Proposal Status | MAIN

Organization: SUNY at Buffalo

Panel Summary #1

Proposal Number: 1640864

Panel Summary:
Panel Summary

Panel Summary for DIBBS Early Implementation Proposals

Proposal # / PI: 1640864/ Kennedy

The proposal is to build a system called Vizier that incorporates, integrates and extends the capabilities of three systems developed previously by the PIs. The proposed system would use the two well known interfaces of code notebooks and spreadsheets, into a versioned data curation system.

Intellectual Merit

Strengths

The proposal leverages prior NSF investments and products by the PIs so there is good value added for those investments and the PIs are well acquainted with the software already. It will fill an important need in the community and add value to the data involved for other uses while enhancing the user workflow within the domain. The spreadsheet/notebook interface design was well received by the panel although there was some discussion about the efficacy of actually putting the two together. There was a good rationale and comparison to existing tools. The panel liked the curation of data while the data are being explored and the general support of scientific workflows in data curation. The integration of curation activities with data exploration and prior history was a strength.

Weaknesses

The actual merging of the two interface paradigms has potential risk as does the proposed schema merging. It would have been nice to have the visualization/curation support some of the normal steps in data use such as identification of outliers.

Broader Impacts

Strengths

This is a generic tool with broad applicability and the final product will go to GitHub. There are user surveys fully incorporated. The educational outreach is broad including students from high school to PhD and the scientific community at large with an emphasis on under-represented groups and a plan to apply for REU support.

Weaknesses

Additional Review Criteria for Early Implementation Awards:

- What are the science outcomes described in the proposal? Are they innovative and made possible by the development? How are outcomes tied to grand challenges, and of interest to and involving multiple science and engineering domains? Are the science outcomes possible given the team and work plan?

The focus here is on the production of tools. There are not really well defined use cases although the panel did agree that the tools actually would apply to a wide variety of use cases. The science case in the social sciences advancements directly is incremental but data science will benefit.

- How does the implementation expand and contribute to the set of resources that serve the community? Are the components extensible and potentially useful to other communities? Is there a clear description of the data, software, or standards that will be produced by the project? (Software is intended in this instance to refer to scientific analysis, visualization or modeling tools necessary to achieve scientific outcomes).

The project does expand the set of useful resources available to the community, especially in the social sciences. The data and software are well described.

- Is the management plan and team appropriate for the goals of the project? What is the plan to demonstrate the proposed capability or resource?

The panel felt that there was a good management plan. The panel recognized and applauded their candor in putting the workplan into terms of time of effort rather than calendar time.

- Characterize the community that will benefit from the project: How many researchers and which domains will directly benefit from the outcomes of the project? How does the project involve and serve more than one research field? Are participants from various communities explicitly identified, and are their roles clear? How does the project clearly demonstrate end user involvement in development and use of a community capability?

They have several brief use cases primarily in the social sciences. The resources being integrated are already popular so we can expect broad acceptance and use.

- Indicate how the community is represented in governance of the resulting capability, including data management and deaccession. A sustainability plan must be included describing how any capabilities developed by the implementation project could be supported beyond the award duration. This may include integration into long-term data or cyberinfrastructure resources either supported by NSF or other institutions, agencies or partners. Sustainability plans will be evaluated on the viability of the sustainable resource, community representation in governance, the fit to the infrastructure being developed, and the likelihood of ingestion into the long-term system.

Their sustainability plan is highly reliant on acceptance and use by the community and there doesn't seem to be a method of governance of the final resource so this part of the proposal could have been stronger.

Overall Panel Recommendation: C

This summary was read by the panel and the panel concurred that this summary accurately reflects the panel discussion.

Back to Proposal Status Detail

National Science Foundation
4201 Wilson Boulevard, Arlington, Virginia 22230, USA
Tel: 703-292-5111, FIRS: 800-877-8339 | TDD: 703-292-5090

Proposal Status | MAIN

Organization: SUNY at Buffalo

Review #1

Proposal Number:	1640864
NSF Program:	DATANET
Principal Investigator:	Kennedy, Oliver A
Proposal Title:	CIF21 DIBBs: EI: Vizier, Streamlined Data Curation
Rating:	Very Good

REVIEW:

In the context of the five review elements, please
evaluate the strengths and weaknesses of the proposal with respect to intellectual merit.

