

***Semantic Enrichment***  
for  
**Enhancing Historical and Cultural Heritage Data**  
to  
**Support Digital Humanities Research**

**Marcia L. Zeng**

School of Information, Kent State University

HELSINKI CENTRE FOR DIGITAL HUMANITIES /

**HELDIG DIGITAL HUMANITIES SUMMIT 2020**

**RESULTS OF TODAY – VISIONS FOR TOMORROW**

Firefox



# Outline

Introduction

Data resources  
needed for DH

Semantic Enrichment Approaches

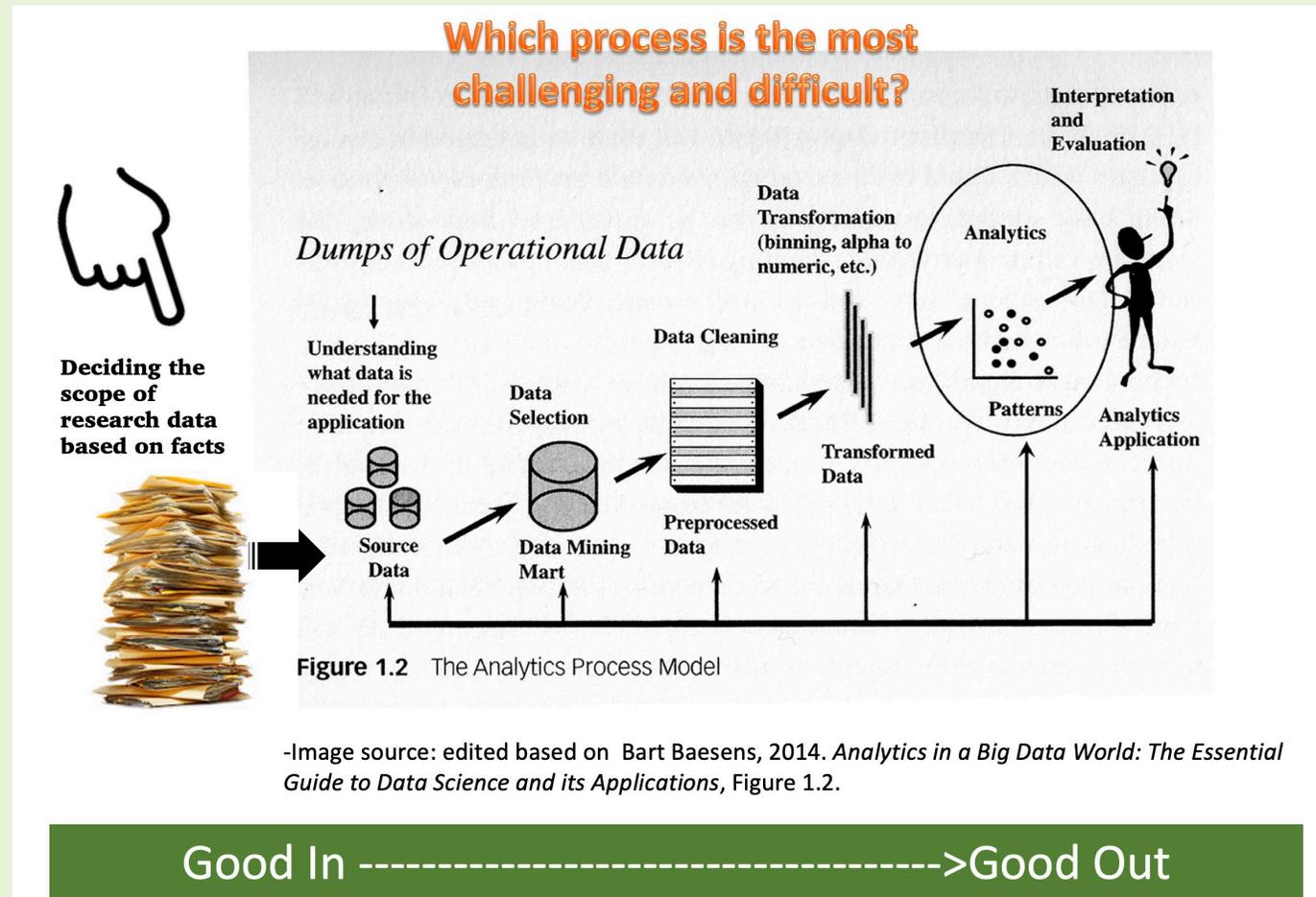
- Structured Data
- Semi-structured Data
- Unstructured Data

Summary and Conclusions



# Introduction

- Demands for historical and cultural heritage data in DH research
- The needs of ensuring the FAIRness of historical and cultural heritage data



Source: Zeng, Marcia L. and James Lee. 2017. Smart Data Approaches to Exploring Independent Datasets across Disciplines, Media, and Perspectives for Research in the Humanities. *Digital Humanities 2017*, August 8-11, 2017, Montreal, Canada. [Slide 26.]





The Digging into Data Challenge (DiD) aims to address how "big data" changes the research landscape for the humanities and social sciences.

<https://diggingintodata.org/>

Round / (year)	DiD Funders	DiD Funder countries	Winner #
Round One / (2009)	<a href="#">NEH</a> , <a href="#">NSF</a> , <a href="#">SSHRC</a> , <a href="#">Jisc</a> (4)	US, Canada, UK.	8
Round Two/ (2011)	<a href="#">NEH</a> , <a href="#">NSF</a> , <a href="#">SSHRC</a> , <a href="#">Jisc</a> , <a href="#">IMLS</a> , <a href="#">AHRC</a> , <a href="#">ESRC</a> , <a href="#">NWO</a> (8)	US, Canada, UK. Netherlands	14
Round Three/ (2013)	<a href="#">NEH</a> , <a href="#">NSF</a> , <a href="#">SSHRC</a> , <a href="#">Jisc</a> , <a href="#">IMLS</a> , <a href="#">AHRC</a> , <a href="#">ESRC</a> , <a href="#">NWO</a> <a href="#">CFI</a> , <a href="#">NSERC</a> (10)	US, Canada, UK. Netherlands	14
Round Four / (2016)  Renamed as the "T-AP Digging into Data Challenge"	<a href="#">NEH</a> , <a href="#">NSF</a> , <a href="#">SSHRC</a> , <a href="#">Jisc</a> , <a href="#">IMLS</a> , <a href="#">AHRC</a> , <a href="#">ESRC</a> , <a href="#">NWO</a> <a href="#">CFI</a> , <a href="#">NSERC</a> <a href="#">MINCyT</a> , <a href="#">FAPESP</a> , <a href="#">FRQ</a> <a href="#">AKA</a> , <a href="#">ANR</a> , <a href="#">DFG</a> , <a href="#">CONACYT</a> , <a href="#">FCT</a> . (18)	US, Canada, UK. Netherlands, Argentina, Brazil, <b>Finland</b> , France, Germany, Mexico, Portugal	14
<b>Total: 4 rounds</b>	<b>18 funders from 11 countries</b>		<b>50 winners</b>



Domains / Areas of Interests	Resources	Approaches
<ul style="list-style-type: none"> <li>activities in humanities &amp; social science</li> <li>ancient language</li> <li>archaeology</li> <li>biodiversity</li> <li>child language development</li> <li>Colonisation of America</li> <li>comparative and epidemiological paradigm</li> <li>criminal intent</li> <li>debating</li> <li>early modern common placing</li> <li>economics</li> <li>English speech</li> <li>epidemiology</li> <li>film and media history</li> <li>financial system</li> <li>history</li> <li>human migration</li> <li>human rights violations</li> <li>information networks</li> <li>information patterns and behaviors</li> <li>journalism</li> <li>language evolution</li> <li>legal structures</li> <li>linguistics</li> <li>literary networks</li> <li>manuscripts provenance</li> <li>music</li> <li>musicology</li> <li>parliaments</li> <li>policy</li> <li>population</li> <li>railroad</li> <li>social science</li> <li>sociological theory</li> <li>standards of living</li> <li>storytelling traditions and story repertoires</li> <li>trading and financial markets</li> <li>vocabularies</li> </ul>	<ul style="list-style-type: none"> <li>audio (music) recordings</li> <li>cuneiform tablets (Mesopotamia )</li> <li>folklore collections</li> <li>GDP per capita</li> <li>geographical data</li> <li>GitHub</li> <li>journals</li> <li>Knowledge Graphs</li> <li>Knowledge Organization Systems</li> <li>letters</li> <li>linguistics databases</li> <li>manuscripts</li> <li>manuscripts (pre-modern European)</li> <li>maps</li> <li>medical images</li> <li>medieval charters</li> <li>multilingual classic text</li> <li>music info</li> <li>news about terrorism</li> <li>newspapers</li> <li>open access publications</li> <li>papyrus documents</li> <li>passages</li> <li>poetry</li> <li>population databases</li> <li>proceedings</li> <li>quotations</li> <li>records in indigenous style</li> <li>records in Spanish</li> <li>signs</li> <li>social media</li> <li>speech datasets</li> <li>speech recordings</li> <li>speeches</li> <li>spoken language collections</li> <li>tweets, political</li> <li>video data</li> <li>writing pieces</li> </ul>	<ul style="list-style-type: none"> <li>annotation</li> <li>comparative analysis</li> <li>computational analysis</li> <li>computing</li> <li>corpus building</li> <li>cross datasets analysis</li> <li>cross-datasets searching</li> <li>cross-linguistic annotation</li> <li>data management</li> <li>data mining</li> <li>image processing</li> <li>indexing</li> <li>linking</li> <li>machine coding</li> <li>machine learning</li> <li>machine translation</li> <li>metadata aggregation</li> <li>metadata analysis</li> <li>metadata auto-generation</li> <li>metadata extraction</li> <li>natural language processing (NLP)</li> <li>protocols development</li> <li>spatial-temporal correlation</li> <li>speech mining</li> <li>text analysis</li> <li>visualization</li> </ul>

Source: Compiled by M. Zeng based on the short descriptions available at <https://dev.diggingintodata.org/awards>



Funders: from more than 10 countries, including NEH, NSF, IMLS

## Resources

- audio (music) recordings
- cuneiform tablets (Mesopotamia )
- folklore collections
- GDP per capita
- geographical data
- GitHub
- journals
- Knowledge Graphs
- Knowledge Organization Systems (KOS)
- letters
- linguistics databases
- manuscripts
- manuscripts (pre-modern European)
- maps
- medical images
- medieval charters
- multilingual classic text
- music info
- news about terrorism

- newspapers
- open access publications
- papyrus documents
- passages
- poetry
- population databases
- proceedings
- quotations
- records in indigenous style
- records in Spanish
- signs
- social media
- speech datasets
- speech recordings
- speeches
- spoken language collections
- tweets, political
- video data
- vocabularies
- writing pieces

Domains / Areas of Interest	Resources	Approaches
<ul style="list-style-type: none"> <li>• activities in humanities &amp; social science</li> <li>• ancient language</li> <li>• archaeology</li> <li>• biodiversity</li> <li>• child language development</li> <li>• Colonization of America</li> <li>• comparative and epistemological paradigm</li> <li>• criminal intent</li> <li>• debating</li> <li>• early modern common placing</li> <li>• economics</li> <li>• English speech</li> <li>• epidemiology</li> <li>• film and media history</li> <li>• financial system</li> <li>• history</li> <li>• human migration</li> <li>• human rights violations</li> <li>• information networks</li> <li>• information patterns and behaviors</li> <li>• journalism</li> <li>• language evolution</li> <li>• legal structures</li> <li>• linguistics</li> <li>• literary networks</li> <li>• manuscript provenance</li> <li>• music</li> <li>• musicology</li> <li>• parliaments</li> <li>• poetry</li> <li>• population</li> <li>• regional</li> <li>• social science</li> <li>• sociological theory</li> <li>• standards of living</li> <li>• storytelling traditions and story repertoires</li> <li>• trading and financial markets</li> <li>• vocabularies</li> </ul>	<ul style="list-style-type: none"> <li>• audio (music) recordings</li> <li>• cuneiform tablets (Mesopotamia )</li> <li>• folklore collections</li> <li>• GDP per capita</li> <li>• geographical data</li> <li>• GitHub</li> <li>• journals</li> <li>• Knowledge Graphs</li> <li>• Knowledge Organization Systems</li> <li>• letters</li> <li>• linguistics databases</li> <li>• manuscripts</li> <li>• manuscripts (pre-modern European)</li> <li>• maps</li> <li>• medical images</li> <li>• medieval charters</li> <li>• multilingual classic text</li> <li>• music info</li> <li>• news about terrorism</li> <li>• newspapers</li> <li>• open access publications</li> <li>• papyrus documents</li> <li>• passages</li> <li>• poetry</li> <li>• population databases</li> <li>• proceedings</li> <li>• quotations</li> <li>• records in indigenous style</li> <li>• records in Spanish</li> <li>• signs</li> <li>• social media</li> <li>• speech datasets</li> <li>• speech recordings</li> <li>• speeches</li> <li>• spoken language collections</li> <li>• tweets, political</li> <li>• video data</li> <li>• writing pieces</li> </ul>	<ul style="list-style-type: none"> <li>• annotation</li> <li>• comparative analysis</li> <li>• computational analysis</li> <li>• computing</li> <li>• corpus building</li> <li>• cross datasets analysis</li> <li>• cross-datasets searching</li> <li>• cross-linguistic annotation</li> <li>• data management</li> <li>• data mining</li> <li>• image processing</li> <li>• indexing</li> <li>• linking</li> <li>• machine coding</li> <li>• machine learning</li> <li>• machine translation</li> <li>• metadata aggregation</li> <li>• metadata analysis</li> <li>• metadata auto-generation</li> <li>• metadata extraction</li> <li>• natural language processing (NLP)</li> <li>• network development</li> <li>• special temporal correlation</li> <li>• speech mining</li> <li>• text analysis</li> <li>• visualization</li> </ul>

**Domains/Areas of Interests || Resources || Approaches**  
Expressed in the Project Descriptions of *Digging into Data Challenge* Round 1-4, 2009-2016





# Digitizing & Documenting → Datafying & Enriching → Contextualizing

Through LAMs: *unstructured data* found in documents and other information-bearing objects:

Transforming unstructured data into → structured data

Resources delivered on the web, including:

- Metadata
- Representative images
- Original documents' transcripts
- Media, etc.



## National Library of Medicine "Turning the Pages" The Edwin Smith Surgical Papyrus

Case 5. A head wound with skull fracture (2,11 - 17)

**Title**  
Practices for a gaping wound in his head that has fractured his skull.

**Examination and Prognosis**  
If you treat a man for a gaping wound in his head, which has penetrated to the bone and split his skull, you have to probe that wound. Should you find that fracture that is in his skull deep and sunken under your fingers, and should the swelling that is on it be high, while he bleeds from his nostrils and his ears, suffers stiffness in his neck, and is unable to look at his shoulders and his chest, then you say about him: "One who has a gaping wound in his head, which has penetrated to the bone and fractured his skull, and who suffers from stiffness in his neck: an ailment for which nothing is done."

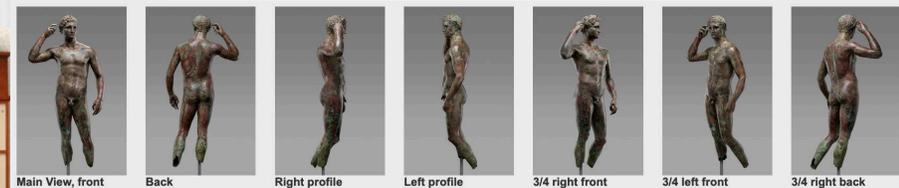
**Treatment**  
You should not bandage him. He is to be put down on his bed until the time of his injury passes.

