Digital Preservation for AI: a flipped perspective

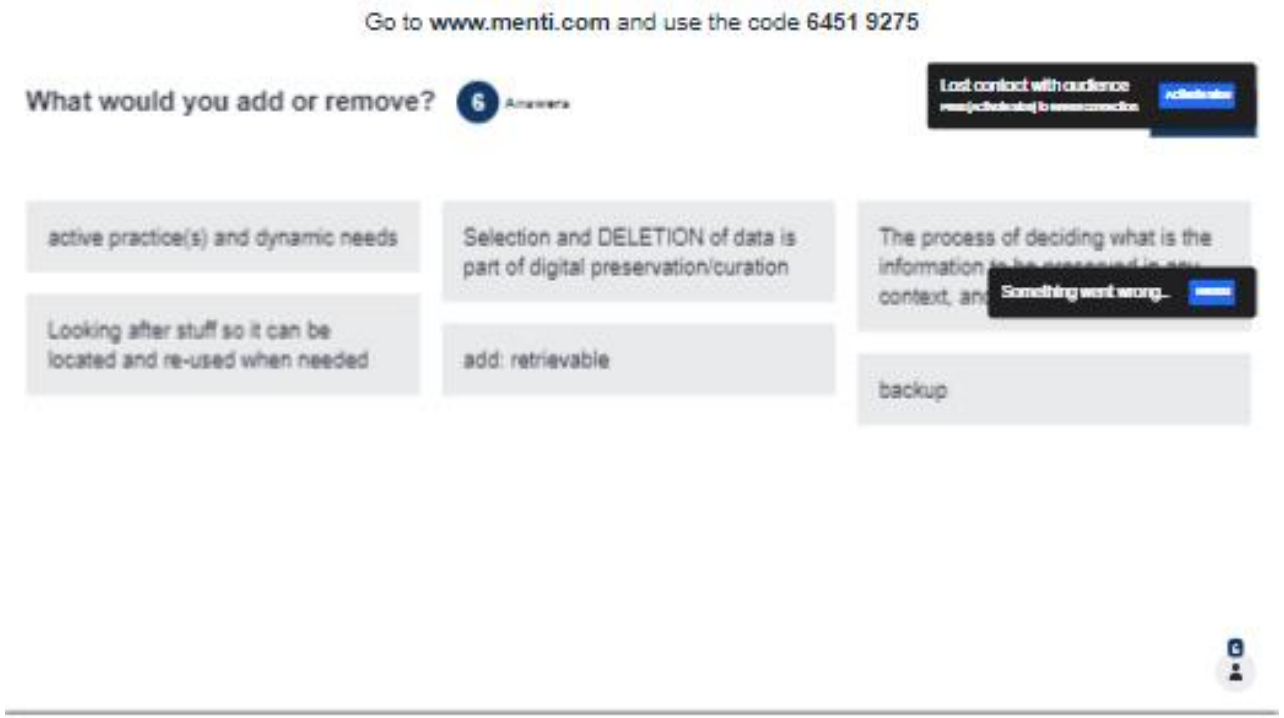
Yunhyong Kim

Agenda

- About me – a brief history
- Setting the scene – AI for digital preservation
- Collaborations going forward

What is digital preservation?

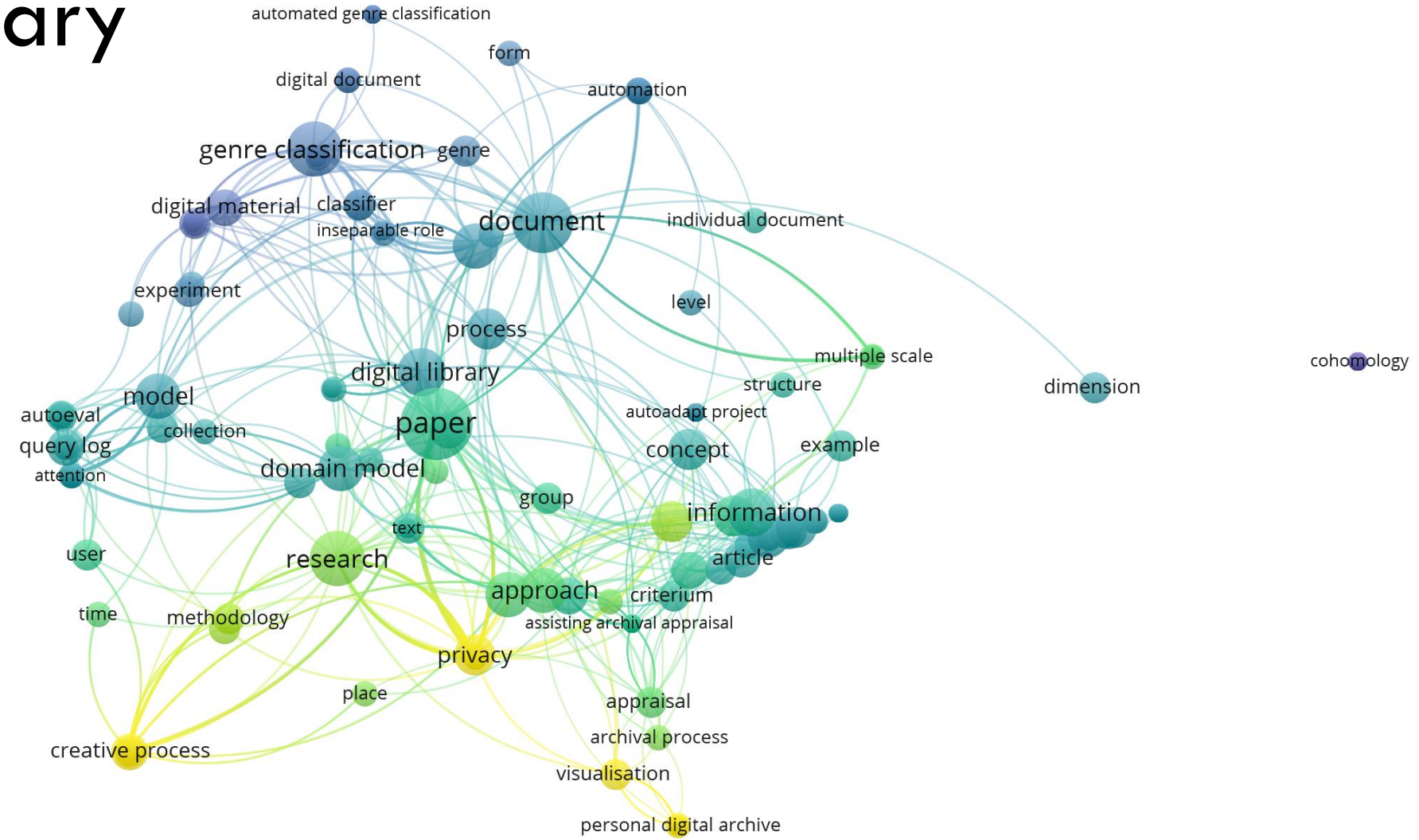
- "Digital preservation is the process of ensuring that digital information remains accessible, usable, and authentic over time. It involves a set of activities and strategies that aim to protect digital assets from loss, degradation, technological obsolescence, and other threats that may affect their long-term accessibility and usefulness."
- "Digital preservation encompasses a range of practices, including the selection of appropriate file formats, the creation of metadata to describe digital objects, the establishment of policies and procedures for long-term storage, and the implementation of backup and recovery strategies. It also involves ongoing monitoring and management to ensure the continued integrity and usability of digital materials."
- "Digital preservation is especially important for organizations and institutions that hold valuable or culturally significant digital collections, such as libraries, archives, museums, and government agencies. By preserving digital materials, these organizations can ensure that important information and cultural artifacts are accessible to future generations."



