Abstract

The Shared Big Data Gateway (*SBD-Gateway*) project harnesses innovative technology developed and piloted at Indiana University to provide a secure environment for data analysis on a national scale. The project will offer an out-of-the-box, open, cloud-hosted solution for libraries and researchers to analyze large datasets. Currently, many research libraries have purchased datasets like Clarivate Analytic's *Web of Science* bibliometric data, but lack the infrastructure, expertise, and resources to process the data in such a way that it is useful for their research communities. The *SBD-Gateway* project offers a low-cost, shared solution to this problem, allowing libraries to forgo the time and effort required to explore potential solutions, identify and hire appropriate experts, clean, parse, and prepare data for hosting, and secure data and monitor access. Membership and sustainability are central to the success of the project. In light of this, support has been secured from Indiana University, Michigan State University, Purdue University, University of Michigan, The Ohio State University, University of Minnesota, Penn State University, University of Iowa, Rutgers, the Big Ten Academic Alliance, all four NSF-funded national Big Data Hubs, Clarivate Analytics, and Microsoft Research. The project has documented interest in phase two and letters of support from the Greater Western Library Alliance, the Private Academic Library Network of Indiana, Northwestern University, and the University of Wisconsin.

During the initial phase of the project, a combination of proprietary licensed and open data identified as key for Big Ten Academic Alliance user communities, will be hosted: 1) *Microsoft Academic Graph (MAG)*, publicly available bibliometric data on over 160 million scientific records; 2) Publicly available patents, intellectual property, entrepreneurship, and innovation data from the U.S. Patent and Trademark Office (*USPTO*); 3) *Web of Science (WoS)* by Clarivate Analytics, proprietary bibliometric data spanning 100+ years 59+ million records and 900 million cited references.

The Indiana University Network Science Institute has created the first in the nation Secure Enclave for Critical Data to host a copy of the *WoS* dataset for use by Indiana University. The *WoS* data in raw and relational forms, which occupies about a Terabyte of space, is currently hosted in the enclave. To improve the efficiency of queries and analysis, we have designed the software to explore and parse the raw data into a relational database, with the flexibility to accommodate data expansions and updates.

The enclave is built to run sophisticated parallelized programs, with support for software like *R, SAS, Stata, SPSS, MatLab, Python, Node.js, open MPI* etc. It includes a user-friendly interface that guides users through the querying process. It also ensures that no data will leave the Enclave without the explicit authorization of the Data Steward. The Secure Enclave for Critical Data has recently achieved the highest security certification by the CDC and NVDRS. Additionally, the IU Network Science Institute has established a cloud-based Data Lake and has worked in collaboration with Microsoft Research to test frequent data updates and dynamic schema analytical tools. Initial experimentation has increased the speed of queries by nine times and spans nearly linearly across available resources.

The platform will host all current and future datasets in multiple formats and instances to ensure maximum versatility and impact of research questions and relevance to a broad range of researchers. This means we need a heterogeneous architecture design that can take advantage of multiple established cloud storage technologies (raw, data lake, relational databases, graph database etc.). We will open the existing enclave and data lake services to a small group of selected institutions for testing to help us better understand the community needs. Building on an already established federated *Single Sign On (SSO)* at IU, we will expand the current system to connect with all the available university and library existing authentication systems. We will iteratively refine the cloud architecture and data selection to improve user experience and sustainability based on user feedback. With the support of the Product Owner Council, the team will deliver a cost-effective cloud gateway for library researchers at the national scale.

The *SBD-Gateway* project will expand the data-mining capacity of libraries throughout the country and make a substantial contribution to the National Digital Platform. Because bibliometric analysis is a critical and longstanding research area in library science, the initial datasets hosted by the *SBD-Gateway* will directly impact practicing librarians as well as library and information science faculty and students conducting citation analysis as part of their research. We have received commitments from 9 libraries in phase one and 8 to continue providing financial support for the platform for at least three years after the grant ends. This provides strong indication of the projects sustainability and lasting success. Moreover, by creating a community of researchers and facilitating communication about the *SBD-Gateway* with partner libraries, we ensure the platform will be user informed and we anticipate a close collaboration between libraries and researchers will evolve from our efforts.

National Leadership Grant Project: Shared Big Data Gateway: A Cloud based infrastructure for Sharing Research Assets and Advancing Library and Information Science

Statement of Need

The Shared Big Data Gateway (*SBD-Gateway*) addresses the IMLS "National Digital Platform" priority by addressing a critical emergent issue faced by academic libraries: providing sustainable, affordable, and standardized data and text mining services for licensed, big data sets, as well as open and non-consumptive data sets too large or unwieldy to work within existing research library environments or with no commercially viable data mining interface. The project team, led by Indiana University, will provide member institutions with a cloud-based, platform solution for making such data available to members with the appropriate security, stewardship, and storage at a fraction of what it would cost them to do so alone. By sharing the cost of this solution across a large number of academic libraries, we will be able to provide a superior solution at a lower cost to members. We will also offer a free tier of basic services for public access. The *SBD-Gateway* will feature standardized data formats, data available in multiple formats including relational and graph database formats as well as flat tables and native formats, shared and custom/private computational resources, a space to share and store queries, algorithms, derived data, results of analyses, workflows, and visualizations. The project seeks $849,339 in grant support for this effort.

This project endeavors to build a community of practice in addition to and adjacent to the SBD-Gateway cyberinfrastructure. We will knit together communities of libraries, researchers, and data providers, seeking out and cultivating relationships between industry partners, researchers who work with the SBD-Gateway hosted data, and member libraries to create a community of stakeholders with mutual interests and investments. Researchers will benefit from *SBD-Gateway* features that facilitate collaboration with peers, and form strengthened connections to libraries and industry partners who will be learning about their needs and how to address them. Industry partners will benefit from an improved understanding of library and researcher needs, and will be able to adapt their products to better suit these needs. Libraries will benefit primarily through substantial cost savings. Additional benefits will accrue through the community of libraries that will emerge - together we will be well-equipped to identify and address other mutual challenges.

The *SBD-Gateway* provides critical infrastructure to address a common problem. Libraries are amassing datasets that do not fit neatly into traditional vendor frameworks or interfaces for making data available to libraries and their patrons. Thus, libraries are licensing datasets directly from vendors and working to provide services to meet analysis, visualization, and machine learning needs. Even when datasets are openly available, they may be so unwieldy to manage that they are virtually unusable to the communities that could otherwise benefit from them. Libraries often lack resources, funds, and expertise to build adequate services around these data. The data then remains inaccessible to the researchers. The *SBD-Gateway* proposes a solution: share expertise, infrastructure, and cost across member libraries and jointly develop solutions for the policy and security implications of these large and sometimes restricted datasets as a community.