**Explanations**  
As for "which has fractured his skull," it is fracturing his skull, with the bones that happen from that fracture sunken to the inside of his skull. The treatise "The Nature of Wounds" has said about it: it is the fracturing of his skull into many pieces, sunken to the inside of his skull.

**Close** **TEXT** **ZOOM**

<https://ceb.nlm.nih.gov/proj/ttp/flash/smith/smith.html>

## J. Paul Getty Museum collection



Provenance Exhibitions Bibliography Education Resources Related Media

▶ 02:35 ◀ Audio: Statue of a Victorious Youth / Carol Mattusch (Highlights)

▶ 01:13 ◀ Audio: Statue of a Victorious Youth - Conservation / J. Podany

▶ 01:44 ◀ Audio: Victorious Youth (Descriptions)

▶ 01:39 ◀ Audio: Statue of a Victorious Youth, feat. Carole Mattusch

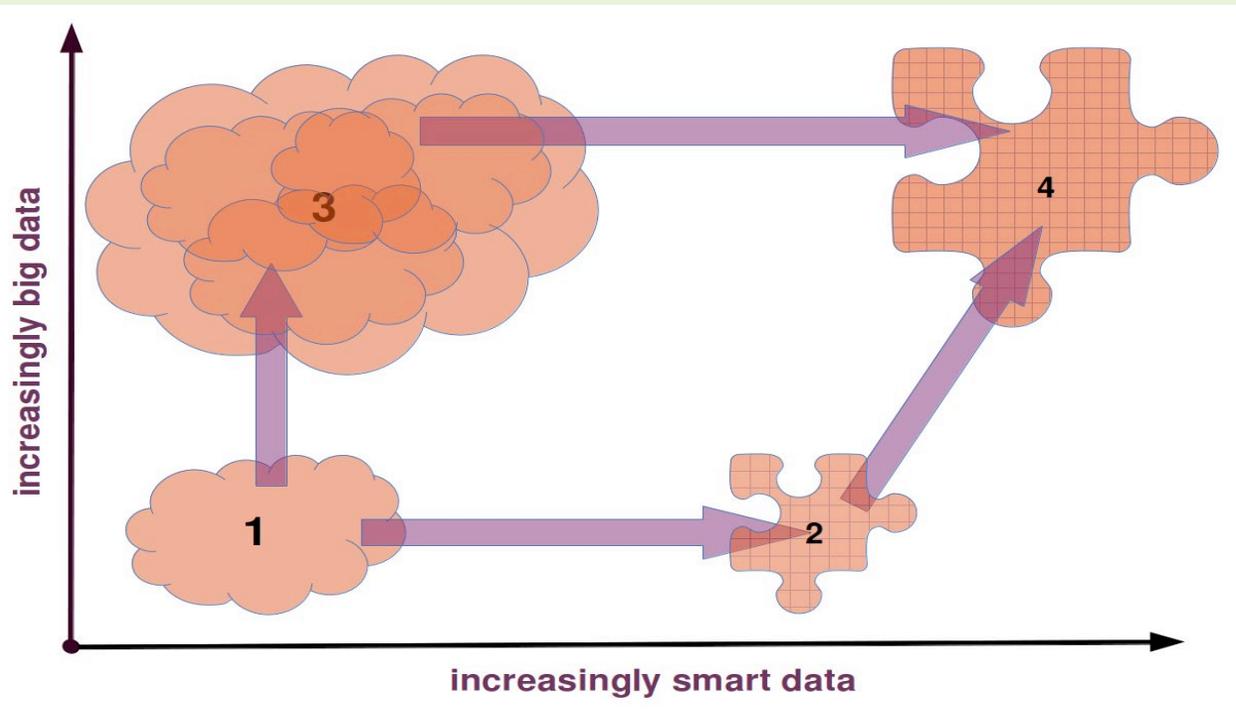
<https://www.getty.edu/art/collection/objects/7792>



# “Big? Smart? Clean? Messy? Data in the humanities”

Schöch, Christof. *Journal for Digital Humanities*. 2(3): 2-13.

Data has to be cleaned, transformed, and analyzed to unlock its hidden potential.



The story of smart and big data.

<http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/>

Marcia L. Zeng. HELDIG DH Summit 2020

Once tamed through organizing and integrating processes, large volumes of unstructured, semi-structured, and structured data are turned into “smart data” that reflect the research priorities of a particular discipline or field.

Smart data inquiries can then be used to provide comprehensive analyses and generate new products and services.



# Smart Data in the context of Big Data

**Big Data**

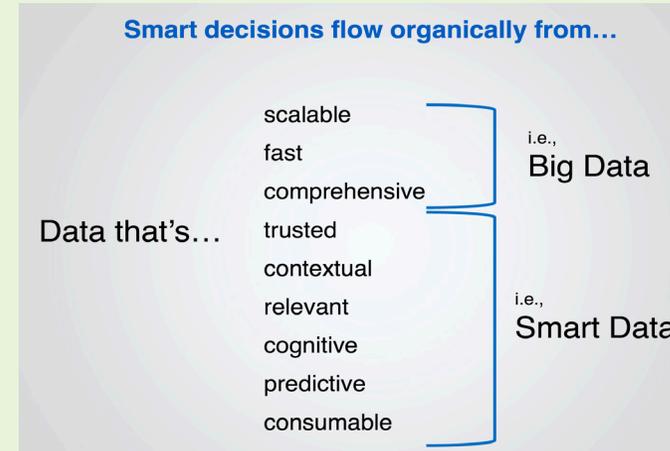
- Volume (data quantity)
- Velocity (data speed)
- Variety (data types & nature)
- Variability (data consistency)
- Veracity (data quality)



**Smart Data**

= The ability to achieve big insights from such data at any scale, great or small.

- trusted
- contextualized
- relevant
- cognitive
- predictive
- consumable



*Kobielus, James. (2016)*

Source: Zeng, M.L. 2017 [DOI: 10.1515/jdis-2017-0001](https://doi.org/10.1515/jdis-2017-0001)  
Compiled based on Kobielus, James. (2016, June). *The Evolution of Big Data to Smart Data*.  
Keynote at Smart Data Online 2016.



# Outline

Introduction



Semantic Enrichment Approaches

1. Structured Data
2. Semi-structured Data
3. Unstructured Data

Summary and Conclusions





# LAM Data Examples

1

Structured

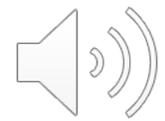
- bibliographies
- indexing & abstracting databases
- citation indexes
- catalogs of all kinds
- special collection portals
- metadata registries
- curated research datasets
- name authorities

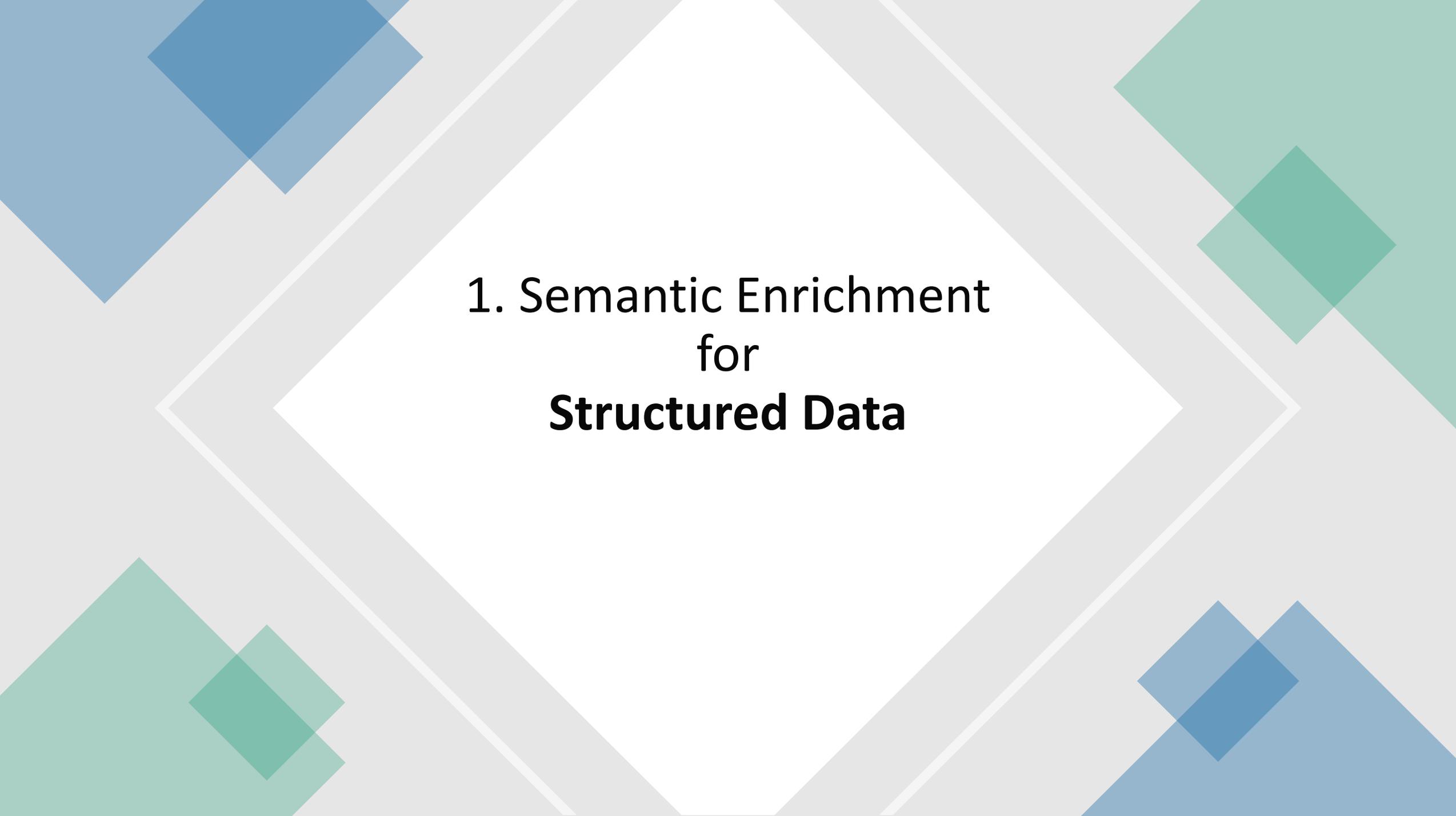
Semi-structured

- ... ..
- Text Encoding Initiative (TEI) files
- archival finding aids
- value added/tagged resources
- unstructured portion within metadata descriptions
- data from Web crawling that need to be cleaned
- ... ..

Unstructured

- documents, cultural artifacts, & original information-bearing objects
  - digitized or not-digitized
  - textual or non-textual
  - in all kinds of formats and media
  - Possibly of undetermined date and/or origin





1. Semantic Enrichment  
for  
**Structured Data**

# Semantic Enrichment for **Structured Data**

➤ A common strategy in LAM data enhancement efforts in order to:

- Overcome challenges relating to data quality and discoverability in the digital age
- Provide more context and multilingual information for cultural heritage (CH) objects

[Refer to the most recent presentations at [SWIB20](#)]

➤ “Enrichment” can be used to refer to

- a process (e.g., the application of an enrichment tool);  
or
- its result (the new metadata created at the end of the process).

- ✓ reconciliation,
- ✓ mapping,
- ✓ alignment,
- ✓ matching,
- ✓ massaging,
- ✓ merging,
- ✓ Interlinking
- ✓ ...

Three main stages

- 1) Analysis
- 2) Linking
- 3) Augmentation

Ref: -- Europeana Task Force on Enrichment and Evaluation. “Report on Enrichment and Evaluation” 29/10/2015



# 1. Semantic Enrichment for **Structured Data**

[Starting point: existing metadata components that are in a controlled/contrrollable form]

## Outline

### Approaches

#### A. Contextualize through entities

Creates typed relationships between resources of different types

#### B. “Massage”

(my) label → to → (their) URI(s)

(my) ID → to → (their) URI(s)

#### C. Connect to real “things”

### Used by (examples)

- **Unified repositories**
  - Europeana
- **Collaborative projects**  
(Suggest to visit):
  - Wikibase “[Project Passage](#)”
  - LD4P (Linked Data for Libraries)
    - Linked Data for Production: Pathway to Implementation ([LD4P2](#))
      - “Knowledge panels”
- **Individual LAMs and institutions**
  - Museum of Modern Art (MoMA)
  - Dictionary of Classic Mayan



# 1-A. Contextualize through entities

- ✓ Creates typed relationships between resources of different types
  - usually on those fields that are in a controlled form

[Europeana enriches xxx by aligning to (xxx)]

agent names →  
places →  
concepts →  
time period → (Semium Time).

- ✓ Relate Objects to concepts, agents, places, etc., using the properties in EDM (e.g., *dc:subject*, *dc:creator*).

### Update from Europeana

by Nov 15, 2020, enriched in this year:

- 1,429,242 for agents
- 15,206,861 for places
- 15,218,899 for concepts
- 18,597,882 for time period

## Europeana Dereferenceable vocabularies

The Getty - Union List of Artist Names (ULAN)	edm:Agent	} agents
Virtual International Authority File (VIAF)	edm:Agent	
Wikidata	edm:Agent	
Gemeinsame Normdatei (GND)	edm:Agent, edm:Place, skos:Concept	} places
Getty Thesaurus of Geographic Names (TGN)	edm:Place	
Geonames	edm:Place	} concepts
The Getty - Art & Architecture Thesaurus (AAT)	skos:Concept	
IconClass	skos:Concept	
Israel Museum Jerusalem Concepts	skos:Concept	
Library of Congress Subject Headings (LCSH)	skos:Concept	
data.europeana.eu WWI Concepts from Library of Congress Subject Headings (LCSH)	skos:Concept	
Europeana Sounds Genres	skos:Concept	
UDC	skos:Concept	
UNESCO Thesaurus	skos:Concept	
YSO - General Finnish ontology	skos:Concept	
Fashion Thesaurus	skos:Concept	
MIMO Concepts	skos:Concept	

- Source: Europeana Semantic Enrichment Framework *Documentation*

Version: 17th November 2016 (updated 2017, 2018, 2020)

Available from <https://pro.europeana.eu/page/europeana-semantic-enrichment> --> [several vocabularies](#) (Compiled by MZ 2020-11-18)



# 1-B. "Massage"

An Example:

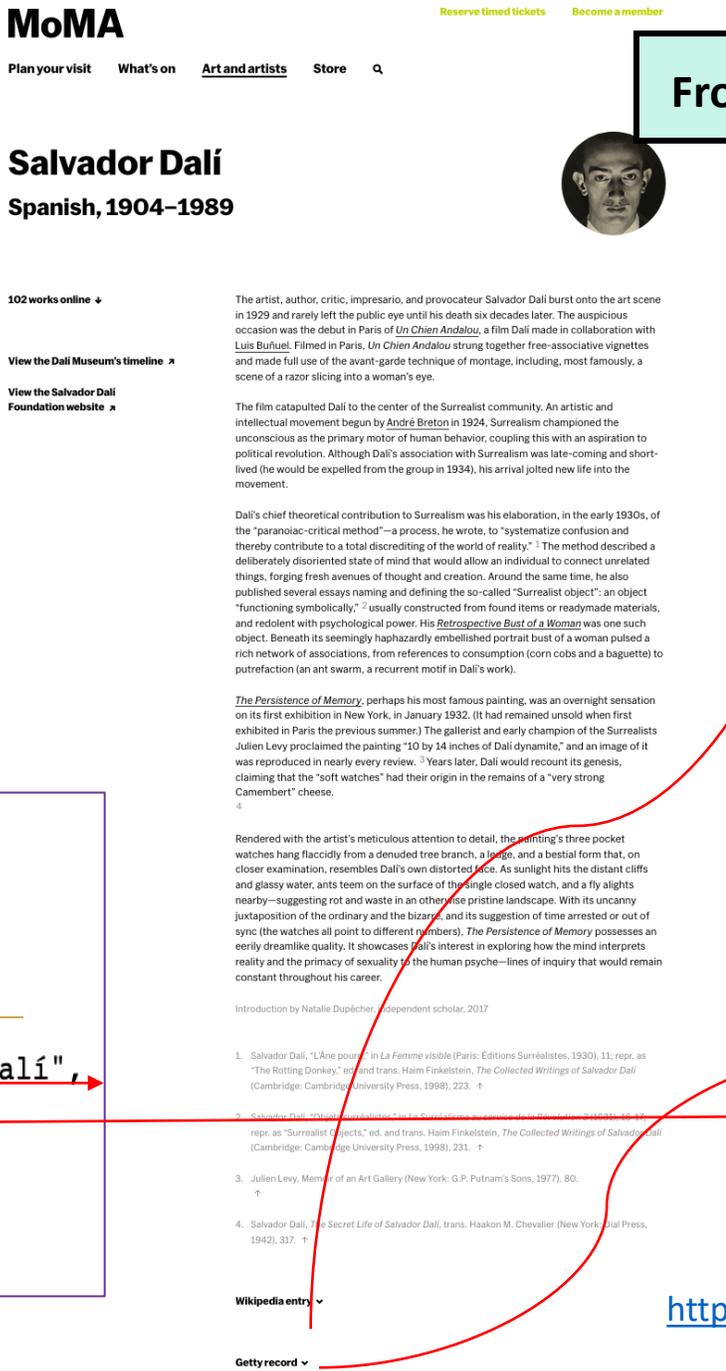
- (my) ID →to→ (their) URI(s)

(my) ID

"sameAs"  
(their) URIs

```
37 <script type="application/ld+json">
38 {
39   "@context": "http://schema.org/",
40   "@type": "Person",
41   "url": "https://www.moma.org/artists/1364",
42   "sameAs": [
43     "https://en.wikipedia.org/wiki/Salvador_Dalí",
44     "http://vocab.getty.edu/ulan/500009365"
45   ],
46   "name": "Salvador Dalí"
47 }
48 </script>
49
50
```

Back-end



The screenshot shows the MoMA website for Salvador Dalí. At the top, there are links for "Plan your visit", "What's on", "Art and artists", "Store", and a search icon. The main heading is "Salvador Dalí" with the subtitle "Spanish, 1904–1989" and a portrait of Dalí. Below this, there are sections for "102 works online", "View the Dalí Museum's timeline", and "View the Salvador Dalí Foundation website". The main content area contains several paragraphs of text about Dalí's life and work, including his debut in Paris, his association with Surrealism, and his chief theoretical contribution. At the bottom, there is a "Wikipedia entry" and a "Getty record" section.

Front-end

Case: MoMA

Wikipedia entry ▾

Tool: OpenRefine

Getty record ^

Union List of Artist Names (ULAN)

Nationalities

Spanish, Catalan

Gender

Male

Roles

Artist, Writer, Illustrator, Painter, Sculptor

Names

Salvador Dalí, Salvador Dalí, Salvador Felipe Jacinto Dalí, Salvador Felip Jacint Dalí Domènech, Salvador Dalí y Domènech, Salvador Felipe Jacinto Dalí y Domènech, Salvador Dalí i Domènech, Salvador Felipe Jacinto Dalí Domènech, Salvator Dalí, Salvador Dalí Domènech, Salvador Dalm y Domenech, Салвадор Дали, Сальвадор Дали, 萨尔瓦多·达利, 达利萨尔瓦多

Ulan

500009365

View the full Getty record ↗

Information from Getty's Union List of Artist Names® (ULAN), made available under the [ODC Attribution License](#)

Images captured 2020-11-18

<https://www.moma.org/artists/1364>





# 1-C. Connect to real “things”

[Other projects using this approach:

LD4P2

“Knowledge panels”

- Cornell Univ. Library,
- Stanford Univ. library, etc.]

Images captured 2020-11-18

<http://fast.oclc.org/searchfast/>

fast.oclc.org/searchfast/?&limit=keywords&facet=all&query=Kennedy%2C John F. (John Fitzgerald)%2C 1917-

searchFAST FAST (Faceted Application of Subject Terminology)

Find FAST Subject Headings (Instructions)

SEARCH FAST

Keywords Kennedy, John F. (John Fitzgerald), 1917-1963 Search

**FAST TERMS**

Search results for: "Kennedy, John F. (John Fitzgerald), 1917-1963"

Limit Results by: All

Displaying 1 to 2 of 2 << Previous Next >>

1 Jump

Heading	Facet	Uses
<a href="#">Kennedy, John F. (John Fitzgerald), 1917-1963</a>	person	13191
<a href="#">Kennedy, John F. (John Fitzgerald), 1917-1963 (Spirit)</a>	person	1

**TERM DETAILS**

**Kennedy, John F. (John Fitzgerald), 1917-1963** [Find in WorldCat](#)

USED FOR:

- Gannaidi, 1917-1963
- JFK (John Fitzgerald Kennedy), 1917-1963
- Kan-nai-ti, 1917-1963
- Kanadī, Jūn Fītz Jīrāld, 1917-1963
- Kanīdī, Jūn F., 1917-1963
- K'enedi, 1917-1963
- Kenedi, Dzhon F., 1917-1963
- Kenedi, Džon Fridžerald, 1917-1963
- Kenedi, G'on F., 1917-1963
- Kenedijs, Džons F., 1917-1963
- Kennedi, Dzhon Fītsžžerald, 1917-1963
- Kennedy, Jack, 1917-1963
- Kennedy, John Fitzgerald, 1917-1963
- Kennedy, Ken, 1917-1963

USAGE:

- LC (2017) Subject Usage: 1,530
- WC (2017) Subject Usage: 13,191

RECORD ID:

fst00035588

SOURCES AND OTHER LINKS:

- [Kennedy, John F. \(John Fitzgerald\), 1917-1963--\(DLC\)n 79055297](#)
- [John F. Kennedy--http://en.wikipedia.org/wiki/John\\_F.\\_Kennedy](#)
- [Kennedy, John F. \(John Fitzgerald\), 1917-1963--https://viaf.org/viaf/68910251](#)

LINKS TO FULL RECORD:

- Permanent Link <http://id.worldcat.org/fast/35588>
- MARC-21 record <http://id.worldcat.org/fast/35588.html>
- MARC-21 UTF-8 (raw data) <http://id.worldcat.org/fast/35588.mrc>
- MARC-21 xml (raw data) <http://id.worldcat.org/fast/35588.mrc.xml>
- RDF record (raw data) <http://id.worldcat.org/fast/35588.rdf.xml>

view MARC

Case: FAST

Front-end

Resources and Other Links



# John F. Kennedy's entry in FAST is enriched with other sources

Case: FAST

**foaf:focus** allows FAST terms (*skos:Concept*) to be connected to URIs that identify real-world entities, to include detailed information that is usually excluded in authority records.

**schema:sameAs** allows FAST terms (*skos:Concept*) to take advantage of all the various string values included in VIAF (containing dozens multilingual name authorities) without having to manually include the values in the RDF triples for the specific term.

correct coding of properties

Front-end

RECORD ID:

fst00035588

SOURCES AND OTHER LINKS:

Kennedy, John F. (John Fitzgerald), 1917-1963--(DLC)n 79055297

John F. Kennedy--[http://en.wikipedia.org/wiki/John\\_F.\\_Kennedy](http://en.wikipedia.org/wiki/John_F._Kennedy)

Kennedy, John F. (John Fitzgerald), 1917-1963--<https://viaf.org/viaf/68910251>

LINKS TO FULL RECORD:

Permanent Link <http://id.worldcat.org/fast/35588>

MARC-21 record <http://id.worldcat.org/fast/35588/marc21.xml>

RDF record <http://id.worldcat.org/fast/35588/rdf.xml>

```
<foaf:focus>
<rdf:Description rdf:about="http://en.wikipedia.org/wiki/John_F._Kennedy">
<rdfs:label>John F. Kennedy</rdfs:label>
</rdf:Description>
</foaf:focus>
```

```
<schema:sameAs>
<rdf:Description rdf:about="https://viaf.org/viaf/68910251">
<rdfs:label>Kennedy, John F. (John Fitzgerald), 1917-1963</rdfs:label>
</rdf:Description>
</schema:sameAs>
```

Back-end

Ref: O'Neill, Ed, and Jeff Mixer 2013. (1) The case for faceting (2) FAST Linked Data mechanics. In 76th Annual Meeting of the American Society for Information Science and Technology (ASIS&T), Montreal, Canada, Nov. 2-6, 2013.



## Possible situations for each of the datasets

- If it has used local controlled vocabularies
  - The terms used or the form representing the concepts and named entities are local.
- If it has used a pre-LOD vocabulary
  - There might be no URIs/IRIs yet.
- If a mapping decision is to be made
  - In a subject domain there could be more than one standard vocabulary.
- If it needs to map its local lists to a standardized LOD KOS\*
  - Human resources and quality control are most critical and could be challenging.
  - In addition to the normal standard vocabularies, other special vocabularies might be needed.
  - Suggest checking:
    - FINTO <https://finto.fi/en/>
    - Mix'n'Match <https://tools.wmflabs.org/mix-n-match/#/>

### For a dataset formed through aggregation

- in addition to the above issues, synonyms and acronyms occur in the data provided by different sources.
- Heavy disambiguation and semantic conflict controls are needed.



## 2. Semantic Enrichment for **Semi-structured Data**





2

# LAM Data Examples

Structured

- bibliographies
- indexing & abstracting databases
- citation indexes
- catalogs of all kinds
- special collection portals
- metadata registries
- curated research datasets
- name authorities

Semi-structured

- ... ..
- Text Encoding Initiative (TEI) files
- archival finding aids
- value added/tagged resources
- unstructured portion within metadata descriptions

Unstructured

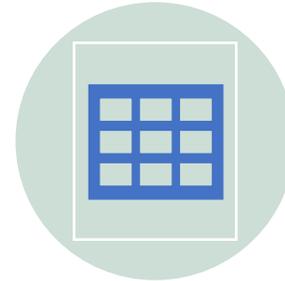
- data from Web crawling that need to be cleaned
- ... ..
- documents, cultural artifacts, & original information-bearing objects
  - digitized or not-digitized
  - textual or non-textual
  - in all kinds of formats and media
  - Possibly of undetermined date and/or origin



# Why use data from semi-structured data resources?



An important feature of semi-structured data resources that should be recognized, is that **they are the products of information processing.**



These semi-structured data **represent the accumulated time, knowledge, and experience of the creators** who generated them through a formal workflow which conforms to professional standards and best practices.



With semantic enrichment processes, the data values in semi-structured data are **contextualized through the metadata elements/fields;** hence, the function and meaning are clearly implied.



By parsing these data through advanced information technologies, these LAM data are dramatically **enriched and are converted into new access points.**





## 2. Semantic Enrichment For **Semi-structured Data** Outline

[Starting point: existing metadata components that are in free-form]

### Investigations and findings (use entity extraction)

- A. MARC 5xx fields, unstructured notes, etc.
- B. Archival Finding aids' descriptions
- C. Special Collections and others
  - 1) Photograph collections' metadata (about the collection)
  - 2) Records of theses and dissertations
  - 3) Museum objects captions

Semi-structured → to → Structured Data



# 2- A. MARC 5xx fields, unstructured notes, TOC

Demo: entity extraction from 5xx fields

Language: English

Topics:

Entertainment Culture	84%
Film (M:H1)	70%
Israel (G:3S)	56%
Music (M:H0)	33%
Marine Port Services (TRBC) (B:64)	33%
Slovakia (G:1C)	11%
Transportation, Ground (TRBC) (B:65)	9%

Entities:

- City
  - London, Greater
- Person
  - Allan Rouse
  - George Harrison
  - John Lennon
  - Kevin Howlett
  - Mike Heatley
  - Paul McCartney
  - Ringo Starr
  - Tony Barrow

Social Tags:

After

500 Program notes by Kevin Howlett, Mike Heatley, and Allan Rouse, and original liner notes by Tony Barrow (18 pages : ports.) inserted in container

500 Includes Please please me mini-documentary

500 Sound recording

505 00 |tI saw her standing there --|tMisery --|tAnna (Go to him) --|tChains --|tBoys --|tAsk me why --|tPlease please me --|tLove me do --|tP.S. I love you --|tBaby it's you --|tDo you want to know a secret --|tA taste of honey --|tThere's a place --|tTwist and shout

511 0 The Beatles (John Lennon, vocals, guitar, harmonica ; George Harrison, vocals, guitar ; Paul McCartney, vocals, bass ; Ringo Starr, vocals, drums, tambourine, maracas) ; with additional musicians

518 Recorded 1962 and February 11, 1963, at Abbey Road Studios

London

Before

Tool used: Open Calais  
 Note: Only for assistant extraction; still need human cleaning process.



# OCLC: "Mining MARC's Hidden Treasures: Initial investigations into how notes of the past might shape our future."

## WHAT

Approximately 19 million records for musical resources in WorldCat were analyzed in 2016.

- Generated during the 45-year history of WorldCat;
- Comprised both musical sound recordings and musical scores;
- Approximately 2.5 million names that can be identified as distinct.

## HOW

- Associating performer names with authority data
- Identifying role terms and phrases with controlled vocabularies
- Extended work has been conducted in multiple languages for
  - the performer roles
  - medium of performance terms
  - associating the name of an instrument with its performer
  - and more.

## WHERE

--from uncontrolled occurrences in notes and/or statements of responsibility in records

The **extra descriptive information** may be found in such fields as:

- 245 subfield \$c (Statement of Responsibility)
- 500 (General Note)
- 505 (Formatted Contents Note)
- 508 (Creation/Production Credits Note)
- 511 (Participant or Performer Note)
- 520 (Summary, Etc.)

Weitz, Jay, Jenny Toves, Diane Vizine-goetz, Nannette Naught, and Robert Bremer. "Mining MARC's Hidden Treasures: Initial investigations into how notes of the past might shape our future." *Journal of Library Metadata* 16, no. 3-4 (2016): 166-180.



# 2-B. Archival Finding Aids' descriptions

## Case: Finding Aids

Portions of a finding aid and explanation of the text used in the semantic analysis process.



**Finding Aid to the Artificial Collection:  
Pearl Harbor Attack (Dec 6 – Dec 8, 1941)**

**Size:** (.5 cu.ft.)

**Dates:** December 6, 1941 – December 8, 1941

**Location of Repository:** Franklin D. Roosevelt Presidential Library

**Name of Finding Aid Author:** Ali Caron & Georgina Garcia

**Date of Creation:** Summer 2011

**Copyright Notice:** The writings of Franklin D. Roosevelt within this collection. The official writings of United States government officials within this collection. The writings of Eleanor Roosevelt within this collection are subject to Mrs. R. Other materials are subject to the United States Copyright law, 17 U.S.C. 101

**Administrative Note:** Franklin D. Roosevelt Presidential Library is the first presidential library used by a sitting president. The library houses documents, photographs pertaining and relating to President Franklin D. Roosevelt (FDR), the events of Pearl Harbor and the library houses materials about the attack

**Scope and Content:** This artificial collection is composed of photocopies of

### Finding Aids to the Artificial Collection: Pearl Harbor Attack (Dec 6-Dec 8, 1941)

**Series Descriptions:** The collection is organized in 2 series:

- Series I: Documents – The items selected for this series remain within December 8, 1941, date range. The items reflect the Pearl Harbor attack that incident. The contents found under the container list portion are title; and folder title, found in quotations.
- Series II: Still Photographs – Images comprising this series are selected from the FDR Library's General Photograph Collection, folder: WWII: Hawaii: Attack on Pearl Harbor. Listed here are original captions taken from the photographs themselves, along with a Library control number. Unless copyright information is stated in the image caption, all of the photographs in this series belong in the public domain. This means that, to the best of our knowledge, the materials may be freely used by the researcher. However, for copyrighted materials, it is the researcher's responsibility to determine the limits of Fair Use as defined by sections 107 to 118 of the copyright law and to obtain permission from the copyright holder for further use.

