INSTITUTE FOR DEFENSE ANALYSES

The Seven C's of Data Curation for the Two C's—Command and Control

Jonathan R. Agre
Marius S. Vassiliou
Karen D. Gordon

February 2015
Approved for public release;
distribution is unlimited.

IDA Document NS D-5441
Log: H 15-000183

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This work was conducted under IDA's independent research program (C6423). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-5441

The Seven C's of Data Curation for the Two C's—Command and Control

Jonathan R. Agre
Marius S. Vassiliou
Karen D. Gordon

20th ICCRTS

"C2, Cyber, and Trust"

Paper 021

The Seven C's of Data Curation for the Two C's – Command and Control

Topics:
(3): Data, Information, and Knowledge

Jonathan R. Agre
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
+1-703-933-6522
jagre@ida.org

Marius S. Vassiliou
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
+1-703-887-8189
+1-703-845-4385
mvassili@ida.org

Karen D. Gordon
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
+1-703-845-2343
kgordon@ida.org

Point of Contact:

Marius S. Vassiliou
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
USA
+1-703-887-8189
+1-703-845-4385
mvassili@ida.org

Abstract

Many important and complex C2 activities require the use of disparate data sources (structured and unstructured) that are time varying, at various levels of quality (completeness, accuracy, etc.), and of ambiguous origins. Currently, dealing with such disparate data is manually intensive and expensive, in large part because of problems with the quality of the data and its ability to be quickly processed. Data curation can enable automated data discovery, advanced search and retrieval, improvement in the overall data quality, and increased data reuse. The process can be described using what we call the "Seven C's" of data curation: (1) *Collect*—Interface to the data sources and accept the inputs; (2) *Characterize*—Capture available metadata; (3) *Clean*—Identify and correct data quality issues; (4) *Contextualize*—Provide context and provenance; (5) *Categorize*—Fit within framework that defines the problem domain; (6) *Correlate*—Find relationships among the various data; and, (7) *Catalog*—Store and make data and metadata accessible with application program interfaces (APIs) for search and analysis. The benefits of the data curation process are a reduction in problem-solving time, improved data quality, increased confidence in solutions, reduced time and manual effort to perform the curation itself, and the ability to solve problems that were previously too complex or time-consuming to solve because of data problems.

1. Introduction

An analysis of Command and Control (C2) failures in military operations, disaster response, and the response and run-up to terrorist attacks has shown that such failures generally manifest themselves as a lack of access to information or an absent, incomplete, irrelevant, delayed, or erroneous transfer of information from those who have it to those who need it [Vassiliou 2013, 2015]. Information of appropriate quality, at the right time, is crucial to the effective conduct of C2. Providing such information is made all the more difficult by the increasing profusion of data caused by growing sets of networked sensors and unmanned systems, social media and online repositories, multiple modes of communication, and ever-cheapening equipment and systems for data storage. All these factors also lead to the crucial problem of operator overload, particularly under the intense time pressure experienced in many operational environments [Shanker 2011].

Automation of data handling is one of the keys to overcoming the pressures of increasing data volume and reducing the severity of the data overload problem. Digital data curation refers to the methods and practices of preparing and managing data for use in computer-based analyses over the life cycle of the data. Data curation enables automated data discovery, advanced search and retrieval, improvement in the overall data quality, and increased data reuse. It may also be a key mechanism in reaching the Department of Defense data sharing objectives as well as enabling the use of widely distributed information that can be required for agile, decentralized command and control [Vassiliou 2015].

C2 sensemaking-related functions and processes—reasoning, inference, planning, decision making, collaborating, etc.—require use of disparate data sources (structured and unstructured) that are time varying, at various levels of quality (completeness, accuracy, etc.), and of ambiguous origins. Currently, dealing with such disparate data is manually intensive, expensive, and time consuming—in large part due to problems with the quality of the data and its ability to be quickly processed. This is especially true in large enterprises such as the U. S. Department of Defense (DoD) [Alberts 2012].

Data curation methods can better support C2 sensemaking by pre-positioning the data so that it can be quickly assembled to answer unanticipated questions, and maintaining the data in a form that facilitates automated processing by computer tools. The original data sources are augmented by metadata (i.e., information about the data) that reduces the burden of understanding and judging the utility of the data for a given purpose.

There are several recent advances that are leading to the increased utility and effectiveness of data curation methods. These include standardization of metadata in various domains, availability of text analytics and natural language processing software, "big data" computational methods [Berman 2013], network availability of domain-specific data repositories, visualization, and improved search capabilities.