This project builds upon theory and practice* in data mining services and data management communities in academic libraries. In particular, it leverages work on the role of non-consumptive reading and data enclaves in academic library collections and work on shared data service models in libraries. There is a strong theoretical basis underpinning the role of academic libraries supporting data and text mining, including roles as data quality hubs on campus.[1] Provisioning large datasets is a modern incarnation of the collection building and stewardship with which libraries have always been charged.[2]

Non-consumptive reading and data enclaves in academic libraries

The mission of the *SBD-Gateway* project is aligned with the HathiTrust Research Center's (HTRC) Data Capsule service, funded in part by the Andrew W. Mellon Foundation, Alfred P. Sloan Foundation, and IMLS (LG-71-17-0094-17). The HTRC Data Capsule has successfully operationalized a model for non-consumptive reading wherein users are able to generate derivative datasets from copyrighted texts in a secure environment. IMLS funding awarded in 2017 supports work extending data capture work beyond copyrighted works to other use cases for accessing restricted data for analysis. Conceptual and workflow elements mapped by Murdock et. al.[3] alongside ongoing work on the Data Capsule service is laying groundwork for technical and policy considerations at stake in operationalizing the *SBD-Gateway*.

* All references are available in the supplementary file 'Supportingdoc2.pdf'

Our work is also informed by the work of Lane and Shipp[4] who have detailed the development of provisioning remote access to the NORC Data Enclave, a collaborative effort between the National Opinion Research Center (NORC) and National Institute of Standards and Technology (NIST). The NORC Data Enclave disseminates business microdata for secure analysis. These efforts include a data sharing model, "a collaborative environment within which researchers can share knowledge" that could inform the *SBD-Gateway* plan to provide a Research Asset Commons.

The proposed project shares some elements with Cloud Kotta, a cloud-based social science data enclave developed with funding from the John Templeton Foundation to the Metaknowledge Research Network, IBM for Computational Creativity, and Facebook[5]. Cloud Kotta can host general datasets, including proprietary datasets such as the Web of Science, and is designed for big data analytic tasks. The *SBD-Gateway* will follow a much more selective data-centered framework, hosting only datasets identified as relevant to library science researchers. A large part of the proposed project is to analyze the real needs of users from participating libraries and institutions, and to foster a community around datasets of interest. The proposed project will also provide a more integrated and user friendly experience, serving not only data science experts, but also users with minimal big data research experience. The Cloud Kotta team have done important preliminary work developing architecture, integrating job management and security, and demonstrating that cloud-based solutions can be orders of magnitude more cost-effective at scale. Their code is also openly available and will serve as an important reference for the *SBD-Gateway* team.

Data management communities of practice

An extensive body of work documents data services in libraries and the development research data management and stewardship models[6,7]. While much of this work emphasizes open data, planning for management, and publication at single institutions, there have been significant advances in approaches to shared or distributed models in libraries. One recent and novel contribution is the Data Curation Network project, funded by the Alfred P. Sloan Foundation and led by Lisa Johnston, which brings together eight research-intensive institutions to build a shared staffing model for cross- institutional research data service support. Significantly, three of the eight participating institutions for that project have also pledged support to the *SBD-Gateway* project: University of Minnesota, University of Michigan, and Penn State University[8]. We will borrow extensively from this prior work to build our community of practice.

The significance of the problem that the *SBD-Gateway* addresses can hardly be overstated. For example, some libraries in the BTAA have purchased Clarivate's *Web of Science* data, which is delivered as raw XML data on physical hard drives. The process of cleaning, reformatting, hosting, securing, and managing these data requires highly skilled professionals with expertise specific to this type of data. This effort has been so daunting to libraries that for some, the purchased dataset has languished unused while a solution is sought and implementation is planned. This is unsurprising considering the complexity of making the data usable to researchers. An experienced data expert must plan, oversee, and manage the process, identifying and acquiring (through hiring if needed) staff with the appropriate expertise. The solution we propose is to pool financial resources and collectively purchase the necessary hardware, software, and human capital. This collaborative approach to provisioning data and text mining services will reduce the cost borne by participating libraries and convey the benefits of the service to many libraries as opposed to just one. Thus, for a fraction of the cost of building these services independently, a library can join the collective and derive a much greater value.

Member libraries and their patrons benefit from the expertise of a team who has hosted the *Web of Science* data and is experienced in cloud-based solutions. They also benefit by forgoing the time and effort required to explore potential solutions, identify and hire appropriate experts, clean, parse, and prepare data for hosting, and secure data and monitor access. Moreover, they gain access to additional datasets (i.e., *Microsoft Academic Graph*[9], *U.S. Patent and Trademark Office* data) that they might not otherwise acquire. The proposed project also builds a community of users who can use the same data with the same tools in the same format, allowing for reproducible research and identification of community needs. We expect the creation of a community of users will strengthen the efficiency, quality, and quantity of research by stimulating collaboration and community standards for the data hosted on the *SBD-Gateway* and the shared products derived from the data, including analyses, visualizations, code, and workflows.

Project Design

The goals of the *SBD-Gateway* project are to: 1) produce the initial *SBD-Gateway* platform with the capability to host as a minimum three datasets (e.g., *Web of Science, Microsoft Academic Graph, U.S. Patent and Trademark* data); 2) engage the library and research community to identify and prioritize additional features, datasets, and other improvements to *SBD-Gateway*; 3) attract enough library partners to sustain the *SBD-Gateway* for at least three years beyond the two year IMLS grant period; 4) create a community of active researchers who use the platform and value its role in facilitating their research and collaborations.

Initial datasets to be hosted on the cloud-based *SBD-Gateway* are those that Indiana University Network Science Institute (IUNI) already hosts on its local servers: 1) *Microsoft Academic Graph* (MAG), publicly available bibliometric data on over 160 million scientific records and three billion citations in JSON format; 2) Publicly available patents, intellectual property, entrepreneurship, and innovation data from the *U.S. Patent and Trademark Office* (USPTO); 3) *Web of Science* (WoS) by Clarivate Analytics, proprietary bibliometric data spanning 100+ years 59+ million records and 900 million cited references in XML format. These datasets have been identified by BTAA libraries as of interest to researchers across a broad range of scientific fields, yet there is no existing institution-based solution to enable shared hosting and data mining solution. IU is currently the only research institution hosting these same datasets for wide use.