The proposal outlines two areas of innovation - integrating a code notebook (Jupyter) as frontend to a workflow engine (VisTrails) and integrating heuristics into the data preparation process to take advantage of the versioned workflows across data projects. The first, with the spreadsheet paradigm incorporated, provides an easy to use and widely adopted interface to systems that are more difficult for researchers to use effectively. In addition, the capture of fine-grained provenance during data preparation is also valuable (and timely). The second, workflow versioning and heuristics, is critical for data quality concerns and reproducibility.

I do wonder if the integration of spreadsheet functionality into the notebook interferes with the notebook's utility. They are different models of interaction, but it's more whether the notebook functionality, predominately through code, will be reduced (and reduced enough to make the notebook irrelevant, ie it's possible to create spreadsheet functionality through a web interface without it), how the code is integrated in the workflow output and how the documentation cells are handled in the same. They discuss integration with data systems but not code. It is a concern in that it is not discussed in the proposal but is very much about how people interact with the Jupyter system.

In the context of the five review elements, please
evaluate the strengths and weaknesses of the proposal with respect to broader impacts.

The system is applicable beyond the initial urban science collaborations and so has the potential to be adopted in any domain that requires data curation of tabular data. A tool providing detailed and legible, provenance for data quality assessment before or after publication should be of interest across a broad set of research activities. Using a known interface such as the Jupyter notebook platform, encourages adoption - it is familiar to users and reduces the barrier to entry.

Although the proposal overall is well-structured, I note that the reference to storing workflows in a cloud system was not described further. It is not clear if that was mean to be part of the platform and thus a development task or a general statement (as a user, you can post your workflow to some cloud environment unconnected to Vizier proper).

Please evaluate the strengths and
weaknesses of the proposal with respect to any additional solicitation-specific review criteria, if
applicable

This may fall under a proof of concept rather than an implementation. The proposal is well-reasoned and was guided by initial requirements gathering. The resources should be adequate given that the collaborators have been involved with most of the components and have developed the timeline and integration activities accordingly.

It lacks much in the way of multiple disciplines as collaborators or as an initial user base to demonstrate the real-world use of the system. This may be less of an issue within the project assuming usability testing is undertaken; however, it is perhaps an issue for the specific RFP. I am unclear on the need for the detailed descriptions of the Airbus data lake and CUSP with the lack of specific research goals related to those efforts.

Summary Statement

The proposal outlines a system to use two known interfaces, code notebooks and spreadsheets, into a versioned data curation system. This encourages "good" data curation practices in collecting detailed provenance information and in the development of repeatable workflows that can then be used, through the recommendation engine, to guide future data preparation activities.

Back to Proposal Status Detail

National Science Foundation
4201 Wilson Boulevard, Arlington, Virginia 22230, USA
Tel: 703-292-5111, FIRS: 800-877-8339 | TDD: 703-292-5090

Proposal Status | MAIN

Organization: SUNY at Buffalo

Review #2

Proposal Number:	1640864
NSF Program:	DATANET
Principal Investigator:	Kennedy, Oliver A
Proposal Title:	CIF21 DIBBs: EI: Vizier, Streamlined Data Curation
Rating:	Very Good

REVIEW:

In the context of the five review elements, please
evaluate the strengths and weaknesses of the proposal with respect to intellectual merit.

Strengths:
o Addresses important issues of data curation and data quality, allowing for informed decisions about flagging or down-weighting bad or suspicious data.
o Recognizes strength of both notebook-style and spreadsheet-style interactions with data, bringing them together in a common interface.
o System will infer curation actions based on prior user selections and edits.
o Iterative and incremental approach to implementation is sound and provides interim deliverables with useful capacity.

Weaknesses:
o At the start of Section 2 it is proposed that the Vizier system will support schema merging, but this is not discussed further except briefly (using Data Tamer).
o I might have expected the tools for curation to include multiple views of data value distributions, for example, as a means for ascertaining likely data entry errors.

In the context of the five review elements, please
evaluate the strengths and weaknesses of the proposal with respect to broader impacts.

Strengths:
o Addresses broadly based challenges in data quality in social/economic/demographic data sources.

Weaknesses:
o It is not clear that the system would be applicable to experimental data.

Please evaluate the strengths and
weaknesses of the proposal with respect to any additional solicitation-specific review criteria, if
applicable

Strengths:
o Implementation plan has been thoroughly thought out. Clear statements of milestones and deliverables. Work plan covers three years, reflecting effort required and not an attempt to propose up to the limit of the program.

