[Slide 1 - Role of standards organizations in the USA and their impact on knowledge representation in the global context]

In the 21st century, data and information and the information technology for managing, publishing, interchanging, and using digital content are key inputs for business development and growth. This talk provides an overview of the basis for standardization in the United States, the roles of the various types of standards organizations, their impact on knowledge representation in the global context, and examples of these activities. The talk focuses on standardization in relation to information management, publishing, reference data, and semantic interoperability.

[Slide 2] [Slide 3] In the United States, Intellectual property and privacy rights are fundamental rights that are defined in the Constitution and its Amendments. Article 1, Section 8, Clause 8 of the Constitution gives Congress the power "to promote the progress of science and useful arts, by securing, for limited times, to authors and inventors, the exclusive right to their respective writings and discoveries." The Fourth Amendment protects the right of privacy against unreasonable searches and seizures by the government; and the Fifth Amendment provides for the right against self-incrimination, which justifies the protection of private information.

[Slide 4] The US government's role in the development and use of standards and conformity assessment is guided by the National Technology Transfer and Advancement Act, OMB Circular A-119, and other federal laws, regulations, and international agreements. OMB CircularA-119 establishes policies on Federal use and development of voluntary consensus standards and on conformity assessment activities.

[Slide 5] There is no US ministry of culture. The US government has limited its involvement in cultural heritage, but there are some surprising exceptions. The US Department of Interior manages one of the largest museum systems in the world with over 73 million objects in its museum collections and 86,000 linear feet of archives. The Library of Congress while not officially the national library of the United States. it often assumes this role. But its primary mission is to research inquiries made by members of Congress.

[Slide 6] [Slide 7] The key types of data and information standards organizations are:

- Government agencies such as the Library of Congress (LoC) and the National Institute of Standards and Technology (NIST);
- Membership non-profits such as the Dublin Core Metadata Initiative (DCMI) and the National Information Standards Organization (NISO);
- Private non-profits such as the Getty Trust; and
- Industry trade associations such as the SEMI (formerly Semiconductor Equipment and Materials International) and Software and Information Industry Association (SIIA).

[Slide 8] The **Library of Congress** has been a global leader in bibliographic description, controlled vocabularies, and classification—all essential components of reference data. This chart identifies some of the standards developed or maintained by the Library of Congress. They include:

- Resource description formats such as the BIBFRAME data model for bibliographic description.
- Digital library standards such as PREMIS (PREservation Metadata: Implementation Strategies) a digital preservation metadata standard.
- Information resource retrieval protocols such as Z39.50 a communications protocol for searching and retrieving information from a database over a computer network.

- Information resource retrieval standards such as ISO 639 that provides a consistent code for languages.
- Controlled vocabularies such as the Thesaurus of Graphic Materials that enables consistent resource description values.

[Slide 9] The **National Institute of Standards and Technology** is the US Department of Commerce agency that develops standard reference materials, and standard reference data for various scientific disciplines. Standard reference materials are used to verify the accuracy of specific measurements and to support the development of new measurement methods. The NIST Text Retrieval Conference commonly referred to as "TREC" has been instrumental in facilitating development of automated categorization of text and image content.

[Slide 10] TREC is an ongoing series of workshops that focus on different information retrieval (IR) research areas. TREC is cosponsored by NIST and the Intelligence Advanced Research Projects Activity (IARPA). Standard data sets that have been used in the TREC workshops include: Legal, Medical, News, Chemical, Genomics, etc.

[Slide 11] Dublin Core is a widely used schema for describing web content, using Resource Description Framework or RDF vocabularies, packaged in application profiles. The Dublin Core originated at an OCLC workshop, and hosted by OCLC (which is located in Dublin, Ohio) until 2009. Dublin Core has become the de facto standard for digital content description in the public web and on private commercial clouds.

[Slide 12] The **National Information Standards Organization**, known as NISO, is a non-profit association accredited by the American National Standards Institute. NISO develops, maintains and

publishes technical standards related to publishing, bibliographic and library applications. NISO has been instrumental in promoting foundational library standards to enable information description and protocols for sharing bibliographic information systems and services.