We have the knowledge, experience, and expertise to provide a secure cloud-based solution to enable broader access. More datasets will be added to the *SBD-Gateway* in the second half of the project, based on suggestions from our Board and our initial user group. Possibilities include: *PubMed, ProQuest NewsPapers* (full text of 11 newspapers), *Oxford English Dictionary*, and *Core Logic* (Census and Property Data). As envisioned, the *SBD-Gateway* will be a cloud based resource featuring centralized, federated, single sign on, using existing university authentication infrastructure (*Shibboleth, Globus, InCommon*, etc.) to grant differential access to platform data according to institutional permissions. An authenticated user may use a web interface to query and retrieve available data from what we refer to as the *Research Asset Commons*. The Commons allows for sharing research assets using user-enabled privacy control, and/or access local tools to compute on *SBD-Gateway* data. All platform functions will access data and tools via an Application Programming Interface (API). The *SBD-Gateway* will host analytic tools and use appropriate data processing/querying software and will allow access to private cloud and local compute resources.

The biggest challenge associated with this project is sustainability, addressed in detail in the Sustainability section. Though the project offers a free tier, we expect that given our extensive support from Big Ten Academic Alliance Libraries, there will be sufficient subscribers to the service to subsidize a free tier for smaller institutions. We have discussed the *SBD-Gateway* project, its value and the annual cost of subscription per subscriber at a recent Board of Directors meeting of the BTAA. We have nine members who have pledged $35K each to support the project (cost share), eight of whom also committed to an agreement to subscribe for a three year period following the IMLS grant period. We have established that the cost savings of joining the collective vs. going alone to host the data is substantial (at least $50k per year saved), so we anticipate this support will continue.

Technical Description: Proof-of-Concept and Technology

The IUNI team has created the first in the nation Secure Enclave for Critical Data to host a copy of Clarivate's *Web of Science* (WoS) dataset for use by all the employees of Indiana University. We have developed the software to explore and parse the raw data into a relational database - an important advantage over the original XML – since it allows much faster queries and computation for analysis employed by heavy users of bibliometric data. The parsing process has to be repeated annually, because of vendor's improvements to previous editions. Using IU's vast supercomputer resources – an annual update can be completed in a matter of weeks. The *WoS* data in raw and relational forms is currently hosted in the enclave - a dedicated node of the IU Karst condominium cluster. The database itself occupies about a terabyte of space and takes advantage of all the power available to the node (currently 48 processing cores and 512 GB of memory). Being part of Karst allows researchers to use the modular software available on the cluster like *R, SAS, Stata, SPSS, MatLab, Python, Node.js, open MPI*, etc., and the Data Steward to submit parallelized requests to hundreds of nodes.

A project in its own right, the Secure Enclave for Critical Data is a secure environment in which users with programming skills can explore the data in their native form, while scientists can interact directly with the database, or use an easy-to-use web interface that guides them through the querying process. Hardened and accessed only through a secure remote desktop connection, no data can leave the Enclave without the explicit authorization of the Data Steward. It has recently achieved its highest security

certification by CDC and NVDRS and is currently also hosting the National Violent Death and Suicide Repository.

For the *Microsoft Academic Graph (MAG)* dataset – the IUNI IT team has set up a cloud-based Data Lake and has worked with Microsoft Research to ensure new versions of the data are deposited on a biweekly basis. At about half a terabyte each, they demand a significant amount of infrequently used storage that is very much in line with cloud storage and archival strategies. Currently the team is exploring cloud native technologies like dynamic schema Data Lake analytics Glue, Athena and U- SQL. Initial experimentation has resulted in query times 9-fold faster and span nearly linearly across available resources. These findings reinforce the idea that combining efforts to create a Shared BigData Gateway for Research Libraries will benefit the entire community and give all its members access to resources not available at individual or even institutional scale. As part of the project, IUNI will open access to its enclave and Data Lake Analytics cloud environment to a group of selected institutions to test current implementations and help better understand the needs of the community while scaling out to a full cloud implementation. We would build on our current achievements and combine established techniques to enable all current and future datasets to be available in multiple formats and instances (raw, data lake, relational databases, graph database etc.) to ensure maximum versatility and impact of research questions and cater to any audience. We are confident in our ability to scale up the current resources thorough a cloud implementation capable of serving a large audience of libraries and general users based on our past implementations of similar projects. One example will be IUNI's *Observatory on Social Media* (*OSoMe* http://osome.iuni.iu.edu) which uses a distributed compute cluster hosting a *No-SQL* database spanning more than 200 Terabyte to answer real time queries by journalists, scholars and the general public about information dissemination in social media.

Strategic Collaborations

We have identified the key stakeholders and partners necessary to succeed in developing the SBD-G and ensure its sustainability - Academic Libraries: Big Ten Academic Alliance (BTAA) Library Initiatives Team, the research libraries of Indiana University, the University of Iowa, the University of Michigan, Michigan State University, the University of Minnesota, the Ohio State University, the Pennsylvania State University, Purdue University, Rutgers University-New Brunswick , the Private Academic Libraries Network of Indiana (PALNI), Northwestern University Medical School Library, Greater Western Library Alliance (GWLA); Industry: Clarivate Analytics Inc. (*WoS* data provider), Microsoft Research (*MAG* data provider), Amazon Web Services (cloud service provider), Microsoft Inc. (cloud service provider); Researchers: International Society of Informetrics and Scientometrics, BigData in Academia: all four NSF-funded Big Data Hubs (Northeast, South, Midwest, and West).

Our combined team of librarians, computational, network science, and science of science researchers, big data experts, and cyberinfrastructure engineers aim to build an *SBD-Gateway* that can acquire, analyze, visualize, and share large proprietary as well as free and open data sets in a non-consumptive way. Non-consumptive sharing allows for access to derived results (e.g. prediction models built from specific or aggregated data sets) without exposing the raw data itself, thereby protecting the privacy of the participants where applicable and adhering to the terms of service of various vendors and institutions.

