

## Establishing Trust in Data Curation: OAIS and TRAC applied to a Data Staging Repository (DataStar)

**Gail Steinhart**  
Cornell University Library

**Ann Green**  
Digital Life Cycle Research  
& Consulting

**Dianne Dietrich**  
Cornell University Library

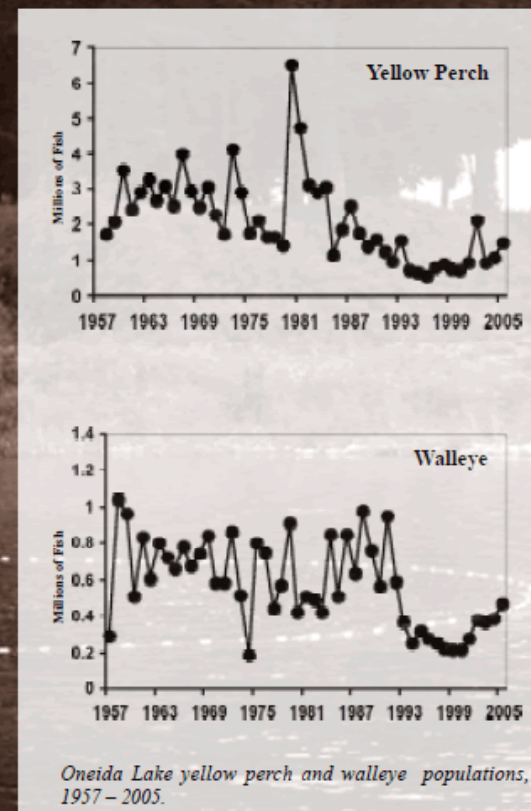
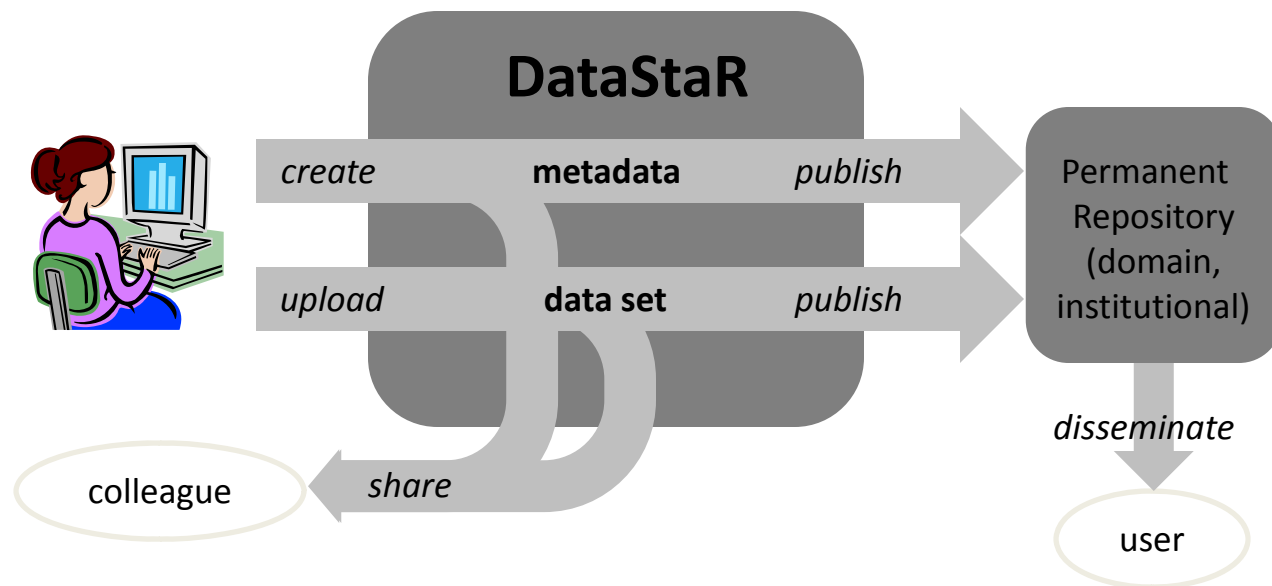
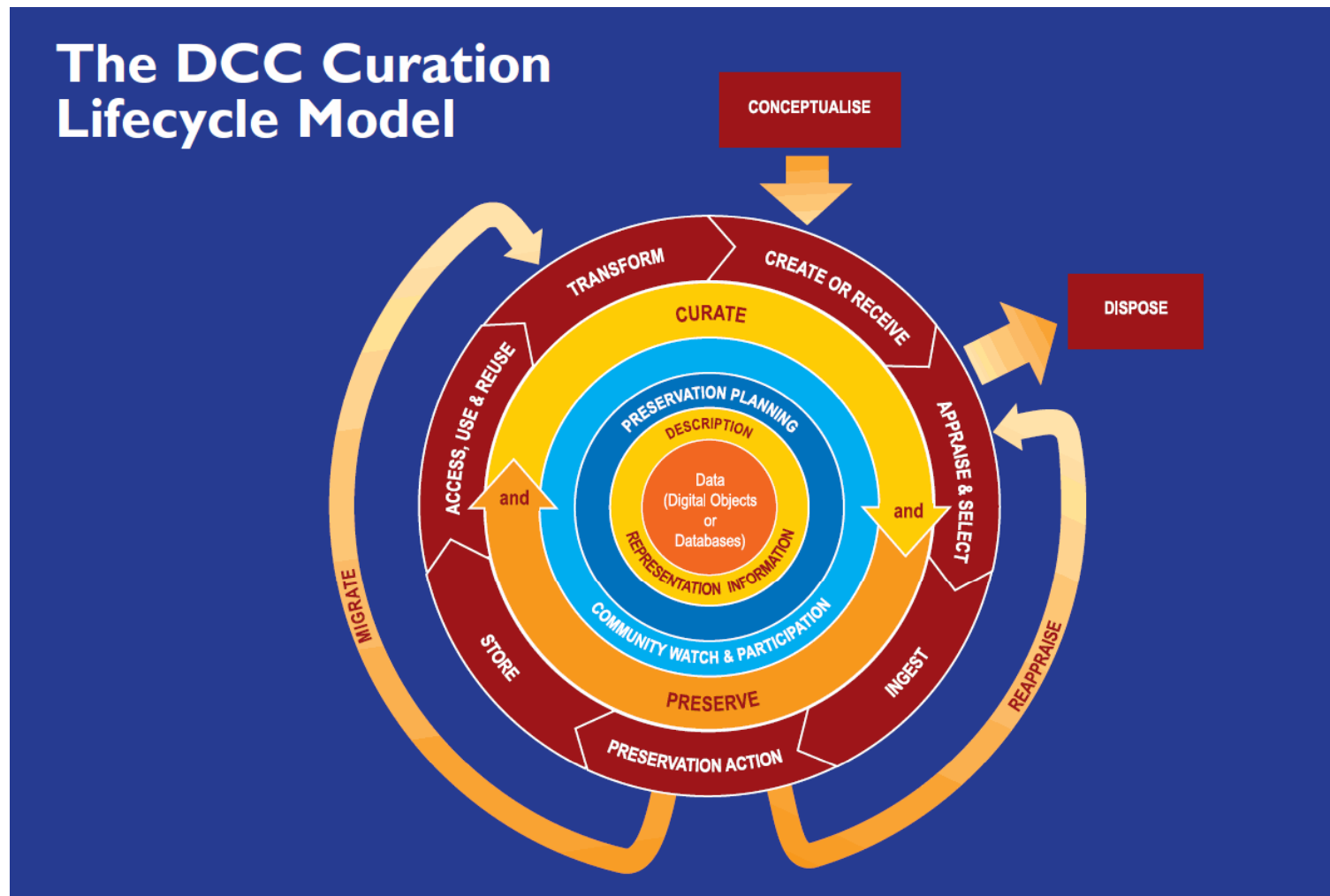


