CIF21 DIBBS: EI: VIFI: **V**irtual **I**nformation-**F**abric **I**nfrastructure for Data-Driven Decisions from Distributed Data (NSF DIBBs Award #1640818) - ***White Paper: Challenges and Risk Mitigation Strategy***

Ashit Talukder, Principal Investigator, University of North Carolina at Charlotte

The stated objective of the **V**irtual **I**nformation **F**abric **I**nfrastructure For Data-Driven Decisions from Distributed Data (VIFI) is to level the playing-field for data-driven innovation by making previously un-shareable data more accessible. Commencing on Oct 1, 2016 and over the course of initial coordination among the collaborating institutions, the VIFI effort has initiated internal mechanisms (Wikis, source code management) to facilitate creation of an open-source toolkit. The team has also commenced exploration of a number of fundamental factors associated with the control, sharing and analysis of voluminous and/or sensitive datasets. While the project has commenced only recently, the team is aware of specific risks and are undertaking risk mitigation strategies to ensure the long term adoption, success and sustainability of VIFI.

Extensible Modular Architecture for Site/Owner Adoption: The infrastructure and use-case teams have identified a High Level Architecture (HLA) which includes a conceptual modularized architecture (comparing the suitability of a distributed, amorphous approach as opposed to a multi-layer system utilizing master/coordinator-slave node architecture) and a preliminary listing of tools (to include orchestration components, rules engines and other elements). Legacy views on software management complexity and its effect on total cost of ownership could pose an impediment to ViFi adoption by data site owners. In order to mitigate this problem, the ViFi platform will be distributed as both core components and additional distribution and provisioning tools that ease the burden of installation and configuration of software components that are optional extensions of the ViFi platform. Such provisioning will be composed to use current best practice for systems administration and security as it pertains to each component. In addition, all provisioning resources will be transparently accessible and distributed using multiple provisioning environment systems (e.g. Puppet, Chef, Ansible).

Data ontology coupling/management standards: Dataset owners may employ fragmented, non-uniform formats optimized towards specific processes and unique, non-standard ontologies may be employed in labeling data attributes. The dataflow capabilities of the ViFi platform are intrinsically tied to data ontology for search, provenance and governance. Management and resolution of ontology conflicts and ambiguities is an ongoing area of research. In order to mitigate problems caused by such ambiguities, the ViFi data ontology will be established using a version centric orientation on which a 'master' ontology branch will be routinely curated as the stable data ontology. Ontology resolution and disambiguation solutions using the team's Cornerstone tools, and Mined semantic analysis will be employed. For point of need additions to the ontology, such as inclusion of a new data domain or additional metadata previously not captured, 'experimental' ontology branches will be used until which time these new ontology entities can be evaluated for transition to the stable master branch. VIFI will demonstrate the benefits of data and ontology standardization through working examples generation through the use of VIFI.

Stakeholder Adoption: Potential users may express skepticism about sharing with others through VIFI. The solution will be to generate and utilize existing successful examples which adequately demonstrate the validity of the security mechanisms. Additional capabilities will be identified through discussions with potential stakeholders then validated, tested, maintained, backed-up and implemented in VIFI. Testing will be conducted through standard methods. VIFI will establish a web distribution capability for the software and the supporting data sets. Information about VIFI can be disseminated through talks, posters and live demonstrations. Groups can then be invited to join in with their datasets or abstractions thereof to try VIFI for themselves. Example avenues are meetings on different types of variables and/or surveys e.g. ZTF and LSST, the annual Astroinformatics conferences, ADASS conferences, Hotwiring the Transient Universe conferences, semi-annual IVOA meetings, semi-annual meetings of the American Astronomical Society, etc. An extensive and rich set of documentation with worked out use cases will be provided. Students can be trained through summer school programs and also assist in developing further VIFI capabilities.

ViFi platform updates and distributed version distribution conflicts: The ViFi team intend to mitigate the conflicts posed by system updates through strict adherence to continuous deployment development practices, RESTful versioning and contentious backwards compatibility.