DEPARTMENT OF DEFENSE
6000 DEFENSE PENTAGON
WASHINGTON, D.C. 20301-6000

CHIEF INFORMATION OFFICER

MEMORANDUM FOR DOD COMPONENT CHIEF DATA OFFICERS

SUBJECT: Federated Data Catalog - Minimum Metadata Requirements

Data managers and data custodians are leading the charge to Make Data Visible. Through their execution of the second Data Decree of the Creating Data Advantage Memo, DoD will publish the metadata of data assets to the Enterprise Service Provider (ESP) to increase speed to data discovery and access. Attachment 1 provides details on DoD's federation model. DoD's federated approach to data cataloging enables Components to develop their catalogs and data governance platforms to support local requirements while increasing the visibility of data assets across the Department through the exchange of descriptive metadata.

Descriptive metadata describes data content, principally for search, discovery, and access control. This type of metadata improves the ability to search, browse, sort, and filter information on data sets and data sources. Attachments 2 & 3 define the descriptive minimum metadata required, with option to provide additional metadata to the DoD Federated Data Catalog (FDC). Guided by the metadata types defined in Attachment 4, the DoD CDO and the enterprise service provider have developed a minimum metadata standard for data sets in the FDC. The exchange of the minimum metadata will enable consistent records of DoD's data assets within the FDC, while recognizing the range of current data maturity and data governance practices across the Department.

Components will begin publishing their data assets with this metadata, as the ESP continues to establish connections between Component catalogs. This memorandum will be reviewed and revised in one year to incorporate lessons learned as Components advance their cataloging programs. DoD will employ a number of different connection patterns for federation, including automated input from cloud object storage, system interfaces, and native coordination through metadata tenancy with the ESP. Additional details on technical connection patterns will be published on data.mil.

Within 14 days of this memorandum, Component data leaders will provide the DoD CDO with a point of contact responsible for data catalog implementation and federation. My point of contact is John Turner at (571) 296-7519 or john.d.turner120.civ@mail.mil.

SPIRK.DAVID.L.JR.1051444562
Digitally signed by SPIRK.DAVID.L.JR.1051444562
Date: 2021.10.01 13:38:17 -04'00'

David Spirk
Chief Data Officer
Department of Defense

Attachments:

1. Data Cataloging Model
2. Minimum Metadata Catalog Requirements: Data Sets
3. Minimum Metadata Catalog Requirements: Data Sources
4. Enterprise Metadata Taxonomy
5. Glossary

ATTACHMENT 1
DATA CATALOGING MODEL

The Department's approach for implementing federated data cataloging is a Hub-and-Spoke model. A Hub-and-Spoke model (Figure 1) is designed to be highly scalable, for searching metadata about data sets and sources.

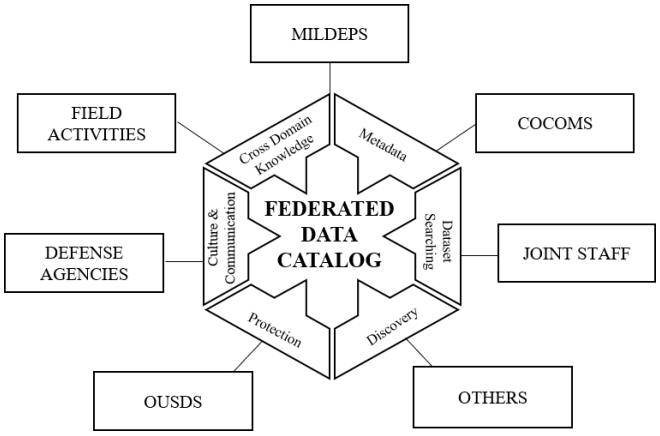

Figure 1: Hub-and-Spoke Model

a) *Hub:* Advana delivers infrastructure for the data cataloging. The hub is responsible for providing a centralized location to search metadata for data sets and sources. The hub encourages data trust and linkage through data provenance and pedigree.

b) Spoke: Ensures the proper quality and trustworthy of its data sets. Spokes are responsible for detecting changes in their data set structures, such as the addition and deletion of columns, changes in column metadata, changes to content and classifications (e.g. detection of new PII), relationship alterations and other irregularities or changes in data quality or characteristics.