Regarding access by researchers, we envision a federated approach across campuses in which current tools, methods, datasets, and algorithms are brought together within a user-friendly, centrally coordinated environment, and are made available for use with both local and remote datasets and compute resources. For example, BTAA has successfully negotiated with the vendor Clarivate Analytics access for its members to a large store of bibliometric data by purchasing the rights to use the *Web of Science (WoS)* for anyone employed by the 14 universities. With over 65 million records and the total number of article/reference links above 1.2 billion, all delivered as raw XML files to the campuses – *WoS* presents a significant challenge to individual researches and a considerable burden to even the largest academic institutions. Other libraries and universities around the country have acquired the same data set negotiating individually with the vendor at an institutional, library or even an individual department or lab level. This requires unnecessary duplication of infrastructure to store that data as well as hiring the personnel required to process, parse, standardize, manage, and moderate and setup the appropriate compute environment. Our preliminary talks indicate many institutions support the idea of creating a centralized data store for this and other datasets based on the already proven implementation at Indiana University and allowing access to the rest of the members based on their collective or individually negotiated level with the vendor. We plan to

utilize the currently available tools and resources institutions nationwide have already built and committed. Through its *Research Asset Commons* this project will bring these resources and tools together to allow all parties access where appropriate while maintaining the restrictions applicable to each university's version and license of the data. This will enable us to provide a template for how future groups can share tools and data within the *SBD-Gateway*.

The *SBD-Gateway* will not stop at the proprietary dataset level, nor will it benefit exclusively a few large libraries. With the introduction of similar datasets like the *Microsoft Academic Graph* – an open source bibliometric dataset that measures more than twice the size of the *Web of Science*, the free tier of *SBD-Gateway* will bring large- scale bibliometric research to every library across the nation. Expanding further with more open datasets like the *US Patent Data (USPTO), PubMed* and others – suggested by researchers and approved by the Product Owner Council, *SBD-Gateway* will transform to serve a dual purpose of providing access to data, analytics, and visualization at size and scale not achievable by smaller institutions, while also providing researchers with the unique opportunity or working with multi-sourced data, combining and analyzing heterogeneous environments and creating multilevel network and correlations allowing for much more complex research questions. Creating a repository of bibliometric data at each institution would be not only costly, but difficult and would not scale easily to future data sources. The proposed approach brings all the data to cloud and creates interfaces between datasets to permit access to distributed datasets in the same query, thereby simplifying the task for the user. Analysis tools will be stored in the cloud as web applications and ready to deploy standalone "docker containers" that allow software to run across platforms. This feature will facilitate the execution of analysis workflows orchestrated by the gateway on any platform, enabling researchers to leverage remote and potentially larger hardware resources than are currently available. The proposed cloud-based approach will also feature a web interface, to access and execute the analytical tools with *Research Asset Commons* to share the results. It will connect to a backend data service to access local and remote data sets through a consistent API. Direct accesses to the API, will ensure that technologically advanced users will be able to use all their available tools, workflows and programming skills in a way not constrained by the graphical interface while taking full advantage of Research Asset Commons features.

The *SBD-Gateway* will be released as an extensible, open-source platform in which collaboration for future development will be welcomed. Through this collaboration, we will build a framework of best practices for sharing of data and tools so that more institutions can join our efforts in the future.

Project Timeline: *Year 1, Q1/Q2, needs assessment, initial design*: IU will immediately make the current IU server-based *WoS, MAG*, and *USPTO* data and tools available to a limited number (10-20) of researchers at partner institutions. We will collect Agile "user stories" from these users as well as from Product Owner Council members. We will design and deploy a test version of *SBD-Gateway* in the cloud. *Year 1, Q3/Q4, development*: We will convene an in-person board meeting to review progress, recommend adjustments, and chart future directions. In September 2019, we aim to convene a workshop at International Society of Scientometrics and Informetrics (ISSI) to introduce additional researchers to the platform, solicit input on design, and select additional datasets. *Year 2, Q1/Q2 testing*: Researchers, including Xiaoran Yan, IUNI Research Scientist, will evaluate the functionality of the test platform by running a wide range of queries and analytics against all databases. The Product Owner Council will review progress in online meetings. *Y2 Q3/Q4 refine platform, add datasets, develop sustainability plan*: We have received support from all four NSF-funded Big Data Hubs (East, South, Midwest, and West) and will improve performance and user experience per inputs from a second user workshop at an all hands meeting of the Midwest Big Data Hub. The AB will meet in person to learn about progress, recommend improvements and data, identify additional user communities, and to develop a sustainable cost model that is affordable and acceptable to the library partners.

Development Plan

Currently, the Secure Enclave for Critical Data holds the raw files of a single bibliometric dataset that has been parsed into a relational database. The Enclave has the ability to perform sophisticated queries via command line tools, SQL statements, and a guided web interface. This service is being used by 48 researchers at Indiana University. Additionally, the Web of Science databases hosted by the Indiana University Libraries have a reported 261,853 queries in 2017. It is possible that some researchers using this service are conducting bibliometric research that could benefit from the *SBD-Gateway* project. The Secure Enclave for Critical Data is essential for ongoing bibliometric research at Indiana University. The proposed project would both enhance the existing set of features and services and make the service broadly available

to research institutions throughout the country. The project will enable researchers to: share results with the community, connect simultaneously, collaborate with other institutions, and cross-examine multiple datasets.

This project will improve upon and enrich the proven Enclave concept, expanding it with technologies that could not be used in a standalone server implementation. The development tasks necessary to scale the *SBD-Gateway* will be as follows: 1) provide access to the Web of Science Enclave and Microsoft Academic Graph databases hosted at IU to *SBD-Gateway* project partners on a temporary basis until a cloud-based *SBD-Gateway* is established; 2) create a federated security login system to utilize each institution's own proprietary authentication systems; 3) develop *Research Asset Commons* – a shared space in which the researchers will be able to save and if desired share their algorithms, data subsets, derived results, tools and methods; 4) create and disseminate Docker container images for the analytics tools; 5) spec out the requirements for adding specific bibliometric and related datasets to the data query service and add one or more as time and cost allow; 6) spec out the requirements for a range of cloud specific technologies to augment existing data query technology and evaluate if they provide added utility at a reasonable cost 7) develop APIs to further extend the usability of data, databases and tools, 8) provide a user-friendly web interface to offer a collection of fast, well known searches as well as guide and suggest the most efficient approach to more complex, user defined queries and (9) develop a native visualization interface as well as APIs, exports and plugins to the most popular external visualization and presentation tools.