2. Types of Data to be Curated

Data curation can be applied to structured, semi-structured, and unstructured data. Unstructured data is generally textual information, such as that found in a technical report, a news article, and so on. Unstructured data can also include photographic and video material. Structured data is information that has been formatted and associated with additional metadata that helps to define its meaning. For example, data in a database has a schema, which describes fields with labels and well-defined semantic meanings, such as "Last Name," so that this information can be processed by computing algorithms. Semi-structured data refers to data content that is labeled, but the data itself can be free-form text. For example, a response to a survey question may fall into this category.

Clearly, structured data has historically been more amenable to computer-based analyses. Techniques for working with unstructured or semi-structured data, enabled by natural language processing and text analytics, are increasingly becoming commercially available, so that analysis of textual data is becoming practical. Other emerging technologies such as machine learning, big data computational methods, and visualization are enabling advances in the analysis of unstructured and semi-structured data.

In general, the more structure or metadata that can be associated with the original data, the easier it is to perform analysis. The intent of data curation is to provide additional structure to the data, particularly the semi- and unstructured data so that automated analyses are possible with less manual effort. At the current stage of development, data curation is not a fully automatable procedure. Subsets of the problem can be automated but others still require a human-in-the-loop. However, both manual and automated data curation can reduce the overall effort significantly, increasing the timeliness of analyses and helping to reduce the operator overload problem. There is a clear relationship between data curation and data quality, and data quality metrics can be used to gauge the effectiveness of data curation. For example, data provenance can be described with metadata and used to help determine how much trust to put in the data.

Wolfram|Alpha [Wolfram 2010, Wolfram|Alpha 2014] is an example of a commercial capability incorporating data curation techniques. The system combines manual and automated methods to curate a variety of subject areas. The data is derived from primary sources such as mathematics, chemistry and physics text books, and then stored in an internal knowledge base. The data is checked by domain experts who periodically review the data items. Real-time data, such as weather or financial, are checked via models for reasonableness.

Applying data curation techniques to the domains of command and control will involve methods developed for the specific problems of C2 and the incorporation of methods developing in other domains, both scientific and commercial. Standardization of metadata formats, as well as terms and their meanings, is a critical aspect of the successful application of curation, regardless of whether it involves automation or manual functions. The adoption of common terminology and concepts facilitates the incorporation of metadata at each of the seven steps of curation providing the foundation for understanding the results of the previous steps.

The problem of data curation can be described using the "Seven C's" model.

3. The Seven C's of Data Curation

The data curation process is a sequence of steps that improves data quality and facilitates further data sharing, processing, and use. The process can be described using what we call the "Seven C's" of data curation:[1]

1. *Collect*—Interface to the data sources and accept the inputs
2. *Characterize*—Capture available metadata
3. *Clean*—Identify and correct data quality issues
4. *Contextualize*—Provide context and provenance
5. *Categorize*—Fit within a framework that defines the problem domain
6. *Correlate*—Find relationships among the various data
7. *Catalog*—Store and make data and metadata accessible with application program interfaces (APIs) for search and analysis

In a data curation process applied to an operational environment, the *collect* step involves automated procedures that capture the data, format it, and store it in an appropriate data repository, such as a relational database for structured data or a NoSQL[2] database for textual documents. The data should be stored in common standard formats such as the Extensible Markup Language (XML) [W3C 2008] or Java Script Object Notation (JSON) [ECMA 2013].

Characterization is applied as data is captured, where additional agreed-upon metadata such as creation time, capture method, sensor description and settings, accuracy, precision, location, and so on, are supplied and recorded along with the data. The appropriate characterization data is dependent on the domain and possible uses of the data. Standardization activities at this level are beginning to emerge for various disciplines, primarily in the medical and biological research domains.

In the *clean* step, basic data quality tools [Agre 2011] are applied to the data to identify and eliminate the issues with the data. Some of the possible problems with the data include erroneous, corrupted, incomplete, duplicate, or unnecessary data. Data cleansing tools will attempt to correct, delete, or replace the problematic data [Rahm 2000]. Many data cleaning techniques are well known in the database community, where they are often implemented as extract, transform, and load (ETL) processes. Methods to clean the

[1] Derived from discussions in Higgins (2008) and Borne (2010).

[2] "Not SQL," or "Not Only SQL," where SQL is Structured Query Language. See Grolinger et al. (2013).

data can vary in complexity from fixing typographical errors, to using artificial intelligence techniques to infer missing relations, to converting to alternative standard representations. It is desirable to clean the data as early as possible in the data curation process, since the cost to fix data problems increases as time passes. These methods are also being extended to unstructured data.

