

Humanities Data Curation



Thomas Padilla, Digital Scholarship Librarian, @thomasgpadilla

Humanities Data Curation

Digital Preservation

Digital Curation

Data curation . . . is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time.

Melissa Cragin et al, “An Educational Program on Data Curation”

DATA CURATION

active and ongoing management

add value + maintain quality

continued management depends on present/possible
interest in use

the above depends on your knowledge of potential
users + understanding of what can be done with
data





[Home](#) » [Academics](#) » [Degree Programs](#) » [Degree Specializations](#) » Specialization in Data Curation

Specialization in Data Curation

Required for the Specialization

LIS562 Metadata in Theory and Practice*

LIS586 Digital Preservation

LIS590DC Foundations of Data Curation

Recommended Electives

Students must take two (2) courses from this list; taking four (4) is advised.

LIS452 Foundations of Information Processing in LIS

LIS453 Systems Analysis and Management*

LIS490DB Introduction to Databases

LIS560 Digital Libraries

LIS561 Information Modeling

LIS590OD Ontology Development

LIS590RO Representing and Organizing Information Resources

Institute Workshops



DH Curation Guide

Digital Humanities Data Curation Institute

A generous grant from the [National Endowment for the Humanities](#) will support a series of **three-day** workshops in 2013 and 2014, on the principles and practices of data curation.

Digital Preservation . . . refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.

- DPC Online, 2012

Digital curation . . . is increasingly being used for the actions needed to maintain digital research data and other digital materials over their entire lifecycle and over time for current and future generations of users. Implicit in this definition are the processes of digital archiving and digital preservation, but it also includes all the processes needed for good data creation and management, and the capacity to add value to data to generate new sources of information and knowledge.

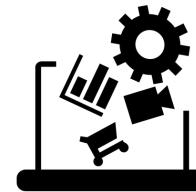
- Beagrie, 2004

Data curation . . . is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time.

- Cragin, 2007

Digital Preservation . . . refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.

Digital Preservation Coalition Online, 2012

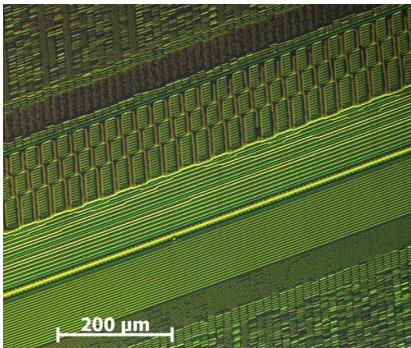




honey badger



honey badger stored on disk



honey badger data on disk

Software - Adobe Photoshop

Lightroom 5.3 (Windows)

Date and Time (Modified) -

2014:07:25 19:42:51

Artist - Yoann JEZEQUEL

Copyright - © Yoann JEZEQUEL

ISO Speed - 400

Sensitivity Type - Recommended
Exposure Index

Recommended Exposure Index -
400

Exif Version - 0230

Date and Time (Original) -
2014:02:16 19:28:34

Date and Time (Digitized) -
2014:02:16 19:28:34

Exposure Bias - 0 EV

Max Aperture Value - 3.5

Metering Mode - Multi-segment

Sub Sec Time Original - 66

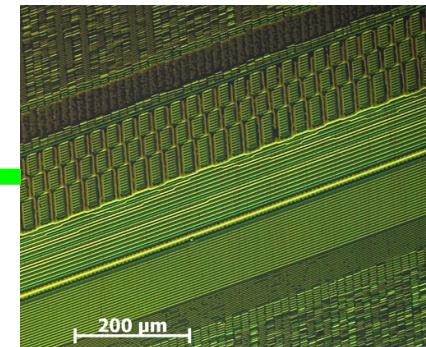
Sub Sec Time Digitized - 66

Focal Plane X-Resolution -
5728.176796

Focal Plane Y-Resolution -
5808.403361

Focal Plane Resolution Unit - inches

Custom Rendered - Normal



METADATA



Metadata is commonly defined as "data about data."

. . . Typically metadata is organized into distinct categories . . . For example, administrative metadata may include the date and source of acquisition, disposal date, and disposal method.

Descriptive metadata may include information about the content and form of the materials. Preservation metadata may record activities to protect or extend the life of the resource, such as reformatting.

Structural metadata may indicate the interrelationships between discrete information resources, such as page numbers.

[EMAIL](#) [PRINT](#) [COMMENTS](#)[Tweet](#)

'We Kill People Based on Metadata'

David Cole*Rick Bowmer/AP Photo*

The National Security Agency's \$1.5 billion data storage facility in Bluffdale, Utah, June 2013

Supporters of the National Security Agency inevitably defend its sweeping collection of phone and Internet records on the ground that it is only collecting so-called "metadata"—who you call, when you call, how long you talk. Since this does not include the actual content of the communications, the threat to privacy is said to be negligible. That argument is profoundly misleading.

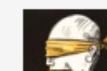
Of course knowing the content of a call can be crucial to establishing a particular threat. But metadata alone can provide an extremely detailed picture of a person's most intimate associations and interests, and it's actually much easier as a technological matter to search huge amounts of metadata than to listen to millions of phone calls. As NSA General Counsel Stewart Baker has [said](#), "metadata absolutely tells you everything about somebody's life. If you have enough metadata, you don't really need content." When I quoted Baker at a [recent debate](#) at Johns Hopkins University, my opponent, General Michael Hayden, former director of the NSA and the CIA, called Baker's comment "absolutely correct," and raised him one, asserting, "We kill people

— NEW POSTS —

NYRBLOG

**Hollywood & Truth**
Francine Prose**China's War on Wit**
Perry Link**In Search of Authenticity**
Tim Parks**China's Democracy Guru**
Ian Johnson

THE GALLERY

Midnight Movies of the Mind
Jed Perl**Brash, Bold, Insolent**
Tomi Ungerer
J. H.



Amanda Visconti

@Literature_Geek



Following

.@edsu's algorithmic poem generation: "If Google mumbled in its sleep, what would it say?" inkdroid.org/journal/2014/0... and inkdroid.org/vogon



Reply



Retweet



Favorite

More



Ed Summers

Google's Subconscious

Can a poem provide insight into the inner workings of a complex algorithm? If Google Search had a subconscious, what would it look like? A few days ago, I ran across these two quotes within hours of...

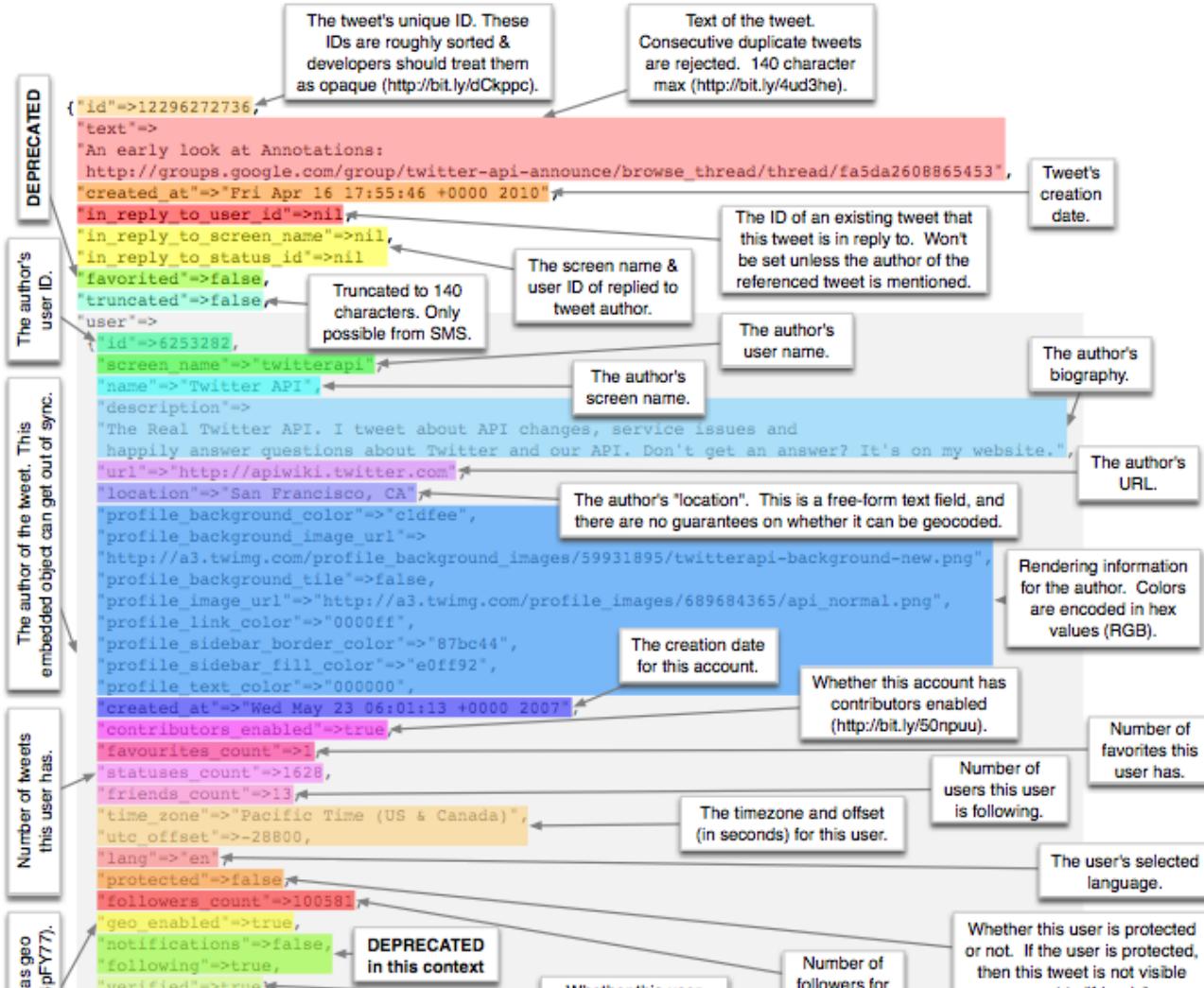
[View on web](#)

10:42 AM - 17 Sep 2014

Flag media



Reply to @Literature_Geek @edsu



FLEETWOOD MAC AT BATTERSEA TOWN HALL



Subject: Fleetwood Mac

Identifier:

Creator:

Source: BAC

Date: 1968

Rights:

Collection: Events (4)

Show others:

Collection

Same year

Description:

Poster of the Fleetwood Mac gig on the 22nd of October 1968

Tags:

Fleetwood Mac

Citation:

"Fleetwood Mac at Battersea Town Hall," *Battersea Arts Centre Digital Archive*, accessed February 12, 2015, <http://www.bacarchive.org.uk/items/show/4796>.



