and Glavic, B. and Hammerschmidt, B. and Kennedy, O. and Lee, S. and Liu, Z. H. and Niu, X. and Yang, Y.. (2017). Adaptive Schema Databases. *CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*. . Status = Deposited in NSF-PAR Federal Government's License = Acknowledged. (Completed by Glavic, null on 12/17/2017) Full text Citation details

Yang, Ying and Kennedy, Oliver. (2017). Convergent Interactive Inference with Leaky Joins. *Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017*. 366-377. Status = Deposited in NSF-PAR doi:10.5441/002/edbt.2017.33 ; Federal Government's License = Acknowledged. (Completed by Kennedy, null on 12/18/2017) Full text Citation details

Arab, Bahareh Sadat and Gawlick, Dieter and Krishnaswamy, Vasudha and Radhakrishnan, Venkatesh and Glavic, Boris. (2017). Using Reenactment to Retroactively Capture Provenance for Transactions. *IEEE Transactions on Knowledge and Data Engineering*. PP (99) 1 to 1. Status = Deposited in NSF-PAR doi:10.1109/TKDE.2017.2769056 ; Federal Government's License = Acknowledged. (Completed by Glavic, null on 12/17/2017) Full text Citation details

Licenses

Other Conference Presentations / Papers

Other Products

Software or Netware.

Community Clean product developed in year 1:

What it is:

A web based front-end for Vistrails/Mimir stack to which data cleaning workflows assembled in Vistrails that may employ heuristics that introduce uncertainty can be deployed
A public facing source for feedback (corrections) on data that is annotated with uncertainty
A facility for accepting feedback to uncertainty in data sets with mapped GPS locations, or in tables of data
Where it is:
https://community.mimirdb.info

Questions it may help with:

How do general users interact with/respond to uncertainty annotations on data sets?
For what type of data sets can cleaning tasks be crowd-sourced?
Is feedback from general users reliable? and how can we prioritize the reliability of user feedback?

Can we generate some interest from local government or organizations to leverage this for existing cleaning tasks?
Technical brief:

Vistrails with the Mimir package is used to create workflows and deploy them to the web front-end.
A Web server (Jetty) dispatches user actions/requests to Mimir (with Sqlite backend).
The Web client uses Scala.js, react, material-ui, d3.js, and google maps for UI and uses web sockets with seamless ajax fallback for communication.
Feedback sources (users) are identified uniquely using a browser fingerprinting technique.

Software or Netware.

The Vizier Server has four components. The main component is the workflow engine. Other components are the data store, a simple file store, and a Web Server.

The workflow engine manages all aspects of maintaining and executing workflows that manipulate tabular datasets. The engine maintains the history of all workflows and allows to create and manage workflow branches. The workflow engine combines concepts from Vistrails and Mimir. At this point we support workflows that contain three type of modules: (1) commands in our VizUAL language to manipulate datasets (i.e., LOAD DATASET, INSERT COLUMN, INSERT ROW, DELETE COLUMN, DELETE ROW, MOVE COLUMN, MOVE ROW, RENAME COLUMN, and UPDATE CELL), (2) a selection of Mimir Lenses (i.e., KEY REPAIR, MISSING KEY, MISSING VALUE, PICKER, SCHEMA MATCHING, and TYPE INFERENCE), and (3) Python scripts.

The data store provides access to all datasets that are generated by Vizier workflows. This includes access to all historic versions of the datasets that were created by different versions of a worklfow. We also provide a simple Python interface to the data store which enables Python scripts in workflow modules to access and manipulate datasets directly. Furthermore, the data store maintains annotations for datasets. Annotations are (key, value)-pairs that are associated with identifiable components of a dataset, i.e., columns, rows, or individual cells.

The file store maintains text files that are used as input to data curation workflows.

The Web Server provides a RESTful API to access the full functionality of the Vizier Server via HTTP requests.

The front end interacts with the Web API and provides a web based User Interface (UI). The UI allows creation and deletion of data curation worklfows. Workflows are manipulated using a notebook-style interface. The interface allows users to insert, modify and delete cells that represent the modules of a workflow. A second component of the UI is a spreadsheet interface. The spreadsheet allows users to directly modify an dataset. Interactions with the spreadsheet are translated into modifications of the underlying workflow. The spreadsheet is currently limited to inserting and deleting column and rows, renaming of columns, and updating of individual cells.