The *contextualize* process is dependent on the context or specific problem domain, as well as on the possible uses of the data. These aspects will inform what additional metadata, such as authentication and other provenance information, is required. For example, an Intelligence Community application may require a higher level of provenance information than a routine logistical request. The domain may also dictate particular formatting or representation of the metadata that is best suited for the data.

Categorization further identifies key properties of interest that can be found in the data. Natural language processing, text analytics, and machine learning can be used on semi-structured and unstructured data to identify and extract concepts from the data that are of interest. Image analysis can be used to identify key features in images or video files. The particular concepts extracted are dependent on the problem domain. For example, sentiment analysis can be used to extract opinions from blog data concerning a new commercial product or a proposed policy initiative.

A *correlation* procedure can be applied across the heterogeneous collection of stored data to match and identify data and concepts. An example would be the temporal or geographic alignment of data for feeding to a target recognition procedure. Other examples from database technology fall under the categories of data integration and entity resolution, and can be quite complicated. For example, gathering all the medical records that belong to a given person is a well-known problem in the healthcare field. This step results in various relations being defined, and these are also stored in a repository. Graph databases or triple stores are considered efficient tools for this purpose [Aasman 2012]. Determining the correlations can be computationally intensive when comparing several large data sets. Parallel or distributed processing techniques such as those employed in big data analyses may be useful at this stage.

Last, in the *catalog* step, the data and its metadata are stored and preserved for their life cycle, and prepared for dissemination (e.g., posting to a data repository, pushing to designated consumers, or indexing for rapid retrieval). Application program interfaces (APIs) can be provided for search, extraction, and basic analyses of the data, typically implemented as web services. Data repositories tailored to the needs of a specific domain are often used at this stage to store and preserve the data and the metadata. Tailored search engines are also often associated with the repository.

The benefits of the data curation process are a reduction in problem-solving time, improved data quality, increased confidence in solutions, reduced time and manual effort to perform the curation itself, and the ability to solve problems that were previously too complex or time-consuming to solve because of problems with the data [Day 2008]. Curated data can be more easily shared, because curation makes it easier for people other than the data's creators to determine its semantic content. Eventually, digitally curated data can be fed to automated reasoning tools (such as Wolfram|Alpha [2014] or IBM's Watson [Fan 2012]) so the analyses can be rapidly accomplished and visualized.

4. Emerging Data Curation Ecosystem

There have been a number of recent activities resulting in large repositories of curated digital data in the domain of scientific publications and scientific datasets. These activities are developing and establishing "best practices" for data curation. Several prestigious scientific journals (e.g., *Science* and *Nature*) are now requiring that authors provide the data used in their articles in an accessible form [McNutt 2014]. A primary motivation is to allow the reproducibility of the research results for validation and reproducibility purposes. Although there are variations, the journals generally are leaving the specific formats, metadata, and repositories up to the various disciplines. For example, genomic data and chemical data follow different guidelines and use different repositories.

In 2013, the White House issued an Executive Order [May 9, 2013] making "open and machine-readable data" a requirement for government information and directing the issuance and implementation of a new Open Data Policy, released as OMB memorandum M-13-13. The U.S. Federal Government has also been moving toward making data resulting from federally funded research activities available to the public, as specified by a White House Office of Science and Technology Policy memorandum requiring each federal agency to create a data management plan [Holdren 2013]. Agencies are currently formulating their plans, but will probably follow the lead of the journals in allowing domain-specific guidelines.

Recently, as demand for access to data has increased, active metadata standards groups and instances of public repositories are being reported. For example, DataCite[3] defines a metadata schema for publication and citation of research datasets. In this domain of identifying scientific publications and data the following organizations are active:

- CrossRef[4] provides Digital Object Identifiers (DOIs) for articles
- DataCite provides DOIs for datasets
- FundRef[5] provides DOIs for funders and organizes funders hierarchically
- ORCID[6] provides IDs for researchers

These organizations maintain repositories of metadata, which facilitate search and enable efficient linking among funders, authors, articles, and data. For example, an analyst can use FundRef to find all the articles resulting from research funded by the U.S. Department of Energy Office of Science.