Image courtesy of the Cornell Biological Field Station

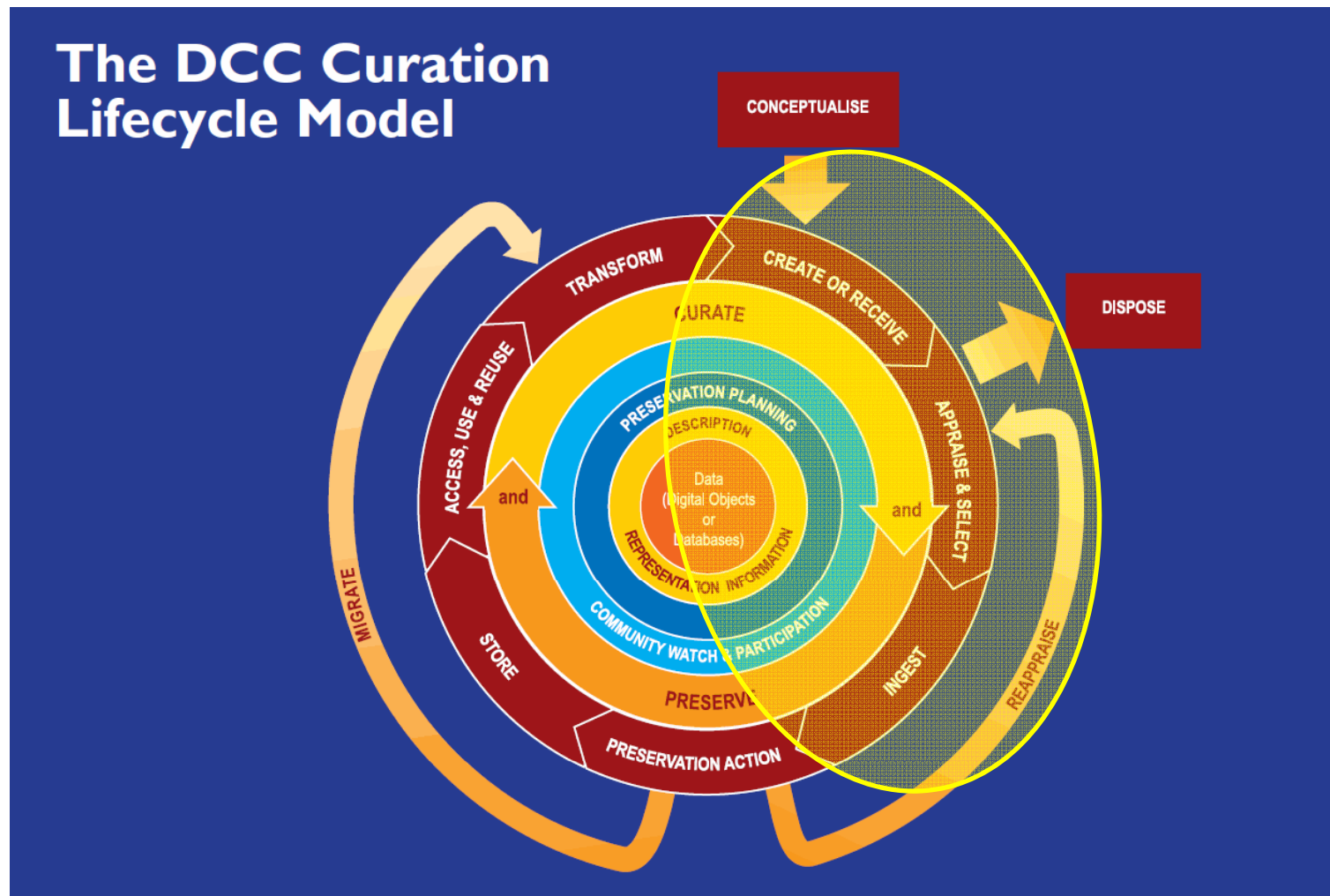
## What exactly is a data staging repository?



## Where does it fit in the life cycle?



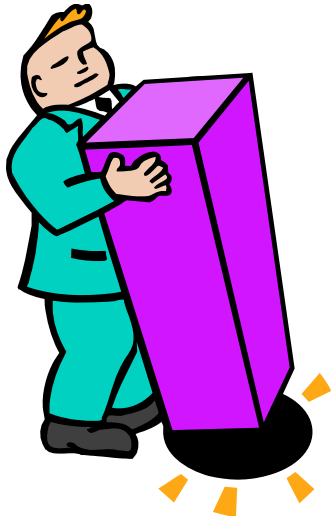
## Where does it fit in the life cycle?



## But DataStaR *isn't* a preservation repository...

"...if repository developers and administrators are guided by a reference model, they are more likely to consider the right issues."