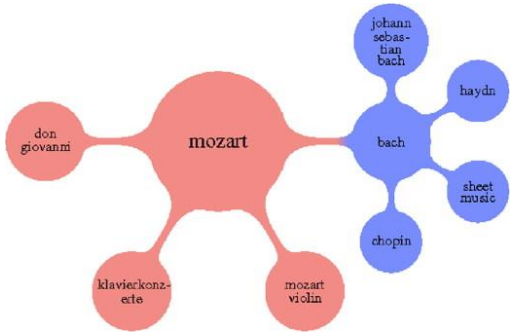
About me

a brief history

Visual Summary



Timeline



https://doi.org/10.1016/j.ipm.2011.07.002

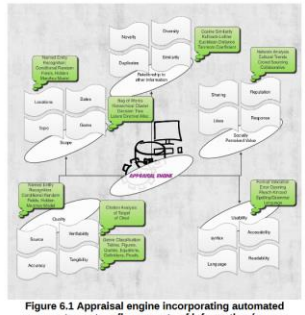
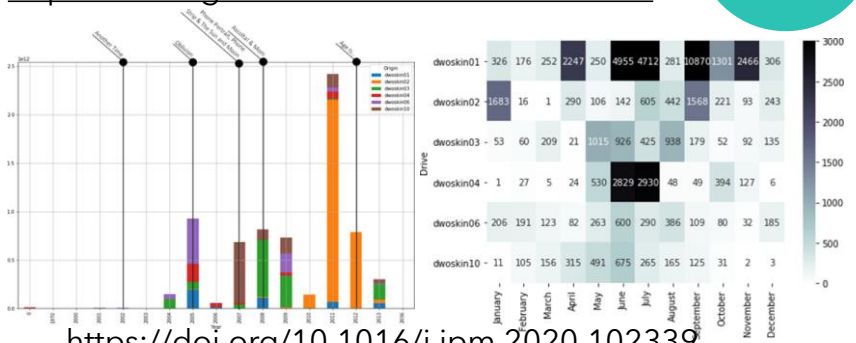


Figure 6.1 Appraisal engine incorporating automated process to capture five aspects of information (scope, quality, relationship to other information, socially perceived value, usability). Green bubbles highlight related automated information processing method.

https://doi.org/10.1002/meet.14505001042



https://doi.org/10.1016/j.ipm.2020.102339

2005 - 2008

1

Automated semantic metadata extraction.

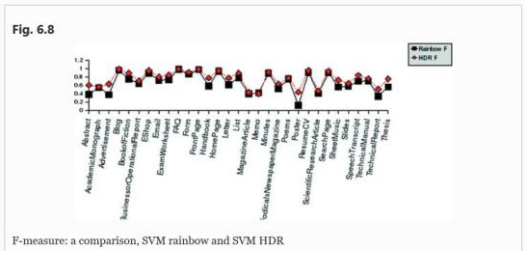


Fig. 6.8

F-measure: a comparison, SVM rainbow and SVM HDR

https://doi.org/10.1007/978-90-481-9178-9_6

2008 - 2010

2

Automatically structuring domain knowledge – taxonomies, term graphs, ontologies.

Relevance criteria of exploratory search.

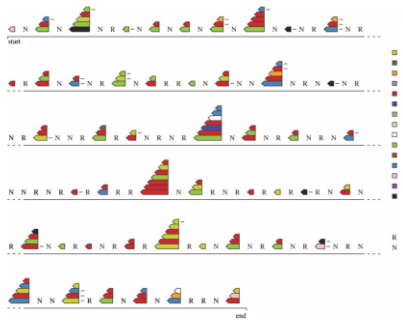


Fig. 10 Visual representation of the search session of participant 2

https://doi.org/10.1007/s00799-011-0067-7

2011 - 2013

3

BlogForever - blog preservation strategy, online designated communities and knowledge graphs

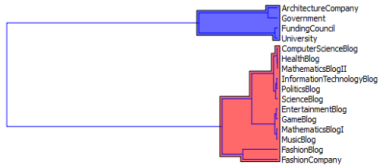


Figure 3. Hierarchical clustering of communities on the basis of preference of content management systems.

https://doi.org/10.2218/ijdc.v10i1.360

2018 - 2022

4

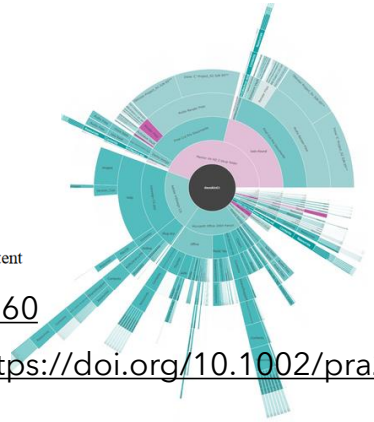
The Legacies of Stephen Dwoskin's Personal Cinema.

Digital preservation of music for Alzheimer's.

2022 - 2023

5

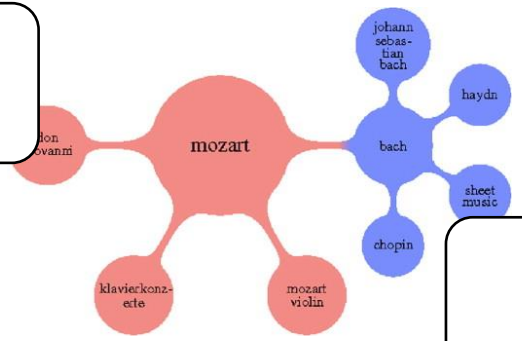
AI for the Arts and Humanities – educating students to get under the hood for responsible AI.