Many data repositories have been established with most being discipline-specific, but some general-purpose ones exist. For example, Data.gov[7] has been implemented for U.S. government data, the Dryad Digital Repository[8] for non-specific fields, and Harvard's DataVerse[9] for social science and other fields. There are

[3] http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf
[4] http://www.crossref.org
[5] http://www.crossref.org/fundref/
[6] http://orcid.org
[7] http://www.data.gov
[8] http://datadryad.org
[9] http://thedata.harvard.edu/dvn/

now registries of data repositories such as Databib[10] and the Registry of Research Data Repositories (re3data.org).[11, 12]
Although these activities fall outside the domain of C2, they represent an important movement toward acceptance of data curation as part of the normal cycle of data handling in domain-specific research and general data sharing.

The U.S. Department of Defense (DoD) has been working to implement a data sharing strategy since 2003 [DoD CIO 2003], including a recently updated version [DoD 2013] defining policies, responsibilities, and procedures. This effort has resulted in the adoption of metadata standards for various DoD functions including the definition of authoritative data sources, data repositories, data exchange standards, discovery metadata standards, and metadata search services.

Experts in military command and control have been attempting to define a standard for military digital data exchange for many years, such as through efforts to define the Universal Core (UCore) and C2 Core [Allen 2009]. Recently, the DoD has adopted the National Information Exchange Model (NIEM) as the foundation for its data exchange standards, dropping further development of UCore and C2 Core [Takai 2013]. A special military operations (MILOPS) domain within the NIEM framework was established that drew from the previous work on UCore and C2 Core [Renner 2014]. NIEM consists of a collection of core components that are common across the communities involved. These communities range from the original law enforcement groups to various other government activities such as maritime, military, and infrastructure protection. Each community is able to define and manage a domain that contains terminology and concepts specific to that group. The MILOPS standard is expected to achieve a slow and steady increase in the number of C2 applications that employ it for data exchange.

As another part of the data sharing effort, the DoD has defined the DoD Discovery Metadata Specification (DDMS) as the common set of descriptive metadata elements that are to be associated with each data asset so that it can be discovered in the DoD enterprise [DISA 2012]. DDMS relies on several existing standards (primarily XML-based) including:

- Data Encoding Specification (DES) for Trusted Data Format (TDF), DES for Enterprise Data Header (EDH), Access Rights Handling (ARH) developed by the Intelligence Community for DDMS structure and security
- DES for Information Security Metadata (ISM) for security markings
- Time, Space, and Position Information (TSPI) from the Open Geospatial Consortium (OGC) Geospatial Markup Language (GML) specifications
- Geopolitical Entities, Names, and Codes (GENC) Standard for codespace and code definitions
- Worldwide Web Consortium (W3C) XLink specification for links within and between documents

The DDMS, in turn, is part of the DoD Data Services environment (DSE), [13] an on-line repository containing the structural and semantic metadata information and the services critical to enabling publication, discovery and search of existing data assets. It holds metadata on a) services such as schemas,

[10] http://databib.org
[11] http://re3data.org
[12] Databib and re3data.org are merging under the auspices of DataCite.
[13] https://metadata.ces.mil/dse-help/en/About DSE

web service description language, stylesheets, and taxonomies, b) descriptive metadata about proposed and approved DoD Authoritative Data Sources (ADSs), and c) the DDMS records. The DSE contains search capabilities that utilize the metadata to discover the data assets. The information in the DSE is overseen and curated by managers and administrators that oversee the metadata in the namespaces, services, and the ADSs.

Currently, the DDMS and DSE are mainly concerned with data that is fairly static, and the curation is primarily manual. The DDMS and DSE represent important steps in the direction of automated data curation; however, there remains much to be done to automate the processes.

5. Challenges

In most current data curation practice, the activities remain largely manual. As a result, the process can take too long to be effective in many tactical situations. Warfighters trying to assess a tactical situation may face too great a volume of digital textual information to process in time to take action. In general, since the data is not prioritized, tagged with provenance, or summarized, the warfighter has no assistance in culling the information and can quickly experience data overload. Natural language processing, as well as text analytics methods for text and image analysis for image and video data, need to be adapted and extended to be useful for this purpose.

The processing of real-time information is a clear case where automation is required. The increased use of video and other sensor data from unmanned air vehicles or other data-intensive sources can produce more data than can be processed within the data's useful lifetime. Image processing, image understanding, and scene analysis methods are still not sufficient to meet the needs.

6. Conclusions

The concept of automated digital data curation can be described in terms of seven steps that augment basic data with descriptive information that increases the utility of the data for analysis purposes and enables increased sharing of the data.