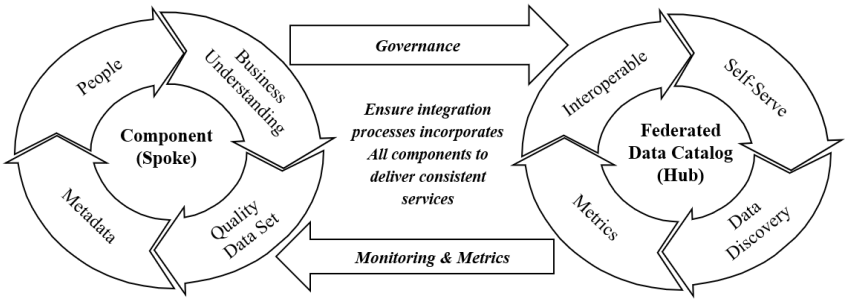

Figure 2: ESP Hub-and-Spoke Operations

As Spokes, each Component is responsible for its own data cataloging capability. This includes cataloging data sets and their associated data dictionaries and the implementation of Department-wide metadata standards.

ATTACHMENT 2
MININUM METADATA CATALOG REQUIREMENTS: DATA SETS

Guided by the metadata types defined in Attachment 4, the DoD CDO and the ESP have developed a minimum metadata standard for data sets. Components can provide additional metadata to meet mission and operational requirements.

Minimum metadata requirements will be available at https://advana.data.mil/#/cdo.

REQ	NAME	DESCRIPTION	EXAMPLE	INPUT TYPE	FORMAT EXAMPLES
DATA SET GENERAL					
Y	Unique Identifier	Unique ID associated with Data Sets.	8359834	VARCHAR	Organization Acronym + Internally unique ID system • AF123 • VA07345 • ARMY123
Y	Data Set Name	Plain language data set name.	NBA 20-21 Player Stats	STRING	PlayerStats
N	Data Set Acronym	Abbreviated name of data set.	NBA PS 20-21	STRING	PS
Y	Data Set Description	A more verbose way to describe what the data contained within a data set is.	Points, Rebounds, and other KPIs for each player in the National Basketball Association	STRING	PlayerStats KPI's
Y	Data Source System	Indicates the data source(s) that the data set was received from. Point of origin of the data.	Basketball Reference System	VARCHAR	Basketball Sys1
Y	Office of Record	Organizational element that is responsible for	NBA Analytics Dept	VARCHAR	NBA Dep123

REQ	NAME	DESCRIPTION	EXAMPLE	INPUT TYPE	FORMAT EXAMPLES
		making decisions related to the data set.			
Y	Legal Authority	Documented legal basis for mission activities associated with the creation, retention and use of a data set(s).	MOA, ICD 501	VARCHAR	MilDept 123
DATA SET ORIGIN					
Y	Data Set URL (if applicable)	Method to request access to the environment that the data set is hosted on.	2020-21 NBA Player Stats: Per Game | espn.com (espn.com)	VARCHAR	www.123.mil
Y	PAI/CAI (Publicly Available/Commercially Available)	Indicates the environment in which data from the data set is stored.	Publicly Available, Commerciall y Available, N/A	VARCHAR	MilEnv123
Y	Date Cataloged	Date on which the data set was ingested into one of the hub/spoke instances of the federated data catalog ecosystem.	4/6/2021	DATETIME	MM/DD/YYYY
DATA SET CONTENT					
Y	Data Dictionary Exists	Selection whether the data set has a data dictionary.	Yes	BOOLEAN	Yes
Y	Data Dictionary Status	The status of data dictionary.	In Development	VARCHAR	Complete, In Development, None
N	Data Source DITPR ID	DITPR ID	NBA2021	VARCHAR	NBASource123
Y	Verification Date of Data Definition	Date on which the data definition was verified.	4/6/2021	DATETIME	MM/DD/YYYY

REQ	NAME	DESCRIPTION	EXAMPLE	INPUT TYPE	FORMAT EXAMPLES
STEWARDSHIP					
Y	Data set Steward name	Indicates the proper point of contact for questions about a data set.	Nate Dullard	VARCHAR	Nate Dullard IV
Y	Data set Steward email	Indicates the proper point of contact's email for questions about a data set.	ndullard@nba.com	VARCHAR	ndullard@domain.mil
SENSITIVE DATA PROFILE					
Y	Exists on Government networks (data sets only)	Boolean True/False does this data set exist on the Government environment.	TRUE	BOOLEAN	True, False
Y	Disclosure & Releasability	Information pertaining to countries, organizations, or communities approved to receive the resource.	Approved for DoD	STRING	NOFORN, FVEY, etc.
N	Distribution Statement	The distribution statement will be reflected in the CUI designation indicator and will be annotated in full on the first page or cover of the document.	Approved for public release Distribution is unlimited	STRING	Distribution is unlimited
Y	Handling Restrictions	Limitations not related to classification such as Controlled Unclassified Information (CUI) designations and Controlled Technical Information (CTI) designations.	Standard CUI and CTI restrictions	STRING	CUI

