Big Data Governance and Metadata Management

Industry Connections Activity Initiation Document (ICAID)

Version: 3.0, 11 February 2020

IC17-006-03 Approved by the IEEE SASB 5 March 2020

Instructions

- Instructions on how to fill out this form are shown in red. It is recommended to leave the instructions in the final document and simply add the requested information where indicated.
- Shaded Text indicates a placeholder that should be replaced with information specific to this ICAID, and the shading removed.
- Completed forms, in Word format, or any questions should be sent to the IEEE Standards Association (IEEE-SA) Industry Connections Committee (ICCom) Administrator at the following address: industryconnections@ieee.org.
- The version number above, along with the date, may be used by the submitter to distinguish successive updates of this document. A separate, unique Industry Connections (IC) Activity Number will be assigned when the document is submitted to the ICCom Administrator.

1. Contact

Provide the name and contact information of the primary contact person for this IC activity. Affiliation is any entity that provides the person financial or other substantive support, for which the person may feel an obligation. If necessary, a second/alternate contact person's information may also be provided.

Name: Wo Chang
Email Address: wchang@nist.gov
Employer: NIST
Affiliation: NIST

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Activity Oversight Committee and Activity officers who are responsible for IEEE work items.

2. Participation and Voting Model

Specify whether this activity will be entity-based (participants are entities, which may have multiple representatives, one-entity-one-vote), or individual-based (participants represent themselves, one-person-one-vote).

Individual-Based

3. Purpose

3.1 Motivation and Goal

Briefly explain the context and motivation for starting this IC activity, and the overall purpose or goal to be accomplished.

Metadata management poses unique challenges with regard to “Big Data” paradigm shift. The governance lifecycle needs to be sustainable from creation, maintenance, deprecation, archiving, and deletion due to volume, velocity, and variety of big data changes and is accumulated whether the data is at rest, in motion, or in transactions. Furthermore, metadata management must also consider the issues of security and privacy for individuals, organizations, and at national levels. From the new global Internet Big Data economy opportunity in Internet of Things, Smart Cities, and other emerging technical and market trends, it is critical to have standard reference architecture for Big Data Metadata Management to support the FAIR (Findability, Accessibility, Interoperability, Reusability) foundation principles.

The goal is to enable data integration/mashup among heterogenous datasets from diversified domain repositories to make data discoverable, accessible, and usable through machine readable and actionable standard data infrastructure. In many cases, it is impossible to transfer big data sets in acceptable timespans, which necessitates both technical and policy decisions concerning what features of a dataset will be exposed and what information will be provided concerning how the data and the features were generated or derived. For example, from a flooding disaster event, it is important to fetch a fraction of needed data fields to visualize geolocation from one data repository while map the selective demographic information from the second data repository and the third data repository for food and water supply to make time sensitive decisions. New metadata management concerns arising from the Big Data paradigm need extensible reference architecture to ensure trustworthy on data quality (“veracity”) throughout the governance lifecycle to meet the ever-global open data FAIR challenges.

“Big Data” poses unique challenges with regard to governance. Lifecycle management faces challenges due to the velocity with which big data changes and is accumulated, as well as due to its volume. Decisions concerning maintenance, deprecation, archiving, and deletion can have significant economic and resource implications for datasets that reach into the terabyte, petabyte, and greater range. In many cases, it is impossible to transfer big data sets in acceptable timespans, which necessitates both technical and policy decisions concerning what features of a dataset will be exposed and what information will be provided concerning how the data and the features were generated or derived. Many big data sets include information that may be considered PII or be subject to privacy policies.

Fundamentally, many of the practices associated with data management and government do not scale from mega- and giga-byte datasets to tera-byte datasets and beyond, and as repositories of big datasets are stood up, policies, practices, and eventually standards are needed to guide everything from how they are publicized and exposed to prospective end users to how they are maintained, how
data quality (“veracity”) is measured, and how their lifecycles are effectively managed.

3.2 Related Work

Provide a brief comparison of this activity to existing, related efforts or standards of which you are aware (industry associations, consortia, standardization activities, etc.).

There are several IEEE societies who have identified areas for pre-standardization and standardization work. Specifically, these include EMBS, CES, Computer Society, and Communications Society. Reference below appendix of standards activities in process. In addition, a number of these standards activities resulted from the work of another Industry Connections activity, "The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems". http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

A first in-person workshop was organized and facilitated by the IEEE Big Data Initiative (BDI). BDI is an IEEE New Initiative Committee (NIC) funded effort, and is leading efforts for big data standardization.