Current practices in the DoD, the Federal Government, and much of the scientific research community involve manual data curation activities. It does not appear that these methods are sustainable given the increasing amount of digital data available and the number of applications that are using these big data sources. Further automation is required to break the bottleneck caused by having humans-in-the-loop to create the desired metadata.

The utility of data curation for command and control is still under investigation. The benefits are clear, but automating the steps so they are accomplished in a timely manner remains a major challenge. Additional research must be directed toward better defining and automating the data curation steps for actual C2 applications under realistic conditions.

References

Aasman, Jan (2012). Graph Databases, Triple Stores and their Uses, NoSQL Conference, San Jose, 2012. http://wilshire.skyworld.com/uploads/handouts/WED_1415_Aasman_Jans_Color_4837.pdf (accessed February 2, 2015)

Agre, Jonathan, M. S. Vassiliou, and Corinne Kramer (2011). Science and Technology Relating to Data Quality in C2 Systems, *Proc. 16th International Command and Control Research and Technology Symposium*, Quebec City, Canada. http://dodccrp.org/events/16th_iccrts_2011/papers/031.pdf (accessed February 12, 2015).

Alberts, David S., M. S. Vassiliou, and Jonathan Agre (2012). C2 Information Quality: An Enterprise Systems Perspective, MILCOM, San Diego, Nov. 2012.
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6415625 (access February 12, 2015).

Allen, M., David, Catherine Macheret, and Mary Ann Malloy (2009). C2Core and UCore Message Design Capstone: Interoperable Message Structure, MITRE, September, 2009.
https://www.mitre.org/sites/default/files/pdf/09_4364.pdf (accessed February 2, 2015)

Berman, Jules K. (2013). *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. New York: Elsevier.

Borne, Kirk (2010). Astroinformatics: Data-Oriented Astronomy Research and Education. *Journal of Earth Science Informatics*, Vol. 1, No. 3, 5–17.

Day, Michael (2008). Current and Emerging Scientific Data Curation Practices. Presentation from 4th Summer School on Preservation in Digital Libraries, Tirrenia, Italy, June 12. http://www.slideshare.net/michaelday/research-data (accessed February 12, 2015).

Defense Information Systems Agency (DISA) (2012). Department of Defense Discovery Metadata Specification (DDMS), Version 5.0, January 11, 2013.

Department of Defense Chief Information Officer (DoD CIO) (2003). Department of Defense Net-Centric Data Strategy," May 9, 2003.
http://dodcio.defense.gov/Portals/0/Documents/DIEA/Net-Centric-Data-Strategy-2003-05-092.pdf (accessed February 12, 2015).

DoD CIO (2013). Sharing Data, Information, and Information Technology (IT) Services in the Department of Defense, DoD Instruction Number 8320.02, August 5, 2013.
http://www.dtic.mil/whs/directives/corres/pdf/832002p.pdf (accessed February 2, 2015)

ECMA International (2013). The JSON Data Interchange Standard, ECMA-404, 1st Edition, October, 2013.
http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf (accessed February 12, 2015).

Executive Order -- Making Open and Machine Readable the New Default for Government Information
http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

Fan, J., A. Kalyanpur, D.C. Gondek, and D.A. Ferrucci (2012). Automatic Knowledge Extraction from Documents. *IBM Journal of Research and Development,* Vol. 56, Nos. 3/4, 5:1–5:10.

Grolinger, Katarina, Wilson A. Higashino, Abhinav Tiwarim and Miriam A.M. Capretz (2013). "Data Management in Cloud Environments: NoSQL and NewSQL Data Stores." *Journal of Cloud Computing: Advances, Systems, and Applications* Vol. 2 No. 22, 1-24. http://www.journalofcloudcomputing.com/content/pdf/2192-113X-2-22.pdf

Higgins, Sarah (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation,* Vol. 3, No. 1, 134–140. http://www.ijdc.net/index.php/ijdc/article/viewFile/69/48 (accessed August 11, 2014).

Holdren, John P. (2013). Memorandum on "Increasing Access to the Results of Federally Funded Scientific Research," Office of Science and Technology Policy (OSTP), February 22, 2013. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf (accessed February 2, 2015)

McNutt, Marcia (2014). Journals unite for reproducibility, Joint Editorial (published online November 5, 2014, by *Science* and *Nature*), *Science*, Vol. 346, No. 6210 (November 7, 2014), 679. http://www.sciencemag.org/content/346/6210/679.full (accessed February 12, 2015).
Also in *Nature*, Vol. 515, No. 7525, 7 (06 November 2014). http://www.nature.com/news/journals-unite-for-reproducibility-1.16259 (accessed February 12, 2015).