**Arrangement:** Series I is arranged alphabetically after the president's papers and Series II is arranged numerically.

**Container Lists:**

**SERIES I: DOCUMENTS**

- OF400: Appointments; Hawaii, 1941
- OF4675: World War II; General, 1941-1942
- PPF200b: Nov. 11, 1941- Jan. 6, 1942; Public Reactions

**Source:**

[http://www.fdrlibrary.marist.edu/archives/pdfs/findingaids/findingaid\\_pearlharborattack.pdf](http://www.fdrlibrary.marist.edu/archives/pdfs/findingaids/findingaid_pearlharborattack.pdf)

Pages

- 1
- 2
- 3
- 4
- 5
- 6

- Text from an archival finding aids
  - Descriptive information
    - creator histories
    - Scope and content notes
    - detailed description of contents, including folder and item titles
  - Abstracts from these descriptions

WHERE



### Research sample **WHAT**

- 43 archival record groups
- from 16 institutions, including:
  - university archives
  - government records archives
  - manuscript/special collections repositories in various LAMs

<http://lod-lam.slis.kent.edu/SemanticAnalysis.html>

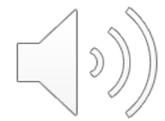
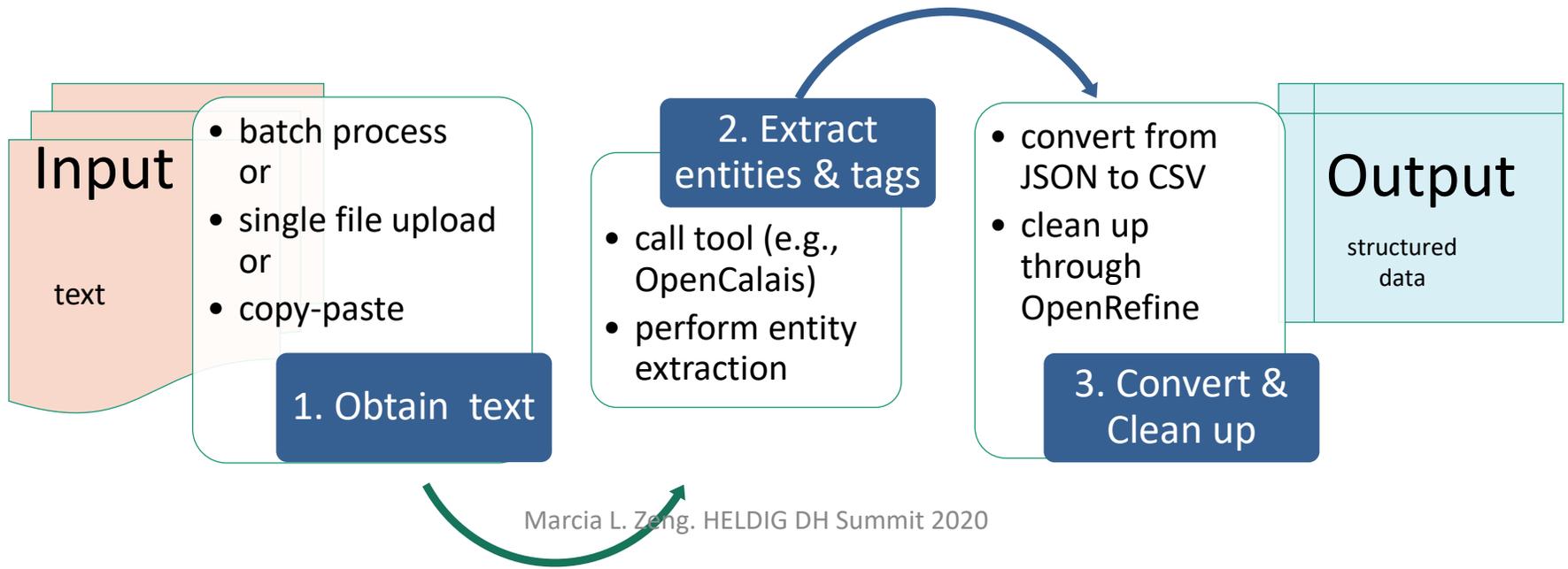
### Tools used: **HOW**

- OpenCalais (demo)
- COGITO Intelligent API (demo)
- MachineLinking
- Zemanta
- OpenRefine

- Functions:
- Entity extraction
  - Tagging
  - Categorization
  - Semantic reasoning
  - Fact mining
  - Data visualization

KSU iSchool  
LOD-LAM team,  
2013-14

### **HOW**



The Calais initiative is about enabling semantic applications by providing a metadata generation web service, sample applications using that service to jumpstart development efforts, and support for developers.

### The Calais Web Service

The Calais web service automatically attaches rich semantic metadata to the content you submit. Using natural language processing, machine learning and other methods

Enter text here:

Before

Finding Aid to the Artificial Collection:  
 Pearl Harbor Attack (Dec 6 – Dec 8, 1941)  
 Size: (.5 cu.ft.)  
 Dates: December 6, 1941 – December 8, 1941  
 Location of Repository: Franklin D. Roosevelt Presidential Library  
 Name of Finding Aid Author: Ali Caron & Georgina Garcia  
 Date of Creation: Summer 2011  
 Copyright Notice: The writings of Franklin D. Roosevelt within this collection are in the public domain. The official writings of United States government officials within this collection are in the public domain. The writings of Eleanor Roosevelt within this

Example from the semantic analysis results using OpenCalais demo tool, indicating the entities and social tags generated.

**Social Tags:**

Hawaii	☆☆☆
Film	☆☆☆
Attack on Pearl Harbor	☆☆☆
Pearl Harbor advance-knowledge conspiracy theory	☆☆☆
National Pearl Harbor Remembrance Day	☆☆☆
USS Arizona Memorial	☆☆☆
Battleship Row	☆☆☆
Pearl Harbor	☆☆☆
Geography of the United States	☆☆☆

**Entities:**

- City
- Company
- Country
- Date
- Facility

Finding Aid to the Artificial Collection:  
 Pearl Harbor Attack (Dec 6 – Dec 8, 1941)  
 Size: (.5 cu.ft.)  
 Dates: December 6, 1941 – December 8, 1941  
 Location of Repository: **Franklin D. Roosevelt** Presidential Library  
 Name of **Finding Aid Author:** **Ali Caron** & **Georgina Garcia**  
 Date of Creation: Summer 2011  
 Copyright Notice: The writings of **Franklin D. Roosevelt** within this collection are in the public domain. **The official writings** of **United States** e public domain. The writings of **Eleanor Roosevelt** within this collection are subject to **Mrs. Roosevelt's** literary estate. All other materials 101, et seq.  
 Administrative Note: **Franklin D. Roosevelt Presidential Library** is the first presidential library and the only presidential library used by a s , and photographs pertaining and relating to **President Franklin D. Roosevelt** (FDR). FDR was in office during the events of **Pearl Harbor**  
 Scope and Content: This artificial collection is composed of photocopies of original documents identified from the holdings of the Franklin D. R specifically the events of three days: December 6, 7, and 8, 1941. The items are a collection of documents gathered from other collections fou The criterion for selecting the historical content is solely based on the date range— December 6, 1941 to December 8, 1941. Selected materials include: documents, diaries, telegrams, letters, memoranda, and photographs. Library staff has endeavored to make this research collection as comprehensive as possible; this collection does not represent the entirety of n nts of **Pearl Harbor**. There is a vast amount of documents relating to the lead up to **Pearl Harbor** attack itself, and the aftermath. To simpl available within this collection. **The Pearl Harbor Guide** is available for researchers seeking additional information, including documents re WWII.  
 Provenance: The Pearl Harbor artificial collection includes: **President's Official File** (OF), **President's** Personal File (PPF), **President's Sec**

After

# Research Findings

From 43 archival record groups from 16 institution  
Extracted **8,096** entities and **336** suggested social tags

- Entities correctly identified via Calais analysis included:
  - personal names (Person)
  - corporate names (Company, Facility, Organization)
  - geographic names (City, Continent, Country, Natural Feature, ProvinceOrState, Region)
  - events (Holiday, PoliticalEvent)
- The scores for each identified entity may be used as a valuable clue about the importance of that entity to the overall scope of the archival collection.
- The “Social Tags,” “IndustryTerm,” and “Products” categorizations were the least reliable in terms of accuracy.

The screenshot displays a software interface with several panels. The main panel, titled 'Entities:', lists various categories with checkboxes and progress bars. The categories include City, Company, Country, Date, Facility, Industry Term, Organization, and Person. The 'Country' category is expanded, showing 'United States (Country)' with a relevance of 54%, a count of 2, and coordinates. A red arrow points from the 'Country' header to this entry. The 'Events & Facts:' panel lists 'Armed Attack', 'Generic Relations', and 'Person Career'. The 'Social Tags:' panel lists various tags with star ratings. A yellow box in the top right corner contains the text 'Case: Finding Aids'. A speaker icon is in the bottom right corner.

**Entities:**

- City
  - Honolulu,Hawaii,United States
- Company
- Country
  - China
  - Japan
  - Philippines
  - United States
- Date
  - 1941-12-07
- Facility
- Industry Term
  - copyright law
- Organization
  - Congress
  - Federal Bureau of Investigation
  - First Army
  - Naval Hospital
  - U.S. Air force
  - United States government
  - United States Navy
- Person
  - Ali Caron
  - Article File
  - Bob Clark
  - Charles Fahy
  - Eleanor Roosevelt
  - Francis Biddle
  - Franklin D. Roosevelt
  - Georgina Garcia
  - Henry Morgenthau Jr.
  - Kirsten Carter
  - Navy Photograph
  - Pearl Harbor Attack John Toland
  - Pearl Harbor Bombing
  - Speech File

**Country** **United States (Country)**  
Relevance: 54%  
Count: 2  
latitude: 40.4230003233  
longitude: -98.7372244786

**Events & Facts:**

- Armed Attack
- Generic Relations
  - The Pearl Harbor Guide, available
  - Francis Biddle, paper
  - Pearl Harbor, damage
  - Franklin D. Roosevelt Presidential Library, the first
  - Charles Fahy, paper
  - Charles Fahy, paper
- Person Career

**Social Tags:**

- Hawaii ★★★★★
- Film ★★★★★
- Attack on Pearl Harbor ★★★★★
- Pearl Harbor advance-knowledge conspiracy theory ★★★★★
- National Pearl Harbor Remembrance Day ★★★★★
- USS Arizona Memorial ★★★★★
- Battleship Row ★★★★★
- Pearl Harbor ★★★★★
- Geography of the United States ★★★★★

Case: Finding Aids

Marcia L. Zeng. HELDIG DH Summit 2020

# Additional notes (about using those tools for entity extraction )

## Suggestions based on the Finding Aids study

- It would be well worth the effort for institutions to experiment with semantic analysis methods
  - as an initial step to suggest key entities and topics,  
or
  - as a final check to ensure that important concepts or entities have not been overlooked.
- For certain types of records, particularly those for which subject indexing is not common, semantic analysis may provide entity-based entry points to archival records that were not previously available.
- Such techniques will enhance subject analysis at the first two levels (description and identification) but are unlikely to be useful for interpretation of the material.

Ref:

Zeng, Marcia Lei, Karen Gracy, and Maja Zumer. 2014. Using a semantic analysis tool to generate subject access points: A study using Panofsky's theory and two research samples. *Knowledge Organization* 41(6): 440-451.

Gracy, Karen and Marcia Zeng. 2015. Creating Linked Data within Archival Description: Tools for Extracting, Validating, and Encoding Access Points for Finding Aids. [poster] *DH 2015*, June 29–July 3, 2015, Sydney, Australia.





## 2. Semantic Enrichment For **Semi-structured Data** Outline

[Starting point: existing metadata components that are in free-form]

### Investigations and findings (use entity extraction)

- A. MARC 5xx fields, unstructured notes, etc.
- B. Archival Finding aids' descriptions
- C. **Special Collections and others**
  - 1) Photograph collections' metadata (about the collection)
  - 2) Records of theses and dissertations
  - 3) Museum objects' captions

Semi-structured → to → Structured Data



## 2-C. a) Theses and Dissertations

Another testing by KSU iSchool LOD-LAM team, 2013-14.

Sample: **WHAT**

44 philosophy theses

- a selected sub-sample (22) from KentLINK; and
- a random sample (22) from OhioLINK.

Focus on: **WHERE**

- **abstracts**
- titles
- keywords
- **introduction paragraphs**

**Result:** New structured data generated from semi-structured data

- Semantic analysis based on the abstracts generated more successful tags than those based on the titles.

Process **HOW**

Submitted to OpenCalais separately to obtain the results.

1. All of the candidate terms were counted according to Agent Names, Geographic Names, Corporate Name, and Topic Terms.
2. They were manually validated (by a philosophy master's student, with MLIS) to determine:
  - ✓ the relevance to the thesis,
  - ✓ the type of a term (e.g., named entity, tag, or general heading),
  - ✓ its availability in
    - LCNAF,
    - LCSH,
    - Wikipedia (as an entry),
    - Stanford Encyclopedia of Philosophy.

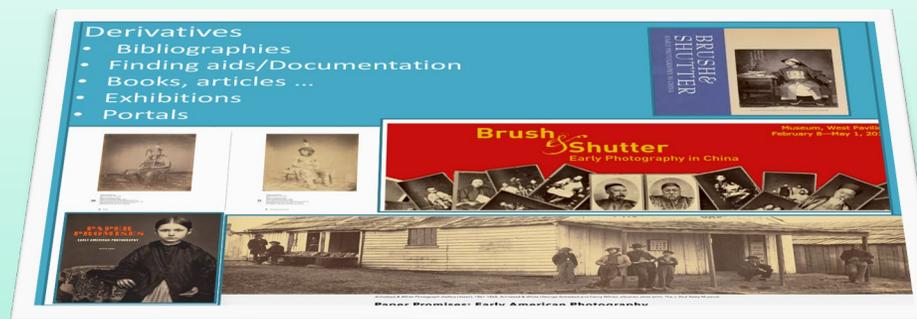


## 2-C. b) Photograph collections' metadata (about the collection)

R&D Derivatives

- Bibliographies
- Finding aids/Documentation
- Books, articles ...
- Exhibitions
- Portals
- Documentaries, media, ...

Collections



Individual images

A photograph collection's metadata record

Semi-structured data [inside of a structured, metadata record]

915 boxes

WHERE

primo.getty.edu/primo\_library/libweb/action/dlDisplay.do?vid=GRI&afte

Books, Journals, Archives, Digital Coll

All items keywords anywhere in the record

**Foto arte minore / Max Hutzel.**  
Hutzel Max.

Available at Special Collections L3 : PHOTO ARCHIVE - HUTZEL (N6911 .H88 )

Request Details

**Title:** Foto arte minore / Max Hutzel.  
**Variant Title:** Max Hutzel photographs of art and architecture in Italy.  
**Author/Creator:** Hutzel Max.  
**Creation Date:** 1960-1990

**Biographical/Historical Note:** German-born photographer and scholar Max Hutzel (1911-1988) photographed in Italy from the early 1960s until his death, resulting in a vast body of photographs referred to by Hutzel as Foto Arte Minore.

