**NIST Big Data Working Group (NBD-WD)**

**NBD-WD-2013/M0035**

|  |  |
| --- | --- |
| **Source:** | **Technology Roadmap Subgroup** |
| **Status:** | **Informational** |
| **Title:** | **Existing Standards Efforts Related to Big Data** |
| **Author:** | **Keith W. Hare**  **Senior Consultant, JCC Consulting, Inc.**  **Vice Chair, ANSI INCITS DM32.2, Databases**  **Convenor, ISO/IEC JTC1 SC32 WG3 Database Languages** |

# Abstract

As the NIST Big Data Working Group Technology Roadmap Subgroup proceeds, one task is to identify “standardization and adoption priorities through an understanding of what standards are available or under development.” While a complete identification of standardization and adoption priorities depends on the Reference Architecture, this document presents an initial view of related standards organization and standards.

Table of Contents

[Abstract 1](#_Toc361930354)

[Introduction 3](#_Toc361930355)

[INCITS, ISO, and ISO/IEC Standards Efforts 3](#_Toc361930356)

[Data Management and Interchange 3](#_Toc361930357)

[Metadata 4](#_Toc361930358)

[Data Storage and Retrieval 5](#_Toc361930359)

[Support for Complex Data Types 5](#_Toc361930360)

[Geographical Information Systems 5](#_Toc361930361)

[IT Security techniques 6](#_Toc361930362)

[Cloud Computing – Distributed application platforms and services 7](#_Toc361930363)

[Coding of Audio, Picture, Multimedia, and Hypermedia Information 7](#_Toc361930364)

[Apache Software Foundation 8](#_Toc361930365)

[W3C 8](#_Toc361930366)

# Introduction

Big Data has generated interest in a wide variety of organizations, including the *de jure* standards process, industry consortiums, and open source organizations. Each of these organizations operates differently and focuses on different aspects.

The following sections describe work currently in planning and in progress in three organizations:

* INCITS and ISO *de jure* standards process
* Apache Software Foundation
* W3C – Industry consortium

As a participant in the USA and international data management standards, I am most familiar with that work. However, I have attempted to identify efforts in other organizations as well. I expect that this document will prompt others to identify areas that I have missed. Ultimately, in this document

# INCITS, ISO, and ISO/IEC Standards Efforts

There are two major international standards development organizations working on computer related standards, ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission). Many of the information technology standards are developed by subcommittees (SC) of an ISO/IEC Joint Technical Committee – JTC1.

Within the USA, there are standards committees that correspond to the ISO and JTC1 committees. The USA committees are under the auspices of INCITS (InterNational Committee for Information Technology Standards), an ANSI accredited standards development organization.

## Data Management and Interchange

Within ISO, the ISO/IEC JTC1 SC32 committee is responsible for Data Management and Interchange standards. Working groups within SC32 are responsible for a variety of standards in the areas of Metadata and Database languages. SC32 currently has a study group in place looking at Next Generation Analytics and Big Data.

SC32 standards efforts focus on interfaces between software layers, not on the underlying physical storage.

The structures of the committees are:

|  |  |
| --- | --- |
| **USA (ANSI) Standards Groups** | **ISO/IEC Standards Groups** |
| **INCITS DM32 Data Management and Interchange**  Chair: Dr. Donald Deutsch, Oracle  Secretary: Mr. Michael Gorman, Whitemarsh Information Systems | **ISO/IEC JTC1 SC32 Data Management and Interchange**  Chair: Mr. Jim Melton, USA  Secretary: Dr. Timothy D. Schoechle, USA |
| **DM32.2 Database**  Chair: Dr. Donald Deutsch, Oracle  Vice Chair: Mr. Keith W. Hare, JCC Consulting, Inc.  Secretary: Mr. Michael Gorman, Whitemarsh Information Systems Corporation  Int'l. Rep: Krishna Kulkarni, IBM Corporation | **Working Group 3 – Database Languages**  Convenor: Mr. Keith W. Hare, USA |
| **Working Group 4 – Multimedia**  Convenor: Kohji Shibano, Japan |
| **DM32.8 Metadata**  Chair: Mr. Daniel Gillman, Bureau of Labor Statistics  Int'l. Rep: Mr. Frank Farance, Farance Inc | **Working Group 1 – e-Business**  Convenor: Wenfeng Sun, China |
| **Working Group 2 – Metadata**  Convenor: Denise Warzel, USA |

The following descriptions of SC32 work were extracted from the June 2013 report from SC32 Next Generation Analytics/Big Data working group:

SC32 N2388b “Report of study group on next generation analytics and big data”, Keith Hare - study group rapporteur, June 2013

<http://www.jtc1sc32.org/doc/N2351-2400/32N2388b-report_SG_big_data_analytics.pdf>

### Metadata

The aim of the metadata standards developed within Working Group 2 of SC32 is to provide an information and software services infrastructure that supports communities that wish to interoperate.