Index of Terms

Properties in the /terms/ namespace	abstract , accessRights , accrualMethod , accrualPeriodicity , accrualPolicy , alternative , audience , available , bibliographicCitation , conformsTo , contributor , coverage , created , creator , date , dateAccepted , dateCopyrighted , dateSubmitted , description , educationLevel , extent , format , hasFormat , hasPart , hasVersion , identifier , instructionalMethod , isFormatOf , isPartOf , isReferencedBy , isReplacedBy , isRequiredBy , issued , isVersionOf , language , license , mediator , medium , modified , provenance , publisher , references , relation , replaces , requires , rights , rightsHolder , source , spatial , subject , tableOfContents , temporal , title , type , valid
Properties in the /elements/1.1/ namespace	contributor , coverage , creator , date , description , format , identifier , language , publisher , relation , rights , source , subject , title , type
Vocabulary Encoding Schemes	DCMIType , DDC , IMT , LCC , LCSH , MESH , NLM , TGN , UDC
Syntax Encoding Schemes	Box , ISO3166 , ISO639-2 , ISO639-3 , Period , Point , RFC1766 , RFC3066 , RFC4646 , RFC5646 , URI , W3CDTF
Classes	Agent , AgentClass , BibliographicResource , FileFormat , Frequency , Jurisdiction , LicenseDocument , LinguisticSystem , Location , LocationPeriodOrJurisdiction , MediaType , MediaTypeOrExtent , MethodOfAccrual , MethodOfInstruction , PeriodOfTime , PhysicalMedium , PhysicalResource , Policy , ProvenanceStatement , RightsStatement , SizeOrDuration , Standard
DCMI Type Vocabulary	Collection , Dataset , Event , Image , InteractiveResource , MovingImage , PhysicalObject , Service , Software , Sound , StillImage , Text
Terms related to the DCMI Abstract Model	memberOf , VocabularyEncodingScheme

CanCore, DDC, EAC-CPF,
FRBR, GEM, IEEE/LOM,
ISAAR(CPF), ISBD, LCC,
MADS, MARC, MARC
Relator Codes,
MARCXML, MathML,
Ontology for Media
Resource, TGN, XMP,
XOBIS

Atom, DwC,
GILS, indecs,
MODS, RSS,
SCORM,
Topic Maps,
Z39.50

AGLS,
DCAM, Linked
Data, METS Rights,
OAI-ORE, OAI-PMH,
ODRL, PREMIS, RDF,
RELAX NG, SGML, SKOS,
SRU, XQuery, XrML

Datasets

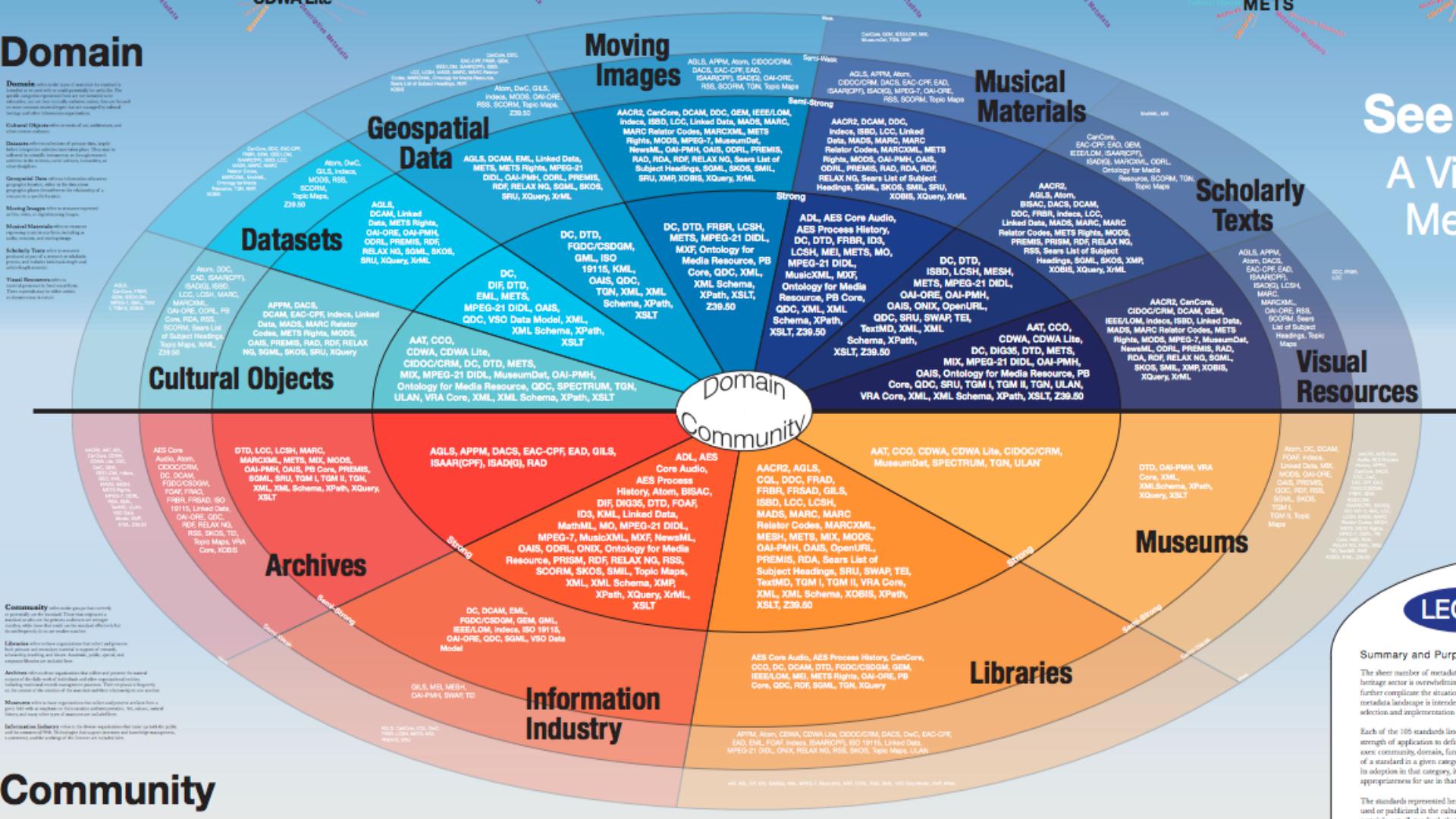
APPM, DACS,
DCAM, EAC-CPF, indecs, Linked
Data, MADS, MARC Relator
Codes, METS Rights, MODS,
OAIS, PREMIS, RAD, RDF, RELAX
NG, SGML, SKOS, SRU, XQuery

AAT, CCO,
CDWA, CDWA Lite,
CIDOC/CRM, DC, DTD, METS,
MIX, MPEG-21 DIDL, MuseumDat, OAI-PMH.

AGLS, DCAM, EML, Linked Data,
METS, METS Rights, MPEG-21
DIDL, OAI-PMH, ODRL, PREMIS,
RDF, RELAX NG, SGML, SKOS,
SRU, XQuery, XrML

DC, DTD,
FGDC/CSDGM,
GML, ISO
19115, KML,
OAIS, QDC,
TGN, XML,
Schema, X
S

DC,
DTD, EML, METS,
MPEG-21 DIDL, OAIS,
QDC, VSO Data Model, XML,
XML Schema, XPath,
XSLT





THERE IS NO
PRESERVATION
WITHOUT ACCESS.