The Vizier Server is available through the GitHub open source software repository: https://github.com/VizierDB/frontend and we will soon deploy a publicly acessible demonstration version.

Software or Netware.

We have produced a release version of Mimir+GProM available at http://mimirdb.info and a corresponding plugin for VisTrails available at https://github.com/VizierDB/Vistrails

Other Publications

Patents

Technologies or Techniques

Thesis/Dissertations

Websites

Participants/Organizations

Research Experience for Undergraduates (REU) funding

Form of REU funding support:	REU supplement
How many REU applications were received during this reporting period?	2
How many REU applicants were selected and agreed to participate during this reporting period?	2
REU Comments:	

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Kennedy, Oliver	PD/PI	1
Freire, Juliana	Co PD/PI	1
Glavic, Boris	Co PD/PI	1
Brachmann, Mike	Other Professional	12
Rampin, Remi	Other Professional	1
Mueller, Heiko	Staff Scientist (doctoral level)	1
Aggarwal, Shivang	Graduate Student (research assistant)	1
Arab, Bahareh	Graduate Student (research assistant)	9
Feng, Su	Graduate Student (research assistant)	12
Huber, Aaron	Graduate Student (research assistant)	12
Niu, Xing	Graduate Student (research assistant)	4
Ota, Masayo	Graduate Student (research assistant)	6
Spoth, William	Graduate Student (research assistant)	3

Name	Most Senior Project Role	Nearest Person Month Worked
Yang, Ying	Graduate Student (research assistant)	3
Alphonce, Olivia	Research Experience for Undergraduates (REU) Participant	3
Alsabbagh, Amer	Research Experience for Undergraduates (REU) Participant	6

Full details of individuals who have worked on the project:

Oliver A Kennedy
Email: okennedy@buffalo.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Organized effort. Supervised UB Team.

Funding Support: n/a

International Collaboration: No
International Travel: Yes, Italy - 0 years, 0 months, 6 days

Juliana Freire
Email: juliana.freire@nyu.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Managed the NYU effort and contributed to the tasks described in the proposal

Funding Support: Moore Sloan Data Science Environment

International Collaboration: No
International Travel: No

Boris Glavic
Email: bglavic@iit.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Managed IIT efforts and GProM backend development.

Funding Support: N/A

International Collaboration: No
International Travel: Yes, Germany - 0 years, 0 months, 13 days

Mike Brachmann
Email: mrb24@buffalo.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 12

Contribution to the Project: Senior Research Developer at UB. Integrated Mimir and GProM, Developed Mimir plugin for VisTrails, and Developed CommunityClean. Oversees development of Mimir.

Funding Support: n/a

International Collaboration: No
International Travel: No

Remi Rampin
Email: remi.rampin@nyu.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 1

Contribution to the Project: Remi is the main developer of VisTrails and has implemented the extensions required by Vizier

Funding Support: NSF (this award) and Moore Sloan Data Science Environment

International Collaboration: No
International Travel: No

Heiko Mueller
Email: heiko.mueller@nyu.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 1

Contribution to the Project: Heiko has contributed to the development of the Vizier backend and frontend.

Funding Support: NSF (this grant) and Moore Sloan Data Science Environment

International Collaboration: No
International Travel: No

Shivang Aggarwal
Email: shivanga@buffalo.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Implemented mechanism for prioritizing sources of uncertainty affecting query results using a linear solver. Work was initially intended for the NPS award listed below, but Shivang also devoted some time integrating the resulting tool into the Community Clean product.

Funding Support: Work partly supported by NPS Award #N00244-16-1-0022

International Collaboration: No
International Travel: No

Bahareh Sadat Arab
Email: barab@hawk.iit.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 9

Contribution to the Project: Developed provenance management for updates using reenactment which is an important functionality for Vizier's regretfree exploration capabilities.

Funding Support: N/A

International Collaboration: No
International Travel: No

Su Feng
Email: sfeng14@hawk.iit.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 12

Contribution to the Project: Development of uncertainty annotation support in GProM and investigated the theory behind uncertainty annotations.

Funding Support: N/A

International Collaboration: No
International Travel: No

Aaron Huber
Email: ahuber@buffalo.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 12

Contribution to the Project: Professional preparation and training. Conducted research on models for annotating queries, in particular with uncertainty-tracking markers, as part of a collaboration with the IIT team.

