# **Metadata Matters**

Christy Chapman Research & Partnership Manager

> Keeping Current November 18, 2020







### The Metadata Universe http://jennriley.com/metadata map/seeingstandards.pdf (2010)



## What is metadata?

"Data about data" is too simplistic a definition. Metadata comprises the tidbits of information that provide contextual understanding about an object, whether that object be digital or analog; text, image or audio; conceptual or concrete, etc.



# **Early Forms of Metadata**



From Wilhelm von Massow, *Die Grabmäler von Neumagen* (Berlin, 1932), fig. 141, p. 243.











# Metadata Today

- Machine readable mark-up makes the metadata searchable by computer systems so that the associated information objects can be found and retrieved by users.
- File types include JSON, HTML, and, XML, which is increasingly the de facto standard.

#### <rdf:RDF

```
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:library="http://purl.org/library/"
    xmlns:bgn="http://bibliograph.net/"
    xmlns:wdrs="http://www.w3.org/2007/05/powder-s#"
    xmlns:schema="http://schema.org/"
    xmlns:genont="http://www.w3.org/2006/gen/ont#"
    xmlns:pto="http://www.productontology.org/id/"
   xmlns:void="http://rdfs.org/ns/void#"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
  <schema:Book rdf:about="http://www.worldcat.org/oclc/559802511">
   <library:placeOfPublication>
     <schema:Place rdf:about="http://id.loc.gov/vocabulary/countries/enk">
       <dcterms:identifier>enk</dcterms:identifier>
     </schema:Place>
    </library:placeOfPublication>
    library:oclcnum>559802511</library:oclcnum>
    <schema:inLanguage>en</schema:inLanguage>
    <schema:datePublished>1750</schema:datePublished>
    <schema:bookFormat rdf:resource="http://bibliograph.net/PrintBook"/>
    <schema:publication>
     <schema:PublicationEvent rdf:about="http://www.worldcat.org/title/-/oclc/559802511#PublicationEve</pre>
        <schema:location>
          <schema:Place rdf:about="http://dbpedia.org/resource/London">
           <schema:name>London</schema:name>
          </schema:Place>
       </schema:location>
        <schema:startDate></schema:startDate>
     </schema:PublicationEvent>
    </schema:publication>
    <schema:exampleOfWork rdf;resource="http://worldcat.org/entity/work/id/430841881"/>
    <schema:name xml:lang="en">A Letter from a Gentleman at Naples, concerning the late discovery of He
eruption of Mount Vesuvius.</schema:name>
    library:placeOfPublication rdf:resource="http://dbpedia.org/resource/London"/>
    <rdf:type rdf:resource="http://schema.org/CreativeWork"/>
    <schema:creator>
     <schema:Person rdf:about="http://experiment.worldcat.org/entity/work/data/430841881#Person/hercul</pre>
        <schema:givenName>HERCULANEUM</schema:givenName>
        <schema:name>HERCULANEUM.</schema:name>
     </schema:Person>
    </schema:creator>
    <wdrs:describedby>
     <genont:ContentTypeGenericResource rdf:about="http://www.worldcat.org/title/-/oclc/559802511">
        <rdf:type rdf:resource="http://www.w3.org/2006/gen/ont#InformationResource"/>
        <schema:about rdf:resource="http://www.worldcat.org/oclc/559802511"/>
        <schema:dateModified>2018-11-10</schema:dateModified>
       <void:inDataset rdf:resource="http://purl.oclc.org/dataset/WorldCat"/>
     </genont:ContentTypeGenericResource>
    </wdrs:describedby>
    <schema:productID>559802511</schema:productID>
 </schema:Book>
</rdf:RDF>
```



# Why Metadata Matters

# Metadata makes good science.



# Scientific data must be FAIR

### In 2016, the <u>"</u> <u>FAIR Guiding Principles for scientific data manage</u> <u>ment and stewardship</u>

" were published in *Scientific Data*. The principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

https://www.go-fair.org/









# **Why Metadata Matters**



If you tag it, they will find it.



If you don't tag it, it doesn't exist.