JEROME MCDONOUGH

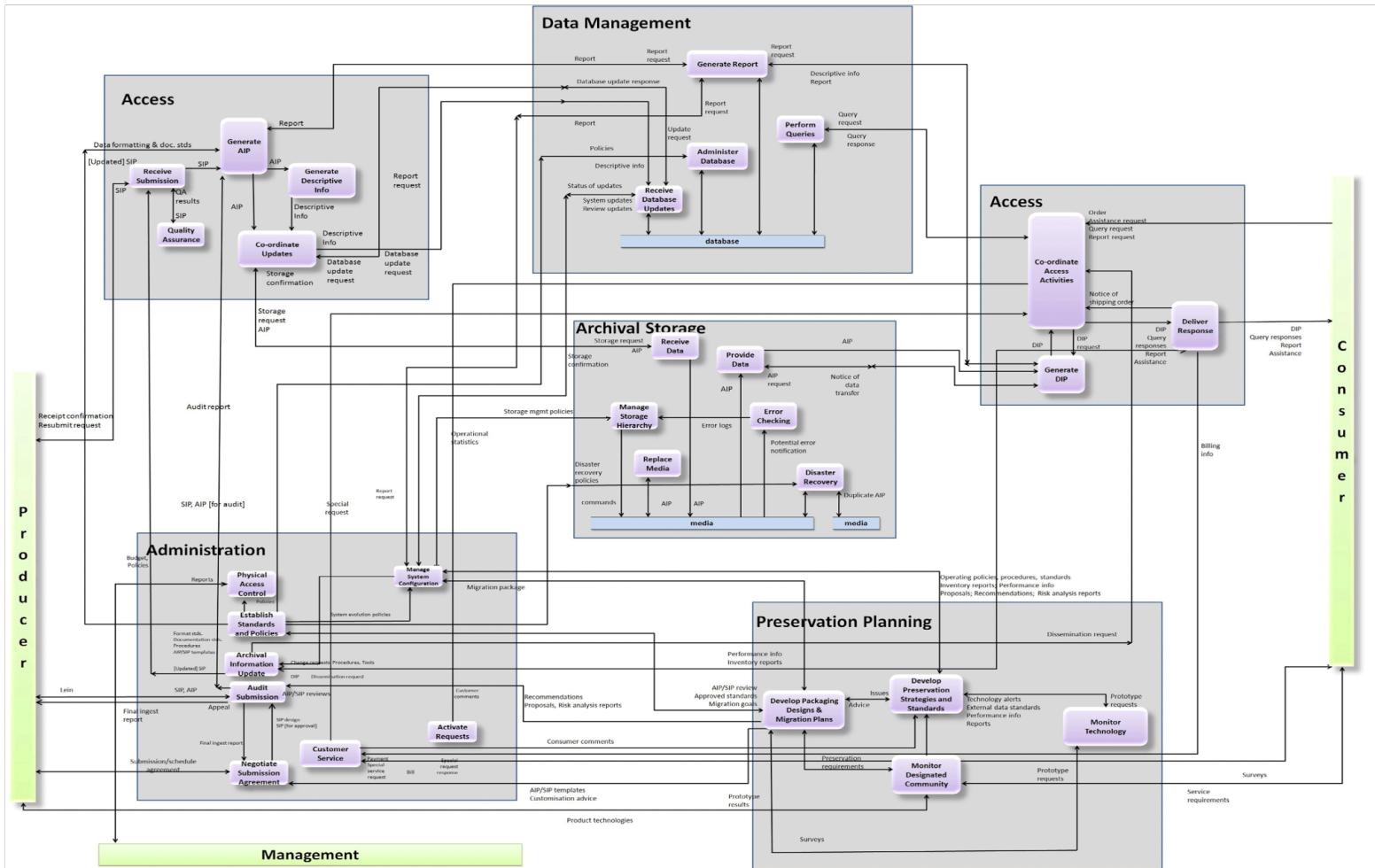


Figure A-1: Composite of Functional Entities

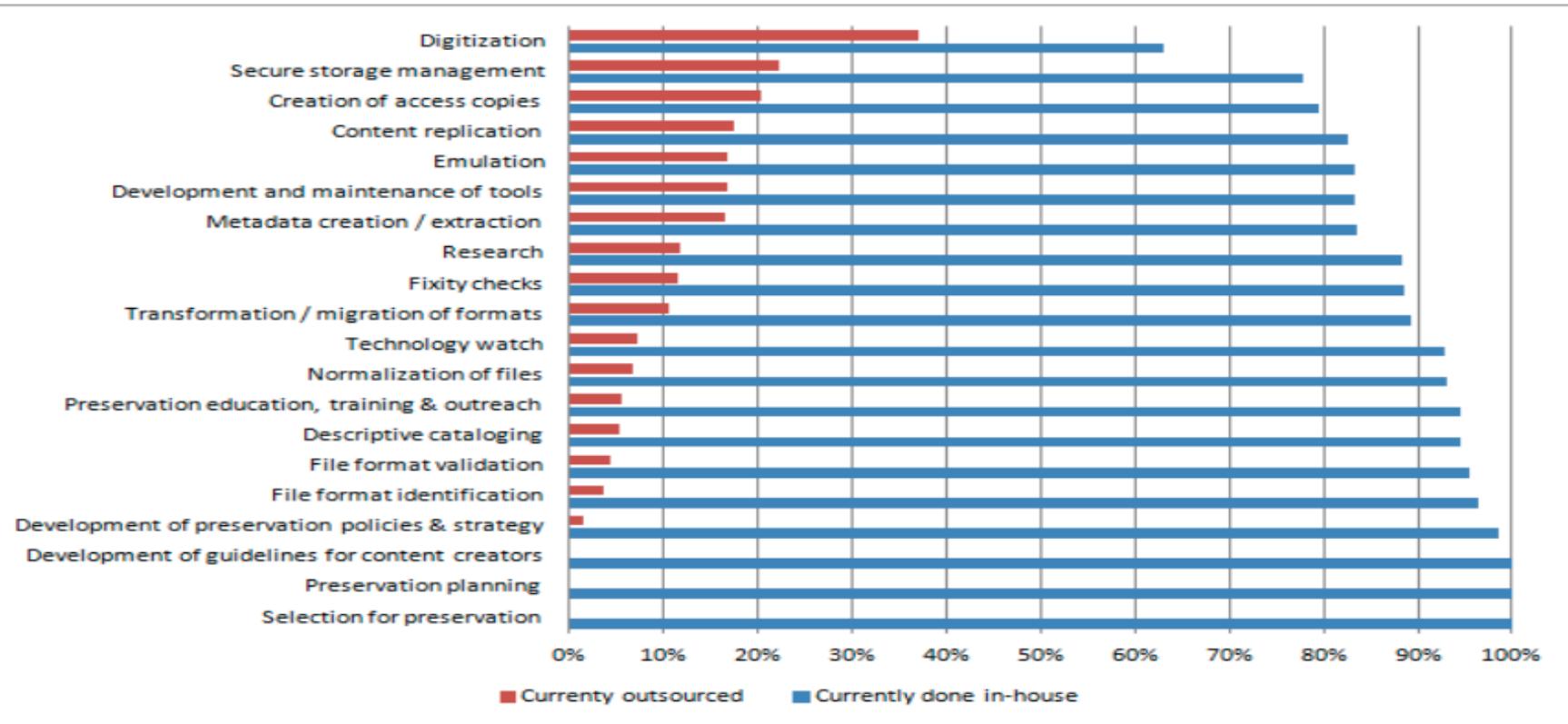
Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	<ul style="list-style-type: none"> - Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system 	<ul style="list-style-type: none"> - At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them 	<ul style="list-style-type: none"> - At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media 	<ul style="list-style-type: none"> - At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	<ul style="list-style-type: none"> - Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content 	<ul style="list-style-type: none"> - Check fixity on all ingest - Use write-blockers when working with original media - Virus-check high risk content 	<ul style="list-style-type: none"> - Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content 	<ul style="list-style-type: none"> - Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	<ul style="list-style-type: none"> - Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files 	<ul style="list-style-type: none"> - Document access restrictions for content 	<ul style="list-style-type: none"> - Maintain logs of who performed what actions on files, including deletions and preservation actions 	<ul style="list-style-type: none"> - Perform audit of logs
Metadata	<ul style="list-style-type: none"> - Inventory of content and its storage location - Ensure backup and non-collocation of inventory 	<ul style="list-style-type: none"> - Store administrative metadata - Store transformative metadata and log events 	<ul style="list-style-type: none"> - Store standard technical and descriptive metadata 	<ul style="list-style-type: none"> - Store standard preservation metadata
File Formats	<ul style="list-style-type: none"> - When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs 	<ul style="list-style-type: none"> - Inventory of file formats in use 	<ul style="list-style-type: none"> - Monitor file format obsolescence issues 	<ul style="list-style-type: none"> - Perform format migrations, emulation and similar activities as needed

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	<ul style="list-style-type: none"> - Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system 	<ul style="list-style-type: none"> - At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them 	<ul style="list-style-type: none"> - At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media 	<ul style="list-style-type: none"> - At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	<ul style="list-style-type: none"> - Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content 	<ul style="list-style-type: none"> - Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content 	<ul style="list-style-type: none"> - Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content 	<ul style="list-style-type: none"> - Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	<ul style="list-style-type: none"> - Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations 	<ul style="list-style-type: none"> - Document access restrictions for content 	<ul style="list-style-type: none"> - Maintain logs of who performed what actions on files, including deletions and preservation actions 	<ul style="list-style-type: none"> - Perform audit of logs

DIGITAL PRESERVATION ACTIVITIES



Digital Preservation Librarian

Indiana University Bloomington - Bloomington, IN

QUALIFICATIONS

Required:

- Knowledge or awareness of . . . OAIS, TRAC, Data Seal of Approval, and ISO 16363
- Experience with digital repository technologies such as Hydra, Fedora, DSpace, and LOCKSS
- Experience with metadata standards such as MARC, EAD, Dublin Core, MODS, METS, PREMIS . .
•
- An understanding of digital objects in various formats
- Experience with multiple computer platforms including Linux, Windows, and/or MacOS X

Preferred:

- Hands-on experience in use of a digital preservation systems in an enterprise-scale environment
- Experience with markup languages such as HTML and XML; style sheets such as CSS and XSL; and other XML-related technologies
- Familiarity with large-scale storage systems such as HPSS and GPFS

OPEN FILE FORMATS | FORMAT NORMALIZATION

Media type	File formats	Preservation format(s)	Access format(s)	Normalization tool
Audio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Email	PST	MBOX	MBOX	readpst
Email	Maildir**	Original format	MBOX	md2mb.py
Office Open XML	DOCX, PPTX, XLSX	Original format	PDF for PPTX	Tool search in progress
Plain text	TXT	Original format	Original format	None
Portable Document Format	PDF	PDF/A	Original format	Ghostscript
Presentation files	PPT	Original format	PDF	Tool search in progress
Raster images	BMP, GIF, JPG, JP2*, PCT, PNG*, PSD, TIFF, TGA	Uncompressed TIFF	JPEG	ImageMagick
Raw camera files/Digital Negative format**	3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F	Original format	JPEG	ImageMagick/UFRaw
Spreadsheets	XLS	Original format	Original format	None
Vector images	AI, EPS, SVG	SVG	PDF	Inkscape
Video	AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, SWF, WMV	FFV1/LPCM in MKV	MP4	FFmpeg
archivematica.org/wiki/Format_policies		<ul style="list-style-type: none"> • ODF (WPD and RTF) • Original format (DOC) 	PDF	Tool search in progress



"I want you to put me in touch
with reality, but be ready to
break the connection fast."

Table 2. File formats identified with positive relative confidence values

FORMAT HETEROGENEITY	es confidence	Occurrence in policies expressing high confidence	Relative Confidence
Comma Separated Values	3	19	73%
Machine Readable Cataloging Record	3	16	68%
Audio Interchange File Format	14	45	53%
Tagged Image File Format	27	88	53%
Plain text document	17	52	51%
Extensible Markup Language	16	47	49%
Portable Document Format/A	9	25	47%
Portable Network Graphics	14	39	47%
Motion JPEG 2000	4	10	43%
Portable Document Format	25	49	32%
JPEG 2000	23	45	32%
HyperText Markup Language	18	33	29%
Postscript	11	19	27%
JPEG	28	42	20%
Waveform Audio File Format	33	47	18%
Graphics Interchange Format	22	31	17%
Rich Text Format	18	24	14%

Kyle Rimkus, Thomas Padilla, Tracy Popp, and Greer Martin.
 "Digital Preservation File Format Policies of ARL Member Libraries: An Analysis".



bit rot

&

fixity

ROFIBOT

FIXITY

Original file

0 1 1 0
0 0 0 1
1 1 1 1
0 0 0 0
0 1 0 1 1 0 0 0
1 1 1 0 0 0 1 1
0 1 0 1 1 0 0 0
1 1 1 0 0 0 1 1
0 1 0 1 1 0 0 0
1 1 1 0 0 0 1 1
0 1 0 1 1 0 0 0
1 1 1 0 0 0 1 1

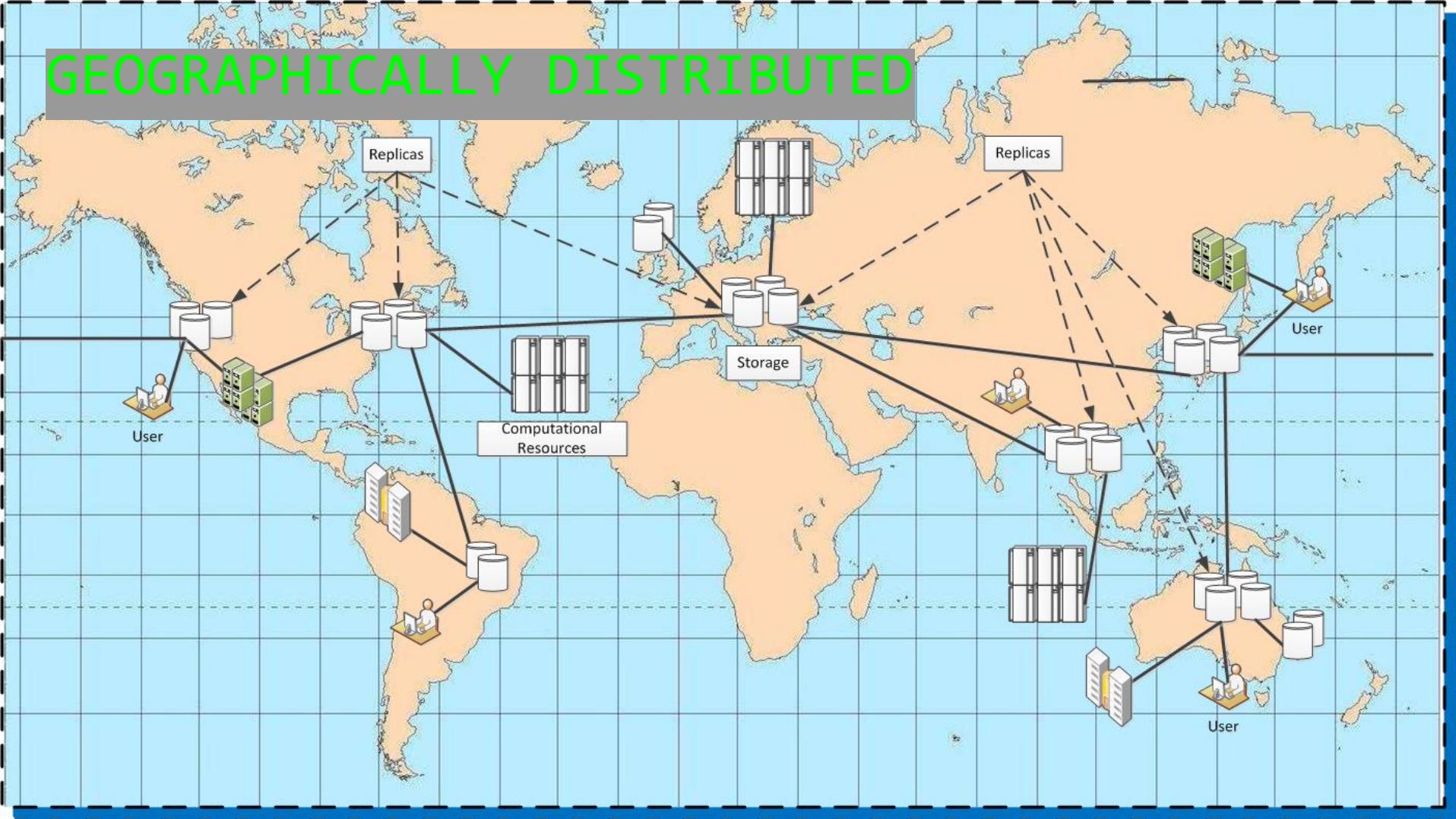
Edited file

0 1 1 0
0 0 0 1
1 1 1 1
0 0 0 0
0 1 0 1 1 0 0 0
1 1 1 0 0 0 1 1
0 1 0 0 0 1 1 0
1 1 1 0 0 0 1 1
0 1 0 1 1 0 0 0
1 1 1 0 0 0 1 1
0 1 0 1 1 0 0 0
1 1 1 0 0 0 1 1

147843254568

- 48789234

GEOGRAPHICALLY DISTRIBUTED



DARK ARCHIVE



[Home](#)[Historical Software Collection](#) | [Classic](#) | [PC Games](#) | [Shareware CD Archive](#) | [Console Living Room](#) | [Tucows Software Library](#) | [Software Library](#) | [Internet Arcade](#)Search:

All Media Types

Advanced Search

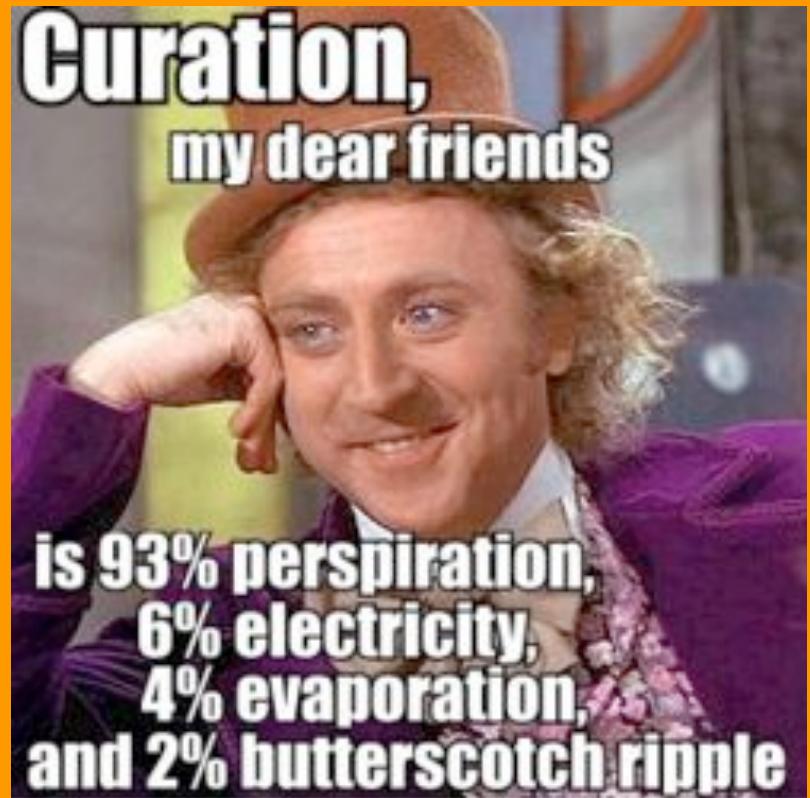
Anonymous User [\(login or join us\)](#)[Upload](#)

Champion Boxing

[Fullscreen](#)[Dark Background](#)[Unmute](#)[Back](#)

No gamepads detected. Press a button on a gamepad to use it.

