AI and Data Science - Machine Learning as Digital Catalyst for Data Curation

KITAMOTO Asanobu

ROIS-DS Center for Open Data in the Humanities (CODH)

National Institute of Informatics

http://codh.rois.ac.jp/ @rois_codh

Humanities Data

ROIS-DS Center for Open Data in the Humanities (CODH)

2016 Pre-center started.

2017 Officially launched.

Member: One director and four project researchers (NII and ISM).

Direction 1: **Innovate humanities research** by computer science and statistical technologies and tools.

Direction 2: **Innovate non-humanities research** by data and questions from humanities.

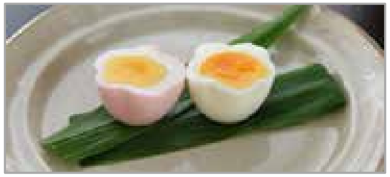
Our team consists of 1 professor, 4 post-docs, and 5 appointed professors.

CODH Datasets

http://codh.rois.ac.jp/dataset/

Dataset of Pre-Modern Japanese Text

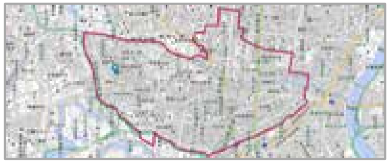
Kuzushiji Dataset



Dataset of Edo Cooking Recipes

Bukan Complete Collection

Collection of Facial Expressions



Dataset of Historical Administrative Boundaries

How to Access Humanities Data?

Humanities data are mainly textual data, but visual and spatial data requires metadata and annotation to enable **deep access** to content.

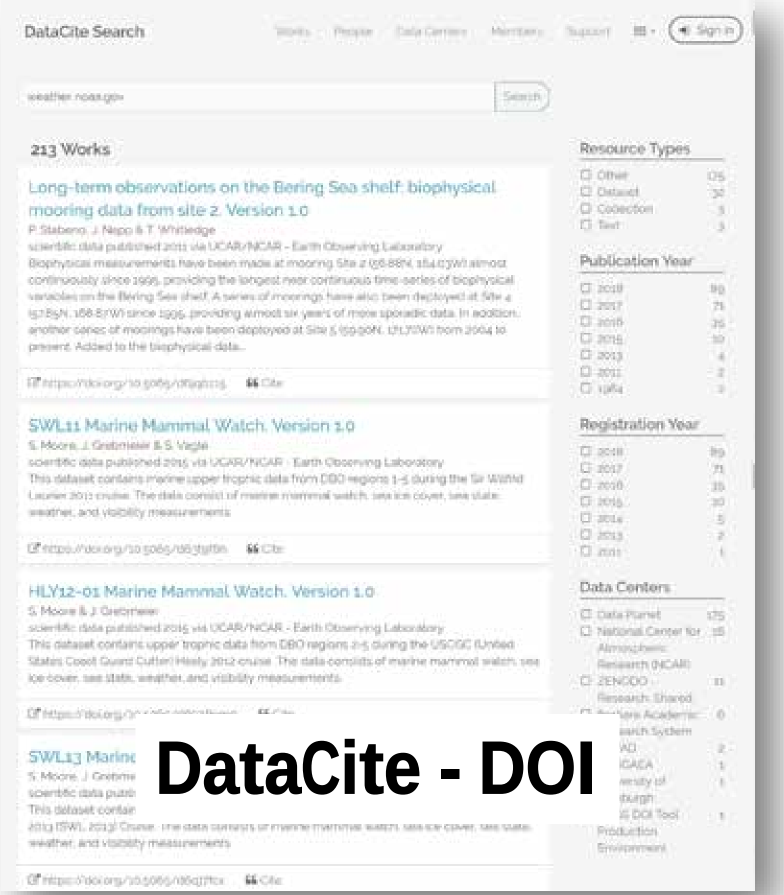
Images

Photographs

Maps

Characters

Knowledge Representation



DataCite Search

weather noaa.gov

213 Works

Long-term observations on the Bering Sea shelf: biophysical mooring data from site 2. Version 1.0

P. Stabeno, J. Napp & T. Whitledge

scientific data published 2011 via UCAR/NCAR - Earth Observing Laboratory

Biophysical measurements have been made at mooring Site 2 (56.88N, 164.03W) almost continuously since 1995, providing the longest near continuous time-series of biophysical variables on the Bering Sea shelf. A series of moorings have also been deployed at Site 4 (57.89N, 168.87W) since 1995, providing almost six years of more sporadic data. In addition, another series of moorings have been deployed at Site 5 (59.90N, 171.70W) from 2004 to present. Added to the biophysical data...

https://doi.org/10.5065/[illegible] Cite

SWL11 Marine Mammal Watch. Version 1.0

S. Moore, J. Grebmeier & S. Vagle

scientific data published 2015 via UCAR/NCAR - Earth Observing Laboratory

This dataset contains marine upper trophic data from DBO regions 1-5 during the Sir Wilfrid Laurier 2011 cruise. The data consist of marine mammal watch, sea ice cover, sea state, weather, and visibility measurements.

https://doi.org/10.5065/[illegible] Cite

HLY12-01 Marine Mammal Watch. Version 1.0

S. Moore & J. Grebmeier

scientific data published 2015 via UCAR/NCAR - Earth Observing Laboratory

This dataset contains upper trophic data from DBO regions 2-5 during the USCGC (United States Coast Guard Cutter) Healy 2012 cruise. The data consists of marine mammal watch, sea ice cover, sea state, weather, and visibility measurements.

https://doi.org/[illegible] Cite

SWL13 Marine [illegible]

S. Moore, J. Grebme[illegible]

scientific data publi[illegible]

This dataset contai[illegible]

2013 (SWL 2013) Cruise. The data consists of marine mammal watch, sea ice cover, sea state, weather, and visibility measurements.

https://doi.org/10.5065/[illegible] Cite

Resource Types

□ Other	175
□ Dataset	30
□ Collection	5
□ Text	3

Publication Year

□ 2018	89
□ 2017	71
□ 2016	35
□ 2015	10
□ 2013	4
□ 2011	2
□ 1964	2

Registration Year

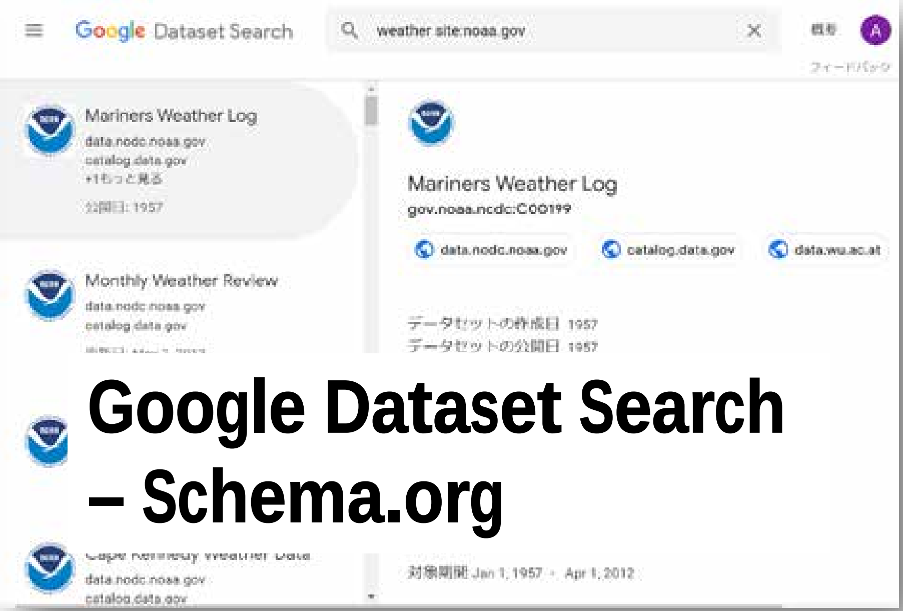
□ 2018	89
□ 2017	71
□ 2016	35
□ 2015	10
□ 2014	5
□ 2013	2
□ 2011	1

Data Centers

□ Data Planet	175
□ National Center for Atmospheric Research (NCAR)	15
□ ZENODO - Research. Shared.	11
□ [illegible]era Academic [illegible]earch System	6
[illegible]AD	2
[illegible]CACA	1
[illegible]ersity of [illegible]burgh	1
[illegible]S DOI Tool Production Environment	1