Recycle Bin	air report pregame.P	Medieval II Total Wa	some scenarios.P	DV-2-linije	DV-11-DRA	ETF21	positive force sheet musi	indeed.PNG	or vaja 2.zip	Microsoft_P	baba_yetu	vmesna oddaja.doc	pete.jpeg	floor exp.png	ponudba-6	48 fun.png	wiring_situ	wire_3_3.PNG	tgstation 2011-07-0	Soldiers - Heroes of	Start Minecraft
test.out	Medieval II Total Wa	Medieval II Total Wa	2 JobList.patch	DV-3-MOS	yellowfloor	monks.png	492.1115_b	cnet_Roller	laser in space.png	porocilo_ra	USCPY_Pri	American Idiot Remake	orvaje_vi2.txt	beatbox.jpg	wiring solars.png	Corgi hat update.zip	wire_1_1.PNG	wire_3_4.PNG	wizards.PNG	Neverwinter Nights Pla	Stronghold Kingdoms
monitor.PNG	shuttle arrives.PNG	DSC00071	view_var_lo	DV-4-CMO	brownfloor copy.png	ognjisce_re	BYOND	player menu1.PNG	snov DV.rar	porocilo 1 rvp.doc	travian.php	Matevž Baloh - Predlog te	or vaje5.zip	it is so true.jpg	admin abbreviatio	Corgi hat update	wire_1_2.PNG	wire_laying	ian tcg.PNG	Run Audiosurf	Company of Heroes
admin verbs 3.PNG	super laggy time.PNG	MEDIEVAL	view_var_lo	. tgstation 2011-09-28	floors.png	ognjisce32	new inv.png	playerpanel	schooling	poročilo1	slika1.PNG	Slike Samo-RD.rar	Slike Samo-RD	SS13Stuff.rar	admin abbreviatio	shk plan.txt	racunalniki	wire_laying	engineerin	Launch Empire Ear	The Sims™ Medieval
admin verbs 4.PNG	sheets.png	MEDIEVAL	₽ . JobListAnd	DV-5-komb	floor light.dmi	ognjisce_re	setup2010	Ğeneric_de	erro.png	RVP - Matevž Baloh - Por	inventory	or vaja 3.tar	atmosexte	tgs	what the station	forum screenshot	wire_1_3.PNG	TGS Statistics 2011-12-19	TGS Statistics 2011-12-24	Play CivCity Rome	Sid Meier's Civilization V
decrypting code.bxt	shuttle leaves.PNG	MEDIEVAL	. spawn_obje.	DV-6-aritm	9432_than	ognjisce40	w19_10_170	. Generic_en	MPP - Matevž	RVP - Matevž Baloh - Por	slika2.PNG	tgstation old	yes it is.PNG	owning a website.html	TGS Statistics 2011-12-16	tgs 2664	wire_2_1.PNG	old sprites	new engineerin	Mall Tycoon 2 Deluxe	Total War SHOGUN 2
air report actual ro	air report laggy.PNG	MEDIEVAL	. ban_unba	DV-7-sekv	view_var_lo	. large storage floor.png	new object tree.PNG	view_var_ne	editor thing.html	RVP - Matevž Baloh - Por	test.html	or vaja 3	pods.PNG	isaidno's stuff	snov DV	1050 tep 2050	wire_2_2.PNG	espeon.jpg	new engineering	GH3.exe - Shortcut	Assassin's Creed Brot
game air report s	uterus bridge.PNG	Medieval II Total Wa	wolf_castle	DV-8-avto	wall siding.png	large storage belt.png	winIDEA	Chemist.png	poorly drawn schematic	n baba_yetu	slika3.PNG	vaja 3a.txt	vaja 3b.txt	rabbit.png	red.png	tgstation13	wire_2_3.PNG	runtime.jpg	new engineering	Napoleon Total War	Farming Simulat
18QJ5.png	picture.PNG	o dog.PNG	abst1.pdf	DV-9-ROM	spacelaw book.PNG	small storage belt.png	testIDEA	castle.PNG	oldbrownfl	baba_yetu	slika4.PNG	Assassin_s	₽ 0001-Updat	Cathode Ray Tube.ppt	🕅 red1.PNG	ç glovesmall	wire_3_1.PNG	captain Ian.PNG	Assassin's Cr Assassin's	HOMEFRO Creed Revelat	Empire Total
	8 4	PSD PS	POF	POF							PDF	15T	*****				1		- Cocation	Courtes	