https://doi.org/10.1002/pra2.329

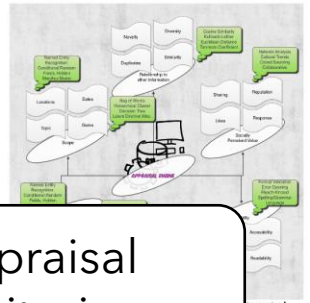
Timeline

Domain Modelling



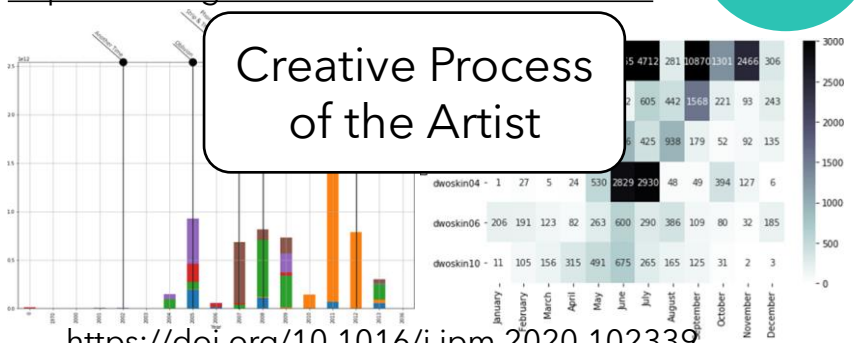
https://doi.org/10.1016/j.ipm.20

Appraisal Criteria



https://doi.org/10.1002/meet.14505001042

Creative Process of the Artist

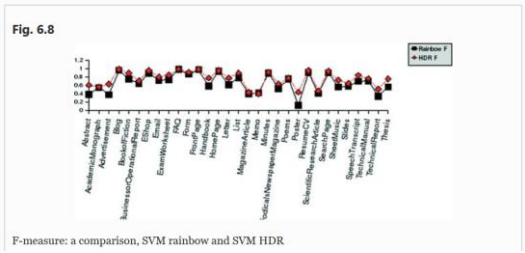


https://doi.org/10.1016/j.ipm.2020.102339

2005 - 2008

1

Automated semantic metadata extraction.



https://doi.org/10.1007/978-90-481-9178-9_6

Classification

2008 - 2010

2

Automatically structuring domain knowledge – taxonomies, term graphs, ontologies.

Relevance criteria of exploratory search.

Relevance Criteria

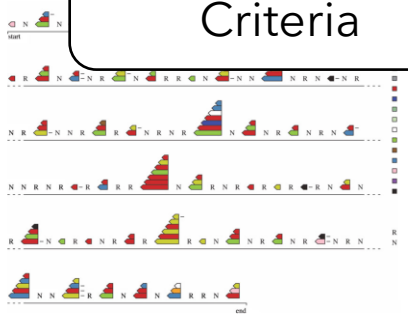


Fig. 10 Visual representation of the search session of participant 2

https://doi.org/10.1007/s00799-011-0067-7

2011 - 2013

3

BlogForever - blog preservation strategy, online designated communities and knowledge graphs

Designated Communities

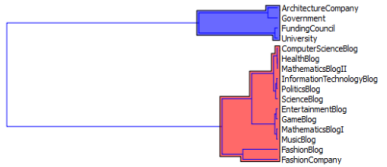


Figure 3. Hierarchical clustering of communities on the basis of preference of content management systems.

https://doi.org/10.2218/ijdc.v10i1.360

2018 - 2022

4

The Legacies of Stephen Dwoskin's Personal Cinema.

Digital preservation of music for Alzheimer's.

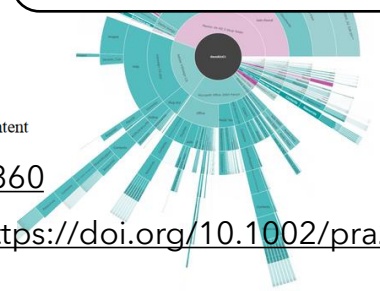
2022 - 2023

5

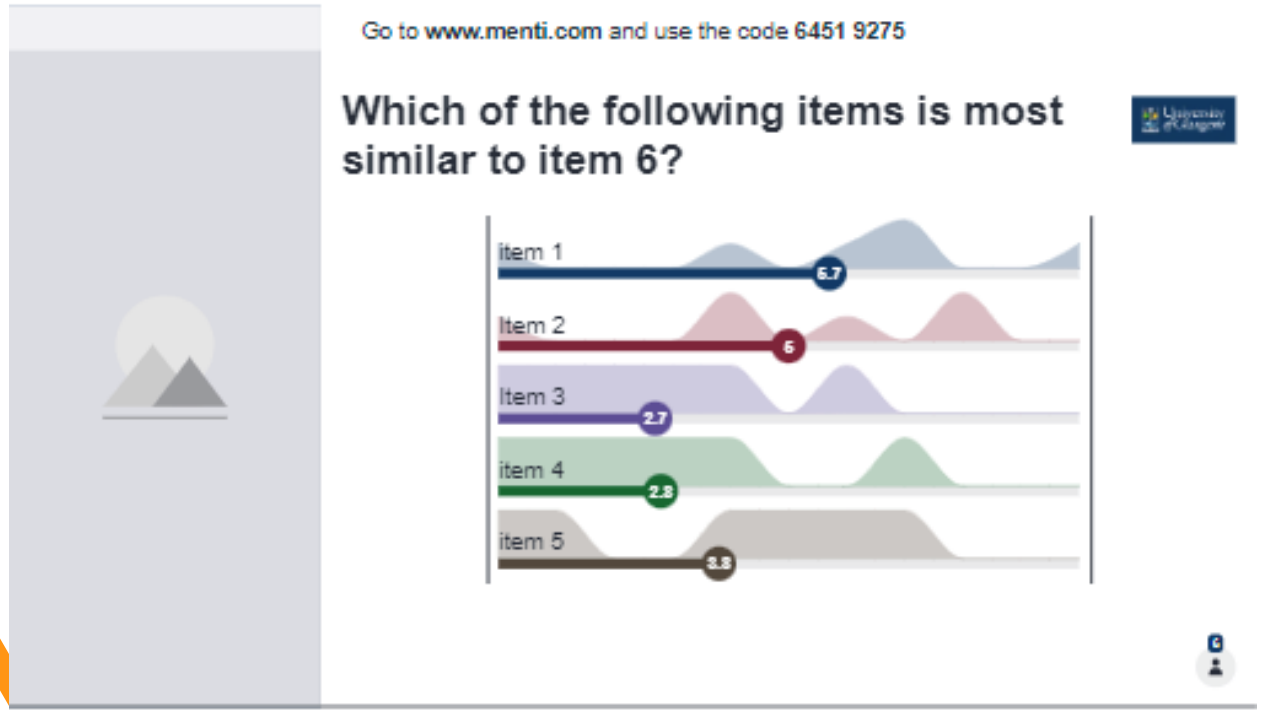
AI for the Arts and Humanities – educating students to get under the hood for responsible AI.