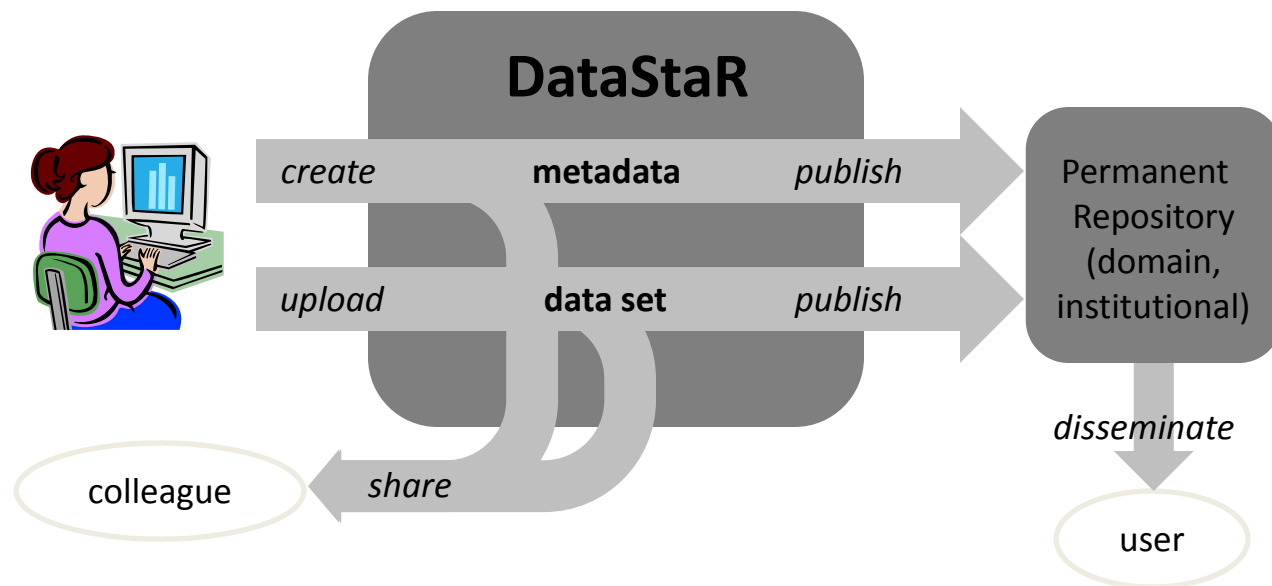
*Allinson 2006: OAIS as a Reference Model for Repositories: An Evaluation*



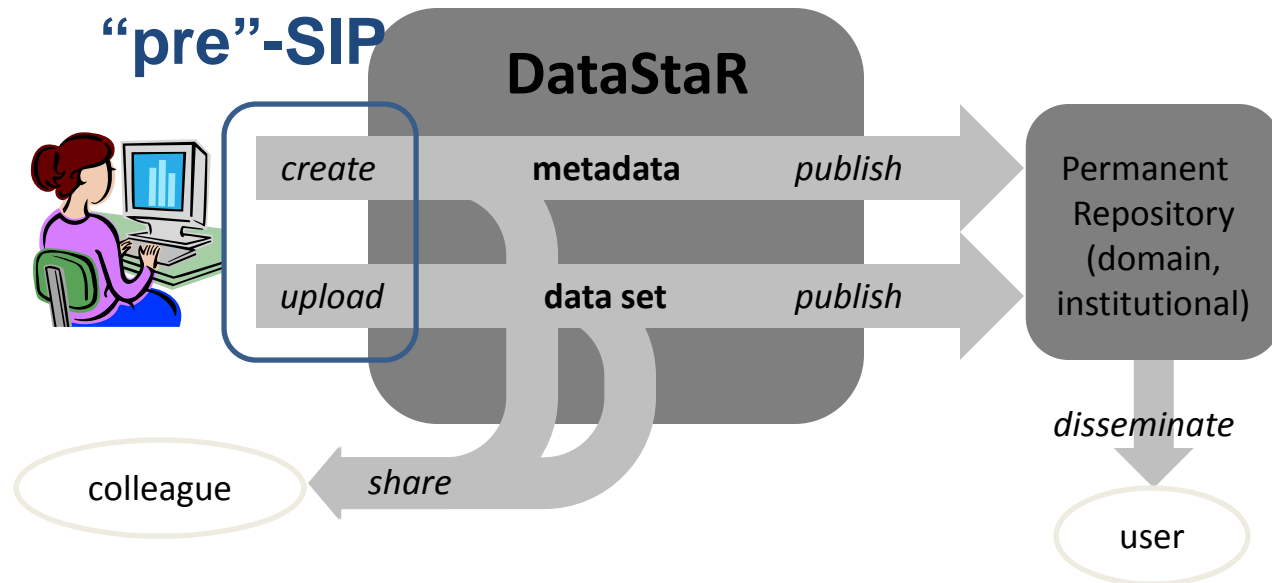
"A repository is Trusted if it can demonstrate its capacity to fulfill its specified functions, and if and if those (...) functions satisfy (...) minimal criteria which all trusted repositories are assumed to require."