DataCite - DOI

https://search.datacite.org/



Google Dataset Search

weather site:noaa.gov

Mariners Weather Log
data.nodc.noaa.gov
catalog.data.gov
+1もっと見る
公開日: 1957

Monthly Weather Review
data.nodc.noaa.gov
catalog.data.gov

Cape Kennedy Weather Data
data.nodc.noaa.gov
catalog.data.gov

Mariners Weather Log
gov.noaa.ncdc:C00199

data.nodc.noaa.gov catalog.data.gov data.wu.ac.at

データセットの作成日 1957
データセットの公開日 1957

対象期間 Jan 1, 1957 – Apr 1, 2012

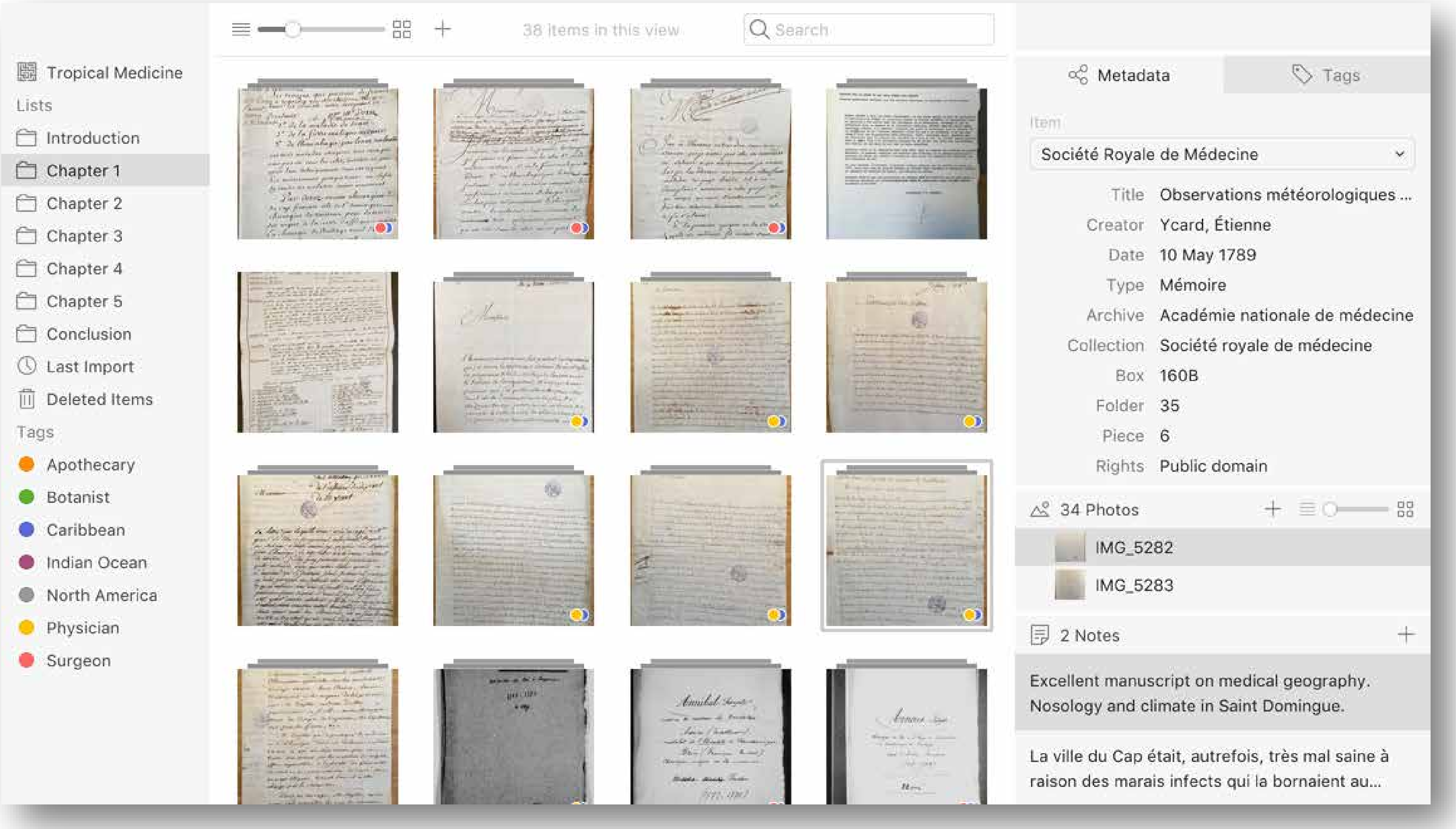
Google Dataset Search – Schema.org

https://toolbox.google.com/datasetsearch

Interoperable metadata and Semantic Web can increase findability.

Manual Image Annotation

https://tropy.org/



Société Royale de Médecine

Title	Observations météorologiques ...
Creator	Ycard, Étienne
Date	10 May 1789
Type	Mémoire
Archive	Académie nationale de médecine
Collection	Société royale de médecine
Box	160B
Folder	35
Piece	6
Rights	Public domain

34 Photos

IMG_5282

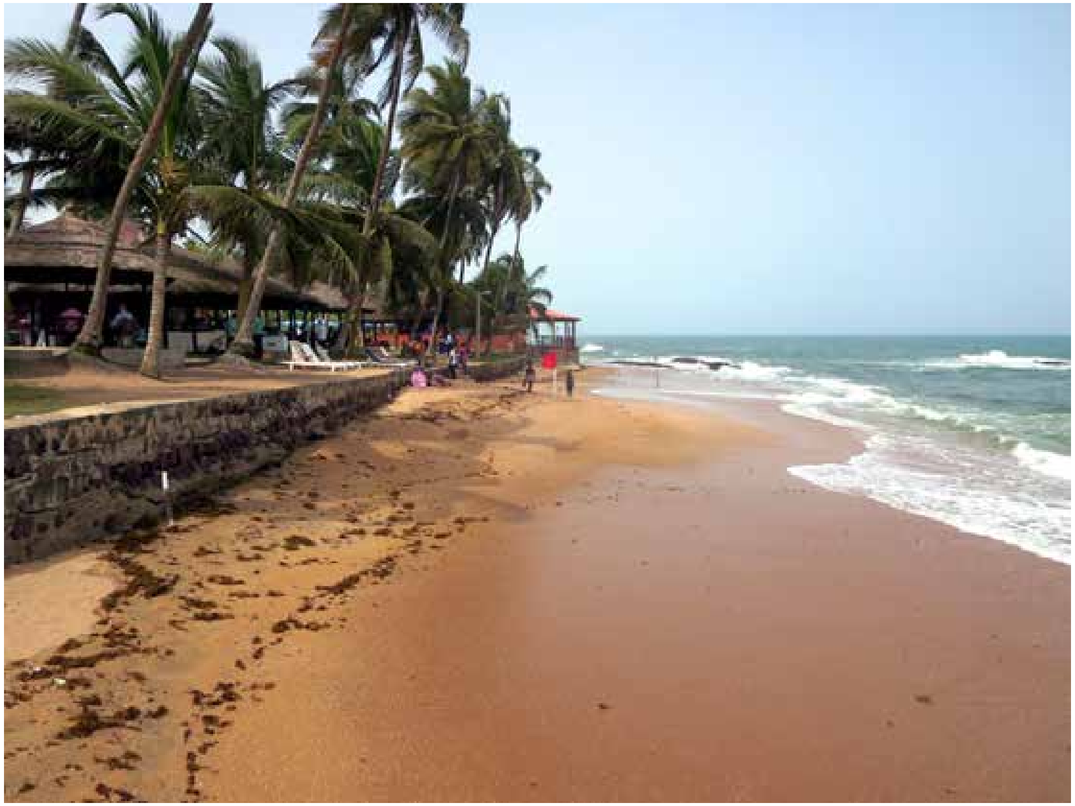
IMG_5283

2 Notes

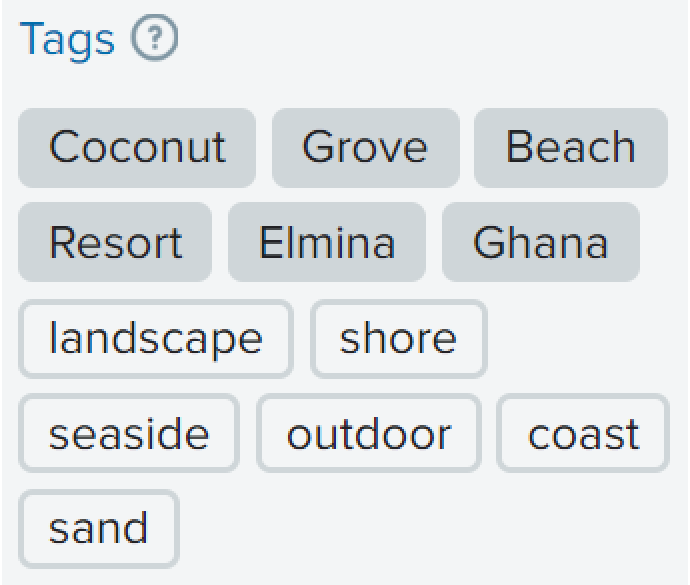
Excellent manuscript on medical geography. Nosology and climate in Saint Domingue.

La ville du Cap était, autrefois, très mal saine à raison des marais infects qui la bornaient au...

Machine Learning (ML)



Credit: David Stanley, https://www.flickr.com/photos/davidstanleytravel/



Tags

Coconut Grove Beach

Resort Elmina Ghana

landscape shore

seaside outdoor coast

sand

Gray: Human annotated tags, White: Machine annotated tags.

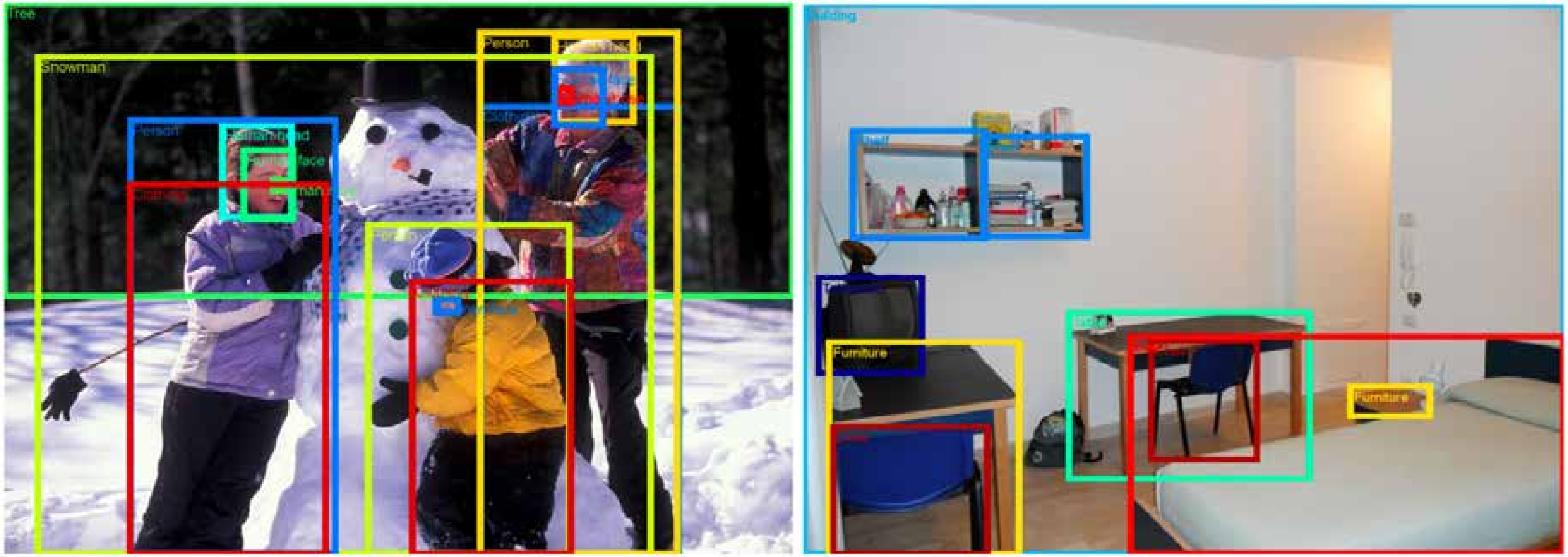
Machine Learning for Photographic Database

Best Practices for AI-assisted Data Curation

1. What could be done by AI, and not by AI? Hype and criticism should be corrected.
2. Machine learning: especially effective for learning patterns from image data.
3. Images, especially photographs: selected as the initial target of the work.
4. General numerical datasets: content-based access is still a challenge.