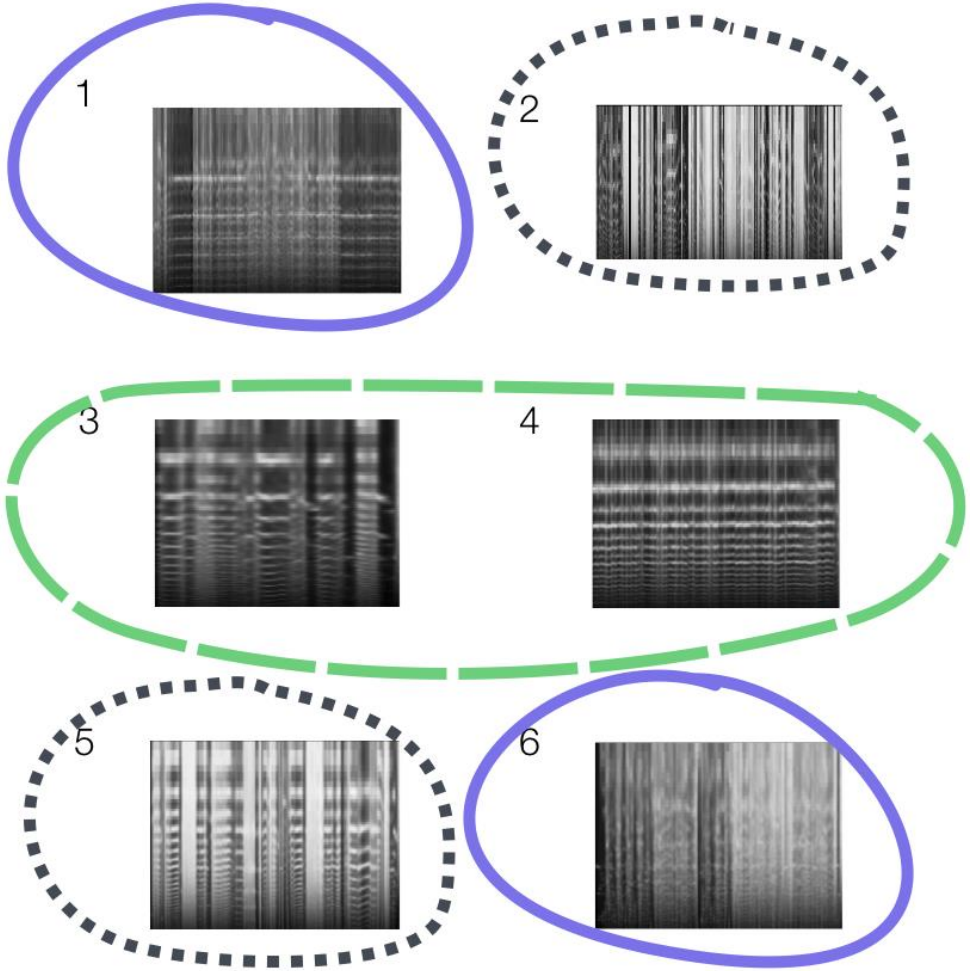
AI in the Arts – Discriminative and Generative

Hard Drive Visualisation



https://doi.org/10.1002/pra2.329





Pairs made by previous participants. 1. Miles Davis “So What”. 2. Arnold Schoenberg, String Trio Op. 45. 3. Bach “Well Tempered Clavier”. 4. Dire Straits “Sultan’s of Swing”. 5. Beethoven “5th Symphony” Op. 67. 6. Adele “hello”

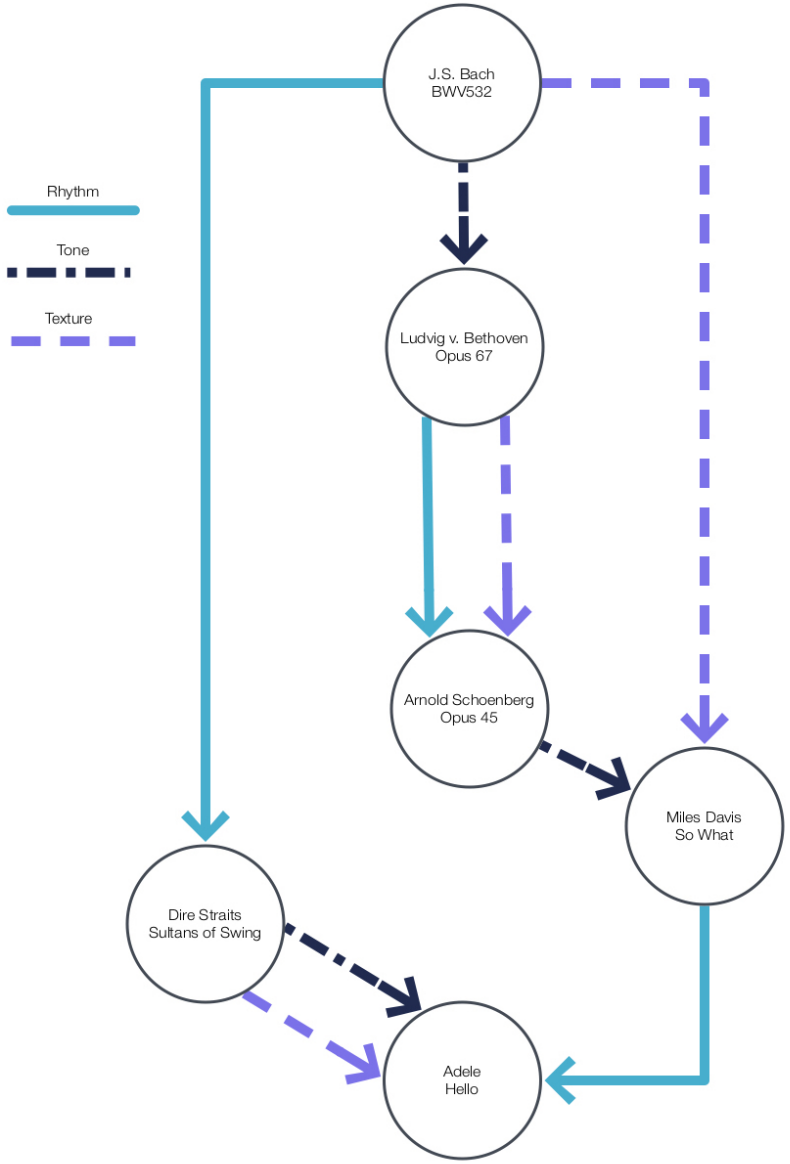


Figure 4: Inferred inherited properties between six pieces of music: colour and style of lines distinguish type of property being influenced and arrows indicate direction of inheritance.

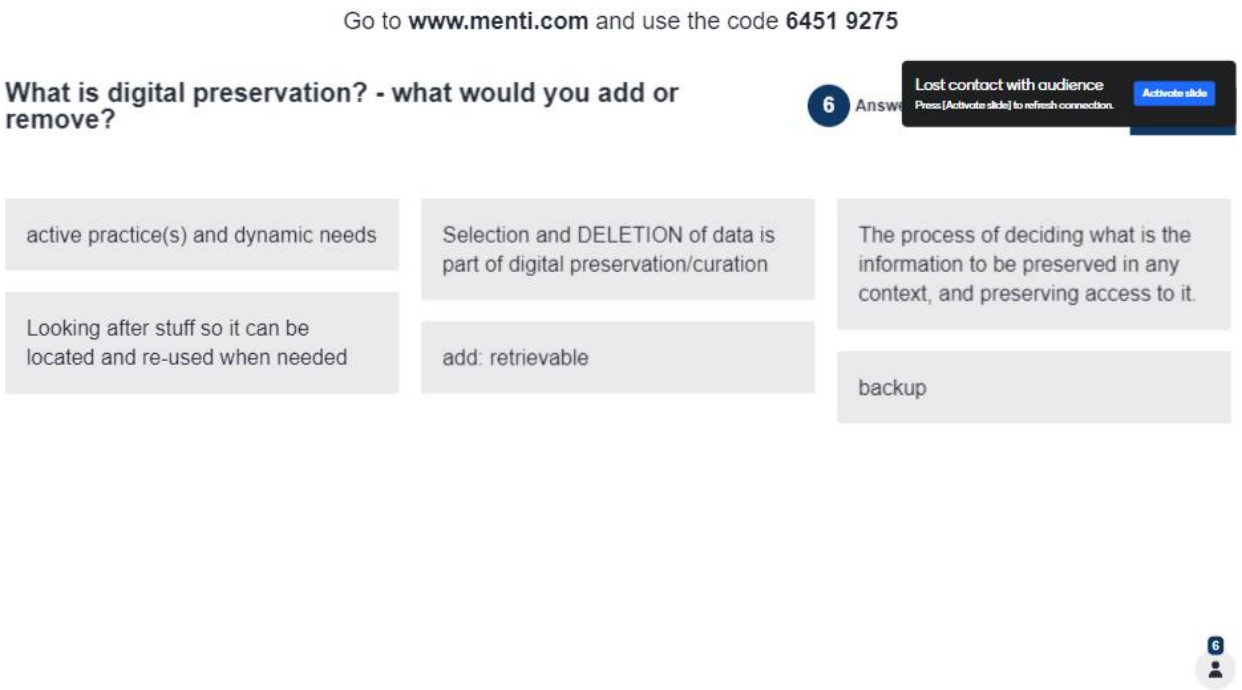
- Some simple inferences made from visual similarity.
- Information science to complement music therapy and/or neuroscience?
- Extended clustering for digital curation of music?