*DigitalPreservationEurope 2008: Repository Planning Checklist and Guidance*

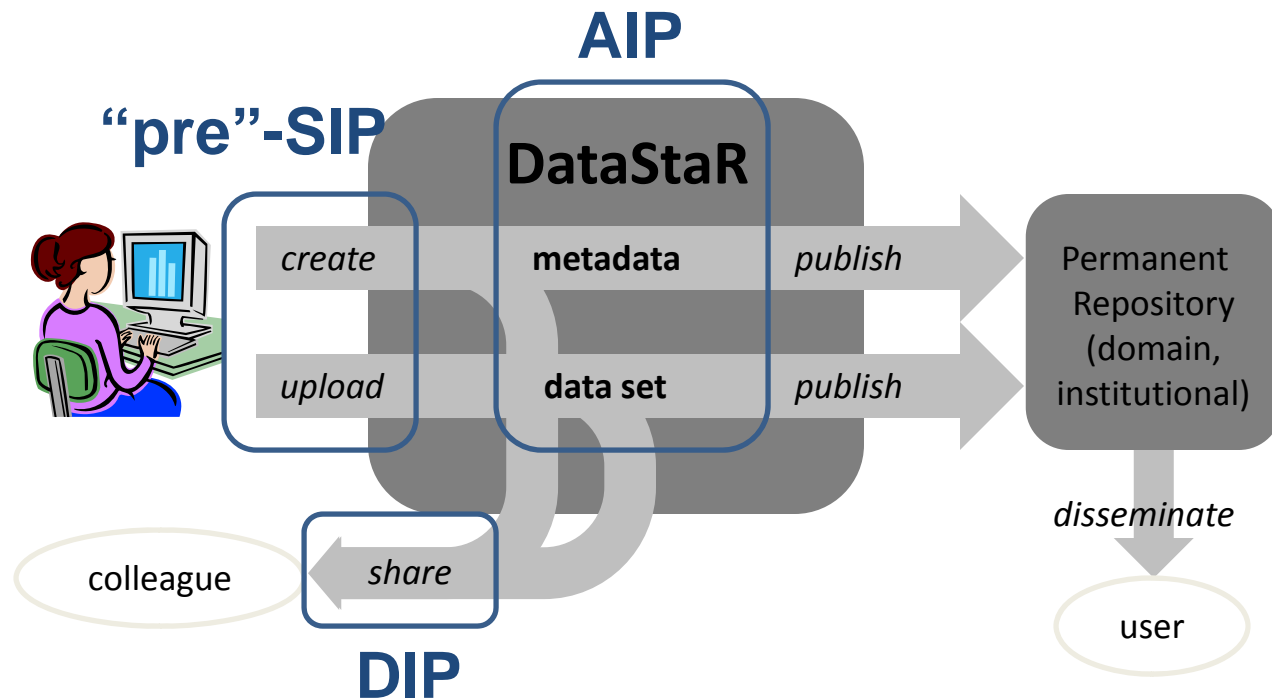
## An OAIS view of DataStaR



## An OAIS view of DataStaR

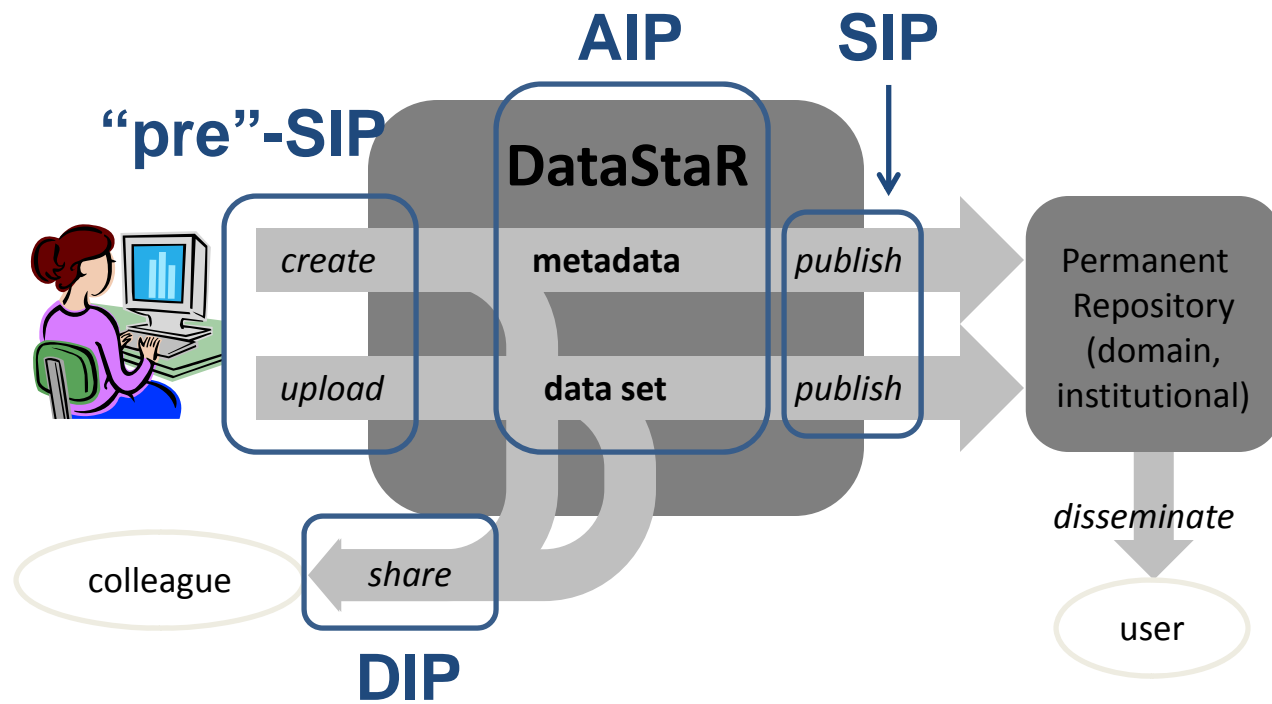


## An OAIS view of DataStaR

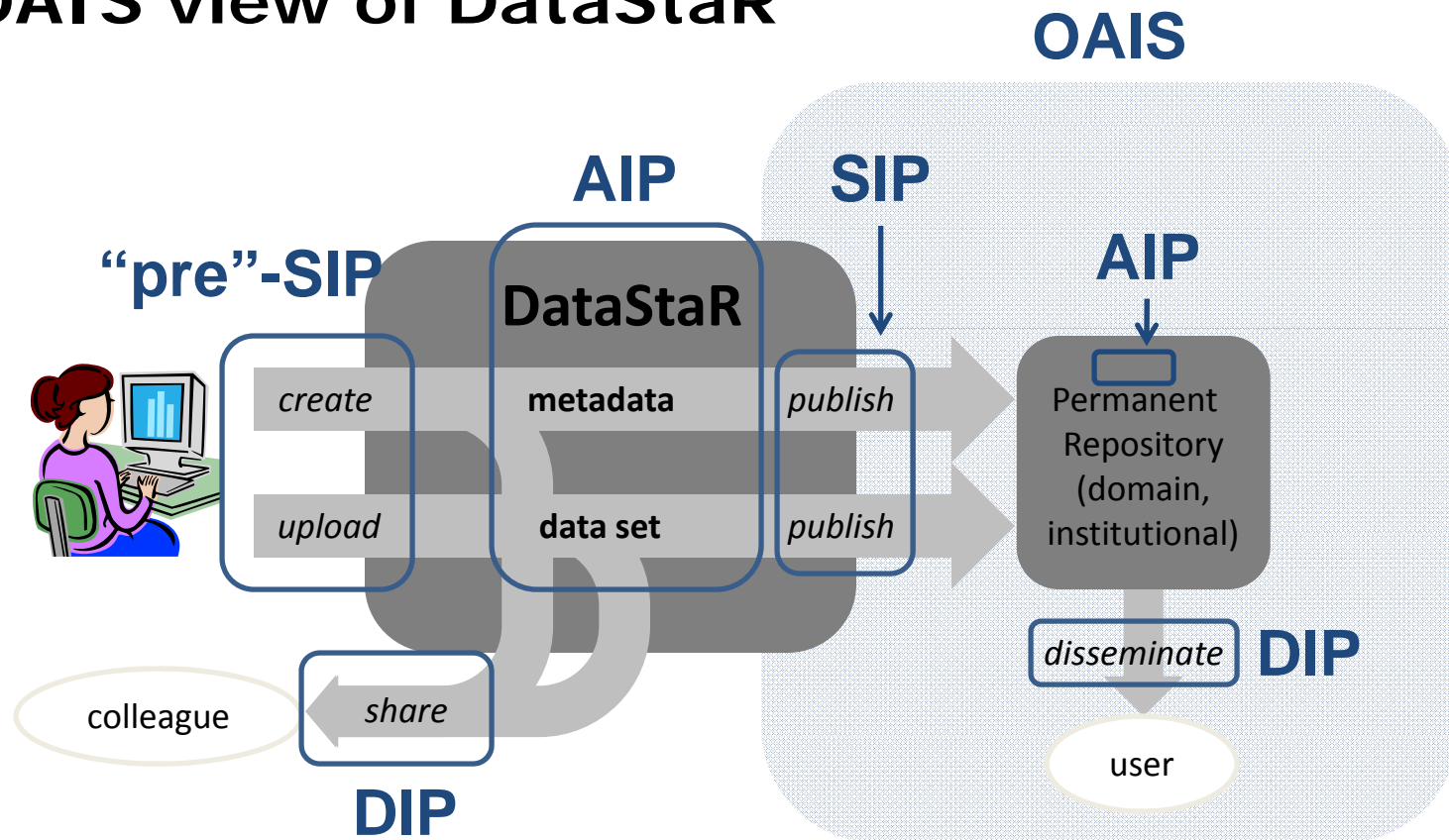




## An OAIS view of DataStar



## An OAIS view of DataStar



# Look at other approaches to implementation

## PLEDGE POLICY LIST

PLEDGE Project Report

Robert Wolfe, MIT Libraries

### Introduction

The objective of this list is to identify and describe the policies that an institutional repository like DSpace would need to define and implement in order to satisfy the requirements of a set of best practices like NARA's TRAC. The descriptions of each policy, specifically the requirements for satisfactory implementation are purposefully abstract, so as to avoid prejudicing the list for or against any one particular institutional repository system.

The TRAC is organized as a checklist for certification or audit. Individual entries in the checklist represent necessary functions or conditions of a trusted digital repository. These functions and conditions are very specifically defined and do not predict the organization and operation of the repository. Often, a single policy declaration is all that is necessary for a repository to accomplish several of the TRAC requirements, sometimes from several different sections of the TRAC.

### Policy Examples

#### *CU-0002 User Privacy (PR02, AR03)*

Repository maintains an explicit definition of user rights to privacy and all allowable uses of end-user data, including usage data. Definitions will include specification of publicly available data and repository's own internal use of data. It will include schedules of data retention and procedures for responding to Homeland Security Agency (HSA) requests for usage data.

Actual DSpace@MIT Policy Example

Descriptive Statement

1. The MIT Libraries have a privacy policy at <http://libraries.mit.edu/dspace-mit/build/policies/privacy.html>

## ICPSR Digital Preservation Policy Framework

Developed by Nancy Y. McGovern, Digital Preservation Officer, ICPSR, last revised June 20, 2007. Related documents to be drafted or made publicly available by ICPSR are indicated by brackets [ ].

### 1. Purpose

ICPSR fulfills its role as a trusted steward of the heritage of the social sciences by capturing the results of past and current social science research for future researchers. The Digital Preservation Policy Framework supports that mission and is the highest level digital preservation policy document at ICPSR. It makes explicit ICPSR's commitment to preserving the digital assets in its collections through the development and evolution of a comprehensive digital preservation program. The framework reflects the goals defined in the *ICPSR Strategic Plan (2005)* and contains references to other relevant ICPSR policies and procedures. The audience for the framework includes ICPSR members, staff, digital content depositors, funders, and users.