Digital curation . . . is increasingly being used for the actions needed to maintain digital research data and other digital materials over their entire lifecycle and over time for current and future generations of users. Implicit in this definition are the processes of digital archiving and digital preservation, but it also includes all the processes needed for good data creation and management, and the capacity to add value to data to generate new sources of information and knowledge.



**Curation,
my dear friends**

**is 93% perspiration,
6% electricity,
4% evaporation,
and 2% butterscotch ripple**

Data curation . . . is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time.

Melissa Cragin et al, “An Educational Program on Data Curation”

DATA CURATION

active and ongoing management

add value + maintain quality

continued management depends on present/possible
interest in use

the above depends on your knowledge of potential
users + understanding of what can be done with
data

Data Curators ... (1) understand relationships among research areas and (2) the potential affordances of data for new applications. Data curators should be able to (3) assess data collections of value, and in which combinations or states they would be most useful (4) identify data that have high priority for both their immediate community and broader domains of interest.

Carole Palmer, Nicholas Weber, Trevor Munoz, Allen Renear. "Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data

Data Curator

Virginia Tech - Blacksburg, Virginia

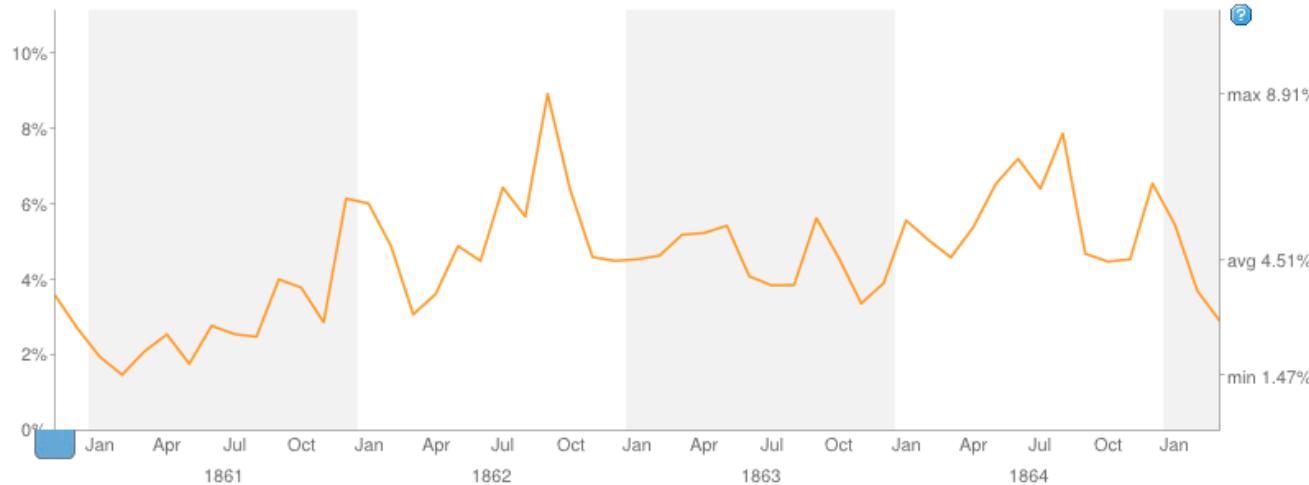
Required Qualifications:

Master's degree or Bachelor's degree and significant experience equivalent to an advanced degree. Relevant experience in working with data and metadata for the purposes of research. Demonstrated use of a scripting language such as Python or Ruby. Demonstrated understanding of database systems. Demonstrated experience in curating, handling, managing, or transforming large datasets. Demonstrated time management and project completion skills. Excellent analytical, interpersonal, oral, and written communication skills. Ability to work independently and collaboratively within a team-based environment.

Preferred Qualifications:

Experience with Geospatial Information Systems and Government Data. Knowledge of federal agency public data policies with a demonstrated understanding of requirements for long-term archiving. Experience using metadata standards for digital preservation, XML, XQuery, and XSLT. Knowledge of issues and developments in repository development. Experience with statistical modeling and data visualization. Experience or interest in linked data and semantic web technologies.

FUGITIVE SLAVE ADS



ADJUST CHART

threshold: chart type: PREDICTIVE WORDS ?

NEGRO YEARS REWARD BOY MAN NAMED

EXEMPLARY ARTICLES ?

Monday, December 23, 1861

RANAWAY--\$10 REWARD.

1

—Ranaway from the subscriber, on the 3d inst., my slave woman PARTHENIA. Had on a dark brown and white calico dress. She is of a



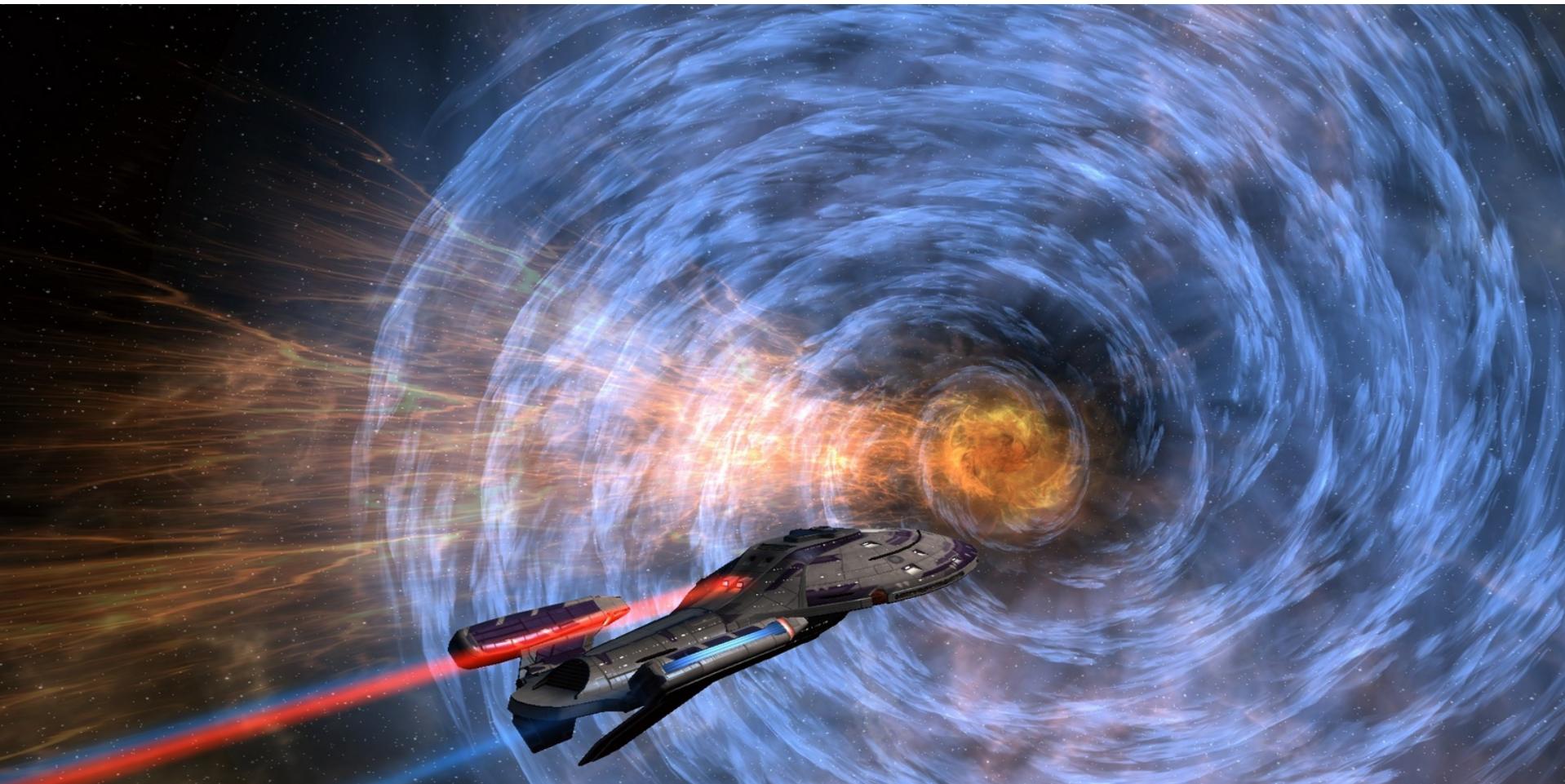
90%



What do you think that green layer is for?



and now a highly speculative Humanities Data Curation Adventure . . .





ProQuest > Find Products & Services > The Vogue Archive

The Vogue Archive

VOGUE ARCHIVE

Every issue. Every page. 1892 to today.

Fast Facts

U.S. *Vogue*, digitized
Complete backfile to 1892
Full page image format

Intended For

Academic

Community College

Get Started

CONTACT US

REQUEST A TRIAL

LOG-IN

 Download .txt Description

 Share / Save   

Significant properties, also referred to as “significant characteristics” or “essence”, are essential attributes of a digital object which affect its appearance, behaviour, quality and usability. They can be grouped into categories such as content, context (metadata), appearance (e.g. layout, colour), behaviour (e.g. interaction, functionality) and structure (e.g. pagination, sections). Significant properties must be preserved over time for the digital object to remain accessible and meaningful.

75¢

NOV. 1

VOGUE

THE
DAZZLE
SEASON:
FASHION
AND BEAUTY
FOR THE
BIG NIGHTS,
THE SNOW
SCENES

GIVENCHY...
AUDREY
HEPBURN'S
CHOICES



Covers are interesting at least in part because they have an indexical relationship to all the pages in that issue – visually representing the contents of the volume, and enticing the consumer to buy the issue at a newsstand.

Lindsay King and Peter Leonard, *Visualizing Vogue*



Robots Reading *Vogue*

Data Mining is in Fashion

Few magazines can boast being continuously published for over a century, familiar and interesting to almost everyone, full of iconic pictures — and also completely digitized and marked up as both text and images. What can you do with over 2,700 covers, 400,000 pages, 6 TB of data? Students, librarians and faculty are excited about the possibilities of working with *Vogue* to explore questions in fields from gender studies to computer science. We highlight some early experiments below:

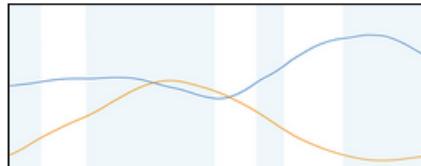
Cover Averages



See continuity and change across the decades.

