



DMVitals: A Data Management Assessment Recommendations Tool

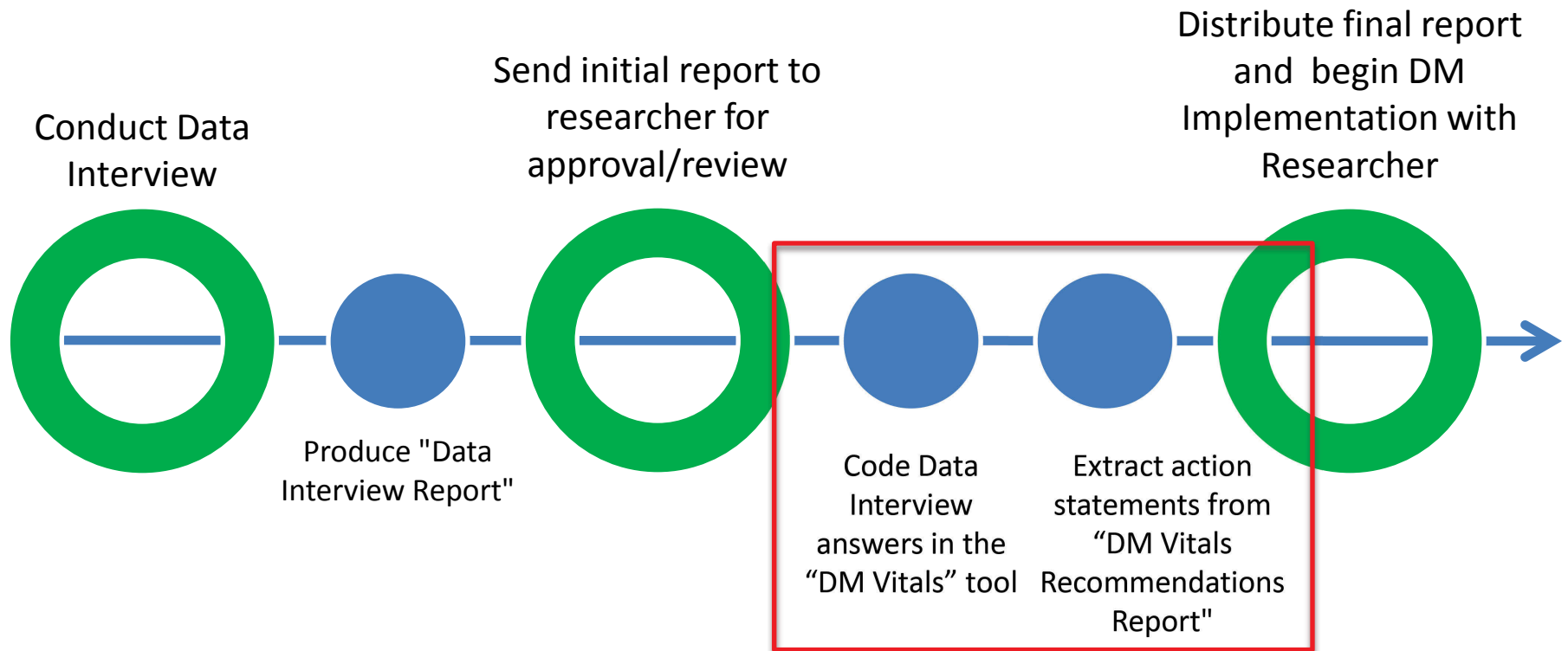
Andrew Sallans, Head of Strategic Data Initiatives
Sherry Lake, Senior Scientific Data Consultant

IASSIST 2012 - June 6, 2012

Interviews/Assessment Preface

- Over past two years we conducted about 25 data interviews
 - Focus on learning about research data practices at UVa and identifying service needs/opportunities
 - Intention of leading into consulting opportunities
- Ended up with conundrum of how to manage “unique” conditions of each research environment against common characteristics of data management within domains and institutional framework