[Slide 13] The **Getty Vocabularies** are a project of the Getty Research Institute which is a pat of the Getty Trust which also operate the Getty Museum in Los Angeles, CA. They contain terminology for describing various aspects of material culture (art, architecture, decorative arts, etc.) used by catalogers, researchers, and information providers. Available as Linked Open Data, XML, Relational Tables, and through APIs, the Getty Vocabularies are widely used as the basis for semantic interoperability in the culture sector. Components for community-based terminology management are provided by the Getty Vocabularies including:

- Editorial guidelines,
- Training materials,
- Contributor authorization, and
- an International Terminology Working Group.

[Slide 14] Semi is the industry association representing electronics manufacturing and the design supply chain. Since electronic components are a part of almost all machinery, the ability to produce and communicate documentation about their use and maintenance is critical for the economic viability of manufacturing. Electronic components have no value without their documentation. Semi produces more than 1,000 standards and guidelines related to all aspects of automated fabs (machinery for manufacturing electronic components) including documentation and training, standardized terminology, and XML schema.

[Slide 15] The Software & Information Industry Association or SIIA is the trade association for the entertainment, consumer and business software industries. SIIA like other trade associations

- Seeks to influence political, economic and social decisions relevant to their industry such as intellectual property protection, privacy and data security;
- Advocates for innovation that build and leverage the influence of the industry such as data driven innovation, artificial intelligence and automation; and
- Develops and promotes best practice recommendations such as derived data products, usage of data, and service levels.

Effective intellectual property policies are critical to economic viability of creating information products and services.

[Slide 16] [Slide 17] This table summarizes the different types of standards with which the organizations we have just reviewed are concerned—primarily

- Standard vocabularies for value sets and schema labels,
- Schema for modeling real world entities in databases,
- Protocols for system-to-system communication, and
- Encodings for reliably converting data from one form to another.

These standards are important for enabling one or more of the critical functions in the information economy—

- Information management in general,
- Publishing content from one or more source systems,
- Reference data management of classifications and hierarchies across systems, and
- Semantic interoperability among systems.

[Slide 18] Most standards organizations in the US are non-profits or industry trade associations. The purpose of standards is seen to facilitate economic activity. The US government tends to get involved when there is sufficient pressure via industry or public advocacy groups or lobbyists, sometimes as a result of litigation. In the US this could result in Congressional legislation, especially if federal funding is required. Policy can be proposed but not sustainably implemented by the executive branch without legislation.

Industry has been most responsive to standardization when it has an economic impact on their operations. For example, most large US companies with EU operations have complied or aim to comply with the General Data Protection Regulation (GDPR). In other cases, a standard or best practice emerges at the right time and becomes widely adopted, for example, the Dublin Core. Finally, some standards are developed by a community of users to facilitate data interchange, for example FpML (Financial products Markup Language) by the International Swaps and Derivatives Association (ISDA).

[Slide 19] [Slide 20] Knowledge graphs are a relatively new form of information representation that is possible as a result of massive amounts of widely accessible digital content and standardization. An enterprise knowledge graph is a representation of an organization's knowledge domain and artifacts that is understood by both humans and machines. It represents an organization's knowledge assets, content, and data—people, places, documents, photos, data in relational databases, etc. **[Slide 21]** —and how these things are related to one another. **[Slide 22]** Those things and relationships are represented using a metadata model specific to the industry or developed for the organization. Typically, this is an 'ontology' that defines classes for the things, properties for the things, and relationships between the things. **[Slide 23]** Simply put, a knowledge graph is an ontology plus instances.

[Slide 24] This is an example of how two different knowledge graphs present the same concept—"Fluid Dynamics". Physics Subject Headings or PhySH is a physics classification scheme developed by the American Physical Society to organize journal, meeting, and other content by topic. At the top level, concepts are organized along two dimensions: Facets (Research Areas, Physical Systems, Properties, Techniques, and Professional Topics) and Disciplines. In addition, the concepts are linked to each other in a hierarchy of broader/narrower relationships, and also via associative relationships that can span across the hierarchies to identify related concepts. The Google knowledge graph is a database which is used to enhance search results based on data that has been collected from web searches and other sources. Information about relevant people, places and things, and the relationships between them is presented in a box next to the search results.