[Explore »](#)

n-gram Search



Search and compare word usage in 400,000+ pages within ads, articles, or all texts.

[Explore »](#)

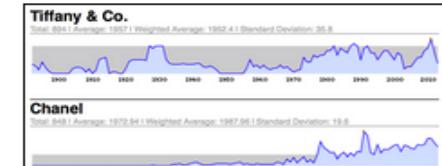
Topic Modeling



Self-organizing themes determined by word cooccurrence.

[Explore »](#)

Advertisements



Sort ads by frequency, date, and industry.

[Explore »](#)



File Format

A standardized way to structure the data stored in a computer file, or a self-contained data-stream or package of related data-streams made available as a discrete entity to a computer's operating system and its programs . . . Essentially, anything that appears to a modern operating system's file browser as a packet of information represented by a character string, a dot, and an extension is considered a representative example of a file format.

Kyle Rinkus, Thomas Padilla, Tracy Popp, and Greer Martin. "Digital Preservation File Format Policies of ARL Member Libraries: An Analysis".

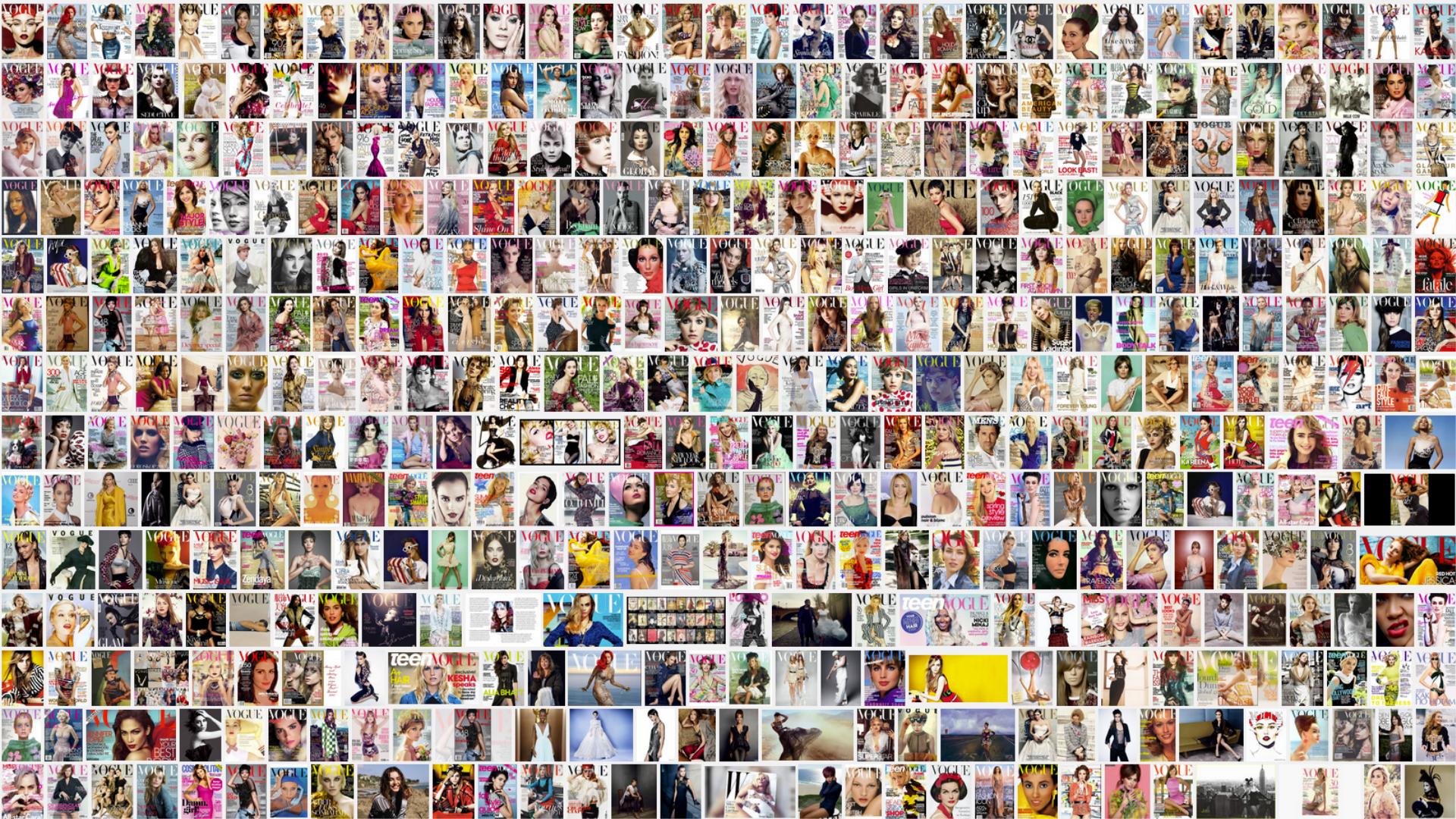


Data Curators ... (1) understand relationships among research areas and (2) the potential affordances of data for new applications. Data curators should be able to (3) assess data collections of value, and in which combinations or states they would be most useful (4) identify data that have high priority for both their immediate community and broader domains of interest.

Carole Palmer, Nicholas Weber, Trevor Munoz, Allen Renear. “Foundations of Data Curation: The Pedagogy and Practice of “Purposeful Work” with Research Data



- JPG/JPEG (Joint Photographic Experts Group)
- still image format
- widely used
- relatively small in size vs. “preservation”
still image formats like TIFF and JPEG2000

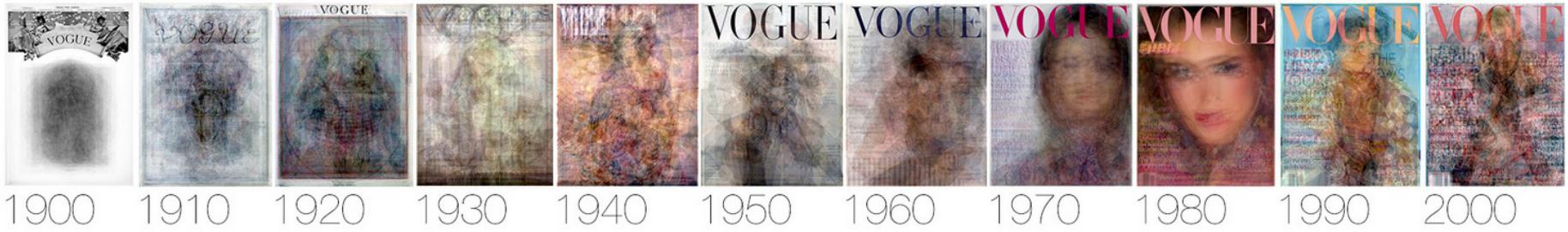


filename

string

scene00001.jpeg
scene0051.jpeg
scene0101.jpeg
scene0151.jpeg
scene0201.jpeg
scene0251.jpeg
scene0301.jpeg
scene0351.jpeg
scene0401.jpeg
scene0451.jpeg
scene0501.jpeg
scene0551.jpeg
scene0601.jpeg
scene0651.jpeg
scene0701.jpeg
scene0751.jpeg
scene0801.jpeg
scene0851.jpeg
scene0901.jpeg
scene0951.jpeg
scene01001.jpeg
scene01051.jpeg
scene01101.jpeg
scene01151.jpeg
scene01201.jpeg
scene01251.jpeg
scene01301.jpeg
scene01351.jpeg
scene01401.jpeg
scene01451.jpeg
scene01501.jpeg
scene01551.jpeg
scene01601.jpeg

Averaging Covers in *Vogue*



Vogue Covers in Colormetric Space

△

22 Oct 2013

▽

2,000 Magazine Covers in d3.js



Coverspace

X min: 0.0
Y min: 0.0



thegogglesdonothing.com/archives/2013/10/vogue_cover_averages



- PDF (Portable Document Format)
- primarily used to represent documents
- widely used



Page 1



Page 2



Page 3



Page 4

Image View

Recognize



MERCANTILE AND INDUSTRIAL EDITION



A WHIT-WILCOX COAL CO. is one of the oldest and largest concerns of their line of business. Established in 1888, they extend this field over a wide area, and have at all times brought them a large and constantly increasing trade. They handle the coal products of the best known companies, prepared to furnish low quotations. Mr. W. H. Whit is president of the company and Mr. C. G. Howell is the general manager.

Mr. E. A. Shaw, was born in Allegany county. He is a man in the middle of life, having been in business now for about ten years. He has had a very difficult time, but has steadily improved his position. They have a very decided advantage in the fact that they are in close touch with the coal companies in the transaction of their business and can in this way do a great deal to help him in his work during good and bad seasons. They have a well equipped office and a team of highly trained men who are always ready to supply a prompt and reliable service.

V. E. REED
Many of our established business men, if asked concerning their striking features, will say that it is the personal elements made by Mr. V. E. Reed. He is a man of great tact and ability. His business sense is strong and he has had a high reputation for reliability. He is a man of great energy and has a large number of friends. He has a well established record and has a large number of clients.

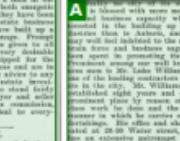
A. SEYFERT & MURRAY
A highly established establishment of Butcher's and General Store located at 100 Clinton street, is among the most successful concerns in the city. They are experienced and high class men, and have a large number of well known, well liked customers. They have a large stock house in local markets. The business will be sold to a man of high character and good business experience. Mr. Reed is a young boy, a tidy body and a good worker. He is a man of average size and has a nervous system.

THOMAS C. SMITH
Author's leading retail stores are once more in the market. The author is engaged in the holding off the packings facilities to the public. The author has a large number of complete stocks to offer.

In addition to his own business, he is also engaged in the sale of stocks to other men. In this connection he has a large number of clients, and is well known at 22 Clark street, and is a man of great tact and ability. Every style of shoe in the market is handled by him, and he also carries a full line of sporting goods. Mr. Smith has recently formed a partnership with Mr. J. C. Smith, who is a well known man in the business. They have a new concern and have a large number of clients.

The firm will be sold with anticipations of success. The firm will be sold to a man of high character and good business experience.

WILLIAM C. SMITH
He was born in Troy, N.Y., graduated at Albany Law School in 1888, and has been a resident of this city ever since. He is a well respected business man and



A WHIT-WILCOX COAL CO. is one of the oldest and largest concerns of their line of business. Established in 1888, they extended this field over a wide area, and have at all times brought them a large and constantly increasing trade. They handle the coal products of the best known companies, prepared to furnish low quotations. Mr. W. H. Whit is president of the company and Mr. C. G. Howell is the general manager.

Mr. E. A. Shaw, was born in Allegany county. He is a man in the middle of life, having been in business now for about ten years. He has had a very difficult time, but has steadily improved his position. They have a very decided advantage in the fact that they are in close touch with the coal companies in the transaction of their business and can in this way do a great deal to help him in his work during good and bad seasons. They have a well equipped office and a team of highly trained men who are always ready to supply a prompt and reliable service.

V. E. REED
Many of our established business men, if asked concerning their striking features, will say that it is the personal elements made by Mr. V. E. Reed. He is a man of great tact and ability. His business sense is strong and he has had a high reputation for reliability. He is a man of great energy and has a large number of friends. He has a well established record and has a large number of clients.

A. SEYFERT & MURRAY
A highly established establishment of Butcher's and General Store located at 100 Clinton street, is among the most successful concerns in the city. They are experienced and high class men, and have a large number of well known, well liked customers. They have a large stock house in local markets. The business will be sold to a man of high character and good business experience. Mr. Reed is a young boy, a tidy body and a good worker. He is a man of average size and has a nervous system.

THOMAS C. SMITH
Author's leading retail stores are once more in the market. The author is engaged in the holding off the packings facilities to the public. The author has a large number of complete stocks to offer.

In addition to his own business, he is also engaged in the sale of stocks to other men. In this connection he has a large number of clients, and is well known at 22 Clark street, and is a man of great tact and ability. Every style of shoe in the market is handled by him, and he also carries a full line of sporting goods. Mr. Smith has recently formed a partnership with Mr. J. C. Smith, who is a well known man in the business. They have a new concern and have a large number of clients.

