Big Data Governance and Metadata Management
Industry Connections Activity Initiation Document (ICAID)
Version: 2.0, 15 March 2017
<mark>IC17-006-01 Approved by the IEEE-SASB 23 March 2017</mark>

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
Phone: Number, including country code
Employer: NIST
Affiliation: NIST

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.

Note: It is anticipated, to start, this activity will have industry participation from AT&T, Huawei, and Verizon Wireless.

3. <u>Purpose</u>

3.1.   <u>Motivation and Goal</u>

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 the "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 a 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 a 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 an extensible reference architecture to ensure trustworthy on data quality ("veracity") throughout the governance lifecycle to meet the ever-global open data FAIR challenges.

3.2.   <u>Related Work</u>

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 at 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, and provides the opportunity for cross collaboration. For example, NIST has a Big Data Public Working Group (NBD-PWG). 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/WG 9 Working Group on Big Data, and works closely with BDI.

This IEEE activity intends to collaborate with 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.

N/A.

3.4.   **Potential Markets Served**

The proposed work has potential societal, 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.

4. **Estimated Timeframe**

Expected Completion Date: 06/2019

5. **Proposed Deliverables**

The deliverables are expected to include:
- Workshops co-located at IEEE sponsored conferences to collect, analyze, 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

The deliverables are expected to include:
- White paper(s) framing the problems, identifying the issues in more detail with use cases, and potential solutions.

- Reference architecture(s) for implementing relevant best practices in big data repositories, including those maintained by participating entities and IEEE Dataport. May be accompanied by reference implementations (stretch goal).
- Workshop co-located at an IEEE sponsored conference to disseminate above.
- Identification and initiation of IEEE standards activities (including recommended practices, guides) related to big data governance, including the development of PARs and recruitment of Working Groups within an appropriate IEEE Standards Committee

6. <u>Funding Requirements</u>

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. <u>Management and Procedures</u>

7.1.   <u>IEEE Sponsoring Committee</u>

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

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

7.2.   <u>Activity Management</u>

If no IEEE sponsoring 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.   <u>Procedures</u>

Indicate what documented procedures will be used to guide the operations of this activity; either a) modified baseline *Industry Connections Activity Policies and Procedures,* or b) Sponsor or Working Group policies and procedures accepted by the IEEE-SA Standards Board. The chosen policies and procedures must be reviewed by ICCom

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

8. <u>Participants</u>

8.1.   <u>Stakeholder Communities</u>

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

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

8.3.　　　Initial Participants

Provide a list 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.

Use the following table for an entity-based activity:

Entity	Primary Contact	Additional Representatives
Entity Name	Contact Name Email Address Phone Number	Name, Email Address Name, Email Address

Use the following table for an individual-based activity:

Individual	Contact Information	Employer	Affiliation
David Belanger	dbelange@stevens.edu	Stevens Institute of Technology	Stevens Institute of Technology, AT&T – retired, IEEE Big Data Initiative (BDI) Chair
Wo Chang - Chair	wchang@nist.gov	NIST	NIST
Mahmoud Daneshmand – Vice Chair	mdaneshm@stevens.edu	Stevens Institute of Technology	Stevens Institute of Technology, AT&T – retired
Kathy Grise	k.l.grise@ieee.org	IEEE	IEEE Big Data Initiative, IBM - retired
Alex Kuo	akuo@uvic.ca	U of Victoria	U of Victoria
Ye Ouyang	ye.ouyang@verizonwireless.com	Verizon Wireless	Verizon Wireless
Robby Robson	robby@computer.org	Eduworks	Eduworks
Yinglong Xia	yinglongxia@gmail.com	Huawei	Huawei

Appendix A:

Big Data Standards – draft

Initiative: BD
Standard Title: Standard for Mobile Health Data
Standard #: P1752
Objective: This standard will define specifications for a mobile health data applications programming interface (API) and standardized representations for mobile health data and metadata. Mobile health data encompasses personal health data collected from sensors and mobile applications.
Status: approved project

Initiative: BD
Standard Title: Data Privacy Process
Standard #: P7002
Objective: This standard defines requirements for a systems/software engineering process for privacy oriented considerations regarding products, services, and systems utilizing employee, customer or other external user's personal data. It extends across the life cycle from policy through development, quality assurance, and value realization.
Status: approved project

Initiative: BD
Standard Title: Algorithmic Bias Considerations
Standard #: P7003
Objective: This standard describes specific methodologies to help users certify how they worked to address and eliminate issues of negative bias in the creation of their algorithms, where "negative bias" infers the usage of overly subjective or uniformed data sets or information known to be inconsistent with legislation concerning certain protected characteristics (such as race, gender, sexuality, etc); or with instances of bias against groups not necessarily protected explicitly by legislation, but otherwise diminishing stakeholder or user well being and for which there are good reasons to be considered inappropriate.
Status: approved project

Initiative: BD
Standard Title: Standard for a Functional Architecture of Distributed Energy Efficient Big Data Processing
Standard #: P1926.1
Objective: This standard specifies a functional architecture that supports the energy-efficient transmission and processing of large volumes of data, starting at processing nodes close to the data source, with significant processing resources provided at centralized data centers.
Status: approved project

Initiative: BD
Standard Title: Standard for Child and Student Data Governance
Standard #: P7004
Objective: This standard provides transparency and social norms quality data governance practices regarding child and student data. It describes specific methodologies to help users certify how they approach accessing, collecting, storing, utilizing, sharing, and destroying child and student data
Status: PAR for review March 22

Initiative: BD
Standard Title: Standard for Transparent Employer Data Governance

Standard #: P7005
Objective: This standard provides transparency and quality data governance practices regarding employee data. It describes specific methodologies to help employers certify how they approach accessing, collecting, storing, utilizing, sharing, and destroying employee data.
Status: PAR for review March 22

Initiative: BD
Standard Title: Standard for Personal Data AI Agent
Standard #: P7006
Objective: This standard describes the technical elements required to create and grant access to a personalized Artificial Intelligence (AI) that will comprise inputs, learning, ethics, rules and values controlled by individuals. This approach will enable individuals to safely organize and share their personal information at a machine-readable level and enable a personalized AI to act as a proxy for machine-to-machine decisions.
Status: PAR for review March 22

Project proposal In Development
- **Wireless data standard**