[Slide 25] There are four types of knowledge graphs—knowledge bases, social networks, data catalogs, and combinations of one or more of these types. **[Slide 26]** Some common uses of knowledge graphs are to enhance and personalize search results, make relationships between entities, recommender applications, ad targeting, and enhancing data analytics or business intelligence. **[Slide 27]** Many ontologies are available to use as knowledge graph frameworks. BARTOC.org lists more than 700 ontologies. **[Slide 28]** In addition to local resources, there are also public data sources such as DBpedia and Wikidata that can be resources for knowledge graph instances. Nevertheless, it takes a lot of work to build the infrastructure for an enterprise knowledge graph, and then establish and implement the policies and procedures to link up a critical mass of an organization's materials.

[Slide 29] Standards are the foundation of knowledge graphs. This table lists some of the key standards—schema, vocabularies, and protocols—that enable this relatively new form of

information representation based on analyzing the relationships implicit in massive amounts of widely accessible digital content.

[Slide 30] This presentation has been a brief overview of the US standardization system with a focus on data, information, and information technology. The US approach to standardization is different from the approach in countries with more centralized governments, yet American organizations and companies are often quick to comply with international standards because that facilitates their participation in the global economy and research environment. Because of this situation in America, Europeans should reach out to thought leaders and key industry partners to work on standardization and other harmonization activities. While the incentives may be different, there is still a shared goal to facilitate effective data and information management, publishing, reference data, semantic interoperability, and use worldwide.

[Slide 31] Are there any questions or comments?

[Slide 32] Standards Bibliography

American National Standards Institute (ANSI). "Overview of the U.S. Standardization System: Understanding the U.S. Voluntary Consensus Standardization and Conformity Assessment Infrastructure." https://www.standardsportal.org/usa_en/standards_system.aspx

Dublin Core Metadata Initiative. https://www.dublincore.org/

Financial products Markup Language (FpML) https://www.fpml.org/

General Data Protection Regulation (GDPR). https://gdpr-info.eu/

Getty Vocabulary Program. https://www.getty.edu/research/tools/vocabularies/

ISO/IEC 2700. https://www.iso.org/isoiec-27001-information-security.html

Library of Congress. https://www.loc.gov/

National Information Standards Organization (NISO). “What is NISO?” https://www.niso.org/what-we-do

National Institute of Standards and Technology (NIST). Standards.gov. https://www.nist.gov/standardsgov

Office of Management and Budget. Circular No. A-119 Revised. “Federal Participation in the Development and Use of Voluntary Consensus Standards and in Conformity Assessment Activities.” (February 10, 1998) https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-119-1.pdf

SEMI. https://www.semi.org/en

Simple Knowledge Organization System (SKOS) https://www.w3.org/2004/02/skos/

Software and Information Industry Association (SIIA). “About SIIA.” https://www.siia.net/About/About-SIIA

US Department of Interior. Interior Museum Program. https://www.doi.gov/museum

Web Ontology Language (OWL) https://www.w3.org/OWL/

Wikipedia. “Library of Congress.” https://en.wikipedia.org/wiki/Library_of_Congress

[Slide 33] Knowledge Graph Bibliography

The Basel Register of Thesauri, Ontologies & Classifications (BARTOC). https://bartoc.org/.

DCMI Metadata Terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

OWL 2 Web Ontology Language. https://www.w3.org/TR/owl2-overview/.

Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. “A Survey on Knowledge Graph-Based Recommender Systems.” arXiv:2003.00911 [cs.IR]

PhySH – Physics Subject Headings. American Physical Society. https://physh.aps.org/.

RDF. https://www.w3.org/RDF/.

RDF Schema 1.1. https://www.w3.org/TR/rdf-schema/.

SKOS Simple Knowledge Organization System Namespace Document - HTML Variant. https://www.w3.org/2009/08/skos-reference/skos.html.

SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/.

Bio

Joseph Busch is the Founder and Principal Consultant of Taxonomy Strategies. Taxonomy Strategies guides global companies, government agencies, and NGO's such as CDW, the Center for Medicare and Medicaid Services, the State Bank of Pakistan, and the Robert Wood Johnson Foundation in developing metadata frameworks and taxonomy strategies to help information achieve its highest value. Before founding Taxonomy Strategies, Mr. Busch held management positions at Interwoven, Metacode Technologies, the Getty Trust, PriceWaterhouse and Hampshire College. He is a Past President of the Association for Information Science and Technology, and a past member of the Dublin Core Metadata Initiative Executive Committee.

facebook.com/joseph.busch1; linkedin.com/in/taxonomystrategies; twitter.com/joebusch