The firm will be sold with anticipations of success. The firm will be sold to a man of high character and good business experience.

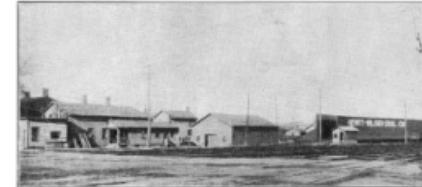
WILLIAM C. SMITH
He was born in Troy, N.Y., graduated at Albany Law School in 1888, and has been a resident of this city ever since. He is a well respected business man and

Text View

Save to



MERCANTILE AND INDUSTRIAL EDITION.



A WHIT-WILCOX COAL CO. is one of the oldest and largest concerns of their line of business. Established in 1888, they extended this field over a wide area, and have at all times brought them a large and constantly increasing trade. They handle the coal products of the best known companies, prepared to furnish low quotations. Mr. W. H. Whit is president of the company and Mr. C. G. Howell is the general manager.

Mr. E. A. Shaw, was born in Allegany county. He is a man in the middle of life, having been in business now for about ten years. He has had a very difficult time, but has steadily improved his position. They have a very decided advantage in the fact that they are in close touch with the coal companies in the transaction of their business and can in this way do a great deal to help him in his work during good and bad seasons. They have a well equipped office and a team of highly trained men who are always ready to supply a prompt and reliable service.

V. E. REED
Many of our established business men, if asked concerning their striking features, will say that it is the personal elements made by Mr. V. E. Reed. He is a man of great tact and ability. His business sense is strong and he has had a high reputation for reliability. He is a man of great energy and has a large number of friends. He has a well established record and has a large number of clients.

A. SEYFERT & MURRAY
A highly established establishment of Butcher's and General Store located at 100 Clinton street, is among the most successful concerns in the city. They are experienced and high class men, and have a large number of well known, well liked customers. They have a large stock house in local markets. The business will be sold to a man of high character and good business experience. Mr. Reed is a young boy, a tidy body and a good worker. He is a man of average size and has a nervous system.

THOMAS C. SMITH
Author's leading retail stores are once more in the market. The author is engaged in the holding off the packings facilities to the public. The author has a large number of complete stocks to offer.

In addition to his own business, he is also engaged in the sale of stocks to other men. In this connection he has a large number of clients, and is well known at 22 Clark street, and is a man of great tact and ability. Every style of shoe in the market is handled by him, and he also carries a full line of sporting goods. Mr. Smith has recently formed a partnership with Mr. J. C. Smith, who is a well known man in the business. They have a new concern and have a large number of clients.

The firm will be sold with anticipations of success. The firm will be sold to a man of high character and good business experience.

WILLIAM C. SMITH
He was born in Troy, N.Y., graduated at Albany Law School in 1888, and has been a resident of this city ever since. He is a well respected business man and

main office is located at No. 12 State street, and a branch office is located at 100 Clinton street. They give a large amount of attention to their business, and have brought them a large and constantly increasing trade. They have won many friends and supporters to their concern, and are well known throughout the state. The managers of the Daily Advertiser are members of the family.

LINKE WILLIAMS
Probably no one of us in the state is blessed with more success in the mercantile and manufacturing industries than is Linke. He is a man of great tact and ability, and has a large number of friends and supporters. He has a very decided advantage in the fact that he is a man of great energy and has a large number of friends and supporters to his concern. The managers of the Daily Advertiser are members of the family.

E. BECK
The military organization of Butcher's and General Store located at 100 Clinton street, is among the most successful concerns in the city. They are experienced and high class men, and have a large number of well known, well liked customers. They have a large stock house in local markets. The business will be sold to a man of high character and good business experience. Mr. Beck is a man of great tact and ability. His business sense is strong and he has had a high reputation for reliability. He is a man of great energy and has a large number of friends. He has a well established record and has a large number of clients.

F. REED
The military organization of Butcher's and General Store located at 100 Clinton street, is among the most successful concerns in the city. They are experienced and high class men, and have a large number of well known, well liked customers. They have a large stock house in local markets. The business will be sold to a man of high character and good business experience. Mr. Reed is a man of great tact and ability. His business sense is strong and he has had a high reputation for reliability. He is a man of great energy and has a large number of friends. He has a well established record and has a large number of clients.

N. R. SMITH
At the same time, we find that a large number of men in the mercantile and manufacturing industries are found in her bosom. And the same reason accounts for the large number of men in the mercantile and manufacturing industries who are found in her bosom. The author has a large number of friends and supporters to her concern, and is well known throughout the state. The managers of the Daily Advertiser are members of the family.

ARTHUR L. SMITH
Arthur L. Smith is the name of the general manager of the Daily Advertiser. He is a man of great tact and ability, and has a large number of friends and supporters to his concern. The author has a large number of friends and supporters to her concern, and is well known throughout the state. The managers of the Daily Advertiser are members of the family.



JOHN H. POST
One of Author's commercial concerns, which is among the most successful. From the records in the files of which Mr. Post is the owner, we find that he is a man of great tact and ability, and has a large number of friends and supporters to his concern. The managers of the Daily Advertiser are members of the family.

J. H. POST
Probably no one of us in the state is blessed with more success in the mercantile and manufacturing industries than is John H. Post. He is a man of great tact and ability, and has a large number of friends and supporters. He has a very decided advantage in the fact that he is a man of great energy and has a large number of friends and supporters to his concern. The managers of the Daily Advertiser are members of the family.

W. H. SMITH
At the same time, we find that a large number of men in the mercantile and manufacturing industries are found in her bosom. And the same reason accounts for the large number of men in the mercantile and manufacturing industries who are found in her bosom. The author has a large number of friends and supporters to her concern, and is well known throughout the state. The managers of the Daily Advertiser are members of the family.

ARTHUR L. SMITH
Arthur L. Smith is the name of the general manager of the Daily Advertiser. He is a man of great tact and ability, and has a large number of friends and supporters to his concern. The author has a large number of friends and supporters to her concern, and is well known throughout the state. The managers of the Daily Advertiser are members of the family.



- XML (EXtensible Markup Language)
- Enables semantic description of data
- Enables computational interaction with tagged features of data at scale
- Widely used

```
251 <recipe class1="meatfishgame" ethnicgroup="mexican" class2="fruitvegbeans">
252 <p><purpose align="center" rend="bold" placement="heading">Tamales</purpose>
253 <list><item>15 dried <ingredient>corn husks</ingredient></item>
254 <item>1 <ingredient>onion</ingredient> and <ingredient>garlic</ingredient></item>
255 <item>1/2 pound <ingredient>pork meat</ingredient></item>
256 <item>2 cups <ingredient>hominy</ingredient></item>
257 <item>4 teaspoons <ingredient>cornstarch</ingredient></item>
258 <item>2 teaspoons <ingredient>salt</ingredient></item>
259 <item>25 <ingredient>almonds</ingredient></item>
260 <item>1 cup <ingredient>raisins</ingredient></item>
261 <item>4 teaspoons <ingredient>lard</ingredient></item>
262 <item>3 1/2 cups <ingredient>hot water</ingredient></item>
263 <item>2 teaspoons <ingredient>baking powder</ingredient></item>
264 <item>3 teaspoons <ingredient>red pepper</ingredient></item>
265 <item>2 tablespns. <ingredient>peanut butter</ingredient></item>
266 </list>
267 </p>
268 <p>Take the <ingredient>hominy</ingredient> and <ingredient>corn starch</ingredient>
and mix with <ingredient>salt,</ingredient> <ingredient>baking powder,</ingredient>
and hot <ingredient>lard.</ingredient> Add the <ingredient>hot water</ingredient> and
beat it with a spoon until it makes a soft light dough, and let it stand for fifteen
minutes.</p>
```

Advertisements in *Vogue*

All

Cars

Cosmetics

Stores

Tobacco

Alpha

Total

Average Year

Weighted Average

Std. Dev.

Virginia Slims

Total: 230 | Average: 1984.59 | Weighted Average: 1984.21 | Standard Deviation: 9.25



Salem

Total: 91 | Average: 1983.05 | Weighted Average: 1981.94 | Standard Deviation: 7.73



Marlboro

Total: 274 | Average: 1967.39 | Weighted Average: 1970.05 | Standard Deviation: 25.13



All

Cars

Cosmetics

Stores

Tobacco

Alpha

Total

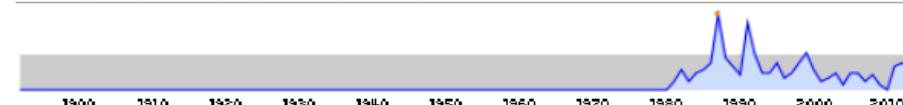
Average Year

Weighted Average

Std. Dev.

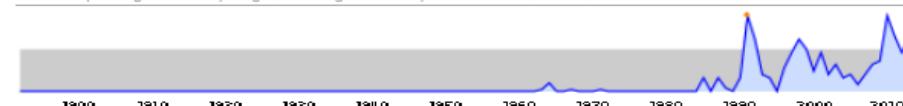
Clarins

Total: 239 | Average: 1996.59 | Weighted Average: 1994.91 | Standard Deviation: 8.7



Covergirl

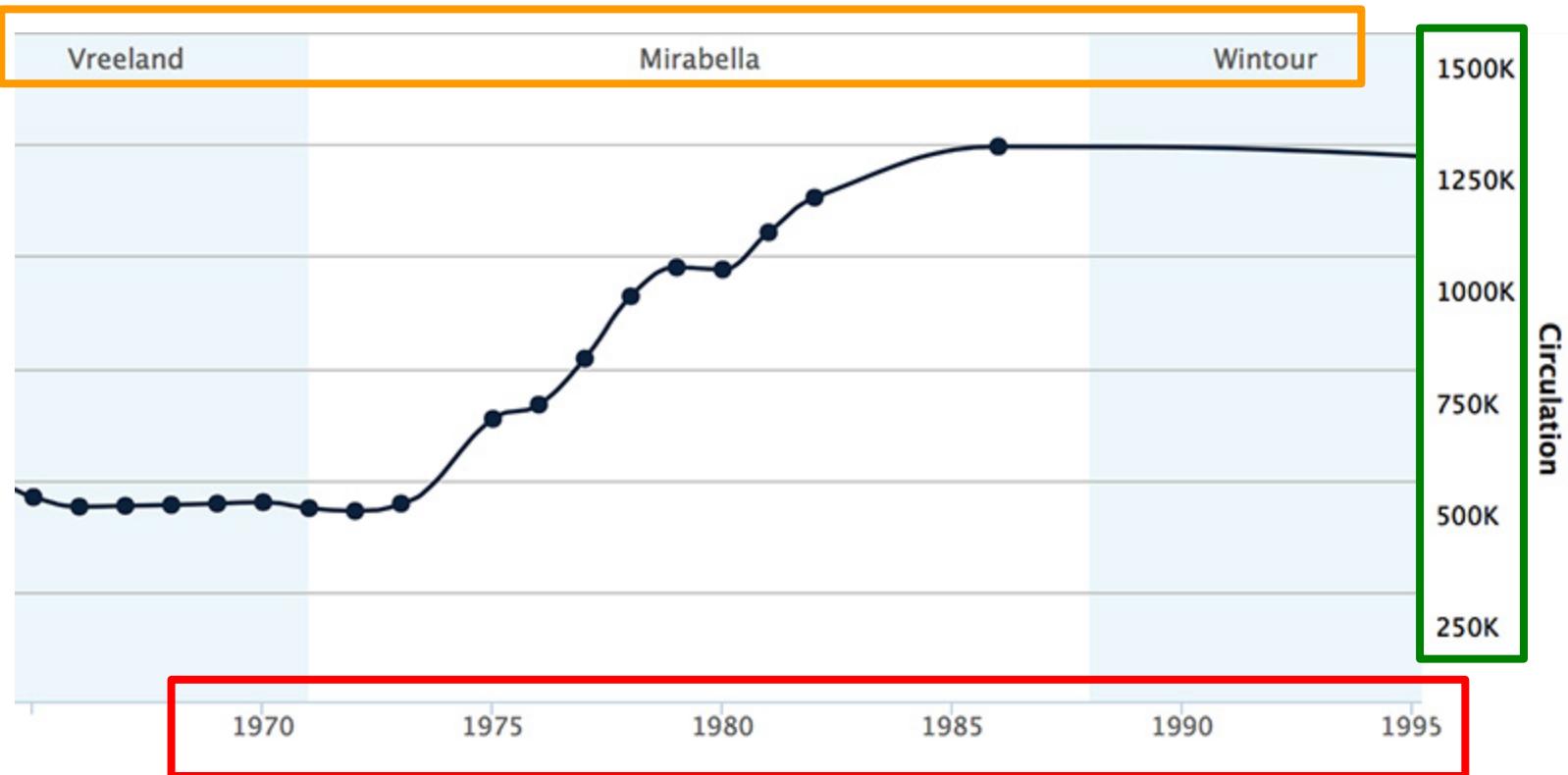
Total: 233 | Average: 1995.53 | Weighted Average: 2000.21 | Standard Deviation: 9.99