The ISO/IEC 11179 series of standards provides specifications for the structure of a metadata registry and the procedures for the operation of such a registry. These standards address the semantics of data (both terminological and computational), the representation of data, and the registration of the descriptions of that data. It is through these descriptions that an accurate understanding of the semantics and a useful depiction of the data are found. These standards promote:

* Standard description of data
* Common understanding of data across organizational elements and between organizations
* Re-use and standardization of data over time, space, and applications
* Harmonization and standardization of data within an organization and across organizations
* Management of the components of data
* Re-use of the components of data

The ISO/IEC 20943 series of technical reports provide a set of procedures for achieving consistency of content within a metadata registry.

The ISO/IEC 19763 series of standards provides specifications for a metamodel framework for interoperability. In this context interoperability should be interpreted in its broadest sense: the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units (ISO/IEC 2382-1:1993). ISO/IEC 19763 will eventually cover:

* A core model to provide common facilities
* A basic mapping model to allow for the common semantics of two models to be registered
* A metamodel for the registration of ontologies
* A metamodel for the registration of information models
* A metamodel for the registration of process models
* A metamodel for the registration of models of services, principally web services
* A metamodel for the registration of roles and goals associated with processes and services
* A metamodel for the registration of form designs

The ISO/IEC 19763 series of standards will also include a technical report describing on-demand model selection based on roles, goals, processes and services and a standard for a registry of registries.

### Data Storage and Retrieval

SC32 WG3 has defined the SQL database language in the 9075 family of standards. Additional work is needed to accommodate new types of objects such as JSON.

Additional work may be needed in Call Level Interfaces (CLI) to better support distributed access

* ISO/IEC 9075-1:2011 Information technology -- Database languages -- SQL -- Part 1: Framework (SQL/Framework)
* ISO/IEC 9075-2:2011 Information technology -- Database languages -- SQL -- Part 2: Foundation (SQL/Foundation)
* ISO/IEC 9075-3:2008 Information technology -- Database languages -- SQL -- Part 3: Call-Level Interface (SQL/CLI)
* ISO/IEC 9075-4:2011 Information technology -- Database languages -- SQL -- Part 4: Persistent Stored Modules (SQL/PSM)
* ISO/IEC 9075-9:2008 Information technology -- Database languages -- SQL -- Part 9: Management of External Data (SQL/MED)
* ISO/IEC 9075-10:2008 Information technology -- Database languages -- SQL -- Part 10: Object Language Bindings (SQL/OLB)
* ISO/IEC 9075-11:2011 Information technology -- Database languages -- SQL -- Part 11: Information and Definition Schemas (SQL/Schemata)
* ISO/IEC 9075-13:2008 Information technology -- Database languages -- SQL -- Part 13: SQL Routines and Types Using the Java TM Programming Language (SQL/JRT)
* ISO/IEC 9075-14:2011 Information technology -- Database languages -- SQL -- Part 14: XML-Related Specifications (SQL/XML)

### Support for Complex Data Types

SC32 WG4 has defined standards for complex data storage and retrieval:

* ISO/IEC 13249-2 SQL/MM Part 2: Full Text provides full information retrieval capabilities and complement SQL and SQL/XML. SQL/XML provides facilities to manage XML structured data while MM Part 2 provides contents based retrieval.
* ISO/IEC 13249-3 Part 3: Spatial provides all the functionalities required to support geo applications. Most big data application now includes processing of GPS data together with geographic information. Thus Part 3: Spatial is also one of the key components of big data applications. This work is carefully coordinated with ISO TC 211 and the Open GIS Consortium.
* ISO/IEC 13249-5 Part 5: Still Image provides basic functionalities for Image data management.
* ISO/IEC 13249-6 Part 6: Data Mining provides all the functionalities required to support statistical data mining applications. SQL/OLAP functionality provide simple online analytic processing while MM Part 6: provides sophisticated statistical data mining functionalities.

## Geographical Information Systems

From the INCITS web site:

Geographic Information Systems form a distinct class of information systems through their unique requirements for collecting, converting, storing, retrieving, processing, analyzing, creating, and displaying geographic data. The generic nature of GIS, organizing information by location, is interdisciplinary and not specific to any application.

The work of L1, Geographic Information Systems (GIS) consists of adopting or adapting information technology standards and developing digital geographic data standards. Digital geographic data standards are concerned with creating, defining, describing, and processing such data.