This framework also addresses a de facto standard of the digital preservation community, *Trusted Digital Repositories: Attributes and Responsibilities*. The organization of the framework reflects the seven attributes of a trusted digital repository:

- OAI compliance
- Administrative responsibility
- Organizational viability
- Financial sustainability
- Technological and procedural suitability
- Systems security
- Procedural accountability.

### 2. OAI Compliance

In achieving its digital preservation objectives, ICPSR recognizes the need to comply with the prevailing standards and practice of the digital preservation community. ICPSR is committed to developing its digital preservation policies, repository, and strategies in accordance with the *Open Archival Information System (OAIS) Reference Model (ISO 14721:2003)*. ICPSR tracks and responds to related OAI initiatives, including developments in digital archives certification, persistent identifiers, preservation metadata, and the producer-archive interface. The mapping of ICPSR's preservation process to OAI is synthesized in *Digital Preservation Requirements Applied to ICPSR* and explained in greater detail in [*OAI Conformance at ICPSR*].

### 3. Administrative Responsibility

As documented in its Constitution, ICPSR was established in 1962 as a data archive within the Institute for Social Research at the University of Michigan. Since then, ICPSR has had a fundamental commitment to and successful track record for acquiring and ensuring long-term access to core digital assets of social science.

#### 3.1 Mandate

## Put TRAC into context

Trustworthy Repositories  
Audit & Certification:  
Criteria and Checklist



preservation repository CRL specifications Certification  
criteria RLG Programs oclc audit digital object management  
NARA trustworthy metadata preservation repository  
CRL specifications Certification criteria RLG Programs  
oclc audit digital object management NARA trustwor-

### TRAC Criteria and Checklist Worksheet 3.4

This worksheet is based upon: Center for Research Libraries. Trustworthy Repositories Audit & Certification: Criteria and Checklist. Version 1.0, February 2007. <<http://www.crl.edu/PDF/trac.pdf>> Source of html file: <http://dljt.org/wp-content/uploads/2007/08/trac-cc-criteria-checklist.html>  
Revisions in the text have been made in accordance with DataStar's mission and implementation parameters.