### April 2, 2019 NOVA https://www.youtube.com/watch?v=T76bK2t2r8g

### Deepfake Videos Are Getting Terrifyingly Real I NOVA I PBS











(

### Is it real, or is it IT black box magic?

### Can I believe that what I see is truly authentic?





# The medical profession has an ethic:

Silicon Valley has an ethos: Build it first and

https://www.nytimes.com/2018/02/12/business/computer-science-ethics-courses.html



<u>Breaking News | Radiolab</u> <u>http://www.wnycstudios.org/story/breaking-news/</u>

IRA KEMELMACHER-SHLIZERMAN, a professor in the computer science department at the University of Washington and also works at Facebook.





### What does this have to do with me?





### arXiv.org > cs > arXiv:1508.06576

#### **Computer Science > Computer Vision and Pattern Recognition**

[Submitted on 26 Aug 2015 (v1), last revised 2 Sep 2015 (this version, v2)]

### A Neural Algorithm of Artistic Style

#### Leon A. Gatys, Alexander S. Ecker, Matthias Bethge

In fine art, especially painting, humans have mastered the skill to create unique visual experiences through composing a complex interplay between the content and style of an image. Thus far the algorithmic basis of this process is unknown and there exists no artificial system with similar capabilities. However, in other key areas of visual perception such as object and face recognition near-human performance was recently demonstrated by a class of biologically inspired vision models called Deep Neural Networks. Here we introduce an artificial system based on a Deep Neural Network that creates artistic images of high perceptual quality. The system uses neural representations to separate and recombine content and style of arbitrary images, providing a neural algorithm for the creation of artistic images. Moreover, in light of the striking similarities between performance-optimised artificial neural networks and biological vision, our work offers a path forward to an algorithmic understanding of how humans create and perceive artistic imagery.

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Neural and Evolutionary Computing (cs.NE); Neurons and Cognition (q-bio.NC) Cite as: arXiv:1508.06576 [cs.CV]

(or arXiv:1508.06576v2 [cs.CV] for this version)

Search...

Help | Advanced







в

### Is it real, or is it IT black box magic?

Can I believe that what I see is truly authentic?

Ethics of the computer scientist demand they consider how their tools will be used and design protections against nefarious uses in the software.



## The Oath recited by an inductee of The Pledge of the Computer Professional is:

I am a Computing Professional.

My work as a Computing Profession affects people's lives, both now and into the future.

e

As a result, I bear moral and ethical responsibilities to society.

As a Computing Professional, I pledge to practice my profession with the highest level of integrity and competence.

I shall always use my skills for the public good.

I shall be honest about my limitations, continuously seeking to improve my skills through life-long learning.

1 shall engage only in honorable and upstanding endeavors.



Students taking the Pledge

By my actions, I pledge to honor my chosen profession.

© 2011-9. The Pledge of the Computing Professional. A 501(c)(3) non-profit organization

#### http://pledge-of-the-computing-professional.org/home-page/the-oath

### The Ten Commandments of Computer Ethics

- 1. Thou shalt not use a computer to harm other people.
- 2. Thou shalt not interfere with other people's computer work.
- 3. Thou shalt not snoop around in other people's computer files.
- 4. Thou shalt not use a computer to steal.
- 5. Thou shalt not use a computer to bear false witness.
- 6. Thou shalt not copy or use proprietary software for which you have not paid.
- 7. Thou shalt not use other people's computer resources without authorization or proper compensation.
- 8. Thou shalt not appropriate other people's intellectual output.
  - Thou shalt think about the social consequences of the program you are writing or the system you are designing.

 Thou shalt always use a computer in ways that ensure consideration and respect for your fellow humans.





# **ACM Code of Ethics**

### https://www.acm.org/code-of-ethics

1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.