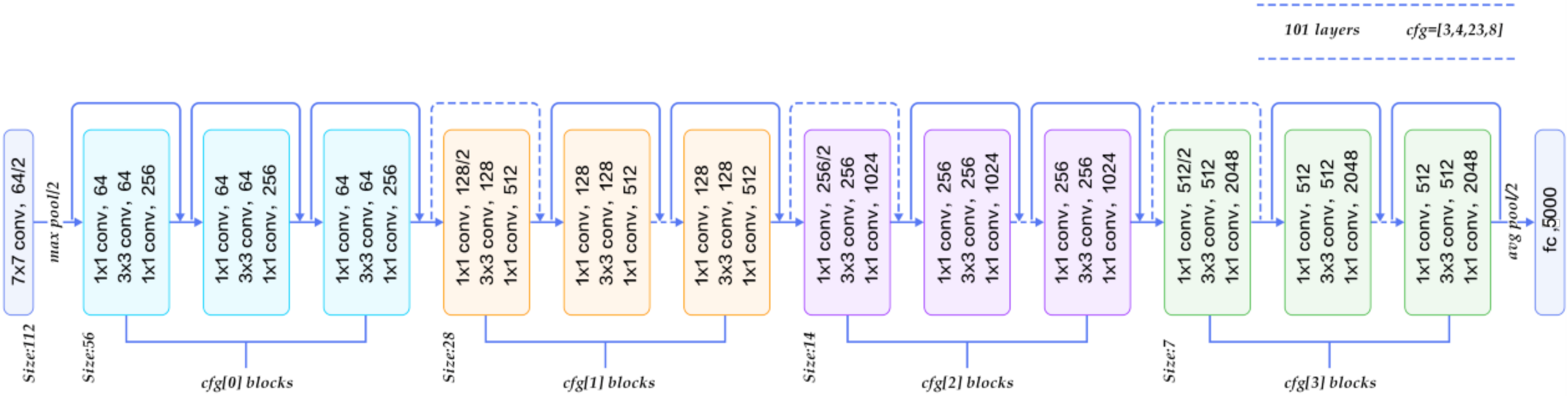
Open Images Dataset V2

https://github.com/openimages/dataset/blob/master/READMEV2.md



Annotated images from the Open Images dataset. Left: FAMILY MAKING A SNOWMAN by mwvchamber. Right: STANZA STUDENTI.S.S. ANNUNZIATA by ersupalermo. Both images used under CC BY 2.0 license.

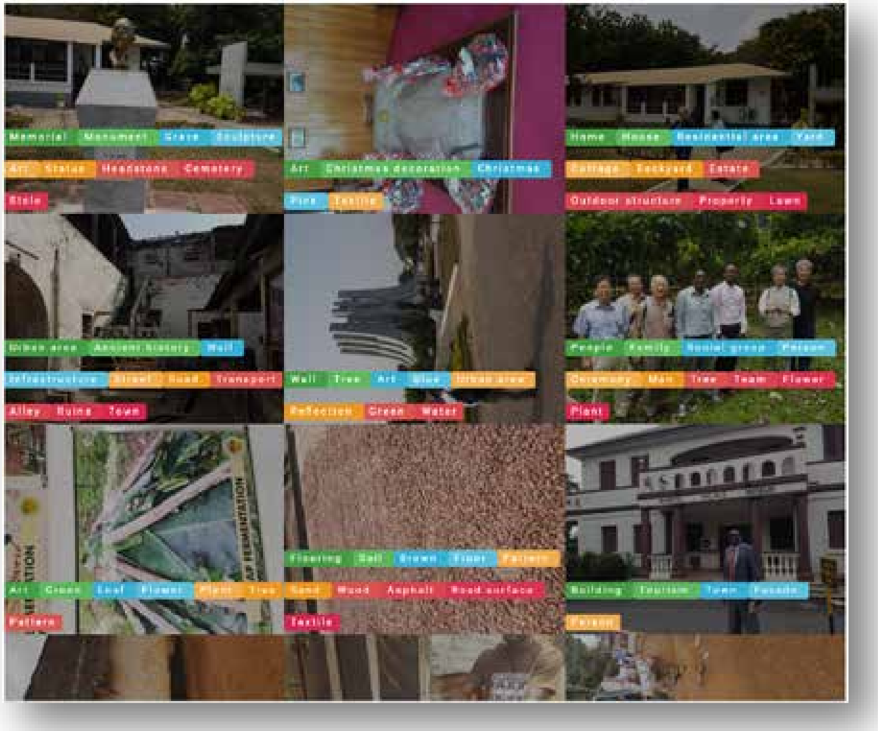
Deep Learning Model



https://medium.com/@siddharthdas_32104/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5

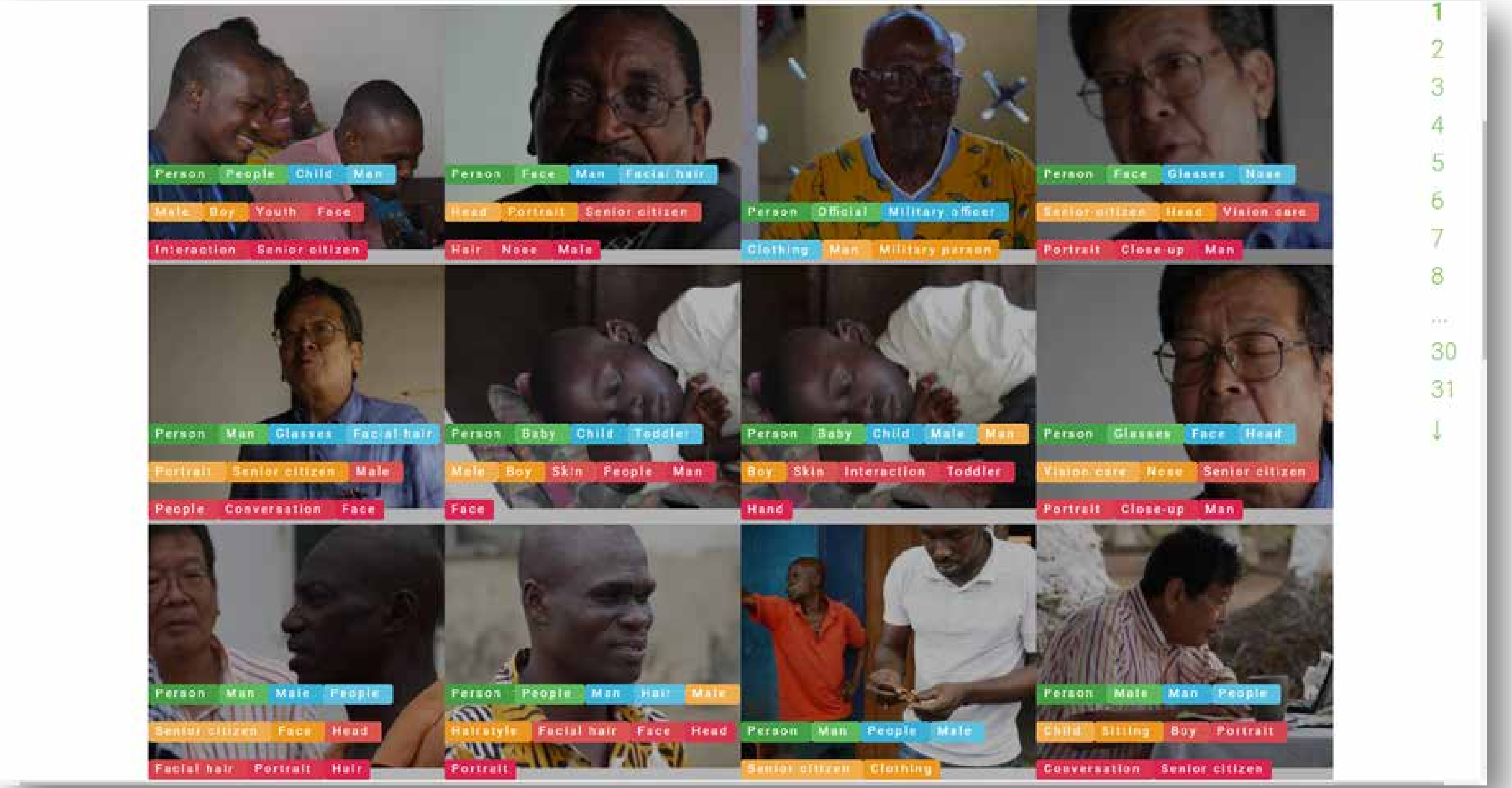
1. ResNet 101 classifier learns 5000 tags from 9 million images (Open Images Dataset V2).
2. We used the model already trained on general photographs, not on our dataset.

Case 1: Ethnology Field Work


Memorial Monument Grave Sculpture
Art Statue Headstone Cemetery
Stele
Art Christmas decoration Christmas
Pink Textile
Home House Residential area Yard
Cottage Backyard Estate
Outdoor structure Property Lawn
Urban area Ancient history Wall
Infrastructure Street Road Transport
Alley Ruins Town
Wall Tree Art Blue Urban area
Reflection Green Water
People Family Social group Person
Ceremony Man Tree Team Flower
Plant
Art Green Leaf Flower Plant Tree
Pattern
Flooring Soil Brown Floor Pattern
Sand Wood Asphalt Road surface
Textile
Building Tourism Town Facade
Person

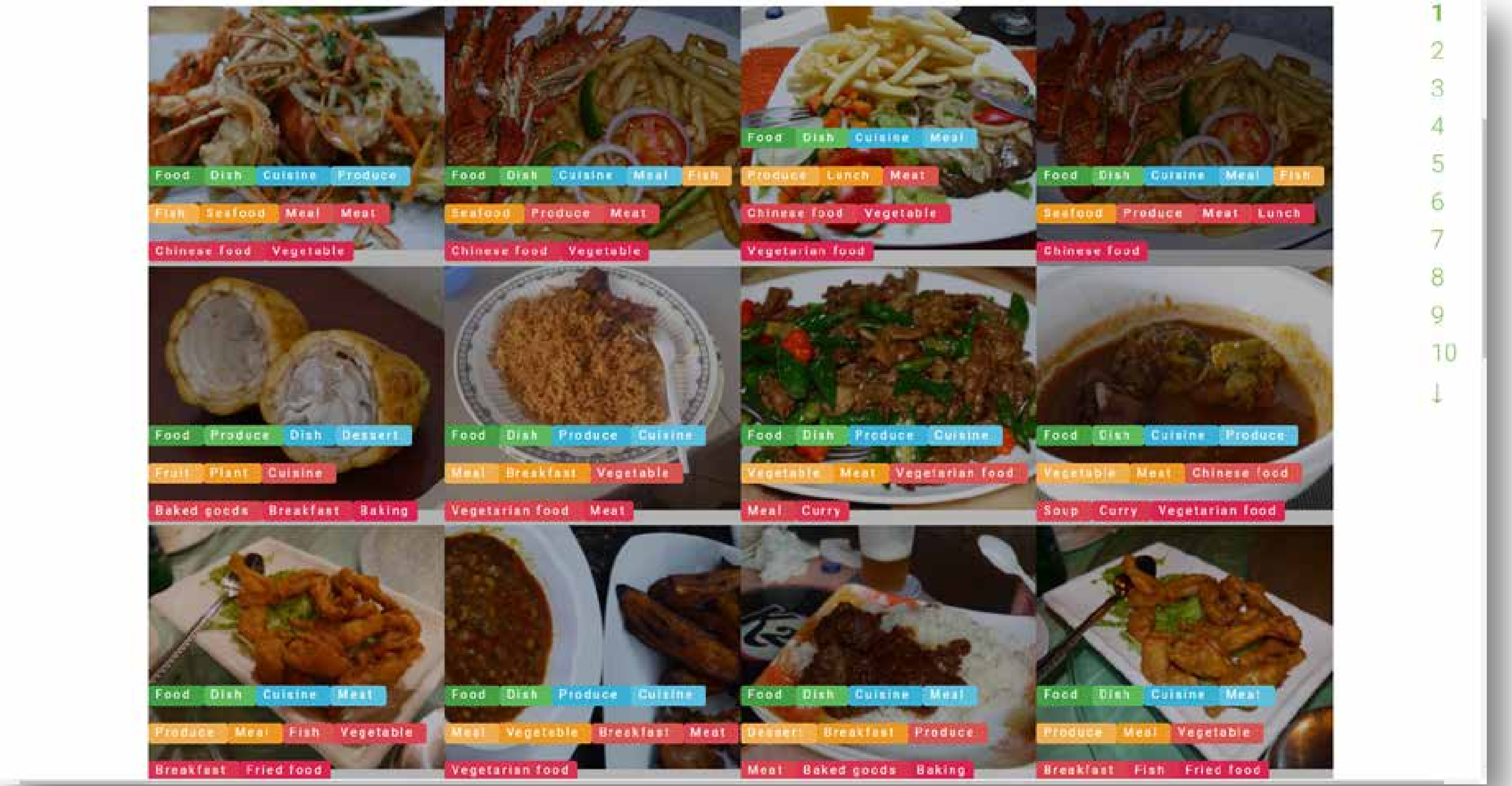
1. Field work in Ghana, in August 2017.
2. About 3,700 photographs, yet to be released to the public.
3. Collaboration with National Museum of Ethnology (Prof. Yoshida, Prof. Iida and others).