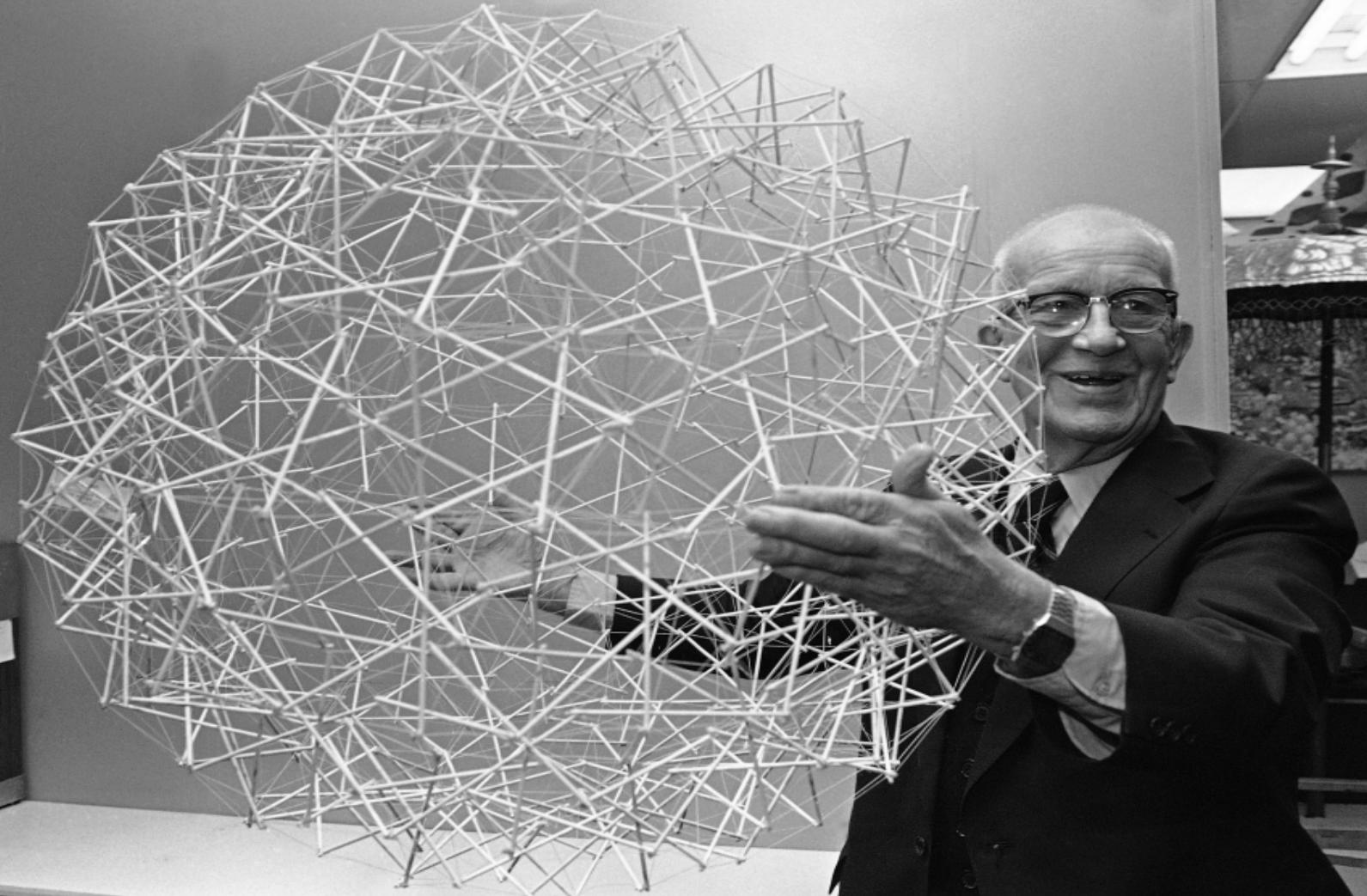


Clinique Laboratories, Inc.

Total: 507 | Average: 1991 | Weighted Average: 1994.15 | Standard Deviation: 11.04







Sharing is caring





At the current time, a key aspect of humanities data curation is thus to ensure that the representations of objects of study in the humanities function effectively as data: that they are processable by machines and interoperable such that they are durably processable across systems and collections while still retaining provenance and complex layers of meaning.

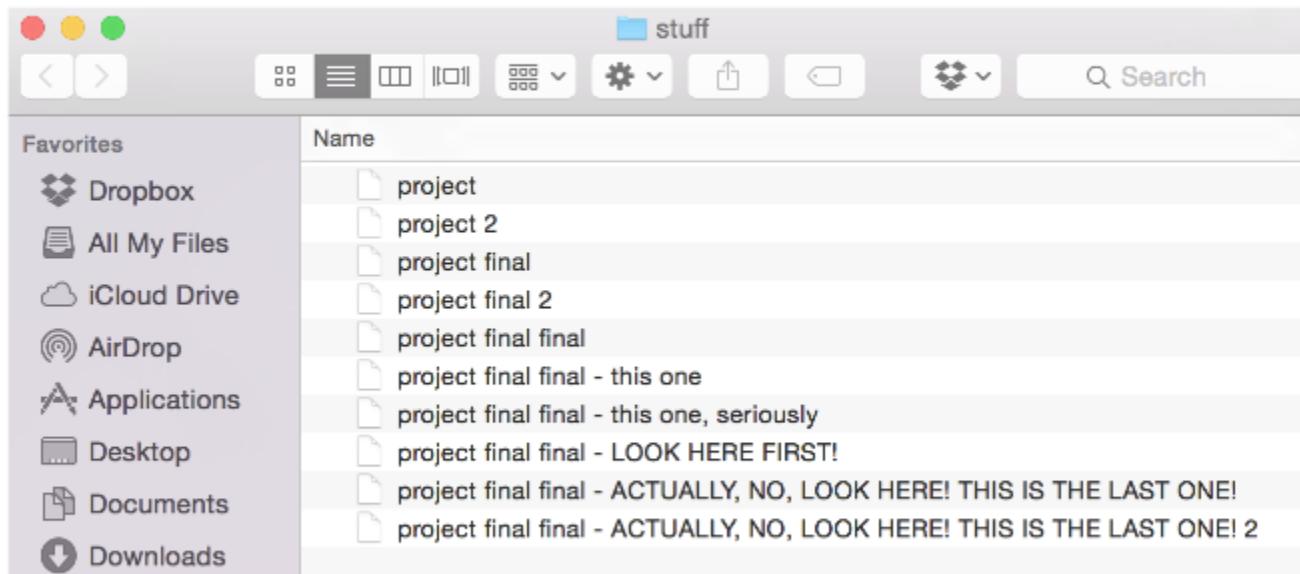
Julia Flanders and Trevor Munoz, “An Introduction to Humanities Data Curation”.

“ORGANIZATION”



FILE NAMING *CONVENTIONS* (nice way to put it)

The Only Naming Convention That Works



DOGHOUSEDIARIES

OPEN FILE FORMATS | FORMAT NORMALIZATION

Media type	File formats	Preservation format(s)	Access format(s)	Normalization tool
Audio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Email	PST	MBOX	MBOX	readpst
Email	Maildir**	Original format	MBOX	md2mb.py
Office Open XML	DOCX, PPTX, XLSX	Original format	PDF for PPTX	Tool search in progress
Plain text	TXT	Original format	Original format	None
Portable Document Format	PDF	PDF/A	Original format	Ghostscript
Presentation files	PPT	Original format	PDF	Tool search in progress
Raster images	BMP, GIF, JPG, JP2*, PCT, PNG*, PSD, TIFF, TGA	Uncompressed TIFF	JPEG	ImageMagick
Raw camera files/Digital Negative format**	3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F	Original format	JPEG	ImageMagick/UFRaw
Spreadsheets	XLS	Original format	Original format	None
Vector images	AI, EPS, SVG	SVG	PDF	Inkscape
Video	AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, SWF, WMV	FFV1/LPCM in MKV	MP4	FFmpeg
archivematica.org/wiki/Format_policies		<ul style="list-style-type: none"> • ODF (WPD and RTF) • Original format (DOC) 	PDF	Tool search in progress

README

The Grange Visitor

Data Prepared:

06/25/2014

Contact:

Thomas Padilla
Digital Humanities Librarian
Michigan State University Libraries
tpadilla@msu.edu

Description:

From 1875-1896, The Grange Visitor served as the official newspaper of the Michigan State publication from 1875-1896 across 429 issues. Collection content relates to general news recipes, advice for homemakers, literary features, obituaries, and advertisements.

Background:

The Grange Visitor collection was digitized with Project Ceres funds allocated by the Un Information Network (USAIN) and the Agriculture Network Information Center (AgNIC).
Data Summary

Format:

The Grange Visitor collection consists of 1200 files, divided as XML metadata records, and text extracted from PDF scans.

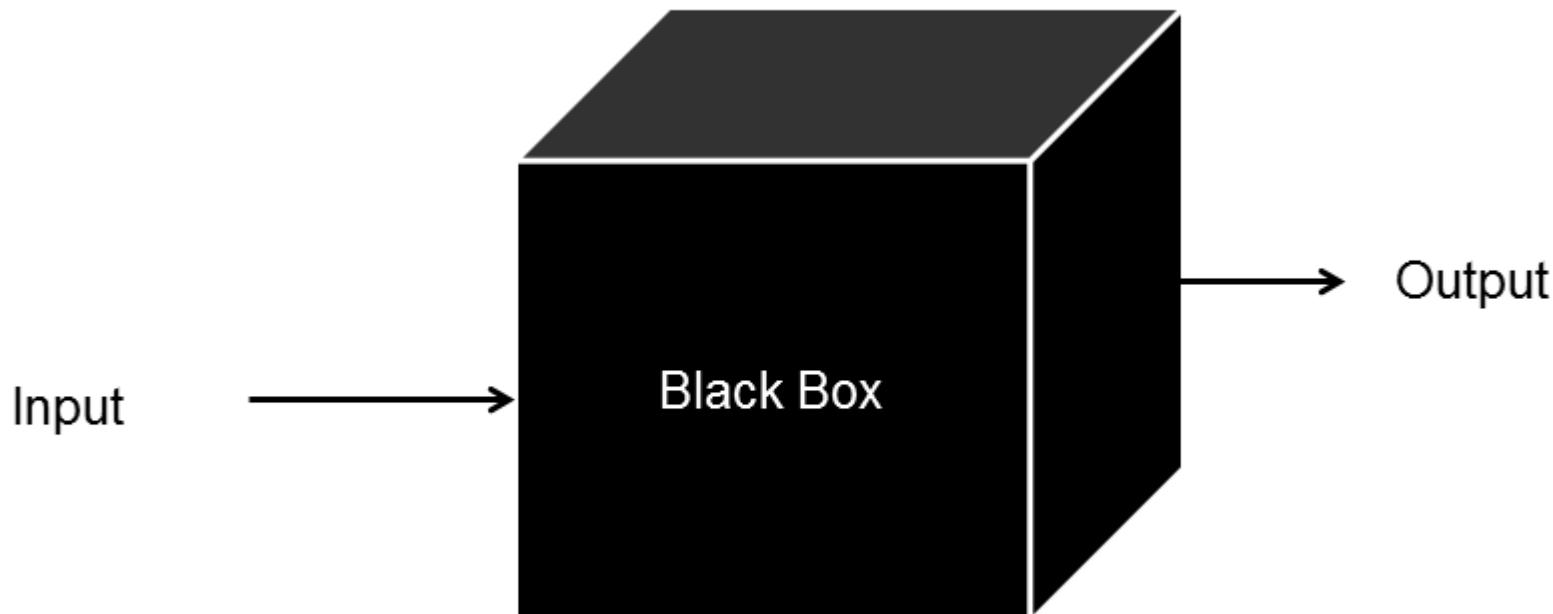
File Naming Conventions:

publication_name_yearmonthday_metadata - e.x. grange_visitor_18750601_MARCXML.xml
publication_name_yearmonthday - e.x. grange_visitor_18750601.txt

Size:

metadata - 1.7 MB (uncompressed)
pdfs - 8.46 GB (uncompressed)
txt - 69.7 MB (uncompressed)

How is the data
captured and
prepared?