This principle, which concerns the quality of life of all people, affirms an obligation of computing professionals, both individually and collectively, to use their skills for the benefit of society, its members, and the environment surrounding them. This obligation includes promoting fundamental human rights and protecting each individual's right to autonomy. An essential aim of computing professionals is to minimize negative consequences of computing, including threats to health, safety, personal security, and privacy. When the interests of multiple groups conflict, the needs of those less advantaged should be given increased attention and priority.

Computing professionals should consider whether the results of their efforts will respect diversity, will be used in socially responsible ways, will meet social needs, and will be broadly accessible.



# **ACM Code of Ethics**

### https://www.acm.org/code-of-ethics

**1.2 Avoid harm.** In this document, "harm" means negative consequences, especially when those consequences are significant and unjust. Examples of harm include unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment. This list is not exhaustive. Well-intended actions, including those that accomplish assigned duties, may lead to harm. When that harm is unintended, those responsible are obliged to undo or mitigate the harm as much as possible.

Avoiding harm begins with careful consideration of potential impacts on all those affected by decisions. When harm is an intentional part of the system, those responsible are obligated to ensure that the harm is ethically justified. In either case, ensure that all harm is minimized. To minimize the possibility of indirectly or unintentionally harming others, computing professionals should follow generally accepted best practices unless there is a compelling ethical reason to do otherwise. Additionally, the consequences of data aggregation and emergent properties of systems should be carefully analyzed. Those involved with pervasive or infrastructure systems should also consider Principle 3.7.\A computing professional has an additional obligation to report any signs of system risks that might result in harm. If leaders do not act to curtail or mitigate such risks, it may be necessary to "blow the whistle" to reduce potential harm. However, capricious



# **ACM Code of Ethics**

https://www.acm.org/code-of-ethics

3.1 Ensure that the public good is the central concern during all professional computing work.

People—including users, customers, colleagues, and others affected directly or indirectly—should always be the central concern in computing. The public good should always be an explicit consideration when evaluating tasks associated with research, requirements analysis, design, implementation, testing, validation, deployment, maintenance, retirement, and disposal. Computing professionals should keep this focus no matter which methodologies or techniques they use in their practice.

# Is it real, or is it IT black box magic? Can I believe that what I see

# is truly authentic?

# Metadata helps us determine the answer.



# Why is metadata important?

### **Digital Provenance**

Tags are like a trail of breadcrumbs, any single element alone isn't too helpful, but together they can lead us to the truth about a digital object.





Yamaguchi, Soken, 1759-1818

Creator

# Sample METS Metadata

.<mix:BasicDigitalObjectInformation> \_<mix:FormatDesignation> <mix(formatName>image/tiff</mix(formatName> </mix FormatDesignation-«mix:Compression» <mix:compressionScheme>Uncompressed</mix:compressionScheme> </mix:Compression> </mix:BasicDigitalObjectInformation> ~mix(BasicImageInformation>) ~~mix:BasicImageCharacteristics> , <mix (PhotometricInterpretation>) <mix:colorSpace>RGB</mix:colorSpace> </mix PhotometricInterpretation> </mix:BasicImageCharacteristics> </mix:BasicImageInformation> ~ <mix:ImageCaptureMetadata> .<mix:GeneralCaptureInformation> <mix:imagerroducer>UC Merced Library</mix:imageProducer> <mix(captureDevice>transmission scanner</mix(captureDevice> </mix:GeneralCaptureInformation> \_<mix(ScannerCapture>) <mix(scannerManufacturer)Fuji</mix(scannerManufacturer) <mix:ScannerModel> <mix(scancerModelName>Lanovia</mix(scannerModelName> </mix:ScannerModel> </mix:ScannerCapture> </mix:ImageCaptureMetadata> wix:ImageAssessmentMetadata> \_<mix:SpatialMetrics> <mix(samplingFrequencyUnit>2</mix(samplingFrequencyUnit> selimersequency> <min:numerator>761</min:numerator> </mix:xSamplingFrequency> smin(ySamplingFrequency>) <mix:numerator>761</mix:numerator> </mix:ySamplingFrequency> </mix:SpatialMetrics> \_<mix:ImageColorEncoding> .<mix:bitsPerSample> <mix:bitsPerSampleValue>8,8,8</mix:bitsPerSampleValue> <mix:bitsPerSampleUnit>integer</mix:bitsPerSampleUnit> </min:bitsPerSample> <mix(samplesPerPixel>3</mix(samplesPerPixel> </mix:ImageColorEncoding> </mix:ImageAssessmentMetadata>