Tag: Person



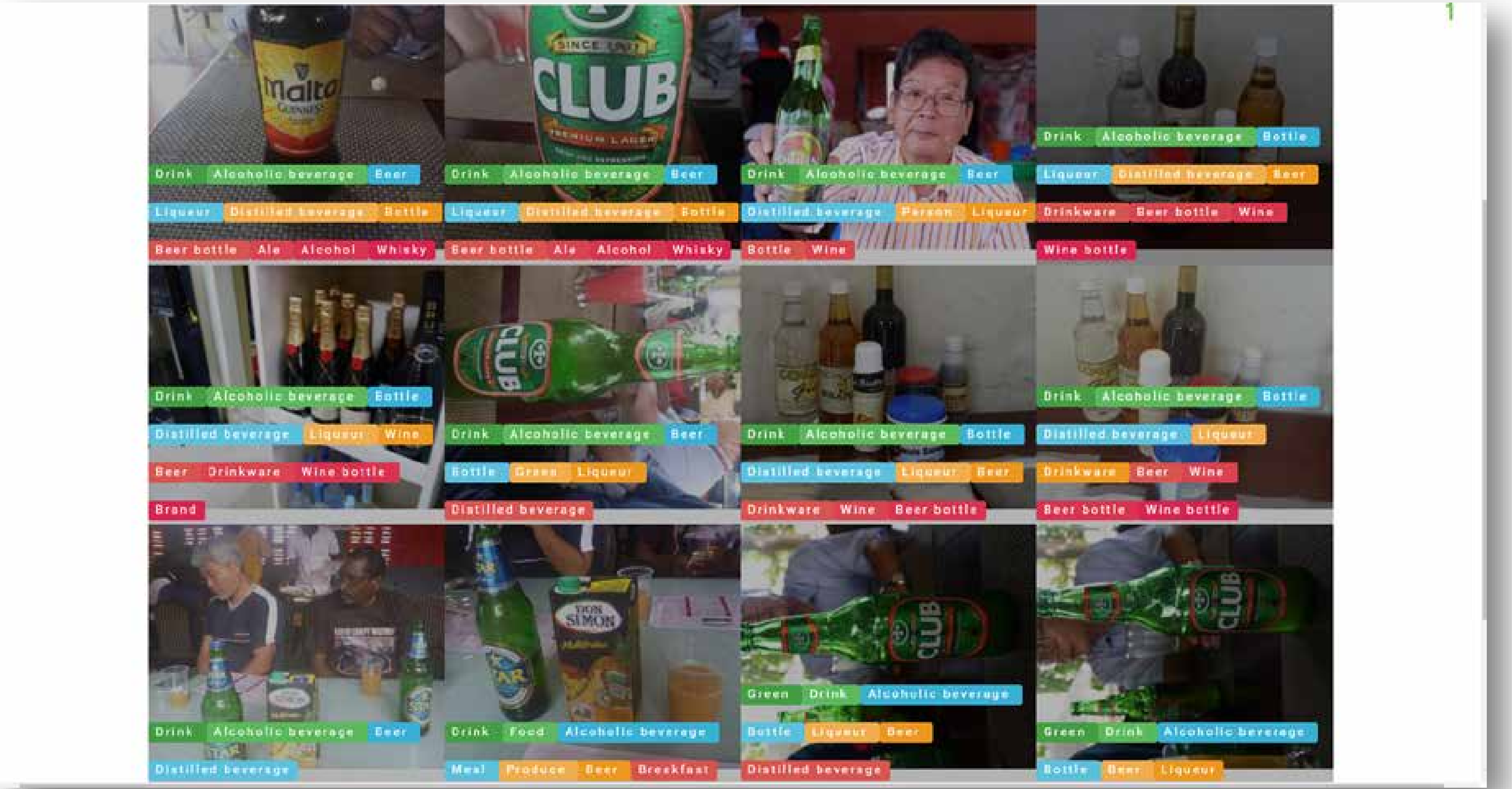
Person People Child Man
Male Boy Youth Face
Interaction Senior citizen
Person Face Man Facial hair
Head Portrait Senior citizen
Hair Nose Male
Person Official Military officer
Clothing Man Military person
Person Face Glasses Nose
Senior citizen Head Vision care
Portrait Close-up Man
Person Man Glasses Facial hair
Portrait Senior citizen Male
People Conversation Face
Person Baby Child Toddler
Male Boy Skin People Man
Face
Person Baby Child Male Man
Boy Skin Interaction Toddler
Hand
Person Glasses Face Head
Vision care Nose Senior citizen
Portrait Close-up Man
Person Man Male People
Senior citizen Face Head
Facial hair Portrait Hair
Person People Man Hair Male
Hairstyle Facial hair Face Head
Portrait
Person Man People Male
Senior citizen Clothing
Person Male Man People
Child Sitting Boy Portrait
Conversation Senior citizen
1
2
3
4
5
6
7
8
...
30
31
↓

Tag: Food

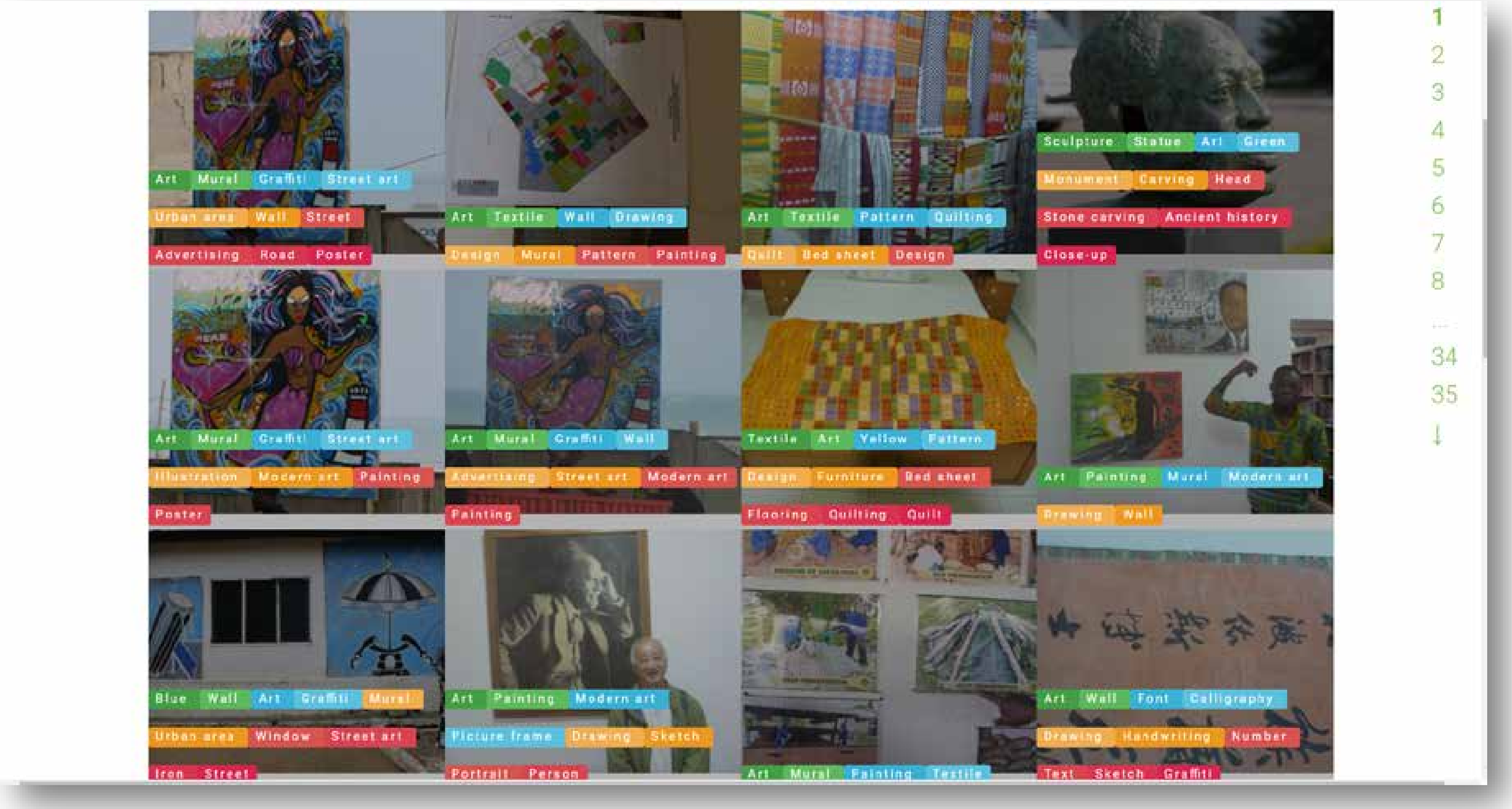


Food Dish Cuisine Produce
Fish Seafood Meal Meat
Chinese food Vegetable
Food Dish Cuisine Meal Fish
Seafood Produce Meat
Chinese food Vegetable
Food Dish Cuisine Meal
Produce Lunch Meat
Chinese food Vegetable
Vegetarian food
Food Dish Cuisine Meal Fish
Seafood Produce Meat Lunch
Chinese food
Food Produce Dish Dessert
Fruit Plant Cuisine
Baked goods Breakfast Baking
Food Dish Produce Cuisine
Meal Breakfast Vegetable
Vegetarian food Meat
Food Dish Produce Cuisine
Vegetable Meat Vegetarian food
Meal Curry
Food Dish Cuisine Produce
Vegetable Meat Chinese food
Soup Curry Vegetarian food
Food Dish Cuisine Meat
Produce Meal Fish Vegetable
Breakfast Fried food
Food Dish Produce Cuisine
Meal Vegetable Breakfast Meat
Vegetarian food
Food Dish Cuisine Meal
Dessert Breakfast Produce
Meat Baked goods Baking
Food Dish Cuisine Meat
Produce Meal Vegetable
Breakfast Fish Fried food
1
2
3
4
5
6
7
8
9
10
↓