Consulting Workflow



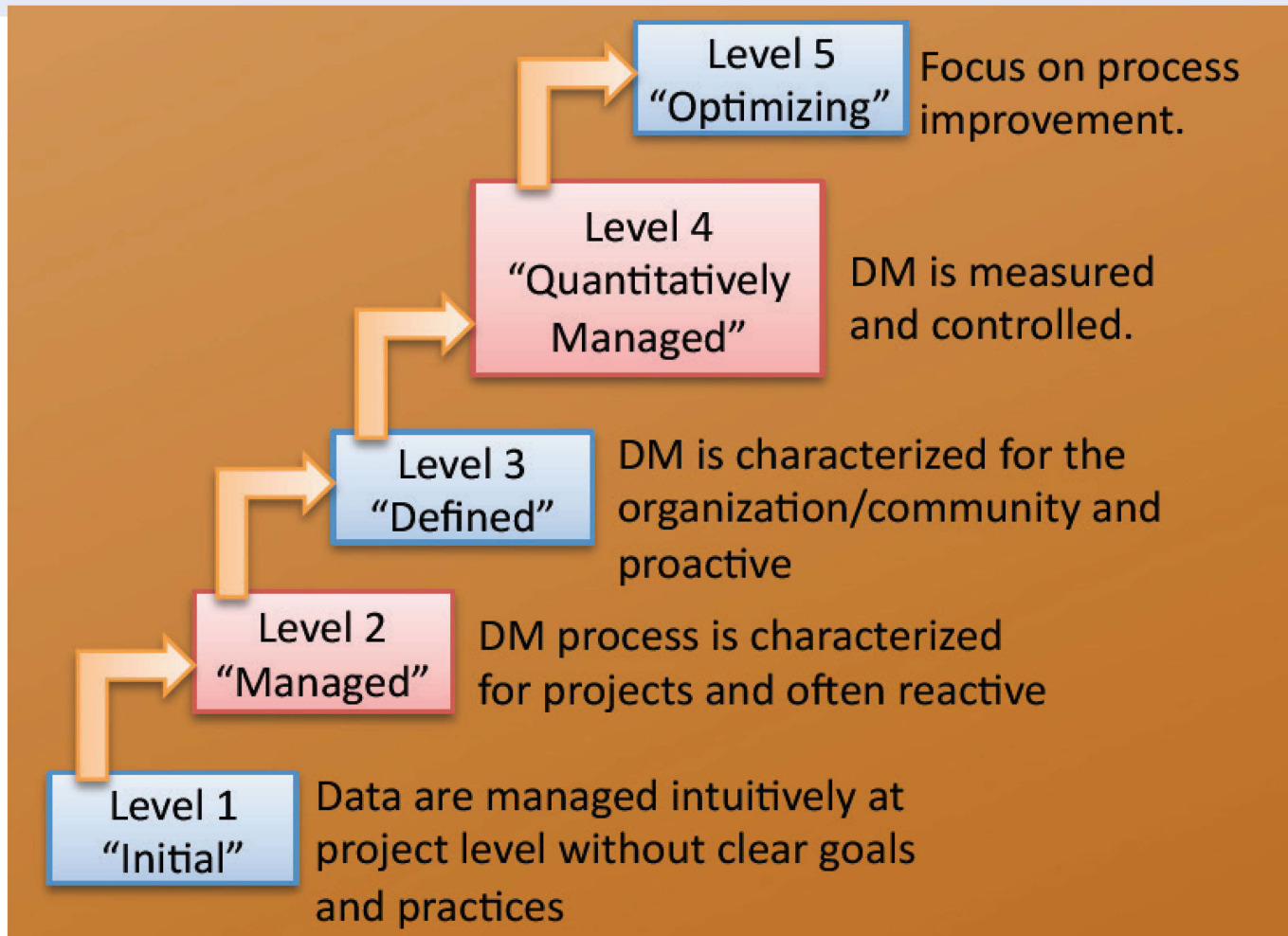
Recommendation Requirements

- Must be a fast process
- Must create actionable and repeatable recommendations
- Must reduce subjectivity
- Must weigh all assessment factors
- Must address present DM condition while showing path for improvement

Components of the DMVitals

- Data management best practice statements
 - UVA sources (ISPRO, SciDaC Guidelines)
 - ANDS long-term sustainability scoring model
- 8 data management categories
- Data interview questions and responses
- Data management maturity index of Crowston & Qin Capability Maturity Model (CMM) for Scientific Data Management (SDM)

Crowston & Qin Capability Maturity Model for SDM



Crowston, K. & J. Qin. (2010). A capability maturity model for scientific data management. In: Proceedings of the American Society for Information Science and Technology, October 24-26, 2010, Pittsburgh, PA. (Poster)