Weaknesses:
o The user survey (Section 1.3) only had 8 people! That's a pretty limited sample upon which to draw conclusions driving the design of a major software system.
o The role of the students is rather ambiguous ("responsible for R&D components of the proposal.").

o The sustainability plan is really just a community engagement plan, and does not really address long-term financial stability.

Summary Statement

This proposal aims to deliver an integrated data curation system that builds upon existing components It will have a user interface that exhibits features of both a notebook/workflow system and a spreadsheet.

Back to Proposal Status Detail

National Science Foundation
4201 Wilson Boulevard, Arlington, Virginia 22230, USA
Tel: 703-292-5111, FIRS: 800-877-8339 | TDD: 703-292-5090

Privacy and Security

Proposal Status | MAIN

Organization: SUNY at Buffalo

Review #3

Proposal Number:	1640864
NSF Program:	DATANET
Principal Investigator:	Kennedy, Oliver A
Proposal Title:	CIF21 DIBBs: EI: Vizier, Streamlined Data Curation
Rating:	Very Good

REVIEW:

In the context of the five review elements, please
evaluate the strengths and weaknesses of the proposal with respect to intellectual merit.

Strengths

The investigators have identified and defined a need in the area of data curation i.e. carrying out data curation strictly as a pre-processing task can be problematic when constraints are discovered at the time of data analysis. They propose to build the Vizier system as a tool for integrating data curation with the data exploration process by using provenance.

Vizier will build on existing tools and leverages NSF investments in CI. The core building blocks, Mimir, GProM, and VisTrails, have been developed by the proposal team. In addition to the primary team there is a strong group of collaborators, representative of the domains, who will provide guidance, feedback, assistance in design and evaluation.
There is already a broad community across domains and they have good plans to grow the community. The ongoing and planned projects represent a broad set of different domains-from urban to business intelligence.

The outreach plan to grow the community is active and includes workshops, publicly released code with good documentation for both users and developers, and demonstrations.

The proposal provides a detailed design rationale and implementation and comparison to existing tools. They have already received and used feedback from the community on the new design. For example, they have selected the spreadsheet metaphor based on user feedback. They have also applied principles of good user interface design in their plans e.g. retaining a sense of user control through both automation and recommendation functions.

Plans for management, evaluation and sustainability are all provided in the proposal. The management plan, along with the timeline and the collaboration plan, provide a detailed list of tasks to be completed and assignment of tasks to the team or to post-docs, students, developers. The different levels of meetings and workshops among the team suggest that there will be clear and timely communication even across these teams that are not co-located.

Weaknesses

It seems that there is a lot of new implementation required to integrate these systems and provide the Vizier functionality. Some details are provided but there is risk due to the need to have all working for the integration to be successful. It is noted, however, that they set out a plan for iterative development to try to mitigate the risk.

The sustainability plan is primarily one of user buy-in. Although there is a broad community and some history of industry support of investigators, this may not be sufficient to keep the resource going.

In the context of the five review elements, please
evaluate the strengths and weaknesses of the proposal with respect to broader impacts.

Strengths

There is potential to impact many different domains even beyond those immediately involved in the proposed work.

The investigators have described educational outreach that is broad including not only students from high school to PhD but also the scientific community at large. There is a stated commitment to involving underrepresented groups and a plan to apply for REU support.

Weaknesses
None noted

Please evaluate the strengths and
weaknesses of the proposal with respect to any additional solicitation-specific review criteria, if
applicable

A scientific outcome will be a better understanding of provenance. There will also be advances in the state of the art in data curation and advances in research in fields in the social sciences, particularly in urban science.

The implementation builds on 3 systems through integration to support curation through provenance.

The various communities will be involved in design, implementation, testing, and through outreach but there doesn't seem to be a method of governance of the final resource. The plan is to put the final Vizier system on GitHub but how will maintenance and collaborations continue beyond the period of the grant?

Summary Statement

Overall, this is a strong proposal based on a well-defined need and a detailed plan. It fits the DIBBS program in that it extends existing resources and there is a good integration of existing and new domains and user communities. There is some risk that not all of the technical components of the implementation which are needed for the integration will be successfully completed in a timely manner. The iterative development plan, however, will likely support a large degree of success. Sustainability will rely on success and strong engagement by end users. The broader impacts are strong for both expanding to scientific domains and inclusive education.

Back to Proposal Status Detail

National Science Foundation
4201 Wilson Boulevard, Arlington, Virginia 22230, USA
Tel: 703-292-5111, FIRS: 800-877-8339 | TDD: 703-292-5090