"mix:BasicDigitalObjectInformation -<mix:FormatDesignation> <mix(formatName>image/jpeg</mix(formatName> </mix:FormatDesignation> ~<mix:Compression> <mix(compressionScheme>JPEG Baseline Sequential</mix(compressionScheme> </mix:Compression> </mix:BasicDigitalObjectInformation> -<mix:BasicImageInformation> "<mix:BasicImageCharacteristics> ,<mix:PhotometricInterpretation> <mix:colorSpace>BlackIsZero</mix:colorSpace> </mix PhotometricInterpretation> </mix:BasicImageCharacteristics> </mix:BasicImageInformation> ~~mix:ImageAssessmentMetadata> \_<mix(SpatialMetrics>) <mix(samplingFrequencyUnit>2</mix(samplingFrequencyUnit> ~mix:gSamplingFrequency> <mix:numerator>300</mix:numerator> </mix:xSamplingFrequency> ~mix(ySamplingFrequency>) <mix:numerator>300</mix:numerator> </mix:ySamplingFrequency> </mix:SpatialMetrics> ~~mix:ImageColorEncoding> \_<mix(bitsPerSample>) <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> <mix:bitsPerSampleUnit>integer</mix:bitsPerSampleUnit> </minubitsPerSample> <mix(samplesPerPixel>l</mix(samplesPerPixel> </mix:ImageColorEncoding> </mix:ImageAssessmentMetadata> \_<mix:ChangeHistory> ~~mix:ImageProcessing> <mix(processingAgency>UC Merced Library</mix(processingAgency> <mix:ProcessingSoftware> <mix:processingSoftwareName>Photoshop CS2</mix:processingSoftwareName> </mix:ProcessingSoftware-</mix:ImageProcessing> </mix:ChangeHistory>









# **Metadata for 3D Models**

model_repository_id	Model	-	Unique id of the Model Record within the repository, assigned when the Model is entered into the repository. Used to reference the Model Record in other repository records.
model_guid	Model	-	GUID of the model, only appible if the records 'model_purpose' element is 'master'
model_file_name	Model	-	Contains the name of the file as it appears in the file system.
model_file_type	Model		Specifies the file type of the model file
date_of_creation	Model	-	Date model was created
derived_from	Model	-	Contains the repository id's of the datasets and/or models that went into creating a model.
creation_method	Model	-	Method used to create model
model_modality	Model	-	Type of information represented in the model file, either a point cloud data or a surface mesh
units	Model	-	Units that should be used to interpret the models scale
model_purpose	Model	-	Use of model. Either identifies a model as an intermediate processing step informative enough to be saved, a master model, or a derivative of a master model for a specific use or delivery method.
registered to			We need a way to track if models are registered to each other and how their transform files should be used
point_count	Model	Point Cloud Attributes	Number of points in a model
has_normals	Model	Point Cloud Attributes	Boolean for if the model has normals
is_watertight	Model	Mesh Attributes	Boolean for if the model is watertight
face_count	Model	Mesh Attributes	Number of faces in a model
vertices_count	Model	Mesh Attributes	Number of vertices in a model
has_vertex_color	Model	Mesh Attributes	Boolean for if the model has vertex color
has_uv_space	Model	Mesh Attributes	Boolean for if the model has uv space
mesh style			Do we need to track if a model is based on triangles, quads, nerbs?
model_maps	Model	Mesh Attributes	Contains the value in the 'uv_map_repository_id" element in the parent UV Map Record. Used to identify the UV Maps to used with a given Model



https://dpo.si.edu/blog/smithsonian-3d-metadata-model

# Metadata for Machine Learning



Figure 2: Simplified version of our schema used to store artifact metadata and lineage information. A detailed version available under Apache license can be found at https://github.com/awslabs/ml-experiments-schema. Bold attributes indicate lineage relationships and every entity can be extended via arbitrary key-value pairs stored in the annotation attribute.