DMVitals Interview Sheet

2	2.1 Describe your day-to-day work with regard to data. What data do you have? What kind of data, how are they created, what type and formats and what software do you use? How much data do you have?		2.2 Are these files yours or do they belong to a wider group or to the institution? Who owns the Intellectual Property rights of the data you create?		3.1 Do you have a data management plan? Who is responsible for managing the data? Are you using any filing or naming conventions for the files? How are the files organized? Is there any documentation on the files and/or data fields?		3.2 How do you share data - among lab group or other colleagues (e-mail, shared drive, removable devices, CD, web pages, other)? Have you had version control issues with many people working on the same data file?		4.1 What challenges have you faced in terms of storage, formats, costs, and continued access to older data?		5.1 Have you been asked to provide or share your data? Could or should your data be reused or repurposed by others, and if so, how and by whom?
3	2.1.1 General Category (experimental, simulation/computational, observational, derived/compiled)	NO	Have you read UVA's Laboratory Notebook and Recordkeeping policy?		3.1.1 Management Plan		3.2.1 File sharing		4.1.1 Do they have older files?		<input type="checkbox"/> 5.1.1 Publisher requirement
4	2.1.2 Creation (sensors, instruments, software)	NO	Have you read UVA's Ownership Rights in Copyrightable Material policy?	NO	DMP only exists in researcher's mind, has not been communicated to research team		3.2.2 Version control issues:	NO	All file formats or data types are current		<input type="checkbox"/> 5.1.2 Funder requirement
5	2.1.3 Data Type (docs, emails, databases, images, videos, etc.)			NO	Basic, informal DMP exists and has been communicated	NO	Versions are managed		4.1.2 Obsolete data formats		5.1.3 Restrictions (Confidentiality, Sensitivity)
6	2.1.4 Data Format (MS Word, Excel, spss, html, jpg, etc.)			NO	DMP has been improved to include all 8 categories.	NO	File changes are recorded	NO	Up-to-date data formats	NO	Data stored securely
7	NO Avoids proprietary software file types			NO	DMP has been reviewed by SciDaC	NO	Record every change to a file, no matter how small (Log files)		4.1.3 Obsolete media	NO	Encrypted sensitive data
8	NO Software agnostic file formats used			NO	DMP is being followed by all research team members.	NO	Uses File Version Control (perhaps SVN)	NO	Data stored on readable media	NO	Data is de-identified
9	NO Open standard representation (ASCII, Unicode, CSV, txt)				3.1.2 Naming Conventions:	NO	Making original document "read only"		4.1.4 Storage space (Also see 3.1.6 & 3.1.7)	NO	Following regulations for protecting and backing up data
10	NO Common software or filetype used by the research community (may be proprietary)			NO	Using file naming conventions				<input type="checkbox"/> 4.1.5 Costs	NO	Encryption is always used when storing and transferring sensitive data.
11	2.1.5 Amount (#files, files sizes, growing?)			NO	Using file naming conventions for specific disciplines					NO	Follows UVA's "Electronic Data Removal" policy
12	2.1.6 Software				3.1.3 File Organization					NO	Faculty/staff who administer sensitive data are following appropriate federal, state, grant agency, or university regulations for protecting and backing up data.
				NO	Files are organized and can be					NO	Cryptology technologies for data storage and transmission of data are based upon