1. IUNI will work with members of the Product Owner Council to identify researchers and library institutions to be included in the initial share of the Web of Science dataset, relational database and web interface in the IU's Secure Enclave for Critical Data. We will also share Microsoft Academic Graph dynamic schema data lake implementation hosted by IU on a commercial cloud platform. We plan to utilize additional cluster nodes provided by IU Research Technologies the and the funds remaining funds in a current $10,000 worth of cloud research credit used to accommodate the additional usage.

2. Building on an already established federated *Single Sign On* (SSO) at IU, we will expand the current system to connect with all the university and library existing authentication systems like *Globus, InCommon, OpenID, OAuth* and others. The project team will work with each institution to ensure compatibility of their internal security systems including dual or multi-factor authentication. This will ensure continual institutional control over user management and satisfy any regulatory or dataset-specific security and privacy requirements and constraints. For example – *Web of Science* is a proprietary dataset to which multiple institutions have negotiated different kind and level of access. A federated identity management system will ensure that users' access level adheres to any vendor's negotiated license restrictions. For open and license-free datasets like the *Microsoft Academic Graph*, it will serve as an evaluation and monitoring mechanism. For smaller entities like rural, community, and tribal colleges that have not developed a central authentication, *SBD-Gateway* will provide direct user registration and authentication that will comply and conform to the established federated logon services.

3. A major part of the project is the development of the *Research Asset Commons* – a space in which users will be able to save their work and share specific sets of results with the community. Using the *SBD-Gateway*'s data management system they will be able to curate analysis results by adding rich metadata and sharing them with specific users, groups or the public, potentially including all the tools, parameters, data, and queries used by the original creator. Datasets in the *Commons* will receive DOIs and will be publicly visible when appropriate. Provenance information will be stored with each published dataset. Existing data visualization tools will be ported to the framework to provide meaningful previews of the results in the browser. This will create a virtual marketplace for sharing research and will promote cross-disciplinary collaboration. Users will be in control of saving data and results for their own consumption or sharing. Using a shared responsibility security model and support for user provided key encryption serve as an extra assurance for users' privacy and control. We envision that the marketplace will evolve in an easy to use sharing platform for multiple tested, proven, ready to use research sub-datasets, algorithms and tools in a tag-based searchable system. Furthermore – metadata will be used in automated, AI based suggestions of the best algorithms and tools to be used on a common queries and datasets.

4. Docker containers and Kubernetes orchestration using private virtual machine clusters during testing (Jetstream, Carbonate) and commercial cloud providers (EKS, AKS) are at the heart of the project to ensure elasticity of resources, scalability, platform independence, and compartmentalization of services. Container services will ensure the Gateway can also deliver reliable and fast self-healing of instances, automated

rollouts and rollbacks and load balancing. As the rest of the project, the container subsystem will be fully open source and transparent. Users will be able to download containers completed with specific tools from the project's repository and run them at their own discretion - on premises or on a cloud platform. Kubernetes scripts and configurations will also be available for more complex implementations.
5. Starting with the described and available datasets (*WoS, MAG, USPTO*, etc.) we will attempt to scale the already proven concept of keeping data in multiple formats that have value to appeal to range variety of research groups, technology, and experience levels and allow for answers of a multitude of scientific questions. This work will pave the way forto the addition of new datasets as well, expanding the use of the Gateway to national libraries. Current formats and access types include raw data storage, relational databases, graph databases and data warehouses; dynamic schema and serverless interactive query services (6) and a programmatically accessed Application Programming Interfaces (APIs) (7). A web query interface (8) and a visualization layer (9) will allow for a quick search of all datasets as well as guided complex query creation and visualization of results. IUNI IT team has already developed the parsers, needed to process raw data from WoS and MAG into relational and graph databases. The process of parsing annual data updates is an ideal candidate for cloud based compute services that can be horizontally and vertically scaled on demand for a short periods. With extended experience with on premises and cloud based implementations of relational and graph databases, like *PostgreSQL, Neo4J, TigerGraph, Agens Graph, Cosmos* and *NeptuneDB* – we are confident in choosing the best database provider for each given dataset, for different types of queries and analytical tasks. This will enable users – to be able to answer a lager variety of scientific questions. The system should be able to suggest the best workflow for each individual query, regardless of its origin (gateway script, API, web interface generated query etc.). We would use the guidance of the Product Owner Council, and usage data from the initial gateway users to establish the best architecture for the most commonly used queries.
6. The use of recently introduced, cutting edge, cloud specific technologies will ensure scalability and will help keep associated costs to a minimum. We plan on keeping all data in a cloud data lake, which not only has unlimited, relatively inexpensive storage, but also allows for computation, machine learning, large scale distributed computing operations and database like searches directly on the uncategorized and unclassified text. By building a database schema atop the raw files using bot-crawlers and metadata tables (Glue, Athena) or defining the relational schema as a part of the querying process (U-SQL), dynamic schema and serverless interactive query services provide a common SQL interface for querying raw text files. They are uniquely in line with the goals of the SBD-Gateway, allowing for complex queries that requiring a lot of processing power to incur charges on demand. Managed, distributed framework (*EMR, HDInsight*)– for skilled users, it offers *Hadoop, Spark, Presto, Flink, Hive, LLAP, Kafka, Storm, R* etc. services that can spin up multi-node clusters automatically, to work with data lake data and be decommissioned when no longer needed. Data lake analytics also provides best-in-class machine/deep learning tools and a simple drag-and-drop interface for novice users, while also offering *Python* and *R* users access hundreds of built-in packages and support for custom code and an elastic, e compute resource to accommodate even the most sophisticated workloads.
7. The envisioned *SBD-Gateway* architecture, calls for the use of microservices - a new and revolutionary way applications are being built and developed. For a microservice architecture to function each individual microservice must be able to communicate with all others as well as with the applications and interfaces they power and with the databases from which they draw real-time information. We plan on building individual service Application Programing Interfaces (APIs) to standardize communication, ensure a higher level of security and allow not only for scalability and growth, but also for potential major changes in the entire architecture. RESTful APIs at lower levels will ensure that upon authentication, developers, and researcher groups will be able to directly access services. They will also be able to build new tools or port their existing ones to the system with ease using well-documented API functions. This will make possible machine learning, text analysis and database analytics scripts and algorithms to be packaged and shared tough the Research Asset Commons, in the form of standalone apps. Users can choose to made them available as part of the Web Interface or exported out via the use of Docker Containers. A high level API Gateway – will ensure that the entire system and all of its parts and major services are available for authenticated calls from users and their own programming or visualization tools. It will also ensure compatibility with campus compute resources.