This worksheet contains all TRAC criteria, annotated to reflect the mission and operations of DataStar. As noted, some requirements are not relevant to DataStar (N), and others will be addressed at a later point as part of mainstream repository development at Mann Library (M).

SECTION	EVIDENCE (examples)	EVIDENCE (actual) and COMMENTS
<b>A. Organizational Infrastructure</b>		
<b>A1. Governance &amp; organizational viability</b>		
A1.1 Repository has a mission statement that reflects a commitment to the management of and access to digital information.	Mission statement for the repository and its organizational context	DataStar repository policies: <a href="https://confluence.cornell.edu/x/fA41BQ">https://confluence.cornell.edu/x/fA41BQ</a> (mission statement, org. context)
A1.2 Repository has an appropriate, formal succession plan in case the repository ceases to operate or the funding institution substantially changes its scope.	Succession plan: explicit and specific statement documenting the intent to ensure continuity of the repository, and the steps taken and to be taken to ensure continuity.	DataStar repository policies: <a href="https://confluence.cornell.edu/x/fA41BQ">https://confluence.cornell.edu/x/fA41BQ</a> (org. commitment)
<b>A2. Organizational structure &amp; staffing</b>		
A2.1 Repository has identified and established the duties that it needs to perform and has appointed staff with adequate skills and experience to fulfill these duties.	A staffing plan; competency definitions; job description; development plans; plus evidence that the repository review and maintains these documents as requirements evolve.	<b>M</b>

Version 1.0  
February 2007

## What did DataStaR need...

Three things:

- Data depositor agreement
- Set of repository policies
- System documentation

# Data deposit agreement

🔴 This page is in development

This document serves as evidence to meet (in part or entirely) the following TRAC requirements:

★ A3.1, A3.3, A5.1, A5.2, A5.3

Specific statements that satisfy a TRAC requirement are noted within the document.

## DataStaR Data Deposit Agreement

DataStaR's purpose is to support collaboration and data sharing among researchers during the research process, and to promote publishing or archiving data and high-quality metadata to discipline-specific data centers, and/or to Cornell's own digital repository (eCommons). DataStaR fulfills its [mission](#) by managing content contributed by depositors as described in its [repository policies](#) and [system documentation](#).

Your use of the DataStaR system requires that you accept the terms of this data deposit agreement. This agreement grants permission for Cornell University Library (CUL), via the DataStaR repository (DataStaR) and other repositories or data centers (including CUL's institutional repository, eCommons), to archive and distribute data and metadata, as described in DataStaR's [repository policies](#) and [system documentation](#), and the terms you agree to below.

By using the DataStaR system, you accept the terms of this agreement, and you (or your proxy, on your behalf) agree to the following:

1. You accept and abide by [DataStaR's repository policies](#), which describe the overall management and operation of the repository and related services.
2. You represent that in utilizing DataStaR for the purposes of sharing your submission privately with selected colleagues, such sharing is not in violation of any applicable laws or policies, including those pertaining to the protection of private and confidential information. ★ A5.4
3. You certify that in preparing this submission for public archiving and distribution, you have removed all information directly identifying the research subjects in these data, and you have used due diligence in preventing information in the collection from being used to disclose the identity of research subjects.
4. You agree to release and hold harmless Cornell University and CUL from any and all liability from claims arising out of any legal actions, including actions concerning identification of research subjects, breaches of confidentiality, or invasions of privacy by or on behalf of said subjects. ★ A5.1
5. If the submission is your original work, and that you have the right to grant the rights contained in this license. You also represent that your submission does not, to the best of your knowledge, infringe upon anyone's copyright or any other applicable laws or licenses. ★ A5.4
6. If the submission is not your original work, you represent that you have obtained the unrestricted permission of the owner to grant CUL the rights required by this license, and that the owner of the material is clearly identified and acknowledged within the text or content of the submission. ★ A5.4
7. If the submission is based upon work that has been sponsored or supported by an agency or organization other than CUL, you represent that you have fulfilled any right of review or other obligations required by such contract or agreement.
8. You grant CUL the non-exclusive right to distribute your data set(s) and metadata ("submission", hereafter), via DataStaR, to authorized users selected by you. ★ A3.3
9. You agree that CUL may make and keep multiple copies of your submission for purposes of security, back-up and preservation. ★ A3.3
10. You represent that you have made a reasonable effort to ensure that data contained in your submission and distributed via DataStaR are accurate.

# Repository policies

This document serves as evidence to meet (in part or entirely) the following TRAC requirements:

★ A3.1, A3.7, A5.1

## DataStaR repository policies v1.0

DataStaR aims to be transparent and accountable in all of its operations. This is partially manifested by creating and maintaining the following publicly viewable documents and policies:

### DataStaR repository policies:

- [DataStaR mission statement](#)
- [Updates and amendments to these policies and related documents](#)
- [Overview for depositors](#)
- [DataStaR collection development policy](#)
- [Data and metadata management policy](#)
- [DataStaR and digital preservation](#)
- [Terms of use](#)
- [Feedback policy](#)
- [References](#)

### Additional documents:

- [DataStaR data deposit agreement](#) (separate document) ★ A3.1, A3.3, A5.1, A5.2, A5.3, A5.4, B1.5, B1.6
- [Preparing data for publication - guidelines](#) (separate document)
- [System documentation](#) (separate document)

## DataStaR mission statement

★ A1.1, A3.1

The purpose of DataStaR is to support collaboration and data sharing among researchers during the research process, and to promote publishing or archiving data and high-quality metadata to discipline-specific data centers, and/or to Cornell's own digital repository; eCommons (<http://ecommons.library.cornell.edu/>).