Interview

Report

FileFmtsDataTypes

OrgFiles

SecStrgBackups

CopyrightPrivConfid

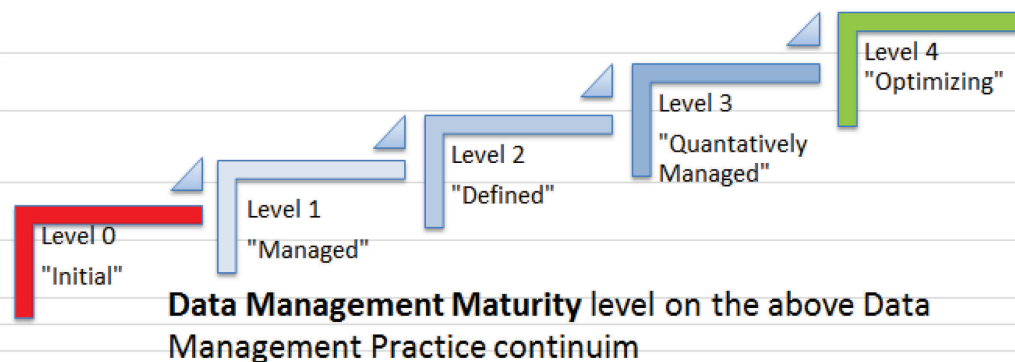
DataDocMetadata

DMVitals Section View

1	Least Sustainable (*1)		Fair (*2)		Satisfactory (*3)		Good (*4)		More Sustainable (*5)			
2												
3			NO	Using a consistent file structure					NO	Record every change to a file, no matter how small (Log files)		
4			NO	Using file naming conventions			NO	Using file naming conventions for specific disciplines	NO	Uses File Version Control (perhaps SVN)	Total YES:	0
5	NO	File changes are recorded					NO	Making original document "read only"	NO	Use Same Structure for Backups	Total Possible:	30
6	NO	Versions are managed									Ratio	0
7												
8	NO	Files are organized and can be found										
9												
10												
11	0	Total YES	0	Total YES	0	Total YES	0	Total YES	0	Total YES	0	Total
12												
13	3	Total Possible	4	Total Possible	0	Total Possible	8	Total Possible	15	Total Possible	30	Total
14												
15												
16												
17												

DMVitals Report View

1				
2		Sustainability Index		
3			Ratio	
4		File Formats Data Types	0%	Least Sustainable
5		Organization of Files	0%	Least Sustainable
6		Security Storage Backups	0%	Least Sustainable
7		Copyright Privacy Confidentiality	0%	Least Sustainable
8		Data Documentation Metadata	0%	Least Sustainable
9		Sustainability Index Average	0%	Least Sustainable
10				
11		Data Management Maturity Level		Initial



	Applicable?	Interview Topic	Sustainability	Phase	Action Statement
16	X	Has documented data sources used	1	1	Properly document data sources used.
17	X	Basic, informal DMP exists and has been communicated	1	1	A DMP is the basis of all data management, and is a critical tool in protecting the continuity of your research process. Once in place, it can continually be updated, provided to new members of the lab as guidelines, and easily be applied to future grant proposals. We will work with you to develop an appropriate plan for managing your data. This is a fundamental first step in improving process.
18	X	Have you read UVA's Laboratory Notebook and Recordkeeping policy?	1	1	Read Uva's "Laboratory Notebook and Recordkeeping" Policy: https://policy.itc.virginia.edu/policy/policydisplay?id='RES-002
19	X	Have you read UVA's Ownership Rights in Copyrightable Material policy?	1	1	Read Uva's "Ownership Rights in Copyrightable Material" Policy: https://policy.itc.virginia.edu/policy/policydisplay?id='RES-001

Consulting Recommendations View

Data Management Consulting Recommendations

Example

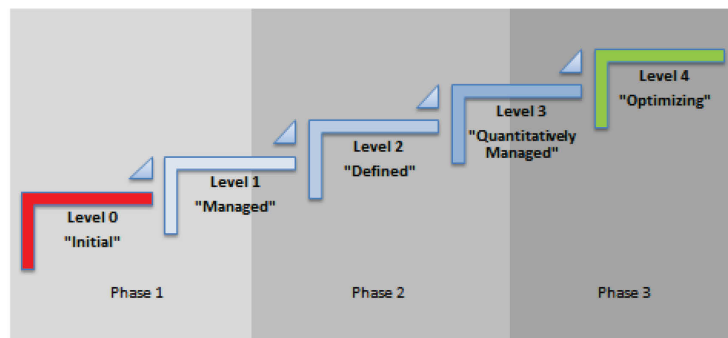
This document is the step-by-step set of instructions for the consulting recommendations.

In response to our discussion regarding your data management practices, we have prepared the following document to outline the primary data management issues and a series of solutions (in phases) that will hopefully help you address these issues. We would be glad to work through implementing these solutions with you and/or your designated data manager.

In addition to these recommendations, we will also use the information collected in the interview as we develop policy and infrastructure recommendations for overall University research data management. Your contributions are most appreciated.