REQ	NAME	DESCRIPTION	EXAMPLE	INPUT TYPE	FORMAT EXAMPLES
Y	Contains PII	Does the data set contain PII?	Yes	VARCHAR	Yes (High/Low/Mod) or No
Y	Contains PHI	Does the data set contain PHI?	No	VARCHAR	Yes (High/Low/Mod) or No
Y	Contains CUI	Does the data set contain CUI?	Yes	VARCHAR	Yes (High/Low/Mod) or No
Y	Security Classification	Selection for the highest classification level of the data set. Marking will follow various specifications that include, but not limited to Information Security Marking Metadata (ISMM) specification, The Trusted Data Format (TDF), NATO STANAG 4774 and 4778.	Unclassified	STRING	Unclassified, Confidential, Secret, Top Secret.

ATTACHMENT 3
MINIMUM METADATA CATALOG REQUIREMENTS: DATA SOURCES

Guided by the metadata types defined in Attachment 4, the DoD CDO and the ESP have developed a minimum metadata standard for data sources on Government networks. DoD CDO recognizes the Department maintains metadata regarding data sources within Information Technology (IT) portfolio management databases (e.g., DITPR). Where available, quality metadata should be leveraged for data cataloging efforts. DoD CDO defers to Components CDOs on specific implementation decisions regarding metadata ingest into Component catalogs.

Minimum metadata requirements are also posted at https://advana.data.mil/#/cdo.

REQ	CHAR NAME	CHAR TYPE	DESCRIPTION	EXAMPLE	INPUT TYPE	FORMAT VALUES
DATA SOURCES GENERAL						
Y	Data Source Full Name	Attr	Plain language data source name.	BoxStats.com Fight Records	STRING	BoxStats
N	Acronym	Attr	Abbreviated Name of Data Source.	BSFR	VARCHAR	BSFR12
Y	Alias	Attr	Alternative Name of Data Source.	Boxing Matches Data Repository	STRING	BSstat
Y	Description	Attr	A more verbose way to describe what the data source contains.	A website data source containing records on World Boxing Association matches	STRING	BoxStats information
Y	PAI/CAI	Attr	Indicates the environment in which the data source is stored.	publicly Available	VARCHAR	Publicly Available/ Commercially Available
N	Notes	Attr	Additional comments on the data source.	Need to refresh API connection	STRING	Bi-weekly refresh
DATA SOURCE ORIGIN & PRODUCTS						
Y	Organization that manages the data source	Relation	Organizational Element responsible for making decisions	World Boxing Association	VARCHAR	WorldBoxing Headquaters123

			related to the data source.			
Y	Provides Data Set(s)	Relation	Indicates data sets that stem from this source.	Mayweather vs. Pacquiao Match Stats	STRING	Mayweather Data Set
N	Data is stored in Database or Repository	Relation	Indicates the database this source is stored in.	PostgreSQL database	VARCHAR	DB2
N	Stored in Data Center	Relation	Indicates the data center this source is stored in.	WBA Data Center	STRING	Data Center ABC
DATA SOURCE CONTENT						
Y	Data Dictionary Exists	Attr	Yes/No verifying if there is a data dictionary or not.	Yes	BOOLEAN	Yes/No
N	DITPR ID	Attr	DITPR ID	928374	VARCHAR	1234567
STEWARDSHIP						
Y	Data Steward Organizatio n	Attr	Indicates organization of Data Steward.	Organizaiton12	VARCHAR	DSOrg123
Y	Data Source Steward Name	Attr	Indicates the proper point of contact for questions about a data source.	Gloria White	VARCHAR	Gloria White
Y	Data Source Steward Email	Attr	Indicates the proper point of contact for questions about a data source.	gwhite@ ufc.com	VARCHAR	gwhite@ ufc.com

ATTACHMENT 4
ENTERPRISE METADATA TAXONOMY

The DoD CDO has adapted an enterprise metadata taxonomy that supports linkage across Components and provide context for data by representing its meaning, business significance and relationships with other data.