Office of Management and Budget (OMB) Memorandum M13-13. "Open Data Policy – Managing Information as an Asset," May 9, 2013. http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf (accessed February 2 2015).

Rahm, Erhard, and Hong Hai Do (2000). Data Cleaning: Problems and Current Approaches, University of Leipzig, Germany. http://dc-pubs.dbs.uni-leipzig.de/files/Rahm2000DataCleaningProblemsand.pdf (accessed February 2 2015).

Renner, Scott (2014). National Information exchange Model (NIEM): DoD Adoption and Implications for C2, *19th International Command and Control Research and Technology Symposium (ICCRTS)*, Alexandria, VA, 18 June 2014. http://www.dodccrp.org/events/19th_iccrts_2014/post_conference/presentations/129.pdf
(accessed February 2, 2015)

Shanker, Thom, and Matt Richtel (2011). In New Military, Data Overload Can Be Deadly. *New York Times*, January 16. http://www.nytimes.com/2011/01/17/technology/17brain.html?pagewanted=all&_r=0 (accessed August 11, 2014).

Takai, Teri (2013). Memorandum on "Adoption of the National Information Exchange Model within the Department of Defense," Department of Defense Chief Information Office, March 28, 2013. http://dodcio.defense.gov/Portals/0/Documents/2013-03-28_Adoption_of_the_NIEM_within_the_DoD.pdf (accessed February 12, 2015).

Vassiliou, Marius, and David S. Alberts (2013). "C2 Failures: A Taxonomy and Analysis." *Proc. 18th International Command and Control Research and Technology Symposium* (ICCRTS).

Vassiliou, Marius S., David S. Alberts, and Jonathan Agre (2015). *C2 Re-envisioned: The Future of the Enterprise*, CRC Press, Taylor and Francis Group, Boca Raton, FL, January, 2015.

Wolfram, Stephen (2010). Making the World's Data Computable. Keynote Speech at the Wolfram Data Summit, Washington, DC, September 9. http://blog.stephenwolfram.com/2010/09/making-the-worlds-data-computable (accessed August 11, 2014).

Wolfram|Alpha (2014). Making the World's Knowledge Computable. Champaign, IL: Wolfram Research. http://www.wolframalpha.com/about.html (accessed August 11, 2014).

W3C (2008). Extensible Markup Language (XML) 1.0, Fifth Edition, W3C Recommendation, 26 November 2008. http://www.w3.org/TR/2008/REC-xml-20081126/ (accessed February 12, 2015).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED (From–To)**
xx-02-2015	Final	Aug 2014 – Feb 2015

4. TITLE AND SUBTITLE

The Seven C's of Data Curation for the Two C's — Command and Control

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Agre, Jonathan R.
Vassiliou, Marius S.
Gordon, Karen D.

5d. PROJECT NUMBER

5e. TASK NUMBER
AE-6-6423 (C6423)

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document NS D-5441
Log: H15-000183

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

10. SPONSOR/MONITOR'S ACRONYM(S)

IDA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Many important and complex C2 activities require the use of disparate data sources (structured and unstructured) that are time varying, at various levels of quality (completeness, accuracy, etc.), and of ambiguous origins. Currently, dealing with such disparate data is manually intensive and expensive, in large part because of problems with the quality of the data and its ability to be quickly processed. Data curation can enable automated data discovery, advanced search and retrieval, improvement in the overall data quality, and increased data reuse. The process can be described using what we call the "Seven C's" of data curation: (1) Collect—Interface to the data sources and accept the inputs; (2) Characterize—Capture available metadata; (3) Clean—Identify and correct data quality issues; (4) Contextualize—Provide context and provenance; (5) Categorize—Fit within framework that defines the problem domain; (6) Correlate—Find relationships among the various data; and, (7) Catalog—Store and make data and metadata accessible with application program interfaces (APIs) for search and analysis. The benefits of the data curation process are a reduction in problem-solving time, improved data quality, increased confidence in solutions, reduced time and manual effort to perform the curation itself, and the ability to solve problems that were previously too complex or time-consuming to solve because of data problems.

15. SUBJECT TERMS

C2, Command and Control, Data, Databases, Data Repositories, Data Curation

16. SECURITY CLASSIFICATION OF:			**17.LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			David Alberts
Uncl.	Uncl.	Uncl.	SAR	14	**19b. TELEPHONE NUMBER (include area code)** (703) 845-2411

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18