Primary Data Management Issues (From Excel Worksheet)

	Ratio	Level
File Formats Data Types	46	Satisfactory
Organization of Files	33	Fair
Security Storage Backups	12	Least Sustainable
Copyright Privacy Confidentiality	50	Satisfactory
Data Documentation Metadata	15	Least Sustainable
Cumulative	31	Fair
Data Management Maturity Level	20% - 40%	Managed



Scientific Data Consulting Group (SciDaC) / February 2012

Page 1

© 2012 by the Faculty and Visitors of the University of Virginia.
This work is made available under the terms of the Creative Commons
Attribution-ShareAlike 3.0 license: <http://creativecommons.org/licenses/by-sa/3.0/>

Data Management Consulting Recommendations

Example

Phase 1 (short-term)

- Read Uva's "Ownership Rights in Copyrightable Material" Policy: <https://policy.itc.virginia.edu/policy/policydisplay?id=RES-001>
- Create and use standardized filename convention schema
- Use non-proprietary or non-software specific file types
- Properly document the context of the data collection

Phase 2 (long-term)

- Use file types and data formats that are commonly used by your research community
- Document structure and organization of data files
- Begin applying appropriate metadata standard to data and other research materials

Phase 3 (future)

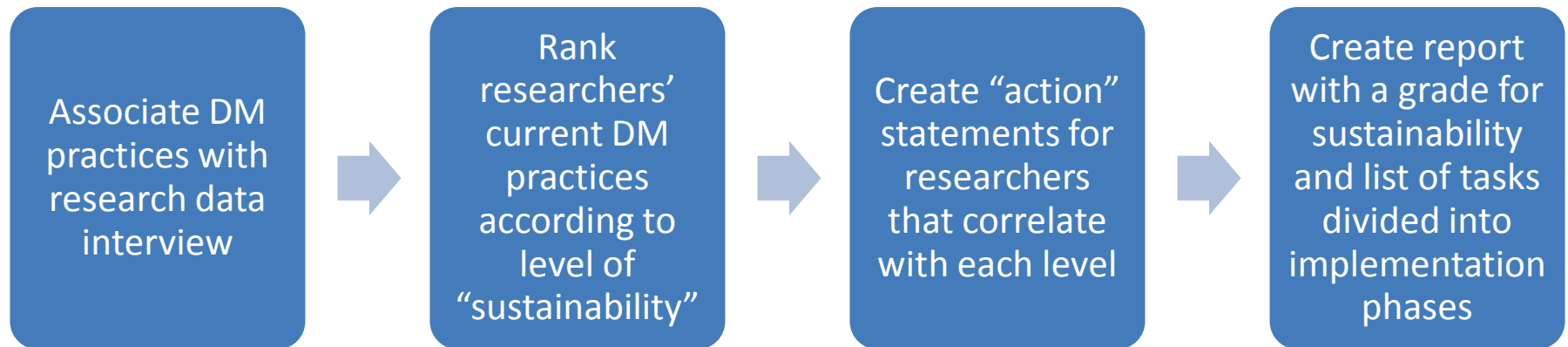
- Have SciDaC review DMP
- Document the analysis process from raw data to "finished" analyzed data.
- Store data storage media and servers in a physically secure environment
- Have data backups formally administered by ITC or some other system administrator