Internal behavior of the code is unknown

WORKFLOW DOCUMENTATION

Journal Articles



Text

OCR

... *The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members...*

NLP

The Namurian Tsingyuan Formation from Ningxia

det

nn

prep

pobj

Table Extraction

Table

Age	Formation
Silesian	Tsingyuan Formation

Relational Features

Entity1	Entity2	Feature
Namurian	Tsingyuan Fm.	nn
Silesian	Tsingyuan Fm.	SameRow

SQL+Python Extractor

Existing Tools

WORKFLOW DOCUMENTATION

4609 records Extens

Extract Operation History

Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

- Recognize named entities in column Title
- Mass edit cells in column DBpedia Spotlight
- Mass edit cells in column DBpedia Spotlight
 - Edit single cell on row 5, column DBpedia Spotlight
- Mass edit cells in column DBpedia Spotlight
- Mass edit cells in column Publisher
- Mass edit cells in column Publisher
- Mass edit cells in column DBpedia Spotlight
- Mass edit cells in column Publisher
- Mass edit cells in column Publisher
- Mass edit cells in column Publisher
 - Edit single cell on row 2062, column Publisher
 - Edit single cell on row 2868, column Publisher
 - Edit single cell on row 854, column Publisher
 - Edit single cell on row 1114, column Publisher
 - Edit single cell on row 2393, column Publisher
 - Edit single cell on row 2206, column Publisher

Select All Unselect All

Close

```
[{"op": "named-entity-recognition/ner", "description": "Recognize named entities in column Title", "engineConfig": {"facets": [], "mode": "record-based"}, "column": "Title", "services": [{"DBpedia Spotlight"}], "parameters": {"DBpedia Spotlight": {"Confidence": "0.5", "Support": "30"}}, "op": "core/mass-edit", "description": "Mass edit cells in column DBpedia Spotlight", "engineConfig": {"facets": [], "mode": "record-based"}, "columnName": "DBpedia Spotlight", "expression": "value", "edits": [{"fromBlank": false, "fromError": false, "from": ["cook book"]}]}
```



The Historic American Cookbook Project



Feeding America

Comments?

[Home](#) [News](#) [The Authors](#) [Browse the Collection](#) [Search the Collection](#) [Museum Objects](#) [Glossary](#) [Partners](#)

MSU Libraries » DMC » Digital Collections » Feeding America

Feeding America: The Historic American Cookbook Project

The Project

The *Feeding America* project has created an online collection of some of the most important and influential American cookbooks from the late 18th to early 20th century. The digital archive includes page images of 76 cookbooks from the MSU Library's collection as well as searchable full-text transcriptions. This site also features a [glossary of cookery terms](#) and multidimensional images of [antique cooking implements](#) from the collections of the [MSU Museum](#).

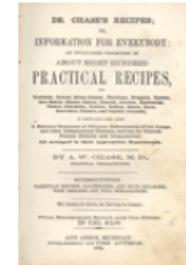
The Feeding America online collection hopes to highlight an important part of America's cultural heritage for teachers, students, researchers investigating American social history, professional chefs, and lifelong learners of all ages.

Feeding America was made possible with funds from a 2001 IMLS National Leadership Grant. The project began September 1, 2001 and was completed on August 31, 2003.

To learn more information about the project please select one of the following:

- [Introductory Essay](#) - An essay by Jan Longone that discusses *Feeding America* and the history of cookbooks in America.
- [Video Tour](#) - A video highlighting some of the key items in the Feeding America project.

Featured Cookbooks



- (3) assess data collections of value, and in which combinations or states they would be most useful
- (4) identify data that have high priority for both their immediate community and broader domains of interest.

Carole Palmer, Nicholas Weber, Trevor Munoz, Allen Renear. “Foundations of Data Curation: The Pedagogy and Practice of “Purposeful Work” with Research Data



Feeding America

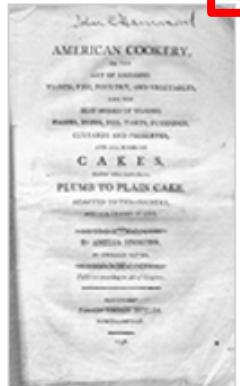
Comments?

[Home](#) [News](#) [The Authors](#) [Browse the Collection](#) [Search the Collection](#) [Museum Objects](#) [Glossary](#) [Partners](#)

[MSU Libraries](#) » [DMC](#) » [Digital Collections](#) » [Feeding America](#) » [Browse](#) » [American Cookery](#)

American Cookery

[Next Book](#)



By [Amelia Simmons](#)

Hartford: Printed for Simeon Butler, Northampton, (1798)

Interest: [The First American Cookbook](#) & [Northeast](#) & [The Great Ladies](#)

[View Page Images](#)

[Read Transcript](#)

[View PDF](#)

Feeding America: The Historic American Cookbook Project

Data Use Survey

Please complete the [Data Use Survey](#)

Download Data

[The Feeding America Dataset](#)

Description

The "Feeding America: The Historic American Cookbook" dataset contains transcribed and encoded text from 76 influential American cookbooks held by MSU Libraries Special Collections. Features encoded within the text include but are not limited to recipes, types of recipes, cooking implements, and ingredients. The 76 texts were chosen among more than 7000 cookbooks that MSU Libraries holds as representative of periods and themes in American cookbook history spanning the late 18th to early 20th century.

Background

The "Feeding America: The Historic American Cookbook" project was made possible with funds from a 2001 IMLS National Leadership Grant. The project began September 1, 2001 and was completed August 31, 2003.

File Naming Conventions

- content_type - e.x. cookbook_text.zip
- bookname - amem.xml

Size

- metadata - 293KB
- plain text - 16 MB compressed, 64 MB uncompressed
- encoded text - 17.6 MB compressed, 78.9 MB uncompressed
- dtd - 21 KB

Data Quality

Quality of text is of high fidelity to original cookbook text given transcription of text rather than application of optical character recognition (OCR). Text data is enhanced by encoding of features of text like recipe, recipe type, ingredient, measurements, and cooking implements.

Index of /dinfo/feedingamerica

	<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
	Parent Directory		-	
	cookbook.dtd.txt	11-Nov-2014 09:33	21K	
	cookbook.records.xml	12-Nov-2014 14:58	293K	
	cookbook.text.zip	13-Nov-2014 17:16	17M	
	cookbook.textencoded.zip	12-Nov-2014 15:33	17M	
	cookbook.titlelist.txt	14-Nov-2014 12:16	3.5K	
	readme.txt	14-Nov-2014 12:19	2.1K	

Data Curators ... (1) understand relationships among research areas and (2) the potential affordances of data for new applications. Data curators should be able to (3) assess data collections of value, and in which combinations or states they would be most useful (4) identify data that have high priority for both their immediate community and broader domains of interest.

Carole Palmer, Nicholas Weber, Trevor Munoz, Allen Renear. “Foundations of Data Curation: The Pedagogy and Practice of “Purposeful Work” with Research Data

Target audience



Multiple disciplines

bioinformatics

ecology

genomics

sociology

digital humanities

neuroscience

geosciences

....

SKILLS

Full computational lab skillset

Good knowledge of programming (R, Python or other), structured data, metadata, proficiency in building workflows and automating tasks

Some knowledge of scripting, using workflow tools, command line

Data + spreadsheets + some statistics and...?

Different levels of career

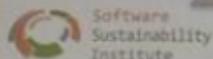
PhD students

postdoctoral researchers

research assistants

researchers in industry

....



LEARNING RESOURCES

At MSU

- DH Workshop Series
- LOCUS
- Thomas Padilla, Bobby Smiley, and Devin Higgins at the Libraries
- Kristen Mapes (DH Specialist & CAL DH Advisor)
- Brandon Locke (LEADR Director)

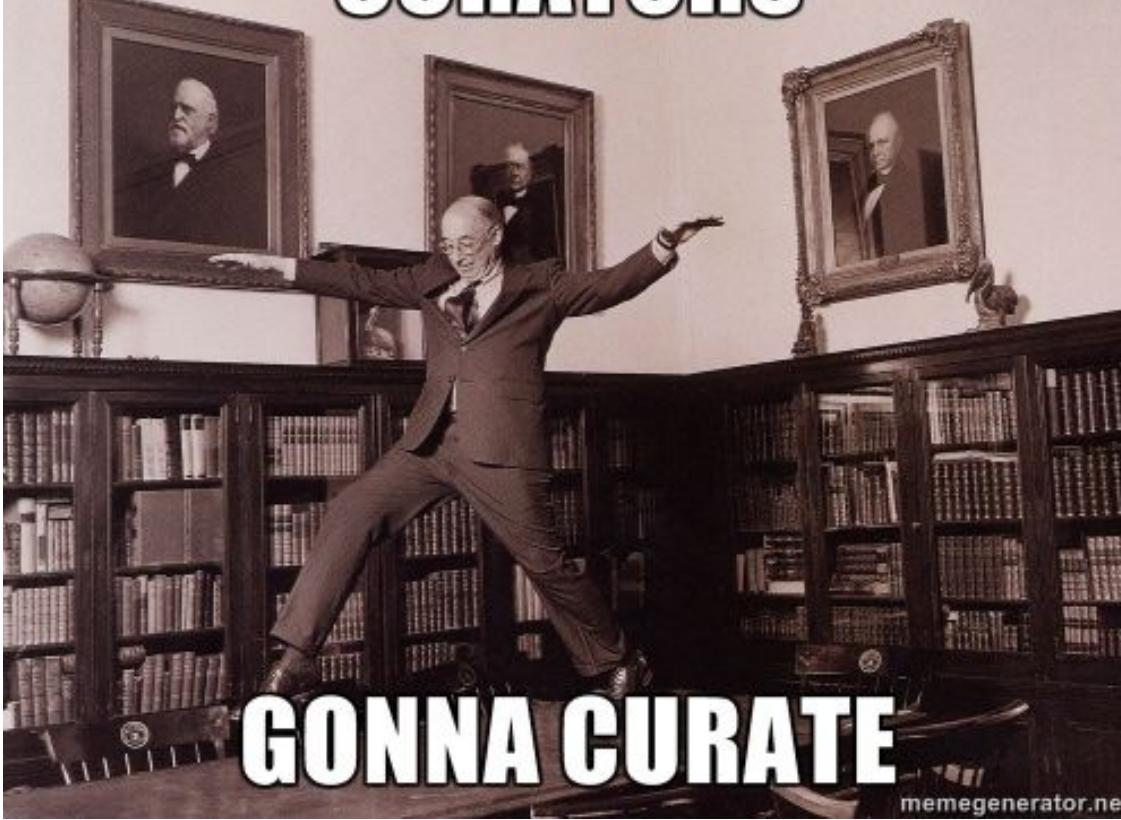
On the Web

- Programming Historian
- The Historian's Macroscope: Big Digital History

Around the Country

- Humanities Intensive Learning and Teaching
- Digital Humanities Summer Institute
- Digital Humanities Summer School
- THATCamp

CURATORS



GONNA CURATE

memegenerator.net