**Arrangement:** Arranged by geographic region in Italy, then by province, city, site complex and monument. Some Ancient and Medieval material have been filed in the core collection of the repository's Photo Archive.

**Physical Desc.:** 915 boxes..  
ca. 67,275 photographic prints : b&w ; 24 x 18 cm..  
ca. 86,400 negatives : b&w ; 10 x 13 cm. or smaller..

**Summary:** Max Hutzel's "Foto arte minore" project comprises thorough photographic documentation of art historical development in Italy up to the 18th century, including objects of the Etruscans and the Romans, as well as early Medieval, Romanesque, Gothic, Renaissance and Baroque monuments. Consonant with Hutzel's belief that throughout Italy there are minor artistic centers that deserve attention, sites depicted are frequently obscure and previously undocumented. Hutzel's work is typified by a feeling for place that goes beyond the purely documentary.

Included are thorough interior and exterior documentation of secular buildings, museum holdings, ancient ruins, and religious institutions covering a broad range of artistic forms and styles, including architecture, painting, frescoes, sculpture, manuscripts, metalwork and other minor arts, ranging in date from Antiquity to late Baroque. The regions most heavily represented are: the Abruzzi, Lazio (including Rome), the Marches, and Umbria. Additional photos cover sites in: Basilicata, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Lombardy, Piedmont, Puglia, Sardinia, Tuscany, and Veneto.

The collection contains more than 67,000 black-and-white prints and approximately 86,400 negatives. Because Hutzel carefully cropped his images for printing, the prints more accurately represent his style than the negatives. Also included are circa 825 photographs of medieval buildings and art in Campania, in the ancient region called "Campi Flegrei," made in 1990 by Roberto Sigismondi, Hutzel's long-time assistant (some 800 additional images from this campaign have been removed and interfiled in the Photo Study Collection's Antiquities section.)

**Cumulative Index/Finding Aid:** Finding aid: Online database HUTZEL, in the repository's STAR system, provides a monument-level finding aid for the collection.

**Notes:** Title from associated documentation.

**Form/Genre:** Black-and-white prints (photographs)  
Black-and-white negatives

**Subjects:** Art, Italian  
Art -- Italy  
Architecture -- Italy  
Architecture, Ancient -- Italy  
Architecture, Medieval -- Italy  
Architecture, Renaissance -- Italy  
Architecture, Baroque -- Italy  
Cities and towns -- Italy  
Streets -- Italy  
Painting, Italian  
Sculpture, Italian  
Decorative arts -- Italy

**Contributors:** Sigismondi Roberto.

**LC Call Number:** N6911 .H88

**ID/Accession Number:** 94-F92  
86.P.8

**Access/Rights:** Open for use by qualified researchers.  
Photographs and permission to publish must be obtained from copyright holder(s).

**OCLC Record Number:** 406995970

**Persistent Link:** http://primo.getty.edu/GRI:GETTY\_ALMA21124005810001551

Back to results list

http://primo.getty.edu/primo\_library/libweb/action/dlDisplay.do?vid=GRI&afterPDS=true&institution=GETTY&docId=GETTY\_ALMA21124005810001551



Before

Original descriptive metadata: →

**Cumulative Index/Finding Aid:** Finding aid: Online database HUTZEL, in the system, provides a monument-level finding aid.

**Notes:** Title from associated documentation.

**Form/Genre:** Black-and-white prints (photographs)  
Black-and-white negatives

**Subjects:**  
 Art, Italian  
 Art -- Italy  
 Architecture -- Italy  
 Architecture, Ancient -- Italy  
 Architecture, Medieval -- Italy  
 Architecture, Renaissance -- Italy  
 Architecture, Baroque -- Italy  
 Cities and towns -- Italy  
 Streets -- Italy  
 Painting, Italian  
 Sculpture, Italian  
 Decorative arts -- Italy

**Contributors:** Sigismondi Roberto.

**LC Call Number:** N6911 .H88

**ID/Accession Number:** 94-F92  
86.P.8

**Access/Rights:** Open for use by qualified researchers.  
Photographs and permission to publish must be obtained from the copyright holder(s).

**OCLC Record Number:** 406995970

**Persistent Link:** http://primo.getty.edu/GRI:GETTY\_ALM,21

[Back to results list](#)

**FOUND IN DOCUMENT**

- Music Album
- Person
- Position
- Province Or State
  - Basilicata,Italy 20%
  - Campania,Italy 20%
  - Emilia-Romagna,Italy 20%
  - Lazio,Italy 20%
  - Lombardy,Italy 20%
  - Piedmont,Oregon,United... 20%
  - Sardinia,Italy 20%
  - Tuscany,Italy 20%
  - Umbria,Italy 20%
  - Veneto,Italy 20%

**DOCUMENT VIEW**

Max Hutzel's "Foto arte minore" project comprises thorough photographic documentation of art historical development in Italy up to the late 19th century. The collection includes photographs of Medieval, Romanesque, Gothic, and Renaissance architecture, painting, fresco, and sculpture. Throughout Italy there are many ancient ruins, and religious architecture, painting, fresco, and sculpture. The collection includes Antiquity to late Baroque. The collection includes photographs of medieval buildings and art in Campania, in the ancient region called "Campi Flegrei," made in 1990 by Roberto Sigismondi, Hutzel's long-time assistant (some 800 additional images from this campaign have been included in the Photo Study Collection's Antiquities section).

ans, as well as early Hutzel's belief that frequently obscure and the purely s, museum holdings, including going in date from including Rome), the

shortname Basilicata  
 latitude 40.5908  
 longitude 16.0933  
 containedbycountry Italy  
 forenduserdisplay false

Relevance 20%

Additional photos cover sites in: Basilicata, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Lombardy, Piedmont, Puglia, Sardinia, Tuscany, and Veneto. The collection contains more than 67,000 black-and-white prints and approximately 86,400 negatives. Because Hutzel carefully cropped his images for printing, the prints more accurately represent his style than the negatives. Also included are circa 825 photographs of medieval buildings and art in Campania, in the ancient region called "Campi Flegrei," made in 1990 by Roberto Sigismondi, Hutzel's long-time assistant (some 800 additional images from this campaign have been included in the Photo Study Collection's Antiquities section).

[Upload Again](#) [View RDF](#)

```
<rdf:Description rdf:about="http://d.opencalais.com/er/geo/provinceorstate/ra1g-geo1539693a2-eb41-f766-dd0e-c8b9595c90a2">
  <rdf:type rdf:resource="http://s.opencalais.com/1/type/er/Geo/ProvinceOrState"/>
  <c:docId rdf:resource="http://d.opencalais.com/dochash-1/dd44e538-f3a0-35fc-8422-7d7f76aa6c04"/>
  <c:name>Basilicata,Italy</c:name>
  <c:shortname>Basilicata</c:shortname>
  <c:latitude>40.5908</c:latitude>
  <c:longitude>16.0933</c:longitude>
  <c:containedbycountry>Italy</c:containedbycountry>
  <!--Basilicata-->
  <c:subject rdf:resource="http://d.opencalais.com/genericHasher-1/00aa24c0-6b83-3042-88a8-9aed5a25efdc"/>
</rdf:Description>
```

Each place name embeds the full name, short name, latitude & longitude, and what country contains this place.

# Demo: Photograph collection

COGITO

Before

Description about an annotated bibliography

[http://www.getty.edu/research/tools/guides\\_bibliographies/photography\\_china/](http://www.getty.edu/research/tools/guides_bibliographies/photography_china/)

After

Using another tool, COGITO Intelligence API

[Paragraph in an annotated bibliography]

<http://www.intelligenceapi.com/>



**Canton** ?  
 GEO Coordinate » 23.12861/113.25889  
 Facts in this document » [China](#)  
 Region » [Guangdong Province](#)  
 Country » [China](#)  
 Continent » [Asia](#)  
 Sentiment » [Neutral](#)

[China](#)

[Beijing](#)

[Shanghai](#)

[Cat](#)

[Southeast Asia](#)

[Europe](#)

[Macau](#)

[Hong Kong](#)

[View Woo-Chow City](#)

[Lai Afong](#)

? This window displays the **Places** entities with their geographic coordinates (left-hand column) along with 2 graphics, the semantic relationships and the Fact Mining details (right-hand column). You can switch between the relationships and the Fact Mining graphs with the buttons below the graphical map.

## Canton



No relations found

Relations  Facts

capital

After

Demo: Photograph collection

COGITO

**China (Asia)** ?  
 GEO Coordinate » 35.0/105.0  
 Continent » [Asia](#)  
 People » [William Saunders](#), [Clark Worswick](#)  
 Organizations » [University Press](#), [View](#), [Getty Research Institute](#), [Qing](#), [Communist Party](#)  
 Principal limited companies » [University Press](#)  
 Places » [China](#), [Canton](#), [Macau](#), [Hong Kong](#), [Beijing](#), [Southeast Asia](#), [Shanghai](#), [Europe](#)

? Fact mining identifies and groups the facts with the relative entities involved. The left-hand column lists the facts extracted from the analyzed text. The different colors depend on the taxonomies of reference (see Categorization tab). For each single fact selected in the left column, the right-hand column displays the related topics and entities above each sentence (atomic knowledge unit). With the new time references extraction (beta version), you can locate facts on a time span providing a new temporal analysis perspective.

? The left column shows the "Old", "Recent", "Present" and "Future" Time References including the retrieved temporal wordings divided by subtypes (i.e. "Absolute", "Relative" and "Range" forms). On the right panel instead, all the sentences respectively containing the temporal collocations are shown with all of the domain entities and topics thus gathered in the phrases. (beta vers.)

**OLD**  
 Absolute » [1911](#), [Feb-8-May-1-2011](#), [2003](#), [1839](#), [After 1949](#),  
 Range » [1852-1892](#), [until the last decade of the twentieth century](#), [1840-1911](#), [in the early twenty-first century](#), [1856-1860](#), [1937-1945](#), [From 1949](#)  
 Relative » [formerly](#), [The time period](#)

?  
 China, [1911](#) (OLD), Main elements (China)  
 History of Photography in China, 1839-ca. 1911: Selected Annotated Bibliography

China, [Getty Museum](#), [Feb-8-May-1-2011](#) (OLD), [Feb-8](#), [May-1-2011](#), Main elements (China)  
 Created in conjunction with the exhibition Brush & Shutter: Early Photography in China at the J. Paul Getty Museum, February 8-May 1, 2011, and the accompanying catalog of the same name

China, [Getty Museum](#), [Feb-8-May-1-2011](#) (OLD), [Feb-8](#), [May-1-2011](#), Main elements (China)  
 It was created in conjunction with the exhibition Brush & Shutter: Early Photography in China at the J. Paul Getty Museum, February 8-May 1, 2011, and the accompanying catalog of the same name.

[Clark Worswick](#), [China](#), [Southeast Asia](#), [Getty Research Institute](#), [2003](#), Main elements (China)  
 Clark Worswick Collection of Photographs of China and Southeast Asia, the Getty Research Institute, 2003.R.22.27

[The time period](#), [1839](#), [after 1839](#), [China](#), [of 1839](#), [1856-1860](#), [1856](#), [1860](#), [treaty port](#), [Qing](#), [half of the nineteenth century](#), [in 1911](#), [1911](#), Main elements (China)  
 The time period for this bibliography - 1839 to ca.1911 - was chosen because it coincides roughly with the blossoming of the photography medium after 1839, the burgeoning power of the photographic process in the West after 1839, and the 1850s-1860s when the

ures » [treaty port](#)  
 Interest » [Getty Museum](#)  
 » [Getty Museum](#), [Getty Research Institute](#), [treaty port](#)  
 old (beta vers.) » [1911](#), [Feb-8-1-2011](#), [formerly](#), [1832-1892](#), [2003](#), [until the last decade of the twentieth century](#), [1840-1911](#), [in the early twenty-first century](#), [The time period](#), [after 1839](#), [of 1839](#), [1856-1860](#), [half of the nineteenth century](#), [in 1911](#), [1937-1945](#), [1945-1949](#), [After 1949](#), [From 1949](#)

[Feb-8](#), [May-1-2011](#), [1832](#), [1892](#), [2003](#), [1840](#), [1839](#), [1856](#), [1860](#), [1937-1945](#)

China, [1911](#) (OLD), Main elements (China)  
 History of Photography in China, 1839-ca. 1911: Selected Annotated Bibliography

China, [Getty Museum](#), [Feb-8-May-1-2011](#) (OLD), [Feb-8](#), [May-1-2011](#), Main elements (China)  
 Created in conjunction with the exhibition Brush & Shutter: Early Photography in China at the J. Paul Getty Museum, February 8-May 1, 2011, and the accompanying catalog of the same name

China, [Getty Museum](#), [Feb-8-May-1-2011](#) (OLD), [Feb-8](#), [May-1-2011](#), Main elements (China)  
 It was created in conjunction with the exhibition Brush & Shutter: Early Photography in China at the J. Paul Getty Museum, February 8-May 1, 2011, and the accompanying catalog of the same name.

Canton, [formerly](#) (OLD), Main elements (Guangzhou, Canton)  
 For example, Guangzhou is the name for the city formerly known as Canton.

Macau, [Hong Kong](#), Main elements (Macau, Hong Kong)  
 However, in the cases of Macau and Hong Kong, we have retained the earlier accepted romanizations.

Hong Kong, [University Press](#), Main elements (Hong Kong)  
 A Chinese-language version of this bibliography may be found in the translation of the catalog published by Hong Kong University Press.

The Writprint of the COGITO estimates the **readability level** of a provided document collecting and forging a full set of readability indexes as well as **grammatical, lexical and semantic analysis scores**.

primo.getty.edu/primo\_library/libweb/action/dlDisplay.do?vid=GRI&aft...

Books, Journals, Archives, Digital Col...

All items keywords anywhere in the record

**Foto arte minore / Max Hutzel.**  
Hutzel Max.

Available at Special Collections L3 : PHOTO ARCHIVE - HUTZEL (N6911 .H88 )

Request Details

**Title:** Foto arte minore / Max Hutzel.  
**Variant Title:** Max Hutzel photographs of art and architecture in Italy.  
**Author/Creator:** Hutzel Max.  
**Creation Date:** 1960-1990

**Biographical/Historical Note:** German-born photographer and scholar Max Hutzel (1911-1988) photographed in Italy from the early 1960s until his death, resulting in a vast body of photographs referred to by Hutzel as Foto Arte Minore.

**Arrangement:** Arranged by geographic region in Italy, then by province, city, site complex and monument. Some Ancient and Medieval material have been filed in the core collection of the repository's Photo Archive.

**Physical Desc.:** 915 boxes..  
ca. 67,275 photographic prints : b&w ; 24 x 18 cm..  
ca. 86,400 negatives : b&w ; 10 x 13 cm. or smaller..

**Summary:** Max Hutzel's "Foto arte minore" project comprises thorough photographic documentation of art historical development in Italy up to the 18th century, including objects of the Etruscans and the Romans, as well as early Medieval, Romanesque, Gothic, Renaissance and Baroque monuments. Consonant with Hutzel's belief that throughout Italy there are minor artistic centers that deserve attention, sites depicted are frequently obscure and previously undocumented. Hutzel's work is typified by a feeling for place that goes beyond the purely documentary.