Scientific Data Consulting Group (SciDaC) / February 2012

Page 2

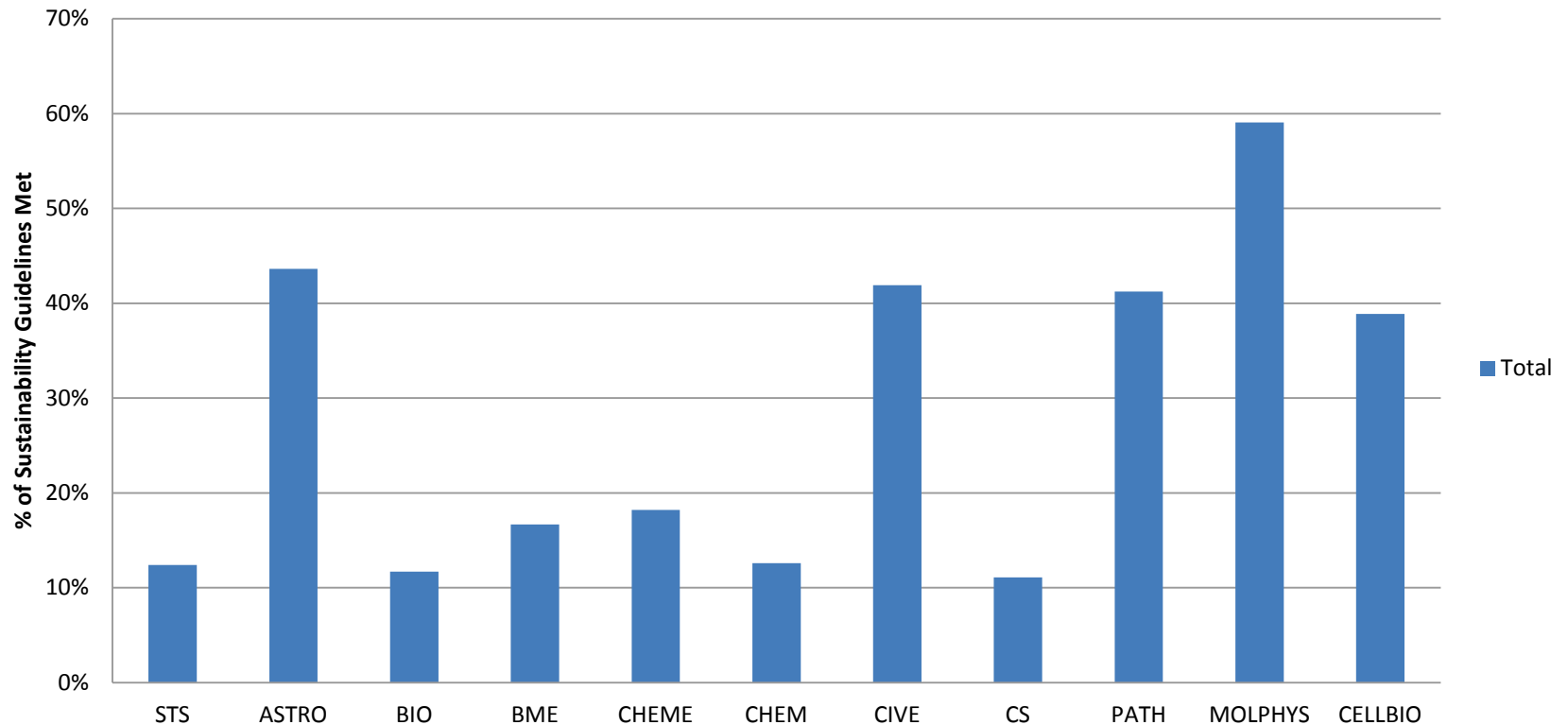
© 2012 by the Faculty and Visitors of the University of Virginia.
This work is made available under the terms of the Creative Commons
Attribution-ShareAlike 3.0 license: <http://creativecommons.org/licenses/by-sa/3.0/>

DMVitals Workflow Recap



DMVitals for Aggregate Learning

**Research Data Management Sustainability
(Example - not real data)**



Major Challenges

1. Assessment tool design, specifically dealing with appropriate weighting, false positives, double negatives
2. Social/ethical implications of giving such focused feedback and criticism to researchers
3. Broader issue of motivations and incentives

DMVitals Next Steps

- Release plan for others to use
 - Starting to develop versions of package
 - Will begin to make stable releases available on our website, along with development roadmap
- Collaboration opportunities for expansion
 - Interested in collaborations to drive further development and integration into services
 - Seeking collaborators now!

References

- Australian National Data Service. (2011) ANDS and Data Storage. Available: <http://ands.org.au/guides/storage.html>. Last accessed May 30, 2012.
- Crowston, K., & Qin J. (2010). A capability maturity model for scientific data management. *American Society for Information Science and Technology Annual Meeting*. Pittsburg, PA. Working Paper available: <http://crowston.syr.edu/content/capability-maturity-model-scientific-data-management-0>. Last accessed May 23, 2012.
- Digital Curation Center. (2011). CARDIO. Available: <http://cardio.dcc.ac.uk/>. Last accessed May 30, 2012.
- Information Technology Security (2010). University of Virginia Information Technology Security Risk Management (ITS-RM) Program. Available: http://its.virginia.edu/security/riskmanagement/docs/ITS-RM_3-0.pdf. Last accessed May 23, 2012.
- University of Virginia Library (2011). Scientific Data Consulting Data Management Home. Web Site available: <http://www.lib.virginia.edu/brown/data/>. Last access May 30, 2012.

Acknowledgements and Contact Information

- Acknowledgements

- Susan Borda

- UVa SciDaC Intern Summer 2011
 - Recent graduate of Syracuse GSLIS program
 - Starting as Digital Curation Librarian at University of California – Merced this summer

- Contact Information

- Andrew Sallans, als9q@virginia.edu
 - Sherry Lake, slake@virginia.edu