DataStaR accomplishes these goals by:

- Providing access-controlled file sharing to Cornell researchers and their collaborators;
- Developing and offering tools for creating metadata records in a variety of formats, and that allow researchers to reuse their personal information and information from previously created metadata records;
- Assisting researchers in identifying discipline-specific data centers for data publication, and in preparing data and metadata to the standards specified by those centers;



# System documentation

This document serves as evidence to meet (in part or entirely) the following TRAC requirements: ★ A3.7

## DataStAR system documentation

### Introduction

The purpose of DataStAR is to support collaboration and data and high-quality metadata to discipline-specific data centers. For further information on DataStAR policies and operations,

#### Sections:

- [Architecture - overview](#)
- [Ingest](#)
- [Access](#)
- [Metadata](#)
- [Publication](#)
- [Administrative and maintenance actions](#)
- [Backup](#)
- [Other functions](#)
- [Hardware inventory and history](#)
- [Software inventory and history](#)
- [Testing and further development](#)
- [Tabled / Items for future development](#)

#### Related documents]

- [Minimum metadata for DataStAR](#)
- [System errors - documentation](#)

### Architecture - overview

The DataStAR system consists of the following applications:

- [Fedora](#) - Open-source digital repository software
- [vitro](#) - Ontology editor and semantic web application development
- [CUWebAuth](#) - Tool to integrate web applications with Cornell University's authentication system
- [DROID](#) - File format identification and validation
- DataStAR-specific code:
  - Code to handle the file upload form.

### Ingest

#### Ingest of Submission Information Package (SIP)

- Authentication: + [DataStAR-32](#) Cornell users are authenticated via [CUWebAuth](#). ★ B1.3
- + [DataStAR-45](#) Non-Cornell users are authenticated via the form-based login available at <http://datastar.mannlib.cornell.edu>. ★ B1.3
- + [DATASTAR-73](#) File size of uploaded files is 1GB, with a total limit of 10GB per account.
- + [DataStAR-40](#) The identity of the uploader of each file is stored as an object property in the DataStAR metadata store. ★ B1.3
- File format verification: + [DataStAR-24 DROID](#) is used for file format identification and validation. File format information is stored as object properties in the DataStAR metadata store ★ B1.4, B2.7, B2.8
- Content verification: + [DataStAR-46](#) Checksums are automatically performed when an object is uploaded to the Fedora repository (by which application?). ★ B1.4
- + [DataStAR-36](#) Checksums are stored (as data properties in DataStAR metadata store, and as properties in fedora) to facilitate routine error checking of content in the repository. ★ B1.4
- + [DataStAR-47](#) A submitted digital object is considered "complete" and "correct" if the upload and verification processes proceed without errors. See also: [System errors - documentation](#). ★ B1.4
- + [DataStAR-39](#) Successful deposits are documented with an automatically generated email receipt, including the checksum of the deposited file(s), to the depositor. See also: [System errors - documentation](#). ★ B1.6, B1.7

#### Construction and maintenance of Archival Information Package (AIP)

- An AIP consists of the file(s) uploaded by a user, and all related metadata - whether supplied by the user, the system, or DataStAR staff. Uploaded files are managed by Fedora (see [Fedora documentation](#) for details). + [DataStAR-40](#) Metadata is managed within the DataStAR metadata store; Fedora content is linked to DataStAR metadata via Fedora identifiers. See [Minimum metadata for DataStAR](#) for description of the minimum required elements. ★ B2.1, B2.3, B2.13, B5.2, B5.3, B5.4
- All submitted objects become part of an AIP immediately upon submission (see also the DataStAR policy on [Data and metadata management](#)). ★ B2.4
- An AIP is considered complete and correct if the upload process proceeds without failure, and the + [DataStAR-40](#) minimum metadata (see [Minimum metadata for DataStAR](#)) required for the data set as a whole, and each file included in the data set, are complete. + [DataStAR-47](#) If user fails to provide complete metadata, the upload fails and the user is notified (see [System errors - documentation](#)). ★ B2.11
- Each AIP is assigned a URI within the DataStAR system at the time it's created. ★ B2.5 Our assumption is that data sets uploaded to DataStAR have not been published elsewhere and do not have a prior unique identifier which must be retained. + [DATASTAR-48](#) Original file names are maintained within DataStAR. ★ B2.6
- + [DATASTAR-26](#) Files associated with an AIP may be updated by a user at any time. + [DATASTAR-49](#) In the case an older file is replaced with a newer one, only the newest is available to DataStAR users. Older versions are available upon request to a DataStAR administrator (accessible via the Fedora management interface). ★ B4.3
- + [DATASTAR-50](#) Deletion of files: data owners may delete content (a single file or an entire AIP) at any time. + [DATASTAR-51](#) The immediate result of this action is to make the content inaccessible to anyone (including the owner), with the exception of system administrators, who will make it available upon request. Content will be retained and remain inaccessible for one year, at which point it will be deleted permanently. + [DATASTAR-52](#) Administrators



# System documentation

This document serves as evidence to meet (in part or entirely) the following TRAC requirements:  A3.7

## DataStar system documentation

### Introduction

The purpose of DataStar is to support the collection and high-quality metadata to support data discovery. For further information on DataStar, see the [DataStar User Manual](#).

#### Sections:

- [Architecture - overview](#)
- [Ingest](#)
- [Access](#)
- [Metadata](#)
- [Publication](#)
- [Administrative and maintenance](#)
- [Backup](#)
- [Other functions](#)
- [Hardware inventory and history](#)
- [Software inventory and history](#)
- [Testing and further development](#)
- [Tabled / Items for future development](#)





































#### Related documents]

- [Minimum metadata for DataStar](#)
- [System errors - documentation](#)

### Architecture - overview

The DataStar system consists of the following components:

- [Fedora](#) - Open-source digital repository
- [vitro](#) - Ontology editor and semantic query engine
- [CUWebAuth](#) - Tool to integrate with CUWebAuth
- [DROID](#) - File format identification tool
- DataStar-specific code:
  - Code to handle the file upload