Tag: Beer

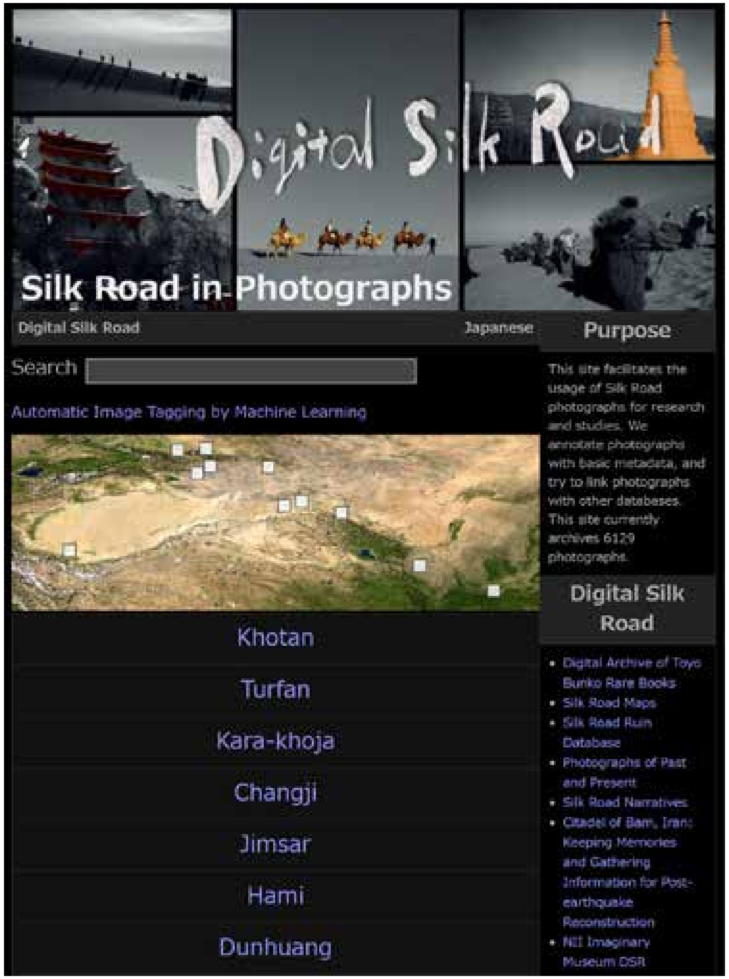

1
Drink Alcoholic beverage Beer
Liqueur Distilled beverage Bottle
Beer bottle Ale Alcohol Whisky
Drink Alcoholic beverage Beer
Liqueur Distilled beverage Bottle
Beer bottle Ale Alcohol Whisky
Drink Alcoholic beverage Beer
Distilled beverage Person Liqueur
Bottle Wine
Drink Alcoholic beverage Bottle
Liqueur Distilled beverage Beer
Drinkware Beer bottle Wine
Wine bottle
Drink Alcoholic beverage Bottle
Distilled beverage Liqueur Wine
Beer Drinkware Wine bottle
Brand
Drink Alcoholic beverage Beer
Bottle Green Liqueur
Distilled beverage
Drink Alcoholic beverage Bottle
Distilled beverage Liqueur Beer
Drinkware Wine Beer bottle
Drink Alcoholic beverage Bottle
Distilled beverage Liqueur
Drinkware Beer Wine
Beer bottle Wine bottle
Drink Alcoholic beverage Beer
Distilled beverage
Drink Food Alcoholic beverage
Meal Produce Beer Breakfast
Green Drink Alcoholic beverage
Bottle Liqueur Beer
Distilled beverage
Green Drink Alcoholic beverage
Bottle Beer Liqueur

Tag: Art


Art Mural Graffiti Street art
Urban area Wall Street
Advertising Road Poster
Art Textile Wall Drawing
Design Mural Pattern Painting
Art Textile Pattern Quilting
Quilt Bed sheet Design
Sculpture Statue Art Green
Monument Carving Head
Stone carving Ancient history
Close-up
Art Mural Graffiti Street art
Illustration Modern art Painting
Poster
Art Mural Graffiti Wall
Advertising Street art Modern art
Painting
Textile Art Yellow Pattern
Design Furniture Bed sheet
Flooring Quilting Quilt
Art Painting Mural Modern art
Drawing Wall
Blue Wall Art Graffiti Mural
Urban area Window Street art
Iron Street
Art Painting Modern art
Picture frame Drawing Sketch
Portrait Person
Art Mural Painting Textile
Art Wall Font Calligraphy
Drawing Handwriting Number
Text Sketch Graffiti
1
2
3
4
5
6
7
8
...
34
35
↓

Case 2: Archaeology Field Work

http://dsr.nii.ac.jp/photograph/



Purpose

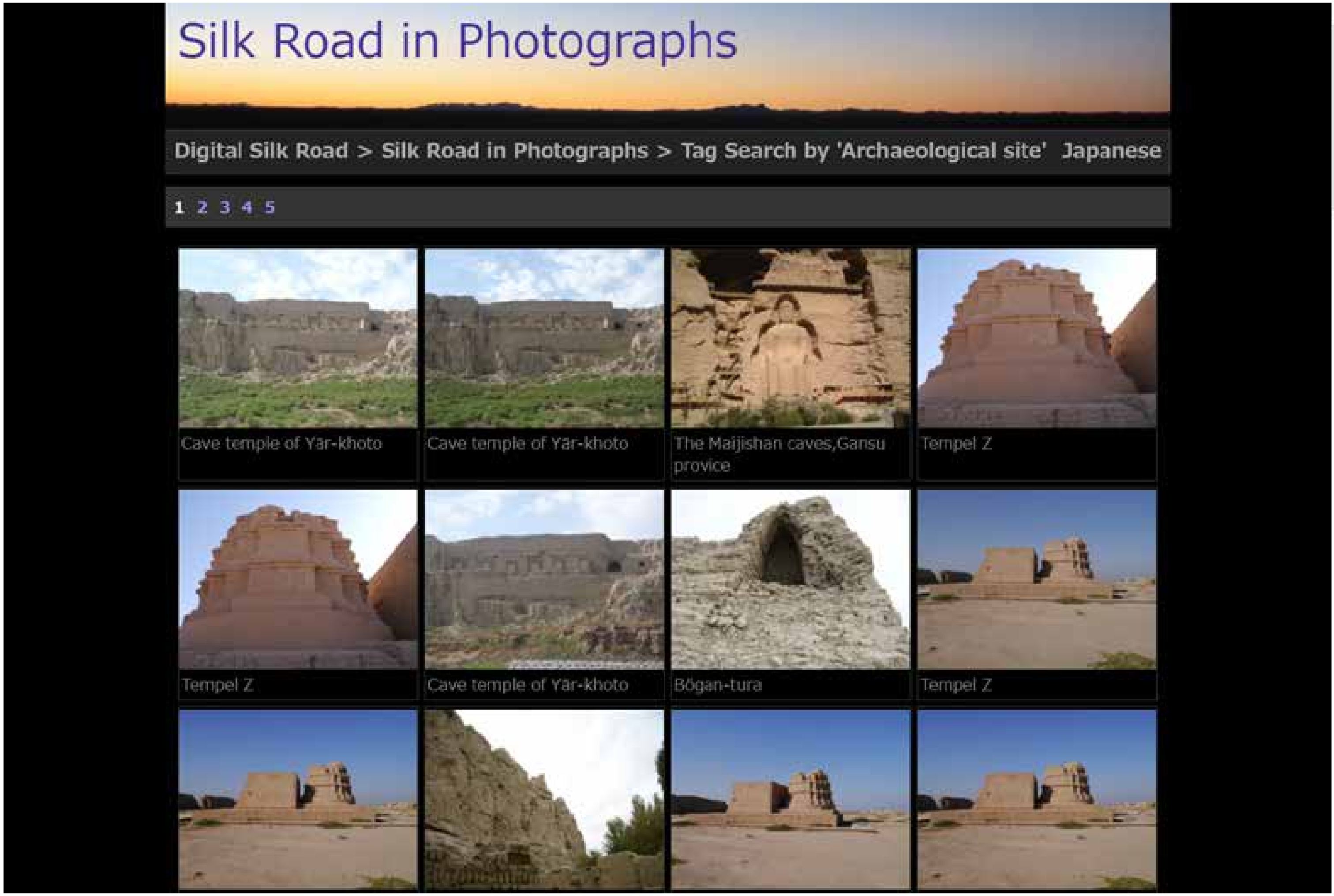
This site facilitates the usage of Silk Road photographs for research and studies. We annotate photographs with basic metadata, and try to link photographs with other databases. This site currently archives 6129 photographs.

Digital Silk Road

- Digital Archive of Toyo Bunko Rare Books
- Silk Road Maps
- Silk Road Ruin Database
- Photographs of Past and Present
- Silk Road Narratives
- Citadel of Bam, Iran: Keeping Memories and Gathering Information for Post-earthquake Reconstruction
- NII Imaginary Museum DSR

1. Photographs of the Silk Road, mainly about old ruins.
2. 6,129 photographs across long time span.
3. Many photographs were taken by Dr. Nishimura in Toyo University.

Tag: Archaeological Site



Silk Road in Photographs

Digital Silk Road > Silk Road in Photographs > Tag Search by 'Archaeological site' Japanese