The structures of the committees are:

|  |  |
| --- | --- |
| **USA (ANSI) Standards Groups** | **ISO Standards Groups** |
| **INCITS L1 - Geographical Information Systems**  Chair: Dr. Liping Di, George Mason University  Int'l. Rep: Dr. Meixia Deng, George Mason University | **ISO/TC 211 – Geographic information/Geomatics**  Chair: Mr. Olaf Østensen, Norway  Secretary: Ms. Bjørnhild Sæterøy, Norway |

## IT Security techniques

From the INCITS web site:

The area of work includes standardization in the following areas:

* Management of information security and systems
* Management of third party information security service providers
* Intrusion detection
* Network security
* Incident handling
* IT Security evaluation and assurance
* Security assessment of operational systems
* Security requirements for cryptographic modules

Protection profiles

* Role based access control
* Security checklists
* Security metrics

Cryptographic and non-cryptographic techniques and mechanisms including:

* confidentiality
* entity authentication
* non-repudiation
* key management
* data integrity
* message authentication
* hash-functions
* digital signatures

Future service and applications standards supporting the implementation of control objectives and controls as defined in IS 27001, in the areas of:

* business continuity
* outsourcing

Identity management, including:

* identity management framework
* role based access control
* single sign-on

Privacy technologies, including:

* privacy framework
* privacy reference architecture
* privacy
* anonymity and credentials
* specific privacy enhancing technologies

The structures of the committees are:

|  |  |
| --- | --- |
| **USA (ANSI) Standards Groups** | **ISO Standards Groups** |
| **INCITS CS1 - Cyber Security**  Chair: Mr. Dan Benigni, NIST  Vice Chair: Mr. Sal Francomacaro, NIST  Int'l. Rep: Mr. Eric Hibbard, Hitachi Data Systems | **ISO/IEC JTC 1/SC 27 IT Security techniques**  Chair: Dr. Walter Fumy  Secretary: Mrs Krystyna Passia |

## Cloud Computing – Distributed application platforms and services

SC38 is responsible for developing standards for interoperable Distributed Application Platforms and Services including:

* Web Services
* Service Oriented Architecture (SOA),
* Cloud Computing

The structures of the committees are

|  |  |
| --- | --- |
| **USA (ANSI) Standards Groups** | **ISO/IEC Standards Groups** |
| **INCITS DAPS38 - Distributed Application Platforms & Services (DAPS)**  Chair: Mr. Steve Holbrook, IBM Corporation  Vice Chair: Mr. Tom Rutt, Fujitsu America Inc.  Secretary: Mr. Joel Fleck II, Hewlett-Packard Company  Int'l. Rep: Mr. John Calhoon, Microsoft Corporation | **ISO/IEC JTC 1/SC 38 Distributed Application Platforms and Services (DAPS)**  Chair: Dr. Donald Deutsch (USA)  Secretary: Marisa Peacock (USA) |

## Coding of Audio, Picture, Multimedia, and Hypermedia Information

From the INCITS web site:

Responsible for the standardization of coded representation of audio, picture, multimedia and hypermedia information - and of sets of compression and control functions for use with such information - such as: audio information, bi-level and limited bits-per-pixel still pictures; computer graphics images; moving pictures and associated audio, multimedia and hypermedia information for real-time final form interchange; and audio visual interactive scriptware.

The structures of the committees are:

|  |  |
| --- | --- |
| **USA (ANSI) Standards Groups** | **ISO Standards Groups** |
| **INCITS L3 - Coding of Audio, Picture, Multimedia, and Hypermedia Information**  Chair: Dr. Arianne Hinds, Cable Television Laboratories  a.hinds@cablelabs.com  Int'l. Rep: Dr. Andrew Tescher, Microsoft Corporation  andytescher@comcast.net | **ISO/IEC JTC 1/SC29 – Coding of audio, picture, multimedia and hypermedia information**  Chair: Mr Kohtaro Asai  Secretary: Ms. Yukiko Ogura |

# Apache Software Foundation

The Apache Software Foundation (<http://www.apache.org/>) provides support for the Apache community of open-source software projects. There are two categories of projects that of obvious interest to the NIST Big Data efforts:

* Big Data – <http://projects.apache.org/indexes/category.html#big-data>
* Databases – <http://projects.apache.org/indexes/category.html#database>

# W3C

The W3C is an industry consortium that develops standard

XML Technology

HTML5 Image Description Extension

The W3C has a “Big Data Community Group”. The website <http://www.w3.org/community/bigdata/> contains the following description:

This group will explore emerging BIG DATA pipelines and discuss the potential for developing standard architectures, Application Programming Interfaces (APIs), and languages that will improve interoperability, enable security, and lower the overall cost of BIG DATA solutions.

The BIG DATA community group will also develop tools and methods that will enable: a) trust in BIG DATA solutions; b) standard techniques for operating on BIG DATA, and c) increased education and awareness of accuracy and uncertainties associated with applying emerging techniques to BIG DATA.

This community group has not published any reports.