Setting the Scene

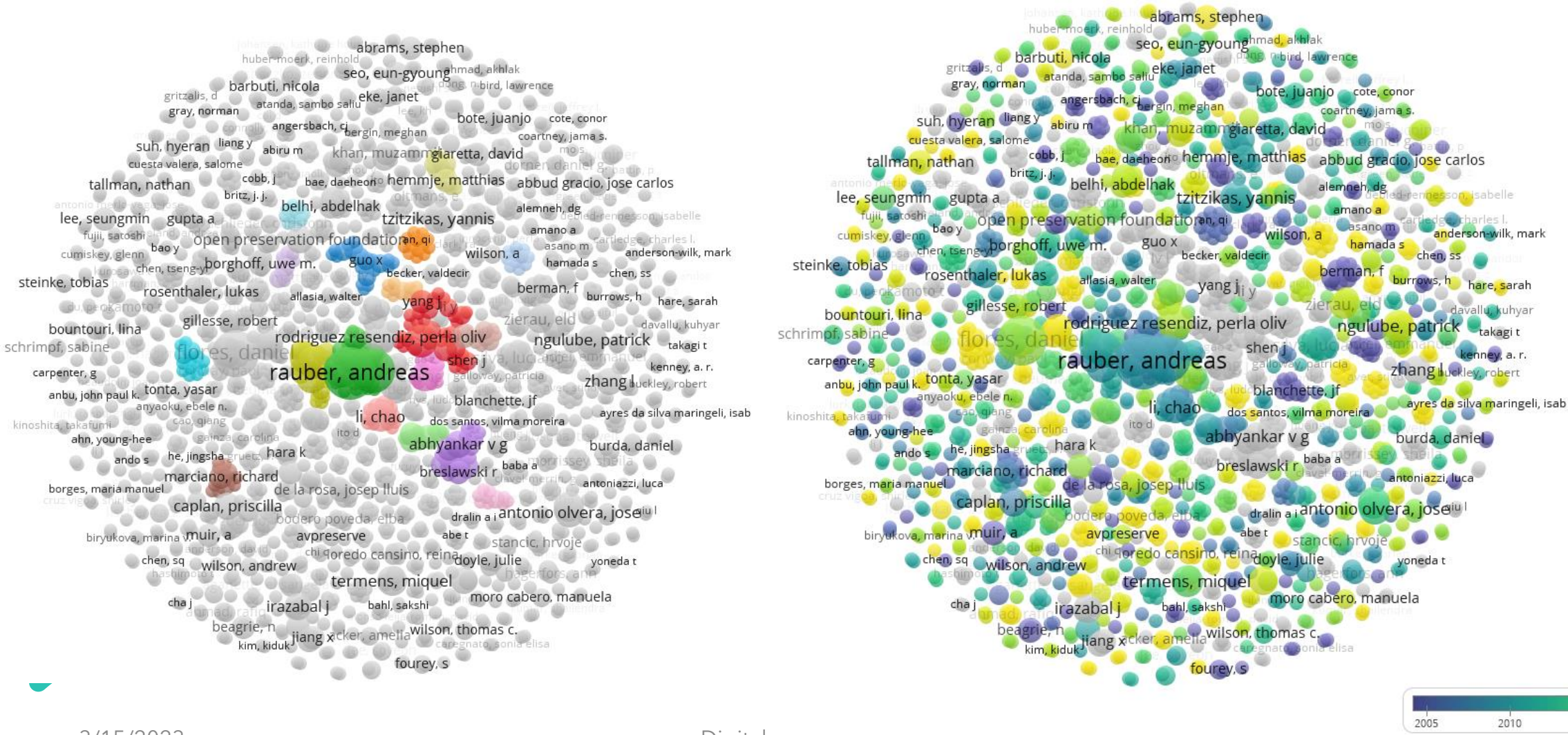
AI for Digital Preservation

What is digital preservation?

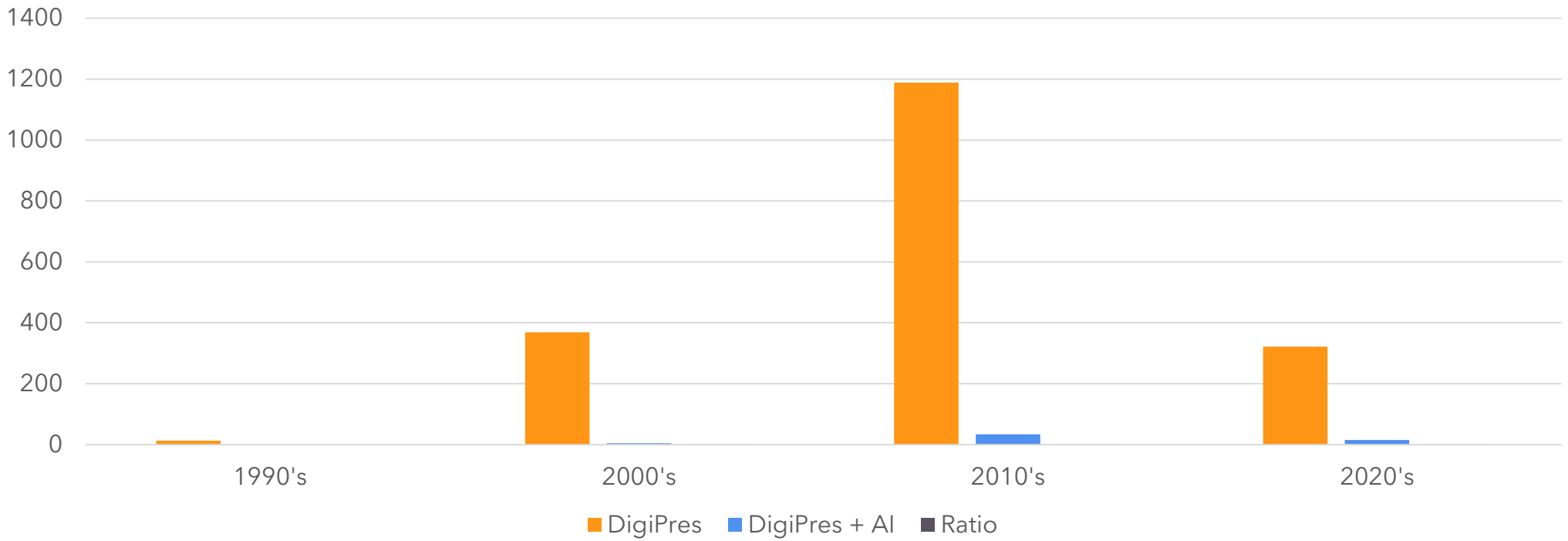
- "Digital preservation is the process of ensuring that digital information remains accessible, usable, and authentic over time. It involves a set of activities and strategies that aim to protect digital assets from loss, degradation, technological obsolescence, and other threats that may affect their long-term accessibility and usefulness."
- "Digital preservation encompasses a range of practices, including the selection of appropriate file formats, the creation of metadata to describe digital objects, the establishment of policies and procedures for long-term storage, and the implementation of backup and recovery strategies. It also involves ongoing monitoring and management to ensure the continued integrity and usability of digital materials."
- "Digital preservation is especially important for organizations and institutions that hold valuable or culturally significant digital collections, such as libraries, archives, museums, and government agencies. By preserving digital materials, these organizations can ensure that important information and cultural artifacts are accessible to future generations."

