National Leadership Grant-Project: Shared BigData Gateway for Research Libraries (SBD-Gateway): A Cloud-based Cyberinfrastructure for Sharing Research Assets and Advancing Library and Information Science.

Introduction: "*Shared BigData-Gateway for Research Libraries*" (SBD-Gateway, aka the "platform") is a two-year project to develop, seed, and maintain a cloud-based, extensible cyberinfrastructure for sharing large academic library data resources and to grow a community of scholars. **We request $868,895 in IMLS funds, and pledge $882,074 cost-share (total costs: $1,750,969)**. Initial datasets include Microsoft Academic Graph (MAG), Web of Science (WoS), and US Patent and Trademark Office (USPTO) data. An Advisory Board of academic library, researcher, and industry partners will provide input on design, features, and datasets. Two workshops will introduce an initial information and library science (ILS) user community to the platform and get user-centered input on design. Our informal poll of libraries suggest support for modest annual membership fees ($25K-$75k/institution) to sustain SBD-Gateway.

Robert H. McDonald, Librarian and Associate Dean for Research and Technologies at Indiana University Libraries will serve as Project Director; Valentin Pentchev, Director of Information Technology at IU Network Science Institute (IUNI) will serve as Co-Project Director for IT. Xiaoran Yan, PhD will serve as Co-Project Director of User Engagement. IU is uniquely qualified to lead the project given their experience developing local sharing solutions for WoS, MAG, and USPTO data for IU and across libraries for the HathiTrust Digital Library[1].

This project leverages IUNI's unique experience in cleaning, parsing, hosting, and securing large proprietary (WOS) and public (MAG, USPTO) datasets. IUNI parsed these data from their original raw text and html files into relational and graphical formats to enable faster searches and whole-dataset network analyses and visualizations that would not otherwise be feasible. IUNI's current solution is server-based and restricted to IU researchers. To serve a larger user base across institutions, each with unique needs and access permissions, we will scale up to a cloud-based, multi-tenant, data mining gateway, capable of serving public patrons all over the world while providing federated access to current and future data according to each institutional library's permissions. The SBD-Gateway will de facto set data sharing standards, thereby enhancing reproducibility of science and enabling research libraries to provide for their researchers' data mining needs. The SBD-Gateway will facilitate sharing of data derivatives, metadata, annotations, visualizations, algorithms, and results according to user preferences in the SBD-Gateway's secure "commons," stimulating collaboration and knowledge discovery.

Statement of Need: The SBD-Gateway addresses the IMLS "National Digital Platform" priority by addressing two of academic libraries' pressing needs. First, *academic libraries are challenged to provide sustainable, affordable, and standardized data and text mining services for licensed, big data sets.* We will provide member institutions a cloud-based, shared platform solution with the appropriate security, stewardship, and storage for proprietary and library-owned datasets at a fraction of what it would cost them to do so alone. Moreover, because member institutions are alleviated of the cost burden of local hosting, the cost barrier will be reduced for small to mid-sized libraries who cannot currently afford to host such data. Some benefits are not possible in "one-off" solutions: standardization of data formats, a team who has already parsed raw data into graph and relational formats, relationships with industry partners, and efficiencies that comes with sharing data custodial tasks, research assets, and computational resources. Second, *for all academic libraries, including those without proprietary data, there is a need for sustainable, affordable, and standardized data and text mining services for open and non-consumptive data sets* that are too large or unwieldy to work within existing research library environments. Computational resource drain becomes especially problematic for network analyses, because citation, co-authorship and other networks are orders of magnitude larger than the original data.

Project Design: Initial datasets to be hosted on the SBD-Gateway are those that IUNI already hosts: 1) *Microsoft Academic Graph (MAG),* publicly available bibliometric data on over 160 million scientific records and 3 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

[1] A complete list of references and project team biosketeches are available at http://go.iu.edu/1PLi

datasets have been identified by BTAA libraries as of interest to researchers across a broad range of scientific fields, yet there is no existing shared hosting and data mining solution. IU is currently the only research institution hosting these same datasets for institution-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: ProQuest NewsPapers (full text of 11 newspapers), Oxford English Dictionary, and Core Logic (Census and Property Data).

As envisioned[2], the SBD-Gateway will be a cloud-based resource featuring centralized, federated, single-sign-on security, using existing university authentication infrastructure (Shibboleth, Globus, InCommon, etc.) to grant access to platform data according to institutional permissions. Once authenticated, a user may use a web interface to query available data and/or retrieve from and populate to a "commons" 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 API. The SBD-Gateway will host analytic tools and use appropriate data processing/querying software and will be designed to permit access to private cloud and local compute resources.

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 Advisory Board (AB) 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 Infometrics (ISSI) to introduce additional researchers to the platform, solicit input on design, and add 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 AB will review progress in online meetings. **Y2 Q3/Q4 refine platform, add datasets, develop sustainability plan:** We 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.

Goals, Outcomes and Impact: SBD-Gateway will create a cloud-based sustainable shared resource that takes advantage of institutional private-cloud and public-cloud infrastructure. It will build cyberinfrastructure for current and future big data text and data mining and analysis affordable to libraries of all sizes, and solve proprietary data access issues for library-licensed data. The platform will be multi-tenant and scalable for nationwide implementation. With stakeholder involvement, it will establish de-facto community data sharing standards, thereby enhancing reproducibility, increasing intellectual exchange, and accelerating knowledge and discovery. Finally, we will commence important conversations aimed at exploring collective regional and national solutions for hosting, curating, and maintaining library owned/leased data collections for libraries of all sizes.

Budget: **We request $868,895 from IMLS for this initiative:** $658,254 for key personnel ($469,410 for salaries; $188,844 for fringe benefits at 40.23%). This includes a team of six IT professionals at .3 to .5 FTE each led by Co-Project Director Pentchev to build the cyberinfrastructure. A Project Coordinator (.50 FTE on grant, .50 FTE cost share) is requested to coordinate all aspects of the project. Indirect costs are applied at 32% ($210,641). **Cost sharing for this project will be provided in the amount of $882,074:** Indiana University - $158,597, including salaries for three librarians, including Project Director as well as for Co-Project Director Yan; Microsoft Research - $170,960; Clarivate Analytics - $102,000; Big Ten Academic Alliance - $354,517. Moreover, we invite cost sharing from libraries at $12,000 each to support a representative to serve on the Advisory Board (this covers time and effort plus $1K travel per year to meetings). So far we have strong indication of cost sharing in this form from five of the BTAA libraries as well as from the Private Academic Libraries Network of Indiana, Northwestern University Medical School Library, Great Western Library Alliance, and the Midwest Big Data Hub.

[2] A figure depicting the initial design of the SBD-Gateway is available at: http://go.iu.edu/1SzZ