Included are thorough interior and exterior documentation of secular buildings, museum holdings, ancient ruins, and religious institutions covering a broad range of artistic forms and styles, including architecture, painting, frescoes, sculpture, manuscripts, metalwork and other minor arts, ranging in date from Antiquity to late Baroque. The regions most heavily represented are: the Abruzzi, Lazio (including Rome), the Marches, and Umbria. Additional photos cover sites in: Basilicata, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Lombardy, Piedmont, Puglia, Sardinia, Tuscany, and Veneto.

The collection contains more than 67,000 black-and-white prints and approximately 86,400 negatives. Because Hutzel carefully cropped his images for printing, the prints more accurately represent his style than the negatives. Also included are circa 825 photographs of medieval buildings and art in Campania, in the ancient region called "Campi Flegrei," made in 1990 by Roberto Sigismondi, Hutzel's long-time assistant (some 800 additional images from this campaign have been removed and interfiled in the Photo Study Collection's Antiquities section.)

Before

Writprint estimates the readability level of a provided document collecting and forging a full set of readability indexes as well as grammatical, lexical and semantic analysis scores. The left-hand column shows a few graphs to concretely assess the document's reading difficulty, the writer's vocabulary level, the minimum schooling grade needed to read the document flawlessly, the spotted slangs and registers, and a table showing accurate and focused text analysis. The right-hand column shows a graph which associates a set of highly validating readability indexes and marks the author's style and efficacy. The two graphs below the former display a narrow grammatical inquiry centered on the verbs' genre and their tense analysis.

**Readability Index**  
50.38  
DIFF 25 50 75 EASY

**Vocabulary richness**  
31.61  
POOR 18 36 54 72 RICH

**Grade**  
Primary school  
Middle school  
Highschool  
College/University

**Registers and Slangs**  
Cyber Illegal Slang 1.79  
Foreign 1.08  
Literary 0.72  
Military Slang 0.36  
Abbreviation 0  
Slang 0  
Social Slang 0  
Criminal Enterprise Slang 0

**Text Statistics**

Index	Value	Standard Values
Sentences Count	9.0	-
Words per Sentence	31.0	10.06-16.92
Characters per Sentence	187.89	55.34-109.92
Short Words Count	14.79	5.20-9.50 %
Different Words Count	79.41	66.10-81.26 %
Uncommon Words Count	7.35	5.22-9.62 %
Common Vocabulary Weight	65.44	42.04-100.00 %
High Vocabulary Weight	7.03	0.96-1.15 %
Technical Vocabulary Weight	3.68	58.70-68.44 %
Nouns Count	23.3	8.32-11.67 %
Verbs Count	8.6	2.94-4.96 %
Adjectives Count	11.83	9.92-12.97 %
Conjunctions Count	4.66	2.36-4.58 %
Adverbs Count	5.02	2.06-4.68 %

**Writprint**  
Short Words  
Lexical Differentiation  
Common Vocabulary  
Adverbs Weight  
High Vocabulary  
Conjunctions Weight  
Nouns Weight  
Adjectives Weight  
standard this text

**Grammatical Tenses**  
Simple Present  
Ed Form  
Simple Past  
Ing Form  
Present Perfect  
Future  
Undefined  
Conditional  
Past Perfect

**Verb Classes**  
Linguistic Communication  
Generic Communication  
Informatica Profession  
Shift Place  
Law  
Movement

After



# 2-C. c) Museum objects' captions

Demo: Museum object captions

## Blue and White plum vase of the four loves in Yuan Dynasty



Before

元青花四爱图梅瓶  
 元代。2006年钟祥市郢靖王墓出土。高38.7厘米，口径6.4厘米，底径13厘米。梅瓶腹部分别绘王羲之爱兰图，陶渊明爱菊图，周敦颐爱莲图，林和靖爱梅、鹤图。

The stories of four ancients and their favorites are painted on the belly of the vase, which are respectively Wang Xizhi loves the orchid, Tao Yuanming loves the chrysanthemum, Zhou Dunyi loves the lotus and Lin Jing loves plum blossom and the crane.

Yuan Dynasty (1271-1368)

After

Tool: Boson

实体识别: [查看文档](#) [结果不正确](#)

更多 较少 **平衡** 准确 更准确

实体类别图示: 时间 地点 人名

← Entity types

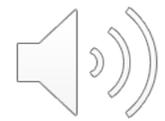
TIME, PLACE, AGENT

Entities →

keywords

关键词提取: [查看文档](#) [结果不正确](#)

名称	权重	名称	权重	名称	权重
图梅瓶	35	郢靖王	24	王羲之	19
梅瓶	33	厘米	24	元代	19
钟祥市	32	周敦颐	23	出土	18
底径	32	青花	22	口径	18
林和靖	24	陶渊明	20	腹部	17



## Blue and White plum vase of the four loves in Yuan Dynasty

Source: hubei.gov.cn 08/26/2016 09:08:27



Time: Yuan Dynasty.

Unearthed from :King Yingjing's Tomb in Chongxiang city in 2006.

Height: 38.7cm,

Surface diameter: 6.4cm,

Bottom diameter: 13cm.

The stories of four ancients and their favorites are painted on the belly of the vase, which are respectively Wang Xizhi loves the orchid, Tao Yuanming loves the chrysanthemum, Zhou Dunyi loves the lotus and Lin Jing loves plum blossom and the crane.

[http://en.hubei.gov.cn/culture/heritage/201608/t20160826\\_889291.shtml](http://en.hubei.gov.cn/culture/heritage/201608/t20160826_889291.shtml)

Before

After

Tool: OpenCalais

FOUND IN DOCUMENT

- ENTITIES
  - Person
    - Lin Jing 80%
    - Tao Yuanming 80%
    - Wang Xizhi 80%
- SOCIAL TAGS
  - Chinese people 100%
  - Chinese culture 100%
  - Tao Yuanming 100%
  - Archaeplastida 66%
  - Prunus mume 66%
  - Chrysanthemum 66%
  - Zhou Dunyi 66%
  - Wang 66%
  - Padma 66%

Tool: COGITO

MAIN ELEMENTS

belly of the vase vase favorite story love the orchid orchid

crane white lotus darling abdomen plum Lin Jing

Tao Yuanming blossom lotus Wang Xizhi ancient

love plum blossom belly Zhou Dunyi love the lotus

love the chrysanthemum chrysanthemum plum tree



# Semantic analysis tools

- taxonomy and ontology-supported
- with machine learning and natural language processing behind

## [Tools tested]

- **Intelligent Tagging** (previously known as **Open Calais**) Demo
  - <https://permid.org/onecalaisViewer>
- **Cogito Intelligence API** free demo version
  - <https://www.intelligenceapi.com/demo/>

## [Other tools]

- **Ambiverse Natural Language Understanding - AmbiverseNLU**
  - <https://github.com/ambiverse-nlu/ambiverse-nlu>
  - → Try the [demo](http://ambiversenlu.mpi-inf.mpg.de) at <http://ambiversenlu.mpi-inf.mpg.de>
- **Gate cloud**
  - <https://cloud.gate.ac.uk/shopfront/sampleServices>
- **Spacy**
  - <https://spacy.io/usage/linguistic-features#entity-linking>

## Cogito Intelligence API:

- **5 specific taxonomies** of terms (in over 1,000 different categories) for Intelligence, Terrorism, Cyber Crime, Crime, and Geographic domains
- **A domain ontology (updated regularly)** with a wide range of diverse topics, for example: weapons, crimes, cyber attacks, points of interest, chemical weapons, controlled substances, terrorist groups, critical infrastructure, world leaders, public companies and more

<https://www.intelligenceapi.com/>





# The take-aways so far

- Semantic analytics, one of the advanced semantic enrichment methods, has been used for analyzing, searching, and presenting information by using **explicit semantic relationships** between known entities.
- The tools used in the experiments are powered by multiple **taxonomies and domain ontologies**, and benefit from **machine learning** and other new **artificial intelligence (AI)** technologies, far beyond normal natural language processing.
- On top of rule-based systems, embedding-based systems **for Knowledge Graph** completion has become a dominating focus in research and development during recent years.



# The take-aways so far (cont.)

- The examples reveal that, ultimately, additional useful data can be derived from **large digital collections** as well as from **individual item-centered information clusters**.
- These activities can be managed **case-by-case**.
  - from the top-down or the bottom-up
  - collectively or independently
  - with or without significant project funding
- Aggregation can be based on the pieces/chunks of information one needs from a dataset, without integrating a whole database or converting full metadata records.

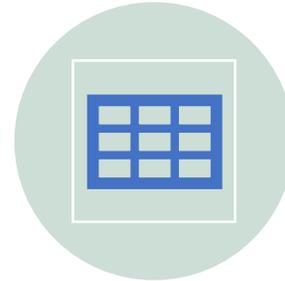


## Revisit

# Why use data from semi-structured data resources?



An important feature of semi-structured data resources that should be recognized, is that **they are the products of information processing.**



These semi-structured data **represent the accumulated time, knowledge, and experience of the creators** who generated them through a formal workflow which conforms to professional standards and best practices.



With semantic enrichment processes, the data values in semi-structured data are **contextualized through the metadata elements/fields;** hence, the function and meaning are clearly implied.



By parsing these data through advanced information technologies, these LAM data are dramatically **enriched and are converted into new access points.**

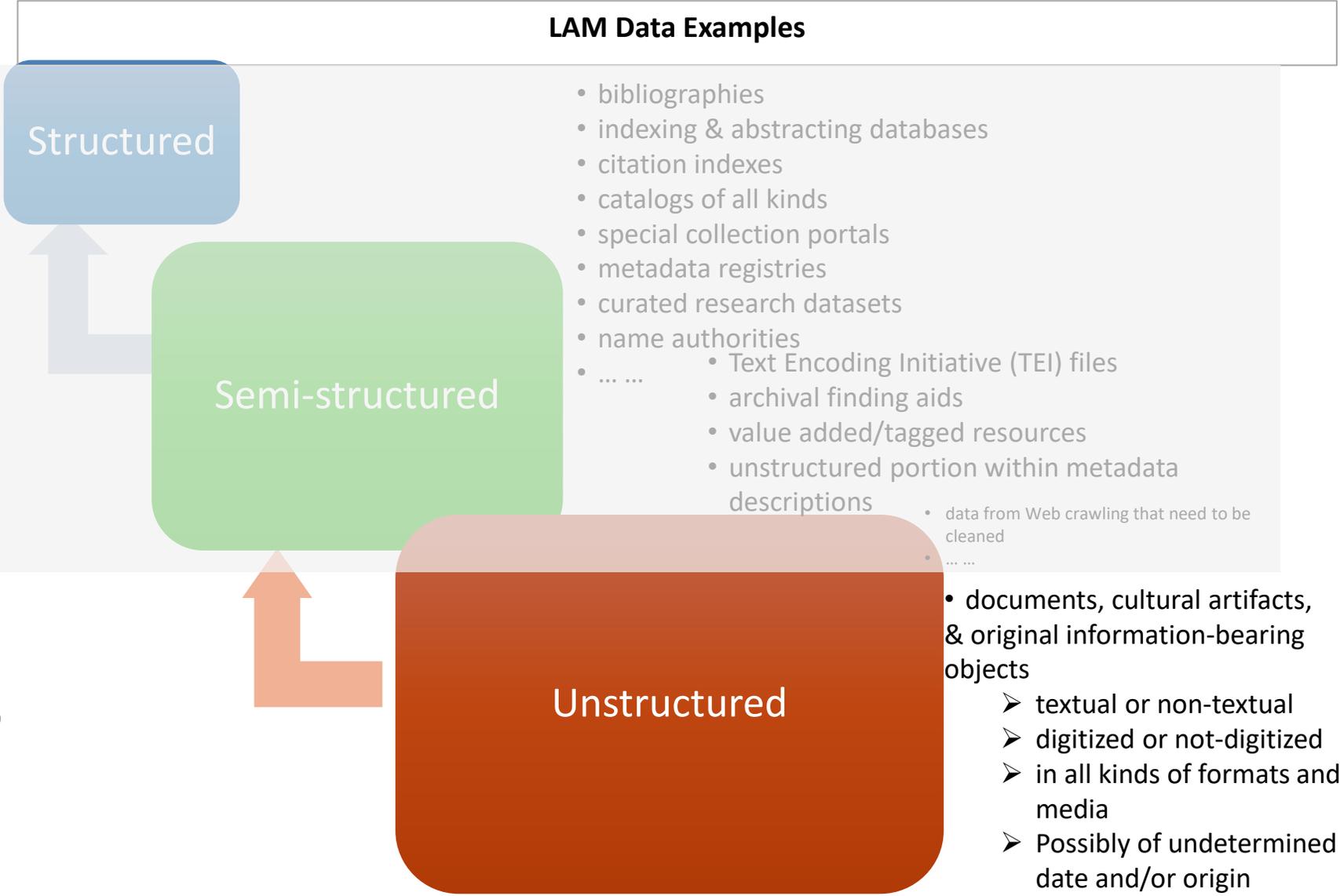


### 3. Semantic Enrichment for **Un-structured Data**





3





### 3. Semantic Enrichment for Un-structured Data

[Outline](#)

#### Examples:

- A. Oral history transcripts
- B. Images
- C. Maps
- D. Murals
- E. Cultural objects
- F. Intangible Cultural Heritages
- G. ...

New structured data generated from **unstructured data** supporting knowledge discovery

Content-based, Semantic-based (vs. word-based)

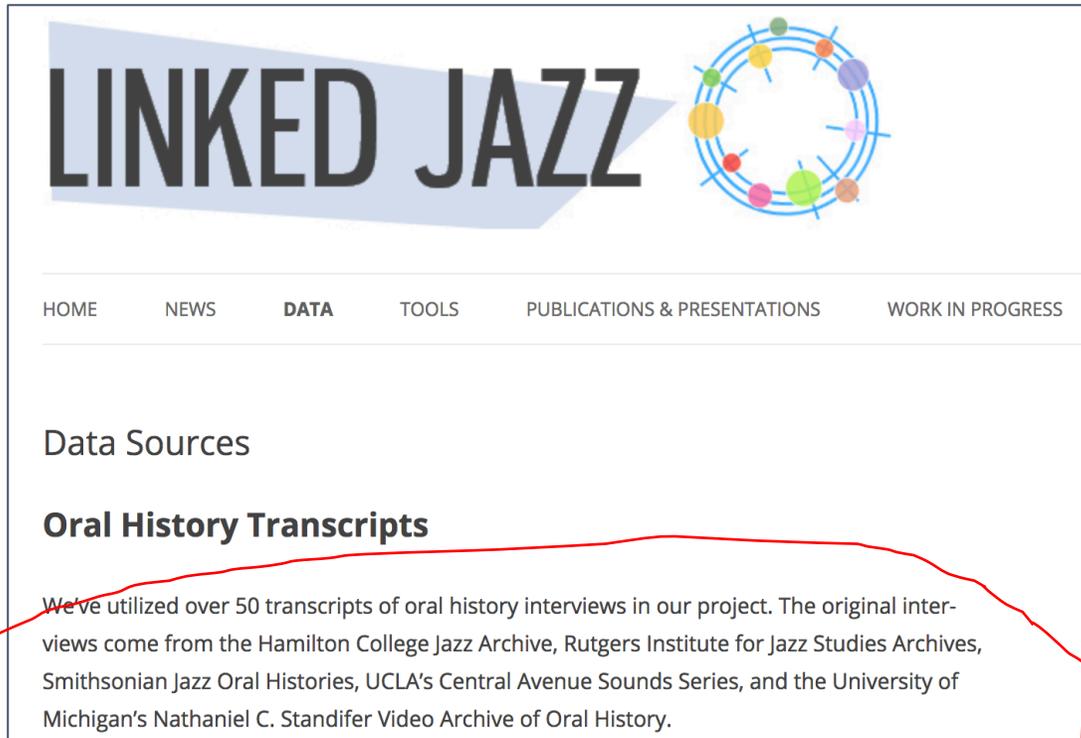


## 3-A. Oral history transcripts

# Case: Linked Jazz

<http://linkedjazz.org/>

The project focuses on digitalized archives of jazz history to expose relationships between musicians and reveal their community's network.