METADATA TYPES	DESCRIPTION	SCHEMA EXAPLES
Business	• The meaning of the data, where its sourced and who is the designated authority.	Dublin Core Metadata Initiative (DCMI) supports best practices across a broad range of business models.
Technical	• API requirements for interoperability. • Format (e.g., JSON, XML, CSV, Parquet). • Encryption and decryption keys. • Timestamp of creation, last update.	MIX— The National Information Standards Organization (NISO) Metadata for Images, AudioMD, VideoMD
Descriptive	• Used for discovery and identification to include such information as title, author, abstract and keywords.	Dublin Core, MODS (Metadata Object Description Schema)
Administrative	• Used to manage information to enable consistency (Status, Expiry date, Content Owner, Access Rights, Intellectual Property).	METS (Metadata Encoding & Transmission Standard), Dublin Core
Structural	• Used to structure and exchange information objects (composition of compound objects) useful for rich snippets which enables Interoperability (Relationships, formats, contents, usage, and volumetric).	METS (Metadata Encoding & Transmission Standard
Reference: Zhang, A.B. and Gourley, D. (2009) Descriptive Metadata. Creating Digital Collections. Available: https://www.sciencedirect.com/topics/computer-science/descriptive-metadata		

ATTACHMENT 5
GLOSSARY

Authoritative Source	A source of data or information that is recognized by members of the Community of Interest to be valid or trusted because its provenance is considered highly reliable or accurate. During the lifecycle process, the authoritative source (or system of use in which it is housed) can evolve according to use. Subject Matter Experts validate that the data is authoritative, and Data Management assures that data from the authoritative source is provided to users, and that it is current. Note: Authoritative Attribute Source, a specialized type of Authoritative Source, is defined in Identity and IAA policy. Its authoritative definition is provided in ICS 500-30.	DAMA Dictionary of Data Management, 2nd Edition, 2011
Data	A representation of facts, concepts or instructions, such as text, numbers, graphics, documents, images, sound or video, in a form suitable for communication, interpretation or processing, which individually have no meaning by and in themselves.	Derived From: (1) DAMA Dictionary of Data Management, April 1, 2011 (2) Newton's Telecom Dictionary, 17th Edition, February 2001
Data Asset	Data maintained and secured as a shared, critical, inexhaustible, durable, and strategic resource with the expectation of future value and benefits. Examples of data assets include databases, documents, data returned as web content, application/system output files and records.	Derived From: (1) Committee on National Security Systems (CNSS), Glossary (2) Peter Aiken, Virginia Commonwealth University; National Association of Chief Information Officers of the states (NASCIO) "Managing Data as a Strategic Asset", 04/28/2015.
Data Attribute	Any distinctive feature, characteristic, or property of a data object that can be identified or isolated quantitatively or qualitatively by either human or automated means. For example, a data object can be made up of one or more data elements, and a data element will typically have data attributes as sub-units.	Derive From: 1) ISO/IEC 27000

Data Catalog	A general term used to describe an environment where metadata is ingested or uploaded and is permanently managed, stored, archived long-term, preserved, and made accessible.	Reference Memo: Designation of Enterprise Services Supporting Federated Data Cataloguing
Data Categorization	A mechanism for establishing order through the grouping of related data, where members of a grouping bear some immediate similarity within a given context. Example groupings include mission INTs, subject, data format, language, and context use.	Derived From: (1) Jacob, E.K. (2004). Classification and categorization: A difference that makes a difference. Library Trends 52(3):515-540. (2) Digital Guardian Article: What is Data Classification? A Data Classification Definition by Juliana De Groot on Monday July 15, 2019 - https://digitalguardian.com/blog/what-data-classification-data-classification-definition.
Data Element	A discrete unit of data that has a unique meaning within a specific model or schema, and may be comprised of sub-units. Example data elements for a person may include last name, first name, and middle initial.	Derived From: (1) DAMA Dictionary of Data Management, 2nd Edition, 2011 (2) NIST Special Publication 800-47
Data Object	An instance of data that is discrete and bounded with an intrinsic, immutable, and unique identity that can persist independently of a system or service. A data object is made up of one or more data elements. For example, a row within a relational database or an image within an image library.	Derived From: (1) ISO/IEC/IEEE 31320-2, First Edition 2012-09-15, Information Technology - Modeling Languages - Part 2: Syntax and Semantics for IDEF1X97 (IDEFobject), p. 16, 18; (2) Newton's Telecom Dictionary, 17th Edition
Enterprise Service Provider (ESP)	Advancing Analytics (ADVANA) program, led by the Office of the Secretary of Defense (Comptroller) / Chief Financial Officer, as an Enterprise Service Provider for data cataloguing on the Unclassified and Secret fabrics, as well as the DoD's singular interface with the Federal Data Catalogue. "Enterprise Services for Federated Data Cataloguing Memo".	Reference Memo: Designation of Enterprise Services Supporting Federated Data Cataloguing
Data Services Environment	Provides an on-line repository enabling developers and data consumers to reuse, understand, and share existing data assets. It includes schemas, web service description language, stylesheets,	Reference: DoDI 8320.02