The IEEE Big Data Initiative Standards Workshop was held in collaboration with the IEEE Reliability Society's ISSRE 2015 conference at NIST Headquarters in Gaithersburg, MD on 2 November 2015. Through this workshop, IEEE BDI identified areas of need and opportunity for standardization of data-related technologies. bigdata.ieee.org/standards

A follow-on workshop and discussion, IEEE Workshop on Big Data Metadata and Management (BDMM '2016), was held in conjunction with the 2016 IEEE International Conference on Big Data, (Big Data 2016 @ http://cci.drexel.edu/bigdata/bigdata2016/) December 5, 2016 in Washington, D.C. bigdata.ieee.org/conferences/bdmm

This ICAID proposal for big data governance is an action resulting from this workshop held in Washington, D.C. Other outcomes from the first IEEE Big Data Initiative Standards Workshop included the creation of study groups and working groups across the IEEE Green ICT Initiative, EMBS, and other areas.

Other important related work within the IEEE includes IEEE DataPort, a web based, AWS big data repository, funded by BDI. DataPort is available now and in beta (bigdata.ieee.org/ieee-dataport). DataPort can host datasets up to 2TB in size. It provides a DOI for each dataset and each data analysis submitted, is integrated with AWS Cloud Services, can store related analysis and documents with datasets, and has been used to support data mining competitions. The proposed IC Activity will use DataPort as a use case, test bed, and proof-of-concept.

Organizations and SDOs other than IEEE are involved in Big Data. For example, NIST has a Big Data Public Working Group (NBD-PWG). NBD-PWG was established in partnership with the industry, academia and government to create a consensus- based extensible NIST Big Data Interoperability Framework (NBDIF) that enables analytics tools to process and derive knowledge through the use of standardized interfaces between swappable architectural components. Its goal is to provide interoperability between big data applications through the NIST Big Data Reference Architecture (NBD-RA). For 2019, NBD-

PWG is happy to announce the Final version of the NBDIF after six years of intensive development by more than 80 contributors from 70 plus organizations. The goals of NBDIF Version 3 (Final) were to enhance the content of Version 2 and the general interface specifications between the NIST Big Data Reference Architecture (NBDRA) components. With that, NIST would especially like to acknowledge NBD-PWG subgroup Co-Chairs and editors (https://bigdatawg.nist.gov/cochairs.php), contributors, and reviewers for their effort and long-term commitment in putting forth this edition. NBDIF Version 3 (Final) is available at: https://bigdatawg.nist.gov/V3_output_docs.php.

These seven volumes of documents were also submitted to the ISO/IEC JTC 1/SC 42/WG 2 Working Group on Big Data and presently served as the foundation document for two standard development projects: (a) Big Data Concept and Vocabulary and (b) Big Data Reference Architecture. Wo Chang is the Digital Data Advisor for the NIST Information Technology Laboratory (ITL), Co-chair of the NBD- PWG, Convenor of ISO/IEC JTC1/SC 42/WG 2 Working Group on Big Data, and is working closely with BDI.

This Activity will coordinate with NIST and other organizations involved in related activities to increase the relevance and reach of the Activity and its deliverables.

3.3 Previously Published Material

Provide a list of any known previously published material intended for inclusion in the proposed deliverables of this activity.

None.

3.4 Potential Markets Served

Indicate the main beneficiaries of this work, and what the potential impact might be.

The proposed work has potential society, economic, and scientific impact in numerous vertical industries, e.g. financial engineering, biomedical, transportation, education, and power utilities. Specifically, the proposed work will guide how big data and big data exchange is governed. It will enable consumers of big data to better understand what is available and how to access it. It will help producers of big data properly set expectations and take steps to ensure that their datasets can be maintained and shared in accordance with their wishes. It will help organizations that store big data make decisions concerning how the big data is stored, curated, exposed, and otherwise governed so as to best serve consumers and producers.

3.5 How will the activity benefit the IEEE?

The deliverables include proposals for new standards.

4. Estimated Timeframe

Indicate approximately how long you expect this activity to operate to achieve its proposed results (e.g., time to completion of all deliverables).

Expected Completion Date: 03/2021

IC activities are chartered for two years at a time. Activities are eligible for extension upon request and review by ICCom and the IEEE-SA Standards Board. Should an extension be required, please notify the ICCom Administrator prior to the two-year mark.