The screenshot shows the Linked Jazz website interface. At the top, the text 'LINKED JAZZ' is displayed in a large, bold, sans-serif font. To the right of the text is a circular logo composed of several colored dots (yellow, green, blue, purple, red, pink) connected by thin lines, forming a network. Below the logo is a navigation menu with the following items: HOME, NEWS, DATA, TOOLS, PUBLICATIONS & PRESENTATIONS, and WORK IN PROGRESS. Underneath the navigation menu, the heading 'Data Sources' is visible, followed by the sub-heading 'Oral History Transcripts'. A red hand-drawn circle highlights the following text: 'We've utilized over 50 transcripts of oral history interviews in our project. The original interviews come from the Hamilton College Jazz Archive, Rutgers Institute for Jazz Studies Archives, Smithsonian Jazz Oral Histories, UCLA's Central Avenue Sounds Series, and the University of Michigan's Nathaniel C. Standifer Video Archive of Oral History.'

<http://linkedjazz.org/network/>

Marcia L. Zeng. HELDIG DH Summit 2020

### Highlights

- Pioneer LOD project
- Powered by OCR and NLP

- 50+ transcripts of oral history interviews
- from 5 archives/ special collections

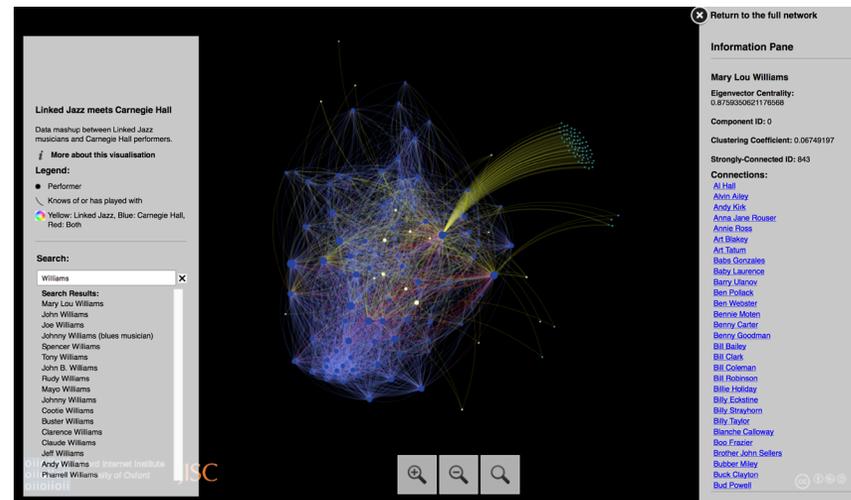




# Data Interlinking: Linked Jazz and Carnegie Hall

- The Carnegie Hall Archives - performance history, Carnegie Hall, 1891-
- [Performance History Search](#) online, 2013 –
  - ~ 50,000 unique events encompassing roughly 90,000 people/performers
  - a significant contribution to researchers of music and cultural history.
- Two data sources selected from ‘jazz’ Carnegie Hall 1912-1955 events, RDF triples describing:
  1. 19197 people and their associated data (instrument played, birth/death date and location, profession)
  2. 154 jazz [top-level]events (performer and group names, date, place (e.g. main hall), and title (top of concert program))

- [Visualized result](#) in Gephi.  
(Data mashup between Linked Jazz musicians and Carnegie Hall performers.)
  - E.g., Mary Lou Williams



[http://pfch.nyc/linked\\_jazz\\_meets\\_carnegie\\_hall/CH-LJ\\_network/index.html](http://pfch.nyc/linked_jazz_meets_carnegie_hall/CH-LJ_network/index.html)

[http://pfch.nyc/linked\\_jazz\\_meets\\_carnegie\\_hall/CH-LJ\\_network/index.html#Mary%20Lou%20Williams](http://pfch.nyc/linked_jazz_meets_carnegie_hall/CH-LJ_network/index.html#Mary%20Lou%20Williams)

## Highlights

- Pioneer LOD project
- Powered by OCR and NLP
- Named entity centered
- Contextualization supported by ontology
- Recognized by music community
- Connected with / reused by Carnegie Hall project



<https://iiif.io/>

IIIF Image API 3.0

IIIF Presentation API 3.0

IIIF Authentication API 1.0

IIIF Search API 1.0

Base URL: {scheme}://{host}/{prefix}/{identifier}

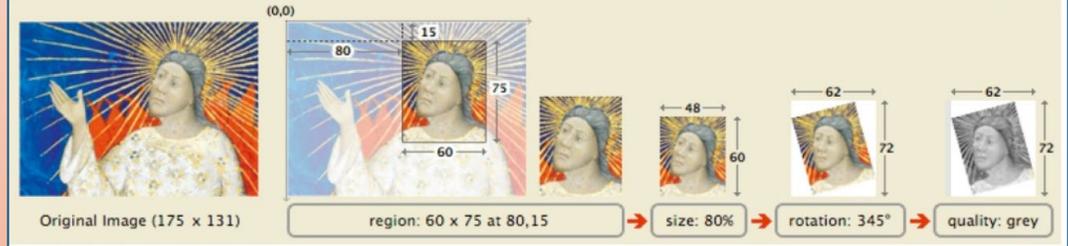
Image Resource:

{base}/{region}/{size}/{rotation}/{quality}.{format}

Order of Implementation

<http://www.example.org/image-service/abcd1234/80,15,60,75/pct:80/345/grey.jpg>

region
size
quality
rotation format



## IIIF: Extend DMS To...



Books



Manuscripts



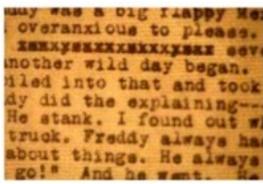
Newspapers



(Sheet) Music



Art / Vis. Resources



Archival Materials



Maps



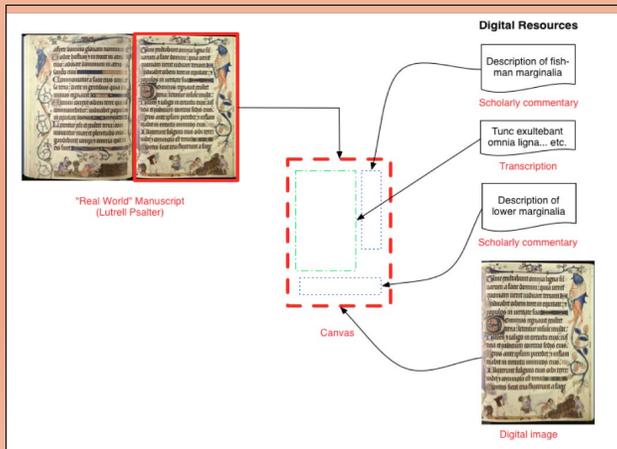
STEM Imagery



Scrolls



Architecture



Using the supported tools, annotations by experts or other contributors can be added to the canvas. Annotations, annotation lists, and content all have unique identifiers that can be processed by machines without confusion.



# Structural and Contextual Views on the WCD (Wooden Slips Character Dictionary) System

Academia Sinica Center for Digital Cultures

- Core functions in WCD platform
  - \* **Metadata Search** by object and by character
  - \* **Image Research** on objects and characters
  - \* **Image annotation** by classified categories
  - \* **Cross-database query** for Chinese characters



Wooden Slips Item H01537

The written characters are living cultural evidence on the

- military and legal systems,
- educational practices,
- economy, beliefs, and
- everyday life

of military personnel and civilians in the Han dynasty (206 BC-AD 220).



Image comparison of the characters “Jia chu” (甲渠), name of a military base

WCD Website: <https://wcd-ihp.ascdc.sinica.edu.tw/woodslip/index.php>

# WCD (Wooden Slips Character Dictionary): Image Annotation



- Annotation under different motivations and sub-types (an extension to IIIF APIs).

## Motivation types

- (1) Tagging
- (2) Describing
- (3) Commenting
- (4) Classifying

- The method of classified annotation in different sub-categories makes the **annotated text more reusable for scholars' needs** by saving object data or interpreting the form and meaning of the characters.

The character “Jia” (甲) can be annotated under different motivations such as tagging, describing, commenting, or classifying (Item H04737)

See on the webpage: <https://wcd-ihp.ascdc.sinica.edu.tw/woodslip/item.php?id1=H04737>



# IIF-based Union Catalog of WCD

歷史文字資料庫統合檢索系統

繁體中文

《歷史文字資料庫統合檢索系統》介紹 使用方法

查詢文字：甲

重新查詢 查詢文字  開始查詢

· 請輸入欲查詢的文字。(限中文單字)

Input field for cross-datasets retrievals

Academia Sinica, R.O.C.

聯珠字典—中央研究院歷史語言研究所|中央研究院數位文化中心

查詢結果：41筆 顯示全部圖檔

Nara Research Institute, Japan

木簡庫—奈良文化財研究所

查詢結果：12筆

University of Tokyo, Japan

草字字典數位資料庫—東京大學史料編纂所

查詢結果：7筆

Character in original database

Display in IIF Manifest JSON-LD

Display on Mirador

- “**Union Catalog**” for searching historical Chinese Characters in cooperation with international research communities in Asia.
- Functions of WCD’s Union Catalog
  - \* **Character retrieval across institutes**
  - \* Redirection to original database
  - \* Access to the IIF Manifest structure of retrieved characters
  - \* Presentation of retrieved characters in Mirador viewer
- System functions based on **IIF APIs** and **customized API for sharing of the search results**

Ref: Chen, S. & Lu, L. 2020, Linked Data as Method for Supporting DH-Research on the Cultural Resources of Chinese Wooden Slips and the Interpretation of Ancient Chinese Characters. DCMI 2020.





# OCLC IIF Explorer (with CONTENTdm)

OCLC ResearchWorks IIF Explorer

Search for images of people, places, events, and more.

Welcome to the IIF Explorer

<https://researchworks.oclc.org/iif-explorer/>

OCLC ResearchWorks has created an index of all of the images in the CONTENTdm digital content management systems hosted by OCLC. The IIF Explorer offers a way to search across those collections. This prototype is called the "IIF Explorer" because the CONTENTdm images are associated with documents that use the IIF (International Image Interoperability Framework) "Presentation Manifest" standard. IIF Presentation Manifests and related images are presented with a viewer provided by Project Mirador. The IIF Explorer is an experimental prototype system, available for testing and evaluation. Please note that data in the IIF Explorer index is not guaranteed to be current with the source CONTENTdm collections, facets for type, audience, creator, etc. are partially representative of the source data based on preliminary metadata analysis and reconciliation, and the user interface is in early stages of development and testing.

Some sample searches

 The Lockheed Vega airplane	 1906 San Francisco Earthquake and Fire	 Illuminated Medieval Manuscripts	 Labor Strike Photographs and Texts	 Maps of Paris, France and its environs
 The Yellowstone Expedition	 Photographs by Carleton Watkins	 Woodcut Works and Illustrations	 Monorail Photographs and Illustrations	 Tanglewood Music Center

The IIF Explorer is an OCLC ResearchWorks prototype application provided under these terms of use.

## Aggregators

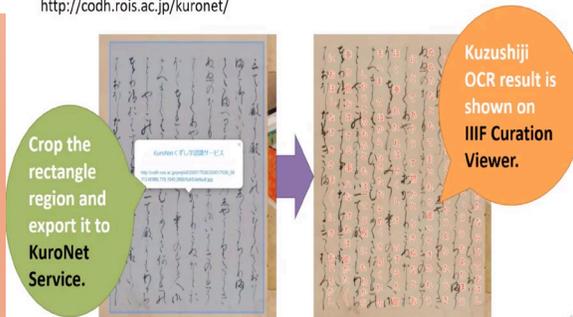
- CONTENTdm
- Artstor
- DPLA (Digital Public Library of America)
- Europeana
- Internet Archive
- Wikimedia Foundation

Endless showcases!

<https://www.youtube.com/channel/UClcQIkLdYra7ZnOmMJnC50A/videos>

(AI OCR)

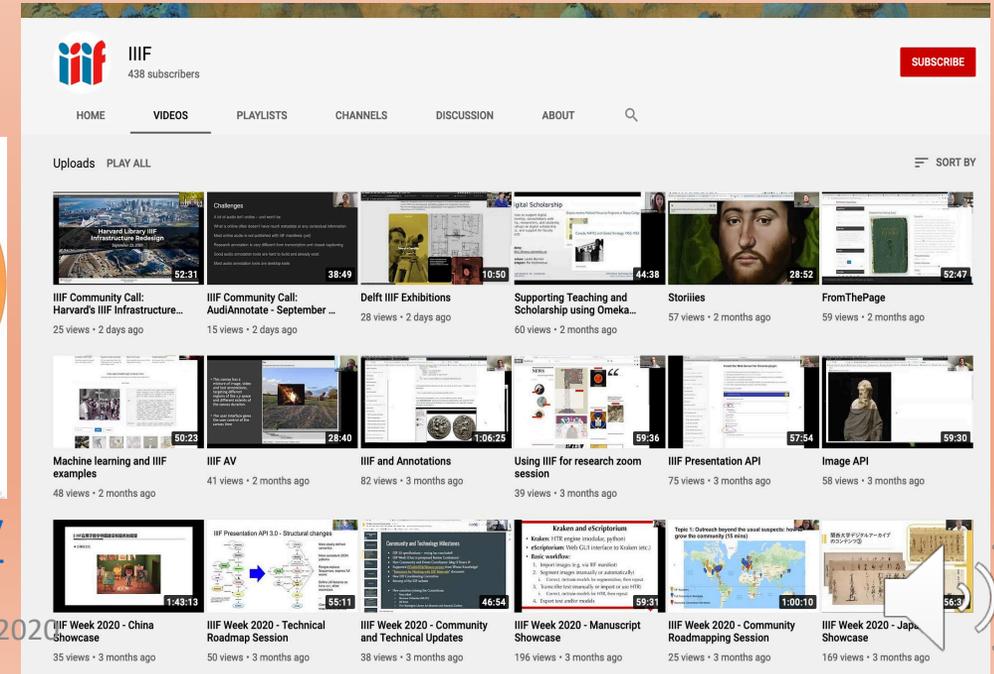
KuroNet Kuzushiji Recognition Service  
<http://codh.rois.ac.jp/kuronet/>



Crop the rectangle region and export it to KuroNet Service.

Kuzushiji OCR result is shown on IIF Curation Viewer.