	taxonomies, descriptive metadata about proposed and approved Authoritative Data Sources, including their relationships and their responsible governance authorities, and descriptive, semantic, and structural metadata about services and other functional capabilities, including service definitions and specifications that can be discovered for subsequent use. The DSE has a Web-based interface with streamlined metadata registration and discovery capabilities that support the visibility of operational capabilities, data standards, and data needs. The DSE provides a number of interfaces supporting both design-time and run-time access to metadata, and it interacts with other registries and repositories through Open Search federation	
Data Set	One or more data objects that share common properties and characteristics and are managed as a unit. For example, the test scores of each military or civilian student in a particular class is a data set.	Derived From: (1) Data for the IC Enterprise Foundational Study (DICE FS) definition, 2017-2018
Data Source	A specific data set or repository from which data can be attained for subsequent use by consumers. A data source may be the combination of multiple, separate data sets or repositories.	Reference: DoDI 8320.07 SUBJECT: Implementing the Sharing of Data, Information, and Information Technology (IT) Services in the Department of Defense
Data Tag	Metadata applied, through tagging to a data asset to help describe characteristics about the data, such as privacy, security, provenance, source, or other information, and can be used to support automated processing. A "tag" is an assertion describing some aspect of a resource, pairing a semantic label with a value (e.g., a document may have a tag name of "Language" with a corresponding tag value of "English"). The tag values may be known a priori (e.g., controlled vocabulary) or not (e.g., folksonomies); marking and distribution statements for controlled technical information.	Derived From: (1) Information Sharing Environment Data Aggregation Reference Architecture (DARA). Prepared by PM, Information Sharing Environment. Version 1.0; December 2014. (2) Priority Objective 3, Data Tagging Functional Requirements, Version 1.0, December 2014.

Data Tagging	The act of associating data tags as metadata to a data asset by identifying, labeling, and describing its information. Typically, tagging supports user interpretation and automated processing.	Derived From: (1) ICS 500-21, Appendix B
Metadata	"Data about data"; administrative or descriptive data attributes that are consistent across mission and business disciplines, domains, and data encodings, and are used to improve business or technical understanding of data and data-related processes.	Derived From: (1) DAMA Dictionary of Data Management, 2nd Edition, 2011 (2) Level 0 IC Core Reference Architecture Integrated Dictionary, 12 June 2013
Structured Data	Content that conforms to a specific, pre-defined schema or data model, or is tagged or otherwise arranged into database tables (rows and columns). Examples include data in relational databases, data in graph databases, call data records, financial transactions, and system audit logs.	Derived From: (1) DAMA Dictionary of Data Management, 2nd Edition, 2011 (2) Gartner report, "Big Content: The Unstructured Side of Big Data - Darin Stewart" - 1 May 2013
Semi-structured Data	Data that has elements of both unstructured and structured data. For example, a MS Word document is generally considered to be unstructured data, but with the addition of metadata tags used to enable discoverability, the data is now semi-structured. Other types of semi-structured data formats include: XML and other markup languages, JavaScript Object Notation (JSON), email, and formats based on Electronic Data Interchange (EDI) standards (ex., X12, EDIFACT, ODETTE).	Derived From: (1) Hills ,Ted, NoSQL and SQL Data Modeling: Bringing Together Data, Semantics, and Software, Technics Publications, April 1, 2016., (2) Pääkkönen, P. & Pakkala, P. (2015). Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems. Big Data Research 2(4): 166-186)
Unstructured Data	Content that does not conform to a specific, pre-defined data model, or is not tagged or otherwise structured into database tables (rows and columns). Examples include documents, presentations, graphics, images, text, reports, videos, or sound recordings.	Derived From: (1) DAMA Dictionary of Data Management, 2nd Edition, 2011 (2) Gartner report, "Big Content: The Unstructured Side of Big Data - Darin Stewart" - 1 May 2013