5. Proposed Deliverables

Outline the anticipated deliverables and output from this IC activity, such as documents (e.g., white papers, reports), proposals for standards, conferences and workshops, databases, computer code, etc., and indicate the expected timeframe for each.

The deliverables are expected to include:

- Workshops co-located at IEEE sponsored conferences to collect, analysis, and identify relevant use cases, requirements, and potential solutions. Document the findings.
- White paper(s) framing the problems, identifying the issues in more detail based from the workshops outlined above.
- Reference architecture(s) concepts and solutions from relevant best practices in big data metadata management to formulate data interoperable infrastructure to enable data integration/mashup between diversified domain repositories, including those maintained by participating entities and IEEE Dataport. A proof-of-concept reference implementation would be welcome.
- Identification and initiation of IEEE standards activities (including recommended practices, guides) related to big data metadata management, including the development of PARs and recruitment of Working Groups within an appropriate IEEE Standards Committee

5.1 Open Source Software Development

Indicate whether this IC Activity will develop or incorporate open source software in the deliverables. All contributions of open source software for use in Industry Connections activities shall be accompanied by an approved IEEE Contributor License Agreement (CLA) appropriate for the open source license under which the Work Product will be made available. CLAs, once accepted, are irrevocable.

Will the activity develop or incorporate open source software (either normatively or informatively) in the deliverables?: No

6. Funding Requirements

Outline any contracted services or other expenses that are currently anticipated, beyond the basic support services provided to all IC activities. Indicate how those funds are expected to be obtained (e.g., through participant fees, sponsorships, government or other grants, etc.). Activities needing substantial funding may require additional reviews and approvals beyond ICCom.

Support for facilitating an in-person workshop co-located at an IEEE conference. With the financial support of the IEEE Big Data Initiative (BDI), the workshop costs would be covered.

7. Management and Procedures

7.1 Activity Oversight Committee

Indicate whether an IEEE committee of some form (e.g., a Standards committee) has agreed to oversee this activity and its procedures.

Has an IEEE committee agreed to oversee this activity?: No

If yes, indicate the IEEE committee's name and its chair's contact information.

IEEE Committee Name: Committee Name
Chair's Name: Full Name
Chair's Email Address: who@where

Additional IEEE committee information, if any. Please indicate if you are including a letter of support from the IEEE Committee that will oversee this activity.

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Activity Oversight Committee and Activity officers who are responsible for IEEE work items.

7.2 Activity Management

If no Activity Oversight Committee has been identified in 7.1 above, indicate how this activity will manage itself on a day-to-day basis (e.g., executive committee, officers, etc).

The Activity will be managed by an Executive Committee as defined in the Activity's Policy and Procedures..

7.3 Procedures

Indicate what documented procedures will be used to guide the operations of this activity; either (a) modified baseline *Industry Connections Activity Policies and Procedures,* (b) Standards Committee policies and procedures accepted by the IEEE-SA Standards

Board, or (c) Working Group policies and procedures accepted by the Working Group's Standards Committee. If option (a) is chosen, then ICCom review and approval of the P&P is required. If option (b) or (c) is chosen, then ICCom approval of the use of the P&P is required.

The Activity will follow the baseline *Industry Connections Activity Policies and Procedures*.

8. Participants

8.1 Stakeholder Communities

Indicate the stakeholder communities (the types of companies or other entities, or the different groups of individuals) that are expected to be interested in this IC activity, and will be invited to participate.

Researchers, technology companies, and government agencies that store and manage big data and make it available to others via public, paid, or private interfaces and services. The stakeholders within those organizations are people responsible for big data policy and governance, with guidance from internal and external consumers of big data. Representatives from Dataport will participate.

8.2 Expected Number of Participants

Indicate the approximate number of entities (if entity-based) or individuals (if individual-based) expected to be actively involved in this activity.

Approximately seven to ten initial individuals. Once started and publicized, we expect significantly more to join.

8.3 Initial Participants

Provide a number of the entities or individuals that will be participating from the outset. It is recommended there be at least three initial participants for an entity-based activity, or five initial participants (each with a different affiliation) for an individual-based activity.

** This is an existing Activity with over 175 participants on the roster.

Individual	Employer	Affiliation
Mahmoud Daneshmand	Stevens Institute of Technology	Stevens Institute of Technology
Wo Chang	NIST	NIST
Alex Kuo	U of Victoria	U of Victoria
Yinglong Xia	Huawei	Huawei