8. The entry point and arguably the most used part of the gateway will be the Web User Interface. It will feature the federated authentication system and upon login will present users with paths to datasets, databases, tools and additional interfaces to guide them through workflows and resources to answer their questions. They would also have access to the *Research Asset Commons* where their own, previous work will be stored along with the possibility to share it, or use some of the already populated tools, solutions and datasets. Having a tag based search system will assure that most appropriate algorithms and resources are easy to find and most relevant ones – automatically suggested. To remove the confusion of the different databases as well as the complexity of technologies like the dynamic schema and serverless interactive query services, we plan on creating a query builder – based upon the one currently used in Secure Enclave for Critical Data – to provide a guided approach to ask scientific questions. It will have the ability to suggest the best format, source and approach - leading the user to a quick search while allowing them to modify selection and parameters and any point. The web interface will also lead to *Jupyter* and *Databricks* notebooks Researchers well versed in scripting languages would be able to access resources directly or through an API to engage big data query and analytics services like *Spark, R, Python, Scala,* or *SQL.*
9. The Shared BigData Gateway will utilize the native cloud visualization systems like *QuickSight* and *Power BI* to make sure users are given an easy and integrated way to visualize results, right from within the web interface. For some of the tools to be ported and created, we will work with their authors or develop in-house individualized visualization interfaces to accommodate a specific need or answer a specific scientific question. Our expertise builds on already established IUNI projects like *Hoaxy* – http://hoaxy.iuni.iu.edu and the tools form the *Observatory on Social Media (OSoMe)* – http://osome.iuni.iu.edu/tools. Through the lower level APIs and the API Gateway – we will ensure compatibility and integration with most commercial visualization software and custom built packages.

Consensus building

The *SBD-Gateway* is a shared project and inherently collaborative. Investment and buy-in from academic libraries, industry partners, researchers, and big data hubs is paramount to the success of the project. In recognition of this, members who are financially supporting the project will have representation on the advisory board which will help shape the development of the platform and provide feedback from their respective communities. Furthermore, we will work with partner institutions during the project's needs assessment phase to identify researchers whose work will be substantially impacted by the SBD- Gateway and solicit user stories from those researchers in order to inform development. Once a test instance of the cloud- based *SBD-Gateway* is deployed, advisory board members will have additional opportunities virtually and in-person to provide feedback and recommendations on modifications. We intend to host a workshop at the International Society of Scientometrics and Infometrics (ISSI) conference as well at the Midwest Big Data Hub meeting to solicit additional feedback on infrastructure development that might benefit user communities.

As a pioneering cyberinfrastructure project with a diverse customer base, our project activities will be driven by practical use cases and empirical analysis of user behaviors. The team will build upon our existing experiences at the university level and adapt to new customer needs and challenges as we scale the service to the national level. We will follow the agile software development principles and iteratively respond to customer feedbacks and user behavior data. As a cloud based big data project, the project design and updates will also be informed by the state of art data analytic, cloud architecture design and database theories.

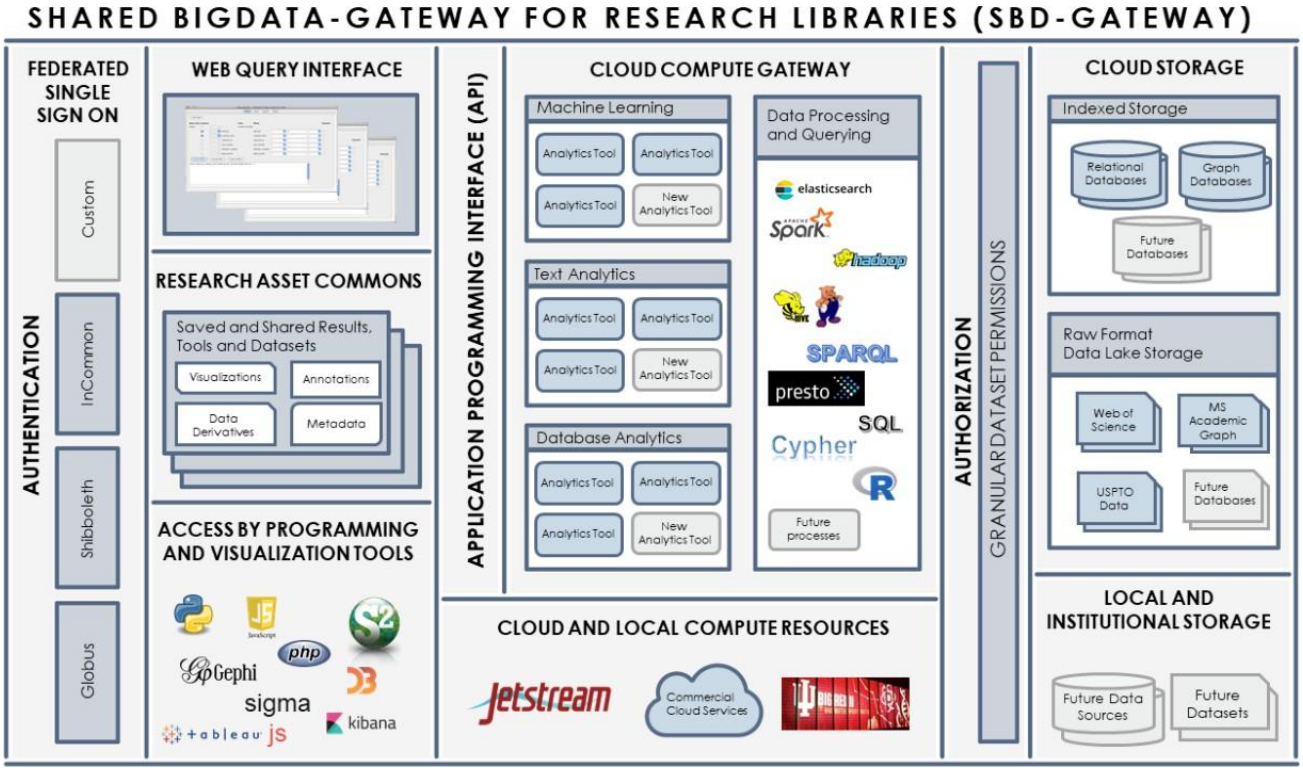
Benchmarks

The unique combination of database service, cloud computing infrastructure and collaborative environment means the project should be assessed from multiple perspectives. These include standard database query benchmarks, cloud efficiency/cost analysis, user behavior/satisfaction analysis and so on. As one of the first big data projects in this area, we will collect these measures at institutional, regional and national level, establishing new benchmarks as we scale up. The ultimate success of this project should be measured by its financial sustainability and more importantly, its systemic impact on the science of science research community. More specifically, how many researchers/institutions are served? Especially those would otherwise unable to access data at this scale. Many scientific outcomes can be conveniently measured using the datasets on the platform itself. For example, we can track the number of new academic publications and collaborations produced through our platform, and evaluate their scientific impact over the years.