1 2 3 4 5

Cave temple of Yār-khoto

Cave temple of Yār-khoto

The Maijishan caves,Gansu provice

Tempel Z

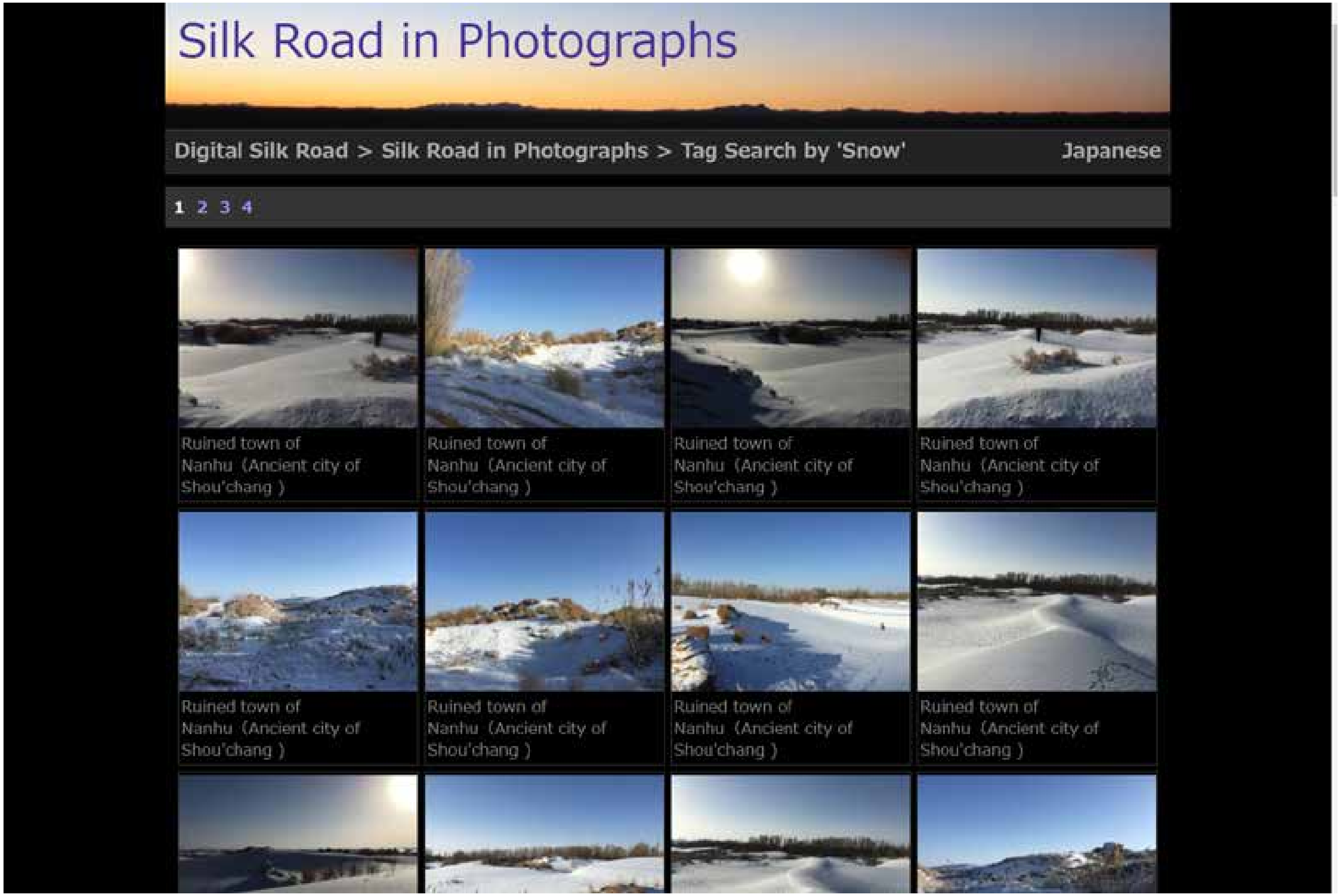
Tempel Z

Cave temple of Yār-khoto

Bögan-tura

Tempel Z

Tag: Snow



Silk Road in Photographs

Digital Silk Road > Silk Road in Photographs > Tag Search by 'Snow'

Ruined town of Nanhu（Ancient city of Shou'chang ）

Ruined town of Nanhu（Ancient city of Shou'chang ）

Ruined town of Nanhu（Ancient city of Shou'chang ）

Ruined town of Nanhu（Ancient city of Shou'chang ）

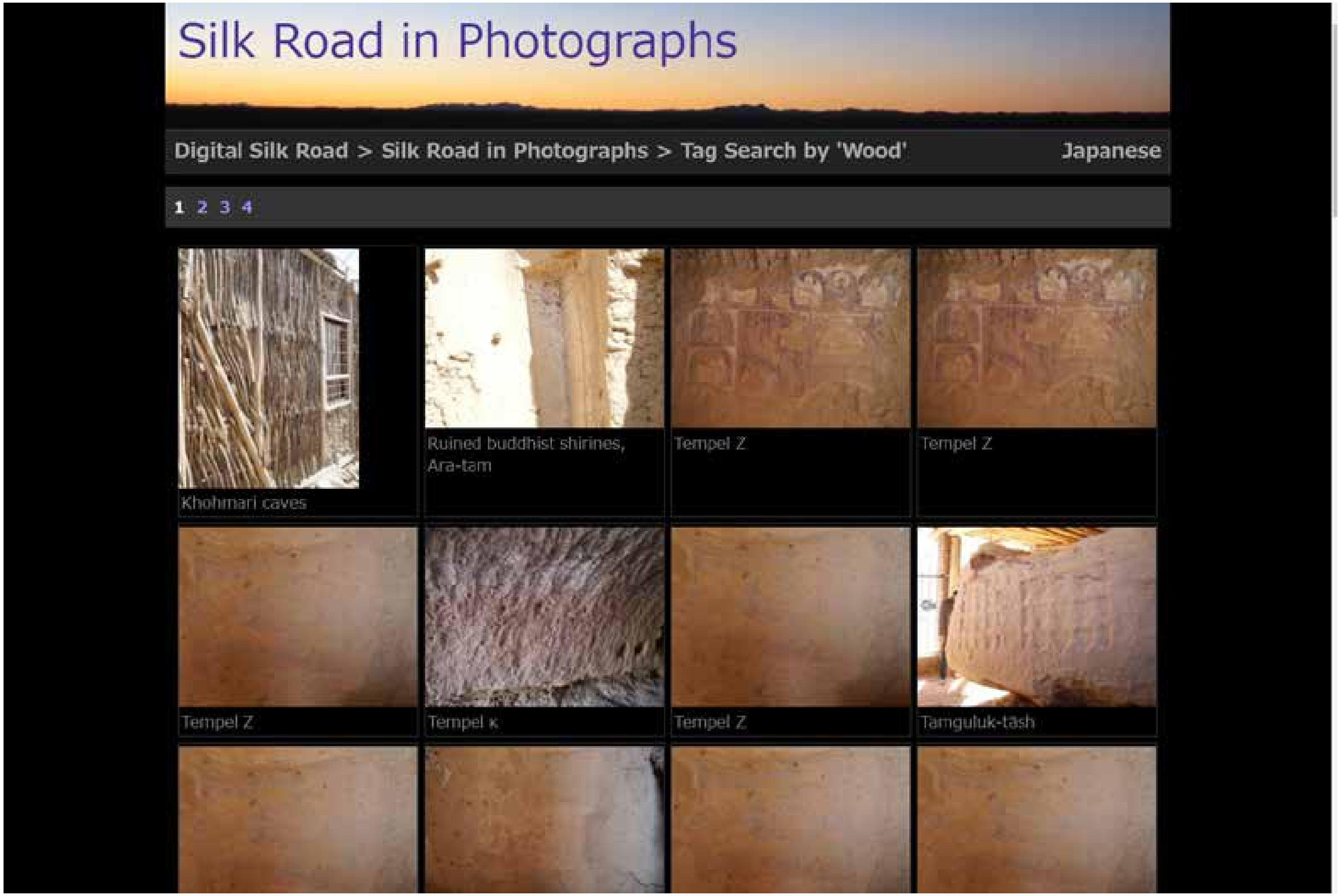
Ruined town of Nanhu（Ancient city of Shou'chang ）

Ruined town of Nanhu（Ancient city of Shou'chang ）

Ruined town of Nanhu（Ancient city of Shou'chang ）

Ruined town of Nanhu（Ancient city of Shou'chang ）

Tag: Wood



Silk Road in Photographs

Digital Silk Road > Silk Road in Photographs > Tag Search by 'Wood' Japanese

1 2 3 4

Khohmari caves

Ruined buddhist shirines, Ara-tam

Tempel Z

Tempel Z

Tempel Z

Tempel κ

Tempel Z

Tamguluk-tāsh

Different Responses from Users

Ethnology Field Work	**Archaeology Field Work**
Image tagging has great potential for grouping photographs by theme.	Grouping by machine-generated tags is less useful than grouping by entity names.
Even if the tag is not correct, it gives some hints about the content.	Some tags are simply wrong due to different training images and domains.

Why Different Reponses?

General Noun Metadata

Ethnology photographs are so diverse that tags of general nouns are effective for grouping.

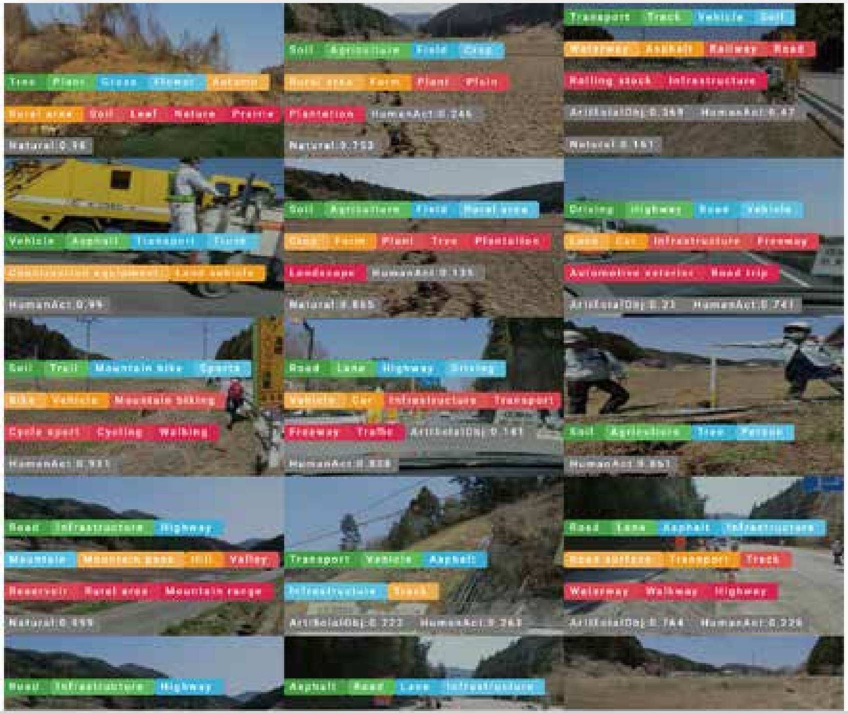
It motivates experts to describe deeper metadata.

Proper Noun Metadata

Archaeology photographs are usually taken with intentions.