Funding Support: n/a

International Collaboration: No
International Travel: No

Xing Niu
Email: xniu7@hawk.iit.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 4

Contribution to the Project: Developed optimizations for annotated query processing which will be deployed in Vizier.

Funding Support: Oracle unrestricted gift

International Collaboration: No
International Travel: No

Masayo Ota
Email: mo1123@nyu.ed
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 6

Contribution to the Project: Masayo is exploring automatic methods for data cleaning

Funding Support: NSF (this grant) and Moore Sloan Data Science Environment

International Collaboration: No
International Travel: No

William Spoth
Email: wmspoth@buffalo.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 3

Contribution to the Project: Prototype implementation of Adaptive Schemas. Schema extraction prototypes. Implemented prototype schema inference knowledge-base. Working on porting Mimir to Spark. Other core contributions to Mimir.

Funding Support: Summer 2016 effort contributing to products related to this proposal supported by a gift from Oracle Effort on NPS Award #N00244-16-1-0022 synergized with efforts on this project. No support charged to this award.

International Collaboration: No
International Travel: No

Ying Yang
Email: yyang25@buffalo.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 3

Contribution to the Project: Developed infrastructure in Mimir supporting incremental model evaluation.

Funding Support: Work relevant to this award entirely supported by a gift from Oracle

International Collaboration: No
International Travel: No

Olivia Alphonce
Email: oalphonc@buffalo.edu
Most Senior Project Role: Research Experience for Undergraduates (REU) Participant
Nearest Person Month Worked: 3

Contribution to the Project: Integrated MatPlotLib into Mimir, making it possible to automatically annotate graphs with uncertainty markers.

Funding Support: n/a

International Collaboration: No
International Travel: No
Year of schooling completed: Junior
Home Institution: University at Buffalo
Government fiscal year(s) was this REU participant supported: 2017

Amer Alsabbagh
Email: aalsabbagh@hawk.iit.edu
Most Senior Project Role: Research Experience for Undergraduates (REU) Participant
Nearest Person Month Worked: 6

Contribution to the Project: Studied automated materialization of versioned data generated by a curation workflows and implemented a prototype of this idea. This work will be incorporated into Vizier to allow the system to decide automatically what (intermediate) results to materialize when the user updates a workflow

Funding Support: N/A

International Collaboration: No
International Travel: No
Year of schooling completed: Sophomore
Home Institution: Illinois Institute of Technology
Government fiscal year(s) was this REU participant supported: 2017

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
Oracle	Industrial or Commercial Firms	California

Full details of organizations that have been involved as partners:

Oracle

Organization Type: Industrial or Commercial Firms
Organization Location: California

Partner's Contribution to the Project:
Financial support
Collaborative Research

More Detail on Partner and Contribution:

What other collaborators or contacts have been involved?

Over the award period, we received unsupported contributions to Mimir from

- Nick Cellino
- Vandit Aruldas
- Sneha Krishnamurthy
- Rakshit Muthappa Padetira
- Anand Sankar Bhagavandas

Impacts

What is the impact on the development of the principal discipline(s) of the project?

- We developed new mechanisms for reasoning about heuristic data cleaning processes, and in particular processes who's schemas depend on heuristics. We proposed a model, called Adaptive Schemas that allows us to generalize query typechecking into a broader class of possible schema lookups.
- We proposed and evaluated a query evaluation pipeline for anytime approximations of graphical inference problems.
- We proposed a sandboxed environment for encoding simultaneous changes to schema and data, and for modeling provenance of records in this new environment.

What is the impact on other disciplines?

Nothing to report as yet, but we will be deploying Vizier to partner researchers in adjacent disciplines in the near future.

What is the impact on the development of human resources?

This project contributed directly or indirectly to the training and professional development of 7 PhD students, 2 undergraduates, and 7 masters students.

What is the impact on physical resources that form infrastructure?

We have produced a public release of Mimir+GProM, integrated Mimir into the VisTrails ecosystem, and developed a prototype notebook+spreadsheet data curation interface.

What is the impact on institutional resources that form infrastructure?
Nothing to report.

What is the impact on information resources that form infrastructure?
Nothing to report.

What is the impact on technology transfer?
Nothing to report.

What is the impact on society beyond science and technology?

Nothing to report.

Changes/Problems

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.