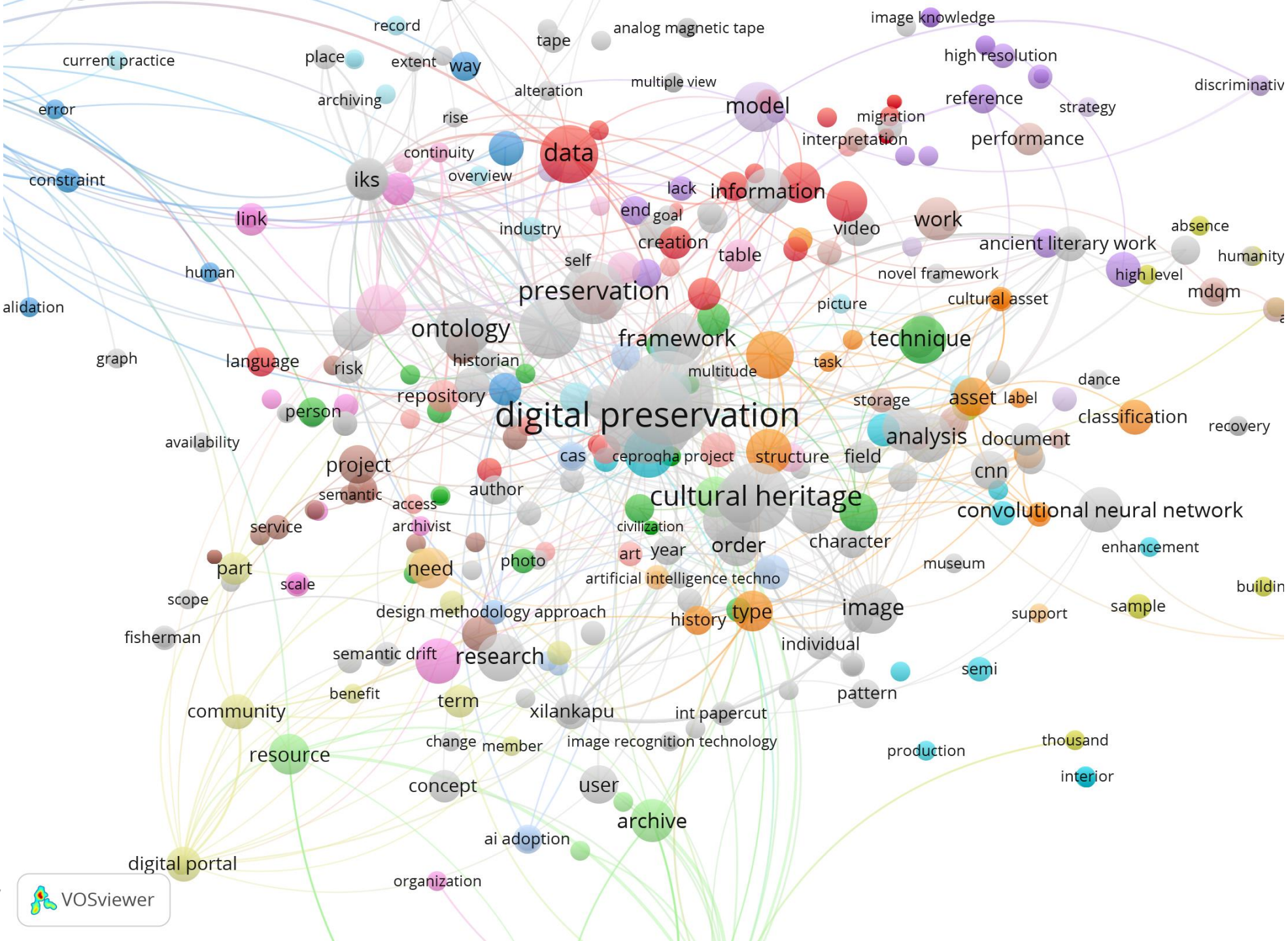


Digital Preservation: Collaboration



Web of Science




record
tape
analog magnetic tape
image knowledge
high resolution
current practice
place
extent
way
multiple view
discriminativ
alteration
archiving
model
reference
strategy
error
rise
migration
interpretation
performance
data
continuity
overview
iks
constraint
lack
information
end
goal
work
video
absence
link
industry
creation
ancient literary work
humanity
table
self
novel framework
high level
human
preservation
mdqm
picture
cultural asset
alidation
ontology
framework
technique
graph
language
risk
historian
task
multitude
dance
repository
storage
asset
label
person
digital preservation
classification
recovery
availability
analysis
document
cas
ceproqha project
structure
field
project
cnn
author
semantic
cultural heritage
convolutional neural network
access
archivist
civilization
service
character
order
enhancement
art
year
photo
museum
part
need
artificial intelligence techno
scale
buildin
scope
design methodology approach
history
type
image
support
sample
fisherman
individual
semantic drift
research
semi
benefit
pattern
term
community
xilankapu
int papercut
change
member
image recognition technology
production
thousand
resource
interior
concept
user
archive
ai adoption
digital portal
organization

Questions

- Is research in digital preservation are shrinking? Or has research moved to industry?
- Is it correct that the digital preservation research landscape tends to be characterized by many small groups working in short bursts?
- Even more dispersed for AI?

What is AI?