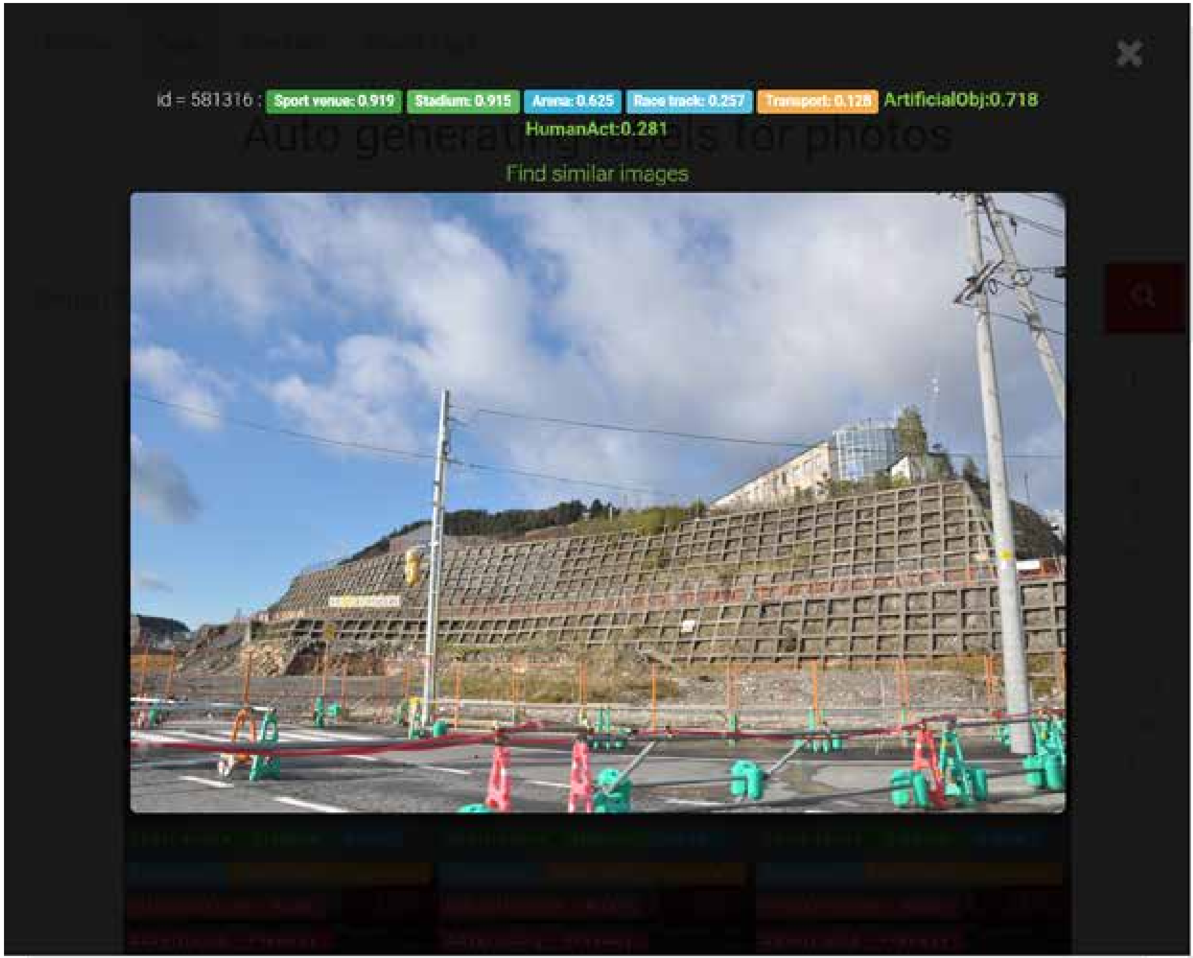
Entity names are difficult to identify by machine learning.

Case 3: Post-Disaster Survey


Tree Plant Grass Flower Autumn
Rural area Soil Leaf Nature Prairie
Natural:0.96
Soil Agriculture Field Crop
Rural area Farm Plant Plain
Plantation HumanAct:0.246
Natural:0.753
Transport Track Vehicle Soil
Waterway Asphalt Railway Road
Rolling stock Infrastructure
ArtificialObj:0.369 HumanAct:0.47
Natural:0.161
Vehicle Asphalt Transport Truck
Construction equipment Land vehicle
HumanAct:0.95
Soil Agriculture Field Rural area
Crop Farm Plant Tree Plantation
Landscape HumanAct:0.135
Natural:0.865
Driving Highway Road Vehicle
Lane Car Infrastructure Freeway
Automotive exterior Road trip
ArtificialObj:0.21 HumanAct:0.741
Soil Trail Mountain bike Sports
Bike Vehicle Mountain biking
Cycle sport Cycling Walking
HumanAct:0.931
Road Lane Highway Driving
Vehicle Car Infrastructure Transport
Freeway Traffic ArtificialObj:0.161
HumanAct:0.838
Soil Agriculture Tree Person
HumanAct:0.861
Road Infrastructure Highway
Mountain Mountain pass Hill Valley
Reservoir Rural area Mountain range
Natural:0.999
Transport Vehicle Asphalt
Infrastructure Track
ArtificialObj:0.722 HumanAct:0.263
Road Lane Asphalt Infrastructure
Road surface Transport Track
Waterway Walkway Highway
ArtificialObj:0.764 HumanAct:0.226
Road Infrastructure Highway
Asphalt Road Lane Infrastructure

1. Photographs of East Japan Earthquake 2011 and Kumamoto Earthquake 2016.
2. More than 10,000 photographs, yet to be released to the public.
3. Collaboration with National Research Institute for Earth Science and Disaster Resilience (NIED).

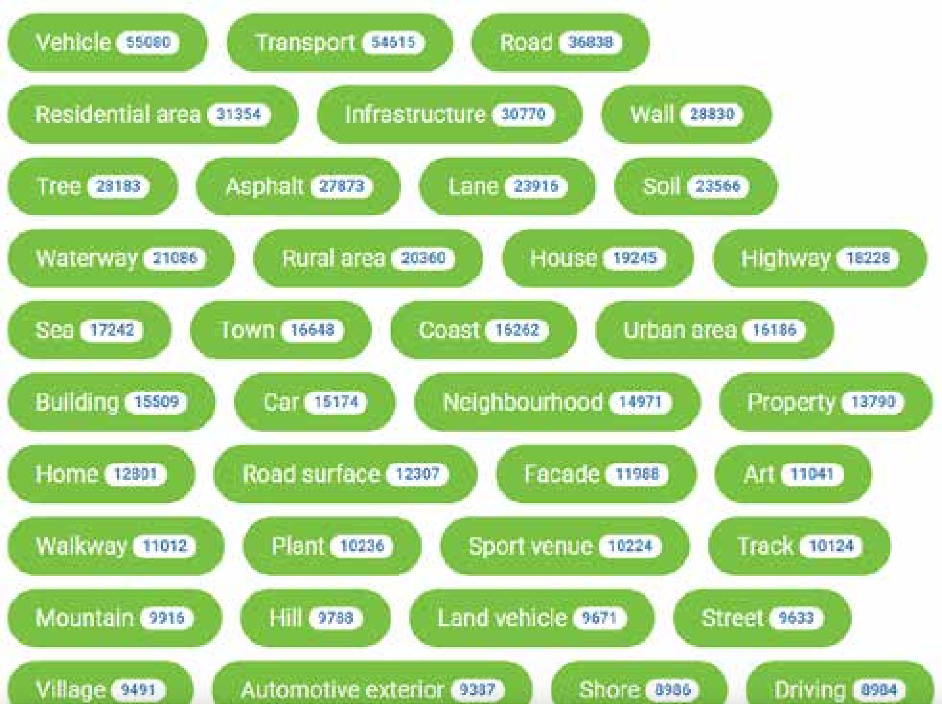
Serendipity
Tag: Stadium



id = 581316 : Sport venue: 0.919 Stadium: 0.915 Arena: 0.625 Race track: 0.257 Transport: 0.128 ArtificialObj:0.718 HumanAct:0.281

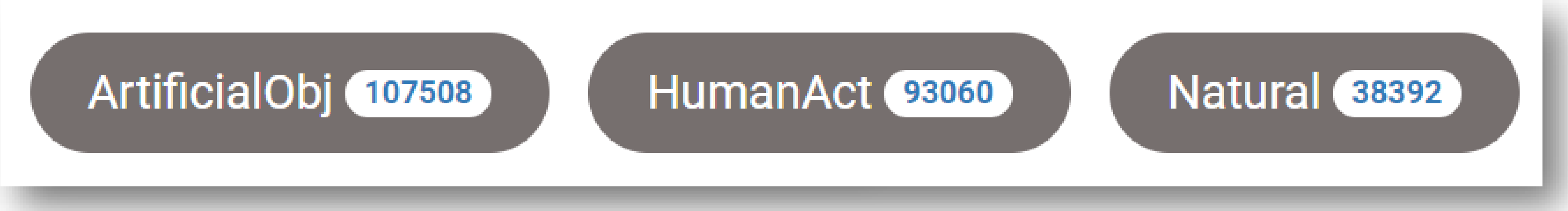
Find similar images

Semantic Grouping of Low-Level Tags



Vehicle 55080 · Transport 54615 · Road 36838
Residential area 31354 · Infrastructure 30770 · Wall 28830
Tree 28183 · Asphalt 27873 · Lane 23916 · Soil 23566
Waterway 21086 · Rural area 20360 · House 19245 · Highway 18228
Sea 17242 · Town 16648 · Coast 16262 · Urban area 16186
Building 15509 · Car 15174 · Neighbourhood 14971 · Property 13790
Home 12801 · Road surface 12307 · Facade 11988 · Art 11041
Walkway 11012 · Plant 10236 · Sport venue 10224 · Track 10124
Mountain 9916 · Hill 9788 · Land vehicle 9671 · Street 9633
Village 9491 · Automotive exterior 9387 · Shore 8986 · Driving 8904

Domain experts need higher-level semantic grouping of low-level tags.



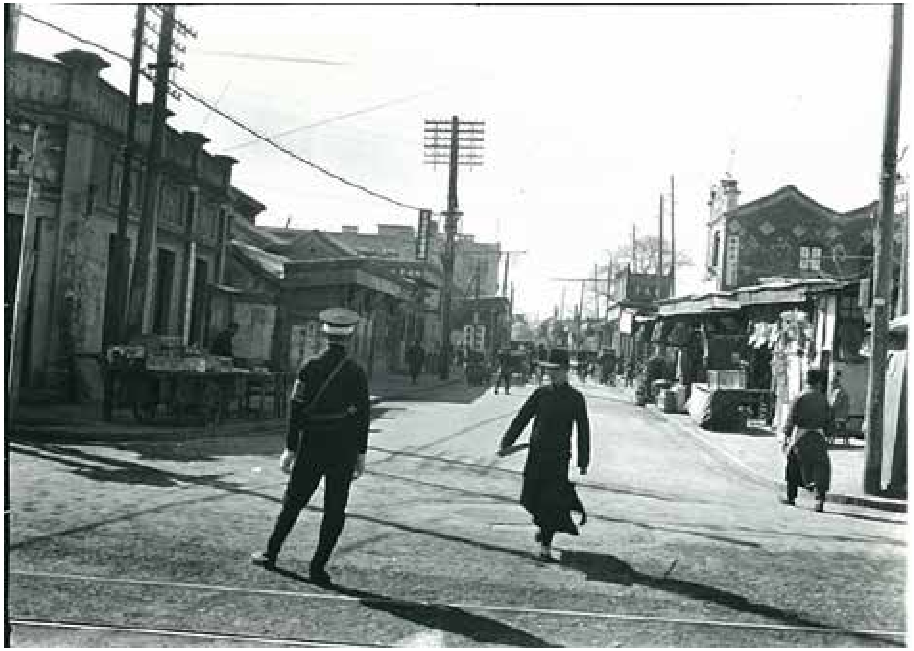
ArtificialObj 107508 · HumanAct 93060 · Natural 38392

Case 4: Historical Photographs

http://codh.rois.ac.jp/north-china-railway/

1. Photographs of North China Railway, a company existed around 1940.
2. More than 35,000 photographs will be released in Feb. 2019.
3. Collaboration with Kyoto University.