Sustainability

The sustainability model for the *SBD-Gateway* project is predicated on annual contributions from partner institutions. We have received commitments from eight BTAA academic libraries to continue providing financial support for the platform for at least three years after the grant ends: Michigan State University, Purdue University, Indiana University, University of Michigan, The Ohio State University, University of Minnesota, University of Iowa, and Rutgers University (Supporting documents 1,2,3,7,8,14,16,17). Additionally, two academic libraries have expressed documented interest in joining the project at the paid membership level once the platform is operationalized: Northwestern and University of Wisconsin-Madison (Supporting documents 13 & 19). Three library consortia have pledged support for phase one (BTAA, supporting document 20) and interest in subscriptions during phase two of the project (Private Academic Library Network of Indiana and Greater Western Library Alliance, supporting documents 11 & 12). The provision of a free tier of service is an important component of the project in that it ensures that the *SBD-Gateway* is a true national resource. It is also beneficial for the sustainability of the platform because it allows researchers to use the *SBD-Gateway* without prior financial commitment from their institutions during the first two years. A larger community of users and diversity of research use cases ensures that the project team has the best possible opportunity to develop *SBD-Gateway* infrastructure that is accessible and adaptable to as many research communities as possible. For researchers using the free tier who find that they have a need for distributed framework and machine learning capabilities, new institutions are able to join as paying members and their annual membership contributions will reduce the cost for all.

The NSF-funded Big Data Hubs have committed to supporting the project with outreach and engagement to their respective communities. Once operational, multiple tiers of memberships based on institution size and computing need will be available. The tiered model will include membership options for universities, individuals and labs, and consortia that wish to use the service as well as the aforementioned free tier.



Diversity Plan

The *SBD-Gateway* serves diverse communities by reducing barriers to accessing infrastructure and service supporting bibliometric data and democratizing access to computing resources needed to analyze these data. Three barriers to access for will be reduced or eliminated. The cost in staffing and resources to ensure that proprietary data are stewarded in such a way that is consistent with the proprietary data use agreements and licenses. *SBD-Gateway* provides access to a cross-institutional community of researchers, data managers, and

librarians using the same datasets with the same infrastructure as well as the expertise and computing power required to parse and update large bibliometric datasets. This shared model for access to a data enclave provisioning both proprietary, licensed datasets and free, community-owned datasets offers a pathway to data-intensive bibliometric research and collection building, especially for institutions with a limited research mission or funding. Libraries at liberal arts institutions and community colleges, for example, will not only have access to the free service tier of the *SBD-Gateway* and the vast bibliometric data associated with it, they will have access to established cloud-based infrastructure that can support their community and consortia needs related to stewarding large, proprietary, or otherwise restricted datasets. This pathway to access for such institutions will enable researchers and students to employ large scale bibliometric analysis and empower members of underrepresented communities to explore data-intensive methods in use of these data.

National Impact

The *SBD-Gateway* project will expand the data-mining capacity of libraries throughout the country and make a substantial contribution to the National Digital Platform. Prior work, such as the IMLS funded project *National Forum: Data Mining Research Using In-copyright and Limited-access Text Datasets* (LG-73-17-0070-17) has laid the groundwork positioning academic libraries as key service providers and partners in data mining research. The *SBD-Gateway* project extends this work by establishing a new national model for cloud-based resource sharing across academic libraries that leverages our institutional and consortial licenses to build cloud-based cyberinfrastructure that supports the data-mining and analysis needs of research libraries of all sizes. This effort will have broad and lasting impact for libraries and researchers, providing a multi-tenant and scalable platform designed to support both open and proprietary datasets critical to big data and text mining research. Alone, most libraries are unable to sustain the cost of such a platform, but working cooperatively, both the costs and redundant efforts to build services and stewardship models to support these data are significantly reduced. By taking advantage of the economy of scale, the *SBD-Gateway* will also bring to all subscribers the collective bargaining power, in terms of acquiring cloud computing resources and future datasets.

Because bibliometric analysis is a critical and longstanding research area in library science, the initial datasets hosted by the *SBD-Gateway* will directly impact practicing librarians as well as library and information science faculty and students conducting citation analysis as part of their research. Furthermore, there is potential for strategic use of the hosted bibliometric datasets for service enhancements in research libraries. For example, citation analysis can be an important tool in evaluating journal relevance[10] and in collection development[11,12]. Access to bibliometric and other text-based datasets hosted in the *SBD-Gateway* will enable academic libraries with appropriate subscriptions to provide a high level of analysis service to data-intensive researchers without the requirements of local expertise or infrastructure. Library-licensed data with nuanced and stringent licensing policies for acceptable use can be made available in the *SBD-Gateway*, which will already have a framework in place for enforcing those policies and ensuring compliance while expediting access for researchers.

Because the *SBD-Gateway* is a shared platform designed with scholarly commons for users, it will establish a community of practice that will foster collaboration and innovation around data mining issues important to research libraries. Data sharing and research collaborations on big data projects such as those that utilize bibliometric data are currently severely hindered by different proprietary data use agreements, data format and workflows at individual level. Through the proposed board meetings and workshops, we aim to identify and foster consensus and common interests of the community. We will facilitate important conversations aimed at exploring collective regional and national solutions for hosting, curating, and maintaining owned or leased data collections, especially for those that have policy use restrictions, for academic libraries of all sizes.

As a platform of bibliometric data, the *SBD-Gateway* will provide an easy-to-use cyberinfrastructure to support standardized data formats, data citation, and an integrated interface allowing users to link between datasets, workflows and publication records. The platform will be designed to be adaptable for fast data updates and new data standards. We will also help promote and disseminate high quality data curation and analytical tools and will publish associated derivative datasets, thereby enhancing reproducibility, increasing intellectual exchange, and accelerating knowledge and discovery.