Artificial Intelligence (AI)

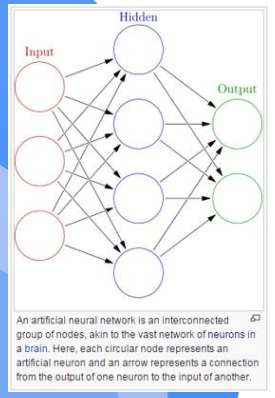
Not Machine Learning

Machine Learning

Not Neural Networks

Neural Networks

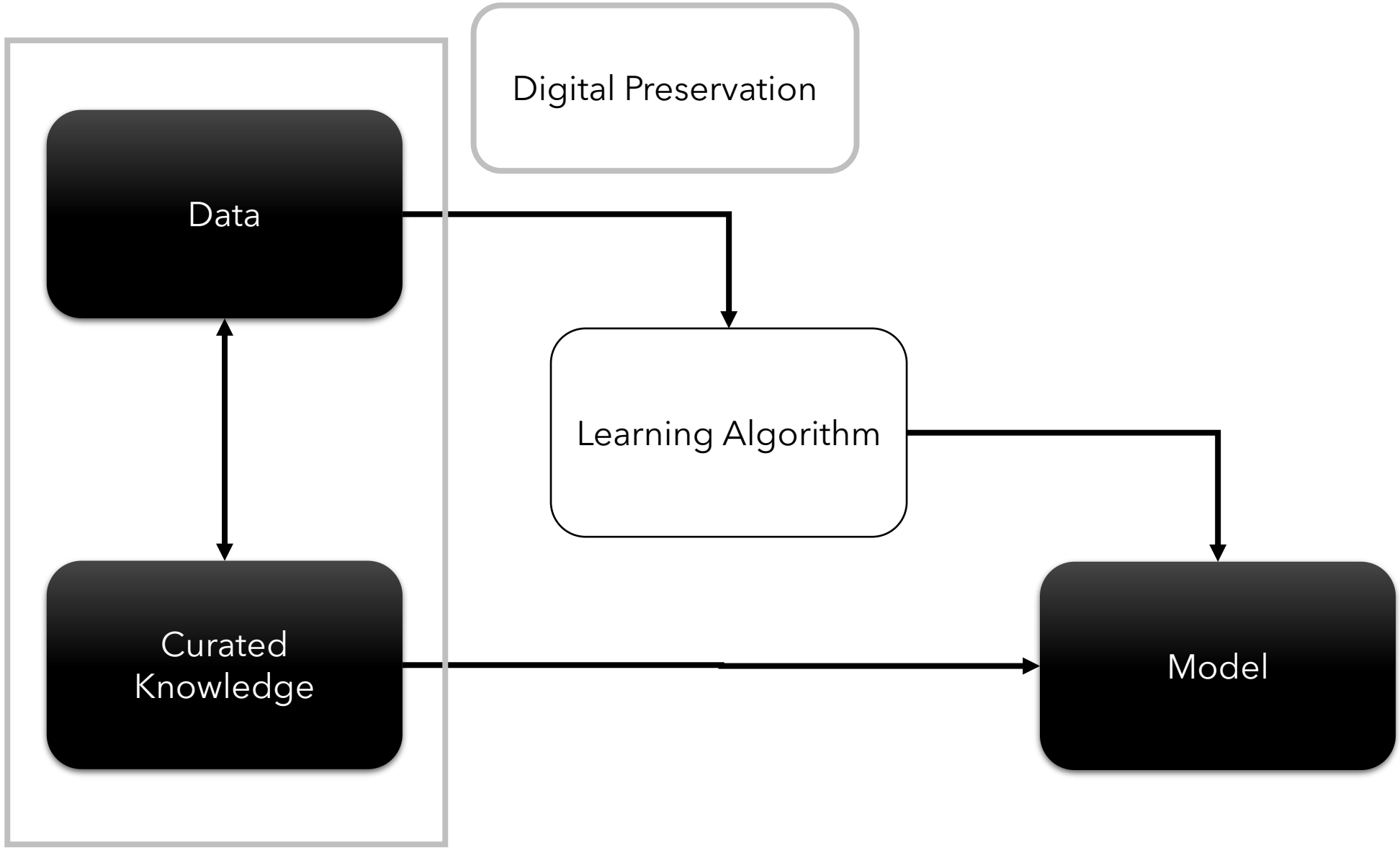
Deep Learning



An artificial neural network is an interconnected group of nodes, akin to the vast network of neurons in a brain. Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another.

ChatGPT

DALL-E 2


Digital Preservation
Data
Learning Algorithm
Curated Knowledge
Model

Questions

- AI is no longer a choice. It is everywhere. But a lot of AI learns from data or curated knowledge. How can digital preservation steer AI in the right direction?
- Increasingly generative AI will impact on born digital material, as new born digital content. It will influence records created by individuals.
- How will we contend with this content for archiving – especially when it is no longer possible to tell where AI stops and humans begin?

Quiz: which is written by an AI?

(Submit answers at menti.com 3603 5563)

Example 1

If tolling bell I ask the cause,
Its somber ring gives me pause.
Did someone pass from this life's stage,
Or does it signal impeding rage?

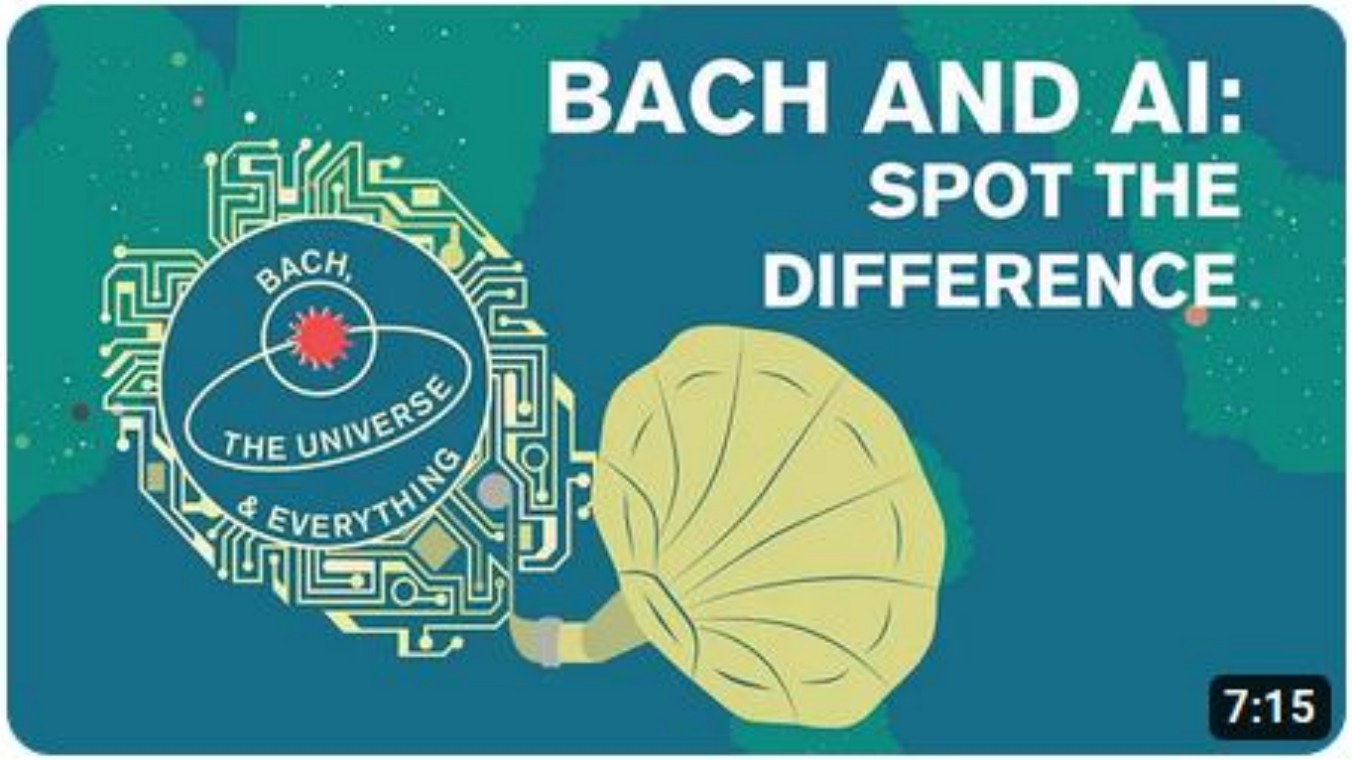
So, let the bell toll its voice to so clear,
For in its sound, we find what we hold dear.
Life, Love, and moments we share,
That in our hearts forever will bear.