Image Tagging



Road, Street, Black-and-white, Monochrome photography, Monochrome, Infrastructure, Transport, Lane, Vehicle, Photograph

Image Colorization

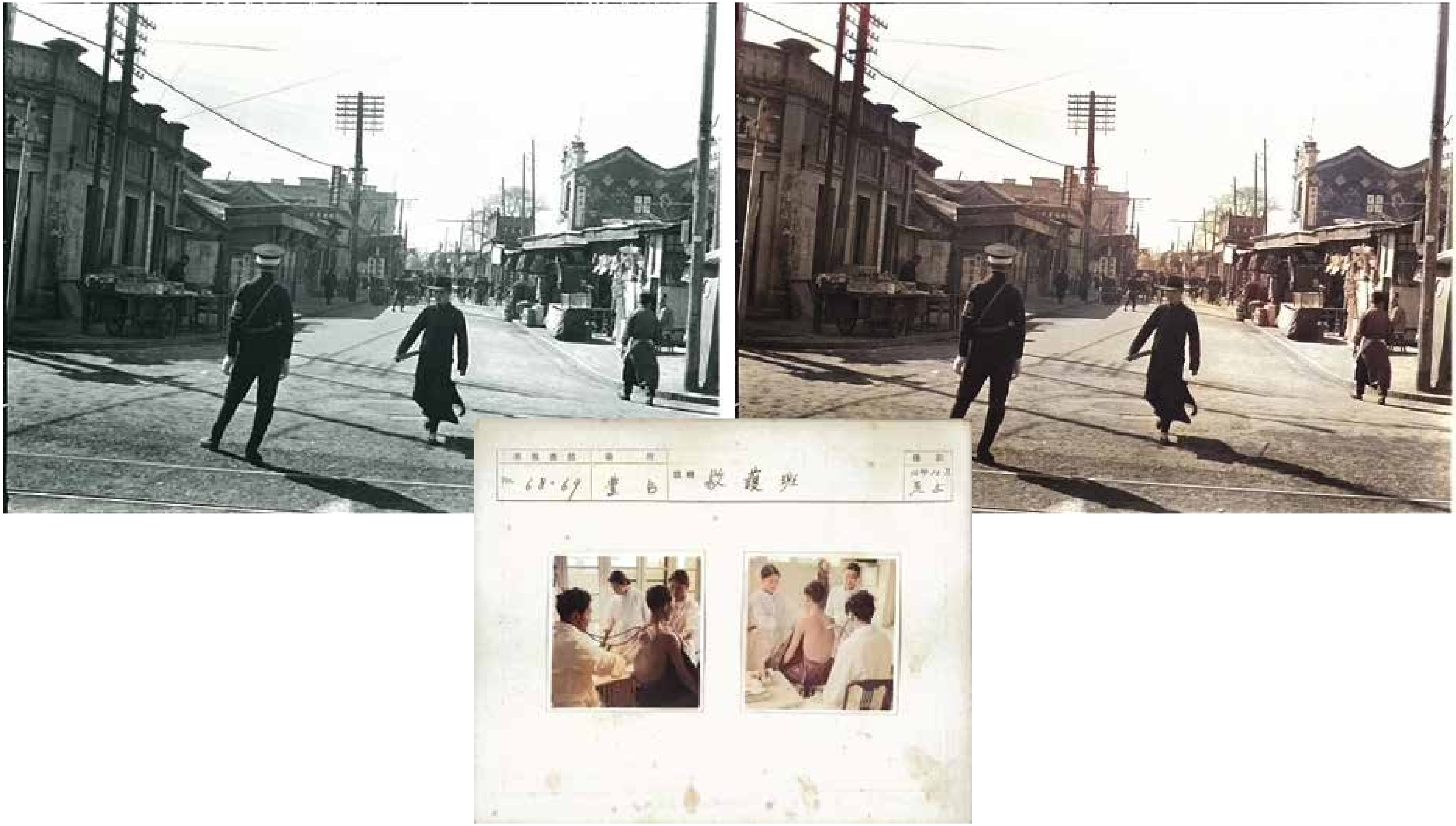
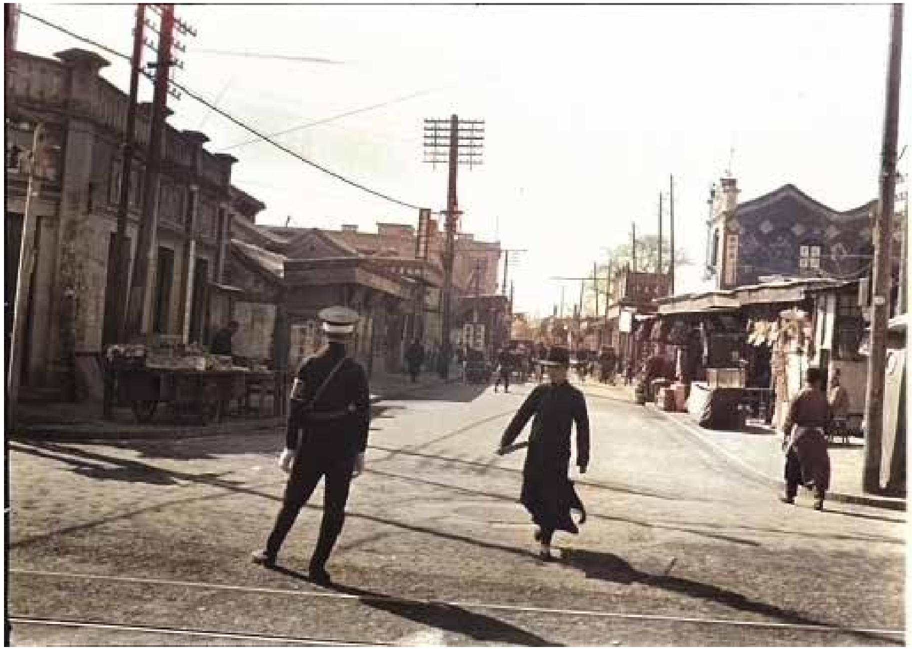


Image Tagging after Colorization



Road, Street, Infrastructure, Town, Transport, Photograph, Urban area, Vehicle, Lane, Pedestrian

Lessons from Two Collections

1. Two photographic collections are too large for humans to annotate one by one.
2. Automatic tagging may be useful as the initial step for improving findability.
3. Statistical research questions, such as thematic distribution may be answered.
4. Other methods can improve findability, such as colorization and object detection.

The Value of Data and FAIR Principle

The Value of Data

1. Intrinsic Value

Raw data

scientists / scholars

2. Basic Value

Organized data

(data) librarians

3. Added Value

Integrated data

(data) curators

4. Persistent Value

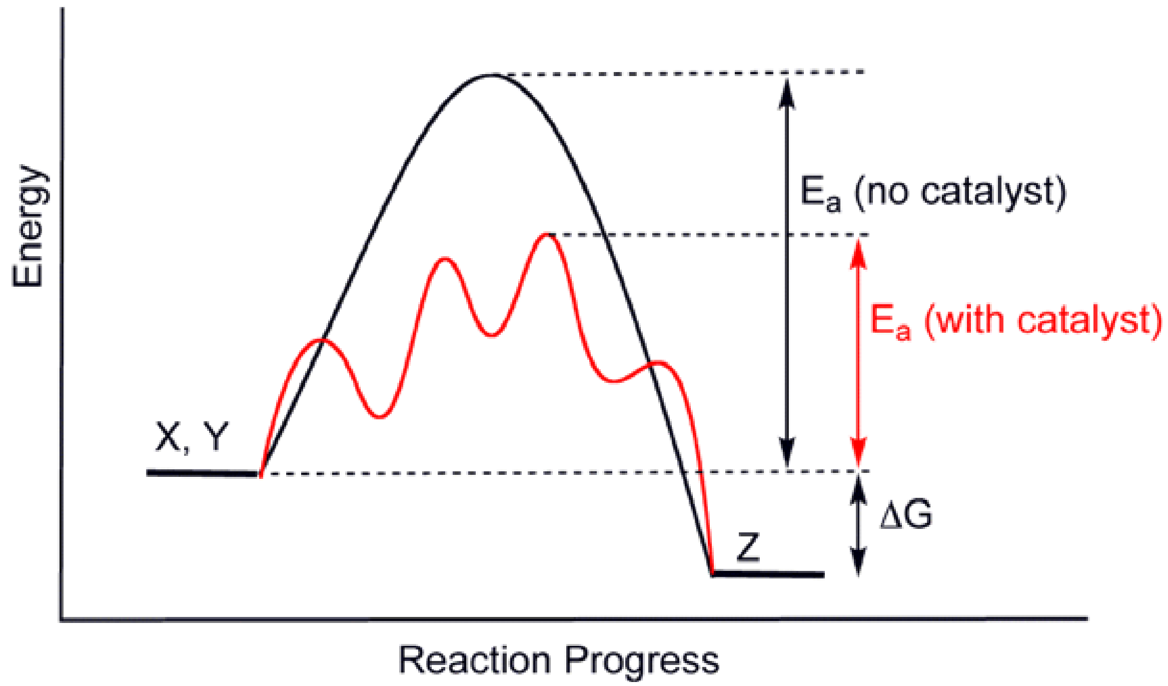
Preserved data

(data) archivists

Machine Learning for Increasing the Value of Data

1. Basic value and added value need high quality metadata for higher value.
2. FAIR (Findable, Accessible, Interoperable, Reusable) principle asks for good metadata.
3. Humans procrastinate in adding metadata, hence the workflow does not start.
4. Use machines to quickly reach **a state which is better than nothing.**

Digital Catalyst



https://commons.wikimedia.org/wiki/File:CatalysisScheme.png

To reach a state of curated data, we need to go beyond the high energy barrier.

Machine learning as digital catalyst reduces the barrier, requiring less human motivation to pass the barrier.

Human-Machine Collaborative Workflow

1. **Machines** can automatically add general noun tags for coarse grouping.
2. **Humans** can manually add proper noun tags for fine meaning as metadata.
3. **Domain experts** can add high-level metadata and semantic grouping.
4. **ML models** can use added metadata as new training data to improve performance.

Conclusion

1. Machine learning, e.g. image tagging, is beneficial for improving findability.
2. General nouns are useful for some apps; other apps require higher level metadata.
3. Better findability (curation) increases the basic value and added value of data.
4. Digital catalyst is a concept of machine-assisted data curation to motivate humans.

Acknowledgment and Links

Photograph collections were provided from the following collaborators:

Dr. Taku Iida in National Museum of Ethnology

Dr. Yoko Nishimura in Toyo University

Ms. Hinako Suzuki in National Research Institute for Earth Science and Disaster Resilience

Dr. Toshihiko Kishi and his colleagues in Kyoto University.

A part of the machine learning workflow was developed by:

Hoàng Văn, Hà (Vietnam National University, HCMC) during NII internship.

- Center for Open Data in the Humanities
  - http://codh.rois.ac.jp/
- Open Science
  - http://agora.ex.nii.ac.jp/~kitamoto/research/open-science/
- Researchmap
  - http://researchmap.jp/kitamoto/