IMLS NLG: Shared Big Data Gateway

Schedule of Completion

Jamie Wittenberg, Val Pentchev, Patty Mabry, Xiaoan Yan Co-PI(s)

	Objectives	2018			2019			2019			2019			2019			2020			2020			2020		
	Schedule of Completion	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
1.0	**Sharing of IU data platforms**																								
1.1	Expansion of curent IU systems to accommodate new users																								
1.2	Identify researchers and library institutions to be included in sharing																								
1.3	Sharing of IU's resources with identified library institutions																								
2.0	**Federated Security Logon**																								
2.1	Creation of a initial Federated Security Login System																								
2.2	Working with individualinstitutions to federate login																								
2.3	Creation of internal uthentication system																								
2.4	Release of production authentication system																								
3.0	**Research Asset Commons**																								
3.1	Creation of a initial Research Asset Commons																								
3.2	Extension of Research Asset Commons based on users' input																								
4.0	**Datataset Storage**																								
4.1	Raw Datastes stored in Cloud																								
4.2																									
4.3	Graph Databases created in cloud																								
5.0	**Dataset Ananysis**																								
5.1	Dynamic schema and serverless interactive query services																								
5.2	Managed, distributed framework (EMR, HDInsight)																								
5.3	Machine/Deep learning																								
6.0	**Containarization**																								
6.1	Creation of initial container servces																								
6.2	Kubernetis orchestration of Docker Containers																								
6.3	Creation of tool and analytics containrs for users' use																								
7.0	**API Development**																								
7.1	Creation of Microservices APIs																								
7.2	API Gateway																								
8.0	**Web Interface**																								
8.1	Logon Web Interafce creation																								
8.2	Research Asset Commons Web Interface																								
8.3	Quaery building Web Interface creation																								
8.4	Quaery building Web Interface apdate based on users' input																								
9.0	**Visualization**																								
9.1	Integration with native visualization platforms																								
9.2	Creation of visualizations for internal tools																								
9.0	**Communications**																								
9.1																									
9.2	Workshops and conferences																								
9.3																									

Projected

SBD-GATEWAY DIGITAL PRODUCT FORM

I. MANDATORY: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS

What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

All digital products that are created as part of this project will be made openly available. Traditional scholarly works generated as part of this project, including scholarly articles, blog posts, websites, and documentation will be published open access with a Creative Commons CC-BY-NC license applied.

What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms and conditions.

Under the terms of the Open Data Commons Attribution License (ODC-BY; http://opendatacommons.org/licenses/by/), users may share, create, and/or adapt these data/databases with proper attribution. All open data used in the project will be subject to this license. Commercially licensed data such as those provided by Clarivate Analytics will be used under the appropriate commercial license agreement.

If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

All data produced during this research will be available freely to the public; we anticipate no sensitive or confidential data. Under the terms of the Open Data Commons Attribution License (ODC-BY; http://opendatacommons.org/licenses/by/), users may share, create, and/or adapt these data/databases with proper attribution."

II. CREATING OR COLLECTING NEW DIGITAL CONTENT, RESOURCES, OR ASSETS

Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

Data collection - we plan on using publicly available or already accuired datasets
Data format - we will preserve the data in their native format - plain text (.txt), comma-separated values (.csv), eXtensible Markup Language (.xml), JavaScript Object Notation (.json) as well as parse it intorelarional and graph databases
Data volume - intitial datasets chosen for the project total between 5 and 10 terrabyte, Tufure expansions wil bring the volum to a few hundred terabytes.

List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider who will perform the work.

We plan on using departamental and University storage and compute systems as well and systems avialble from leading cloud providers.
On premises - data will be stored on DC2 - https://kb.iu.edu/d/avvh and copmuted by dedicated nodes on the Carbonate Cluster https://kb.iu.edu/d/aolp. It will be also parsed in PostgerSQL, Neo4j, TigerGraph and Agens Graph databases
In cloud, data willbe stored in S3 ar BLOB data lakes and be parsed in relational and graph databases. We also plan to use serverless interactive query services like Athena nad U-SQL

List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, pixel dimensions).

We will preserve the data in their native format - plain text (.txt), comma-separated values (.csv), eXtensible Markup Language (.xml) and avaScript Object Notation (.json)

III. WORKFLOW AND ASSET MAINTENANCE/PRESERVATION

Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

The project uses a dedicated Data Steward who will be responsible for data quality controls and monitoring. Users will not be allowed to upload data to the Gateway allowing to create a pristine, controlled environment. Data Steward work will be subject to review by the Technical Development Team Lead and teh Project Owners Commitee

Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. 200.461)

We plan to sustain the project indefinetely via members' contributions. Still data archival plans will be setup from teh very begingin. Data migration and archival will vary based on the cloud system chosen, but will include hot, cool and cold storage as well as an archival and backup system utilizing the The Scholarly Data Archive (SDA) at Indiana -University https://kb.iu.edu/d/aiyi

IV. METADATA

Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g. MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g. thesauri).

Based on input from Project Owners Commite and research libraries, we will choose the appropriate standard for metadata structure to use in the Shared BigData Gateway and the Research Asset Commons

Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

Metadata will be archived on a regullar basis utilizing the The Scholarly Data Archive (SDA) at Indiana - University https://kb.iu.edu/d/aiyi

Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

Metadata will be avilable via Application Programming Interface (API)

V. ACCESS AND USE

Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

Shared BigData Gateway will include publicly open as well as proprietary datasets. We plan to develop a federated, single sign on security, using existing university authentication infrastructure (Shibboleth, Globus, InCommon, etc.) to grant differential access to platform data according to institutional permissions. All platform functions will access data and tools via an Application Programming Interface (API).
All data products that need to be published for access and re-use will be published in the IU Scholarworks Insitutional Repository with appropriate datacite DOIs and atribution (Example Jetstream VM image -https://doi.org/10.5967/P9S94Q).

Provide URL(s) for any examples of previous digital collections or content your organization has created.

Observatory in Social Media http://osome.iuni.iu.edu/tools hosts a 200TB HBase database on a 15 node Hadoop cluster making it avilable tothe public in real time.
Web of Science Enclave hosts more than a billion objects - http://iuni.iu.edu/resources/web-of-science
Human Connectome Peoject's IUrepository hosts more than 100TB of high resolution MRI images of 1200 subjects https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release; http://iuni.iu.edu/resources/hcp
Graphical User Interface for Legislative Data (GUILD) is a graph database of legislators, Committees, and Bills for the past three (and growing) legislative sessions of the Indiana General Assembly (IGA) http://iuni.iu.edu/resources/guild