Example 2

If tolling bell I ask the cause.
„A soul has gone to God,"
I'm answered in a lonesome tone;
Is heaven then so sad?

That bells should joyful ring to tell
A soul had gone to heaven,
Would seem to me the proper way
A good news should be given.

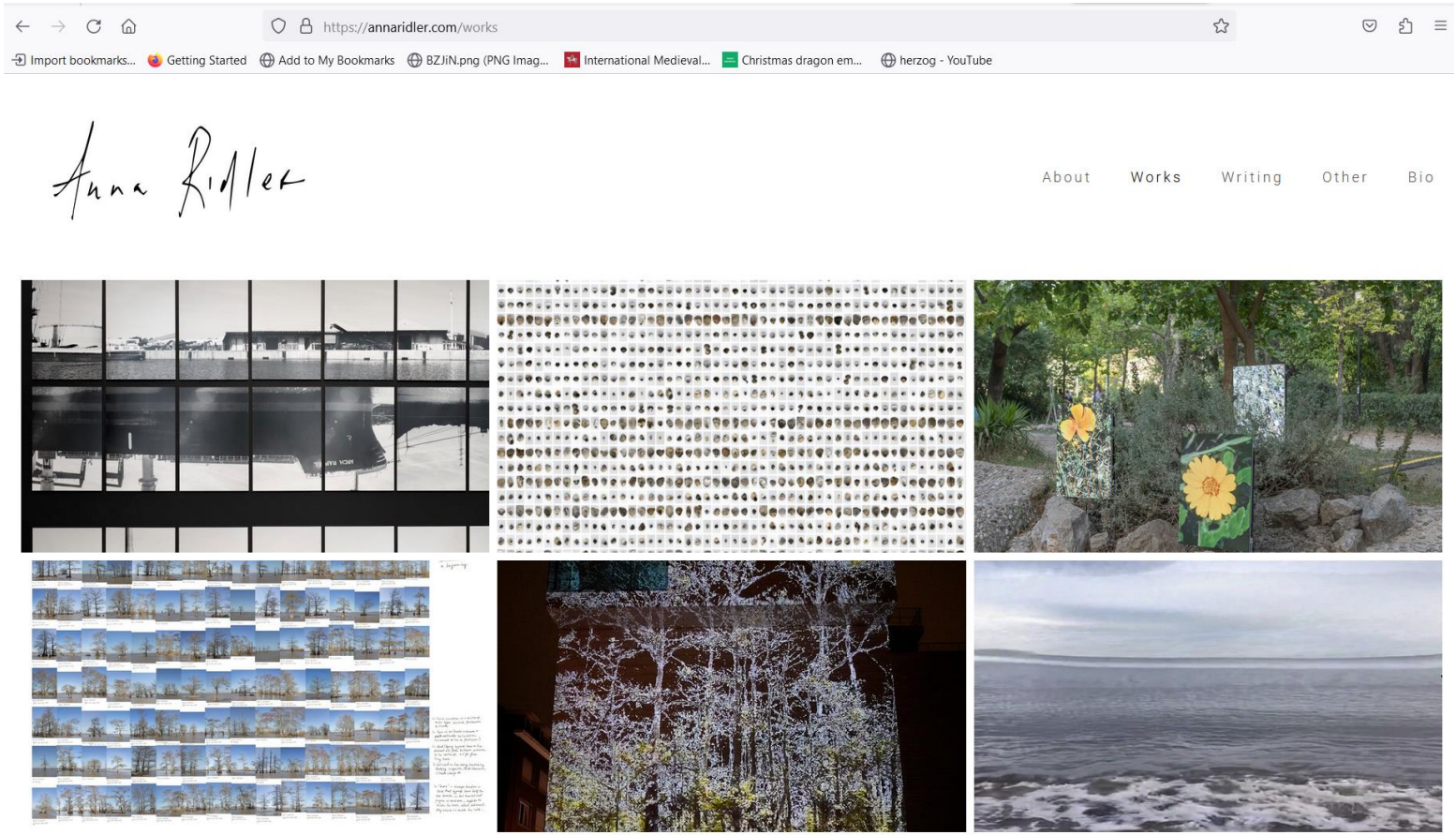
Bach or not Bach?



https://www.youtube.com/watch?v=lv9W7qrYhbk

AI and Art

The Shell Record is both a dataset of shells gathered by that artist from the foreshore of the river Thames in early 2021 and moving image pieces generated by using a GAN trained on this imagery. It explores **ideas around collecting, value and trade and links** to the history of shells being used as stores of value as one of the earliest forms of currency. It is also a work that is **a record of the shells of the Thames**, forever written into the blockchain, at this particular moment. A recent scientific paper shows that shells that have been in the river since the end of the last Ice Age have now become rare, replaced by other non-native species and that this will eventually become a time-marker for the Anthropocene as they become fossils in the strata.



Collaborations going forward: AI for the Arts and Humanities

Contribute to Increasing AI literacy

- Bias is not only about data but representation across the workflow.
- Some data are underrepresented, especially in cultural heritage.
- Archives, libraries, museums, gallery as part of AI education.

Help cultivate "presented code"

- Elaborating on context of the code
- Methodology (vs. method) matters!
- Code presented for the Arts and Humanities
- Unpack the process and interaction – creative, archival, data.

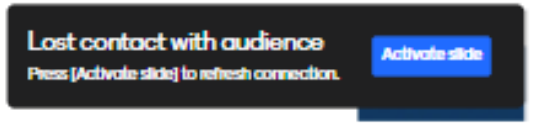
Toward collaborative research

- Calls such as UKRI Centres for Doctoral Training in AI.
- Need archives, libraries, museums and government org as partners.
- Escaping the disciplinary and sectoral silos and culture.

The way to get started is to quit talking and begin doing.
Walt Disney

Go to **www.menti.com** and use the code **6451 9275**

What do you use AI for? 2 Answers



tracking down stats (and references) to support research

metadata enhancement . transcription and translation

2

Thank you

Yunhyong Kim

yunhyong.kim@glasgow.ac.uk

https://www.gla.ac.uk/schools/humanities/staff/yunhyongkim/