T	Key	Summary	Components	Assignee	Reporter	Pr	Status
	<a href="#">DATASTAR-98</a>	XML Schema datatype validation for auto-generated forms	Metadata - UI	Brian J Lowe	Brian J Lowe		 Resolved
	<a href="#">DATASTAR-97</a>	Find a nice date entry widget	Metadata - UI	Brian J Lowe	Brian J Lowe		 Open
	<a href="#">DATASTAR-58</a>	Implement Glaze reification to displayRank conversion for method steps, paragraphs, data table attributes, etc.	Metadata - infrastructure	Brian J Lowe	Brian J Lowe		 Open
	<a href="#">DATASTAR-57</a>	Inference for smushing on EML reference IDs	Metadata - infrastructure	Brian J Lowe	Brian J Lowe		 In Progress
	<a href="#">DATASTAR-46</a>	<a href="#">DATASTAR-34</a> ↳ perform checksum at time of upload	Ingest	Brian Caruso	Gail Steinhart		 Open
	<a href="#">DATASTAR-44</a>	Improved handling of multiple graphs and reasoners	Metadata - infrastructure	Brian J Lowe	Brian J Lowe		 In Progress
	<a href="#">DATASTAR-43</a>	Automatic form generation	Metadata - UI	Brian J Lowe	Brian J Lowe		 In Progress
	<a href="#">DATASTAR-42</a>	EML custom rendering	Metadata - UI	Brian J Lowe	Brian J Lowe		 In Progress
	<a href="#">DATASTAR-40</a>	Review and implement minimum data to gather on a file upload	Ingest	Brian Caruso	Brian Caruso		 Open
	<a href="#">DATASTAR-39</a>	Email confirmation of file upload.	Ingest	Brian Caruso	Brian Caruso		 Open
	<a href="#">DATASTAR-38</a>	<a href="#">DATASTAR-34</a> ↳ display a checksum of a dataset to user on dataset zip download	DIP creation	Brian Caruso	Brian Caruso		 Open
	<a href="#">DATASTAR-37</a>	<a href="#">DATASTAR-34</a> ↳ have a script that verifies	AIP maintenance	Brian Caruso	Brian Caruso		 Open

 B1.3

B1.3  
stored as object properties in the repository (by which application?).  
facilitate routine error checking of proceed without errors. See also:  
sum of the deposited file(s), to the  
DataStar staff. Uploaded files are metadata store; Fedora content is required elements.  B2.1, B2.3, metadata management).  B2.4  
minimum metadata (see [Minimum DataStar-47](#) If user fails to provide  
uploaded to DataStar have not names are maintained within  
older file is replaced with a newer (accessible via the Fedora  
DATASTAR-51 The immediate result of s, who will make it available upon DATASTAR-52 Administrators

# How did we do?

TRAC SECTION	DA	POLICIES	SYSDOC	N	M	TRAC SECTION	DA	POLICIES	SYSDOC	N	M	TRAC SECTION	DA	POLICIES	SYSDOC	N	M
A 1.1		X				B 1.1	X					C 1.1			X		
A 1.2		X				B 1.2		X	X			C 1.2			X		
A 2.1					X	B 1.3	X		X			C 1.3			X		
A 2.2					X	B 1.4			X			C 1.4			X		
A 2.3					X	B 1.5	X					C 1.5			X		
A 3.1	X	X				B 1.6		X	X			C 1.6			X		
A 3.2		X			X	B 1.7	X	X	X			C 1.7					X
A 3.3	X					B 1.8			X			C 1.8					X
A 3.4					X	B 2.1			X			C 1.9					X
A 3.5		X				B 2.2				X		C 1.10					X
A 3.6		X	X			B 2.3			X			C 2.1					X
A 3.7		X	X			B 2.4			X			C 2.2					X
A 3.8			X			B 2.5			X			C 3.1					X
A 3.9					X	B 2.6			X			C 3.2					X
A 4.1					X	B 2.7			X			C 3.3					X
A 4.2					X	B 2.8			X			C 3.4					X
A 4.3					X	B 2.9				X							
A 4.4					X	B 2.10		X									
A 4.5					X	B 2.11			X								
A 5.1	X	X				B 2.12				X							
A 5.2	X					B 2.13			X								
A 5.3	X	X				B 3.1				X							
A 5.4	X					B 3.2				X							
A 5.5				X		B 3.3				X							
						B 3.4				X							
						B 4.1		X									
						B 4.2		X									
						B 4.3		X	X								
						B 4.4			X								
						B 4.5				X							
						B 5.1		X	X								
						B 5.2		X	X								
						B 5.3			X								
						B 5.4			X								
						B 6.1		X									
						B 6.2		X	X								
						B 6.3		X	X								
						B 6.4		X	X								
						B 6.5			X								
						B 6.6		X	X								
						B 6.7			X								
						B 6.8			X								
						B 6.9			X								
						B 6.10			X								

## How did we do?

Number and percentage of TRAC criteria addressed by  
(agreement, policies, system)

TRAC SECTION	Depositor agreement	Repository policies	System doc/ requirements
A (24 criteria)	6 (25%)	9 (38%)	3 (13%)
B (44 criteria)	4 (9%)	14 (32%)	30 (68%)
C (16 criteria)	0	0	6 (38%)

## How did we do?

Number and percentage of TRAC criteria addressed by (agreement, policies, system)

TRAC SECTION	Depositor agreement	Repository policies	System doc/ requirements
A (24 criteria)	6 (25%)	9 (38%)	3 (13%)
B (44 criteria)	4 (9%)	14 (32%)	30 (68%)
C (16 criteria)	0	0	6 (38%)

- Section A criteria (organizational infrastructure) addressed mainly by policy

## How did we do?

Number and percentage of TRAC criteria addressed by (agreement, policies, system)

TRAC SECTION	Depositor agreement	Repository policies	System doc/ requirements
A (24 criteria)	6 (25%)	9 (38%)	3 (13%)
B (44 criteria)	4 (9%)	14 (32%)	30 (68%)
C (16 criteria)	0	0	6 (38%)

- Section A criteria (organizational infrastructure) addressed mainly by policy
- Section B and C criteria (digital object management and technologies, technical infrastructure and security addressed mainly (but not exclusively) by system

## What didn't we do?

TRAC SECTION	Address at transition to production system	Not relevant to DataStar
A (24 criteria)	11 (46%)	1 (4%)
B (44 criteria)	0	8 (18%)
C (16 criteria)	10 (63%)	0
TOTAL (84 criteria)	21 (25%)	9 (11%)

*We are making an effort to address 64% of the TRAC criteria, in the pilot phase.*

## Some observations

- Understanding/interpreting the criteria is a lot of work.
- The right tools might simplify policy development.
- The right software might simplify system specification.
- Compiling /presenting evidence: for auditors, or for users?
- Picking your partners...

*TRAC has a lot to offer, even if long-term preservation isn't your focus.*

**Thank you.**



**Gail Steinhart**  
[GSS1@cornell.edu](mailto:GSS1@cornell.edu)

**Ann Green**  
[green.ann@gmail.com](mailto:green.ann@gmail.com)

**Dianne Dietrich**  
[dd388@cornell.edu](mailto:dd388@cornell.edu)

This material is based upon work supported by the National Science Foundation under Grant No. III-0712989. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.