<http://codh.rois.ac.jp/kuronet/>



IIF 438 subscribers

HOME VIDEOS PLAYLISTS CHANNELS DISCUSSION ABOUT

Uploads PLAY ALL SORT BY

- IIF Community Call: Harvard's IIF Infrastructure... 25 views • 2 days ago
- IIF Community Call: AudiAnnotate - September... 15 views • 2 days ago
- Delft IIF Exhibitions 28 views • 2 days ago
- Supporting Teaching and Scholarship using Omeka... 60 views • 2 months ago
- Stories! 57 views • 2 months ago
- FromThePage 59 views • 2 months ago
- Machine learning and IIF examples 48 views • 2 months ago
- IIF AV 41 views • 2 months ago
- IIF and Annotations 82 views • 3 months ago
- Using IIF for research zoom session 39 views • 3 months ago
- IIF Presentation API 75 views • 3 months ago
- Image API 58 views • 3 months ago
- IIF Week 2020 - China Showcase 35 views • 3 months ago
- IIF Week 2020 - Technical Roadmap Session 50 views • 3 months ago
- IIF Week 2020 - Community and Technical Updates 38 views • 3 months ago
- IIF Week 2020 - Manuscript Showcase 196 views • 3 months ago
- IIF Week 2020 - Community Roadmapping Session 25 views • 3 months ago
- IIF Week 2020 - Japan Showcase 169 views • 3 months ago



New: 3D Community

[iif.io/community/groups/3d](http://iif.io/community/groups/3d)

# Maphub

## a historical map annotation portal

- Annotation
- Multilingual searching
- Geo-referencing
- Map overlays



### Highlights

- Annotation through mash-up culture
- Supporting multilingual retrieval
- Beyond documenting maps
- Contextualization supported geo-referencing



**Add Control Point**

Enter a location that belongs to this point:

- Havana, Ciudad de La Habana, Cuba
- Havana, Florida, United States

maphub maps users Search maphub... user23 my account sign out

new and accurate chart of the West Indies with the...

Map Satellite

Regular Map Google Maps Google Earth Control Points

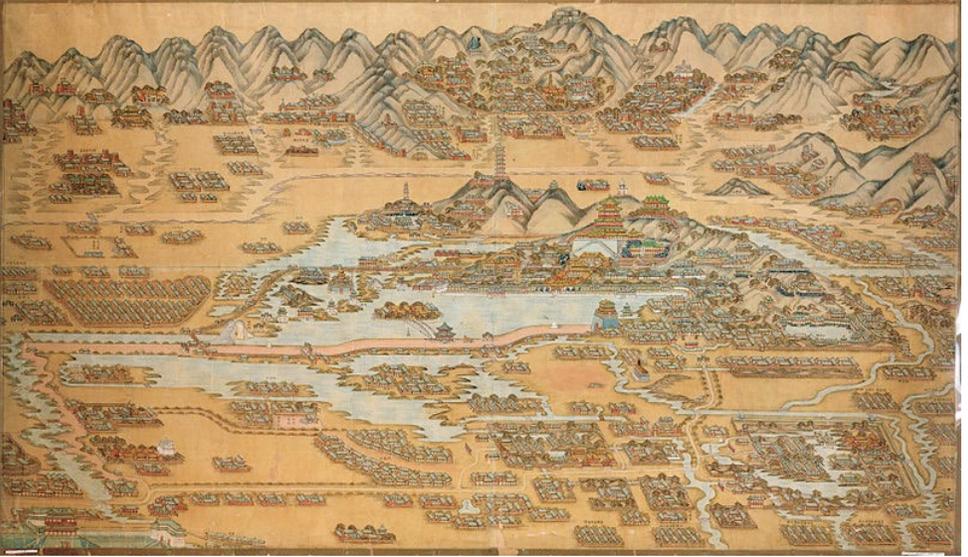
Title: new and accurate chart of the West Indies with the...  
 Subject: Maps  
 Author: Bowen, Emanuel  
 Date: 1720?

© maphub 2012 | contact | help | service status

<http://maphub.github.io/>

Marcia L. Zeng. HELDIG DH Summit 2020

My testing, 2013



"Eight Banners" Brigade barracks, temples, villages, bridges, mountains, and the Summer Palace in Beijing. Produced in Pen-and-ink and watercolor at later Qing Dynasty.

maphub maps users **北京**

After my annotation, it is searchable by Chinese.

Search Results Your search for '北京' returned 1 results.

**[Beijing Yi he yuan he ba qi bing ying]**

Date: [After 1888]  
 Author:  
 Subject: Maps, Manuscript  
 Updated: 25 days ago  
 1 Annotation (show all)

Highlights

- Annotation through mash-up culture
- Supporting multilingual retrieval
- Beyond documenting maps
- Contextualization supported geo-referencing

**Add Annotation**

Shows the "Eight Banners" Brigade barracks, temples, villages, bridges, mountains, and the Summer Palace in Beijing. Produced in Pen-and-ink and watercolor at later Qing Dynasty.

颐和园八旗兵营图  
 收藏于美国国会图书馆的《颐和园八旗兵营图》，绘制于晚清时期。全图以颐和园为中心，着重绘出了遍布于颐和园周边的八旗兵营。当中也对所涉及地区的园林、寺庙有所描绘。

Tags  
 Summer Palace Beijing Qing Dynasty Eight Banners  
 Click on a tag to accept it. Click once more to reject it.

Save annotation

**Add Control Point**

Enter a location that belongs to this point:  
 Beijing

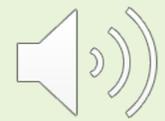
- Beijing, Beijing, China
- Beijingzi, Liaoning, China
- Beijing, Jiangxi, China
- Beijing, Guangdong Province, China
- Beijing, Chongqing Shi, China
- Beijing, Shanxi Sheng, China
- Beijingzui, Hubei, China
- Beijingtou, Shaanxi, China
- Beijingshan, Anhui Sheng, China
- Beijingtang, Guangdong Province, China
- Beijingling, Hainan, China
- Beijinggang, Hunan, China

Save control point

**Add Control Point**

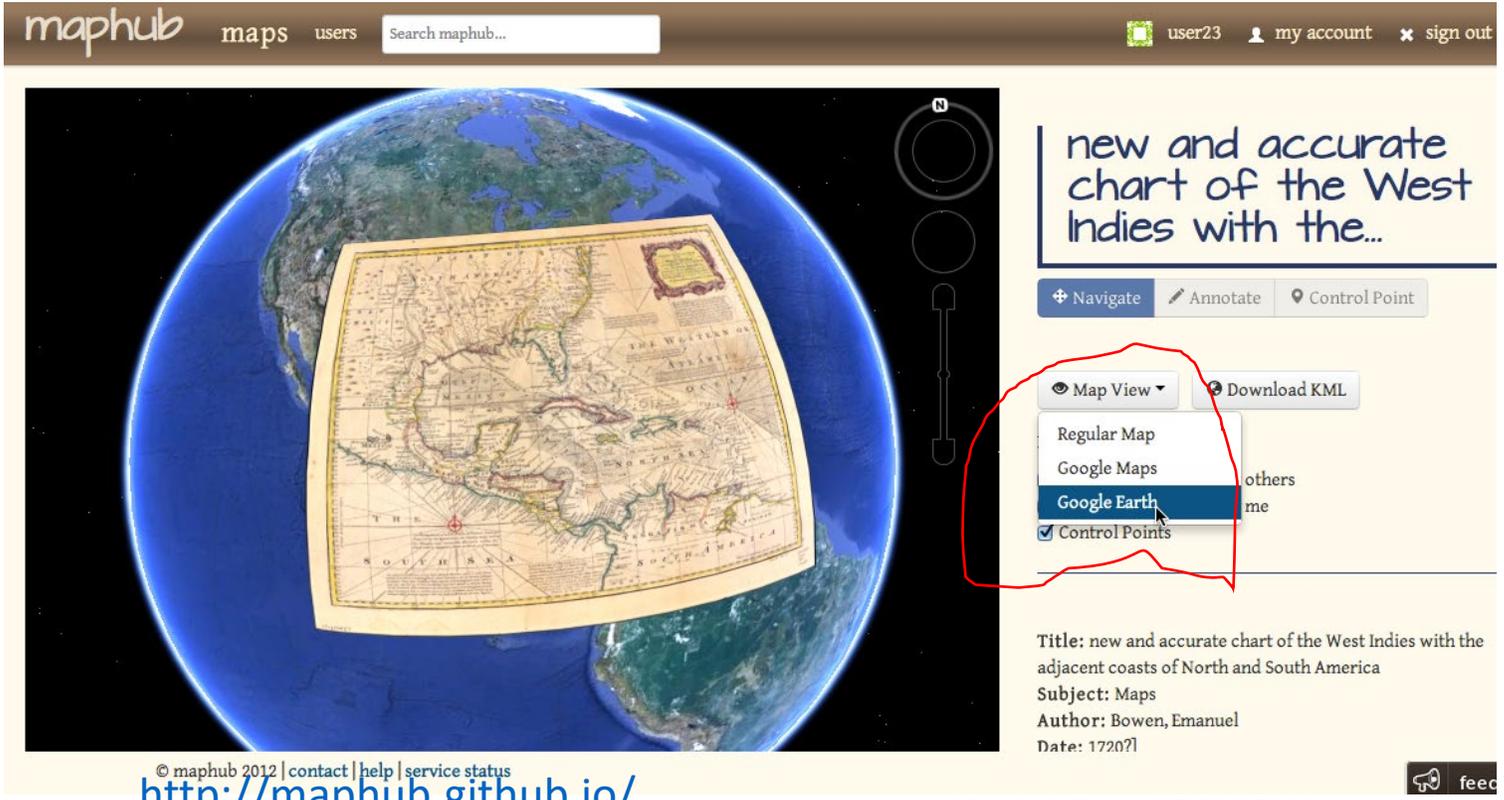
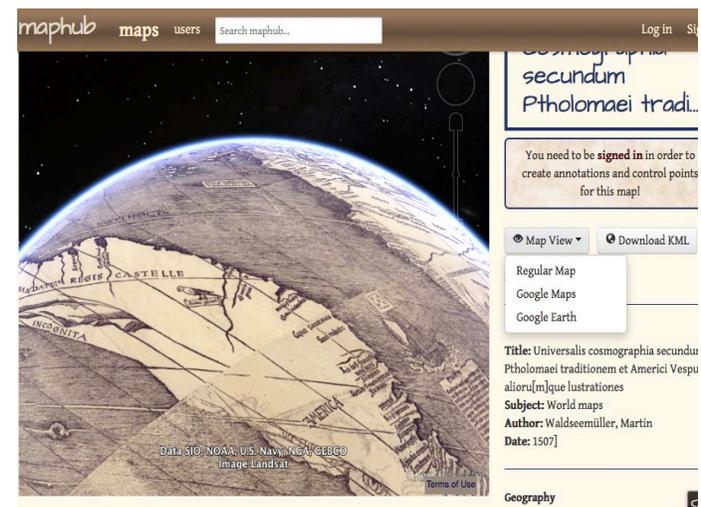
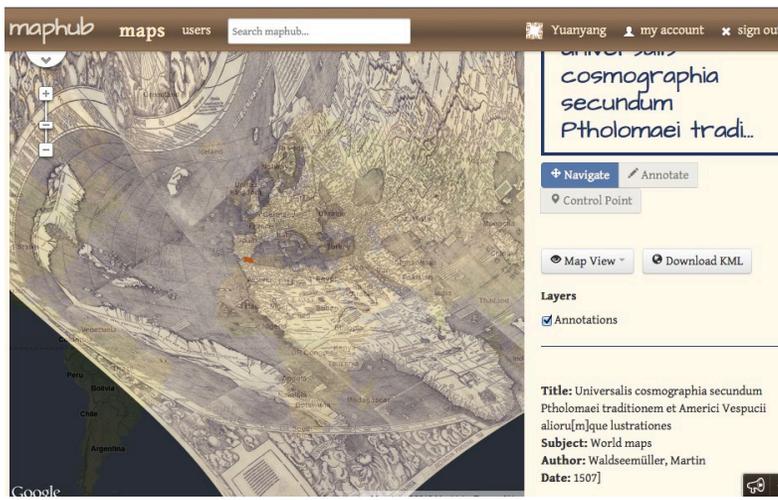
Enter a location that belongs to this point:  
 Kunming Lake, Beijing, Beijing, China

Save control point



### Highlights

- Annotation through mash-up culture
- Supporting multilingual retrieval
- Beyond documenting maps
- Contextualization supported geo-referencing
- Revealing history through map overlays



## 3-D. Murals

# Dunhuang Mogao caves

(also known as the  
Thousand Buddha Grottoes)

## Semantic Enrichment and Thesaurus projects



Center for Digital Humanities  
Wuhan University, China



- Located south-east of the Dunhuang oasis on the Silk Road, in Gansu province, China.
- Started in AD 366.
- With 492 caves, the total size of the murals reaches more than 45,000 square meters .



Dunhuang murals depict various aspects of

- medieval politics
- economics
- culture
- arts
- religion
- ethnic relations
- daily dress in Western China.





The rich content of the Dunhuang caves makes image representation a challenge.

- Identification of an image's internal objects—its “ofness”, is often ignored or lacks sufficient granularity;
- Content contained in these images continues to evade adequate semantic disclosure and connection ;
- The neglect of non-expert and novice users' needs during the process of developing and using CH digital images.

Source: Wang, et. al. 2020.



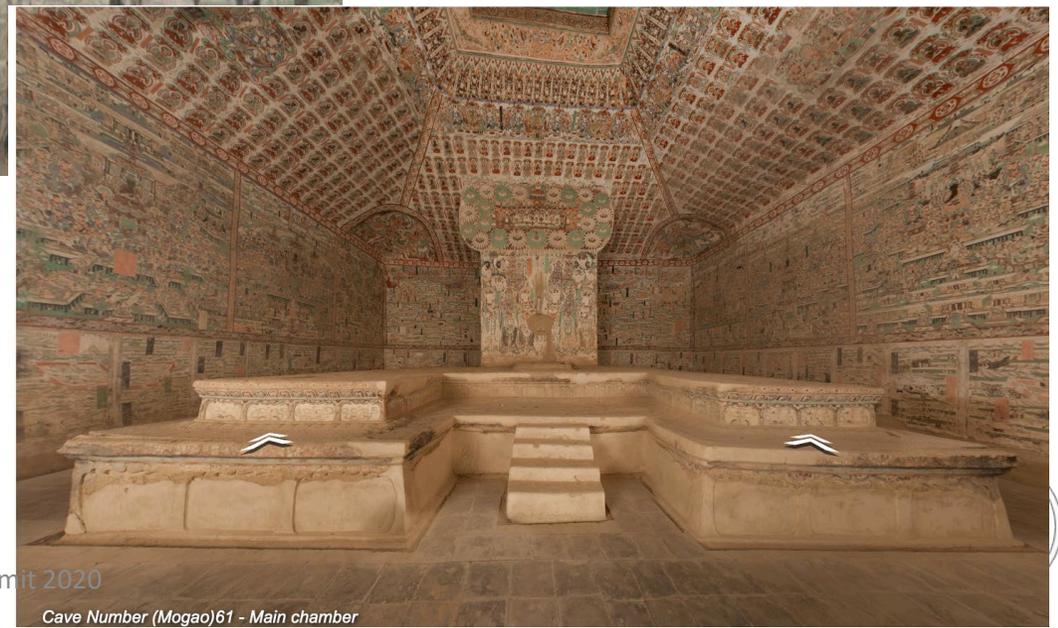
Zoom in:



**Wutai Mountain Map** 《五台山图》  
13.45 meters long x 3.42 meters wide.

Located on the west wall of the main chamber of cave  
61 of the Dunhuang Mogao Grottoes →

Image source: <https://www.e-dunhuang.com/cave/10.0001/0001.0001.0061>





WUHAN UNIVERSITY

标注图片
POI Annotation
✕

The annotation on Mural Content

选区属性(Region) -

图片属性(ImageFile) +

The annotation on Digital image of mural

URI for Each POI

ID

编辑	1	http://dh.whu.edu.cn/IDAMSservice/adcd121ea5a2496287701b6a	河东道山门西南
编辑	2	http://dh.whu.edu.cn/IDAMSservice/adcd121ea5a2496287701b6a	资福和尚庵
编辑	3	http://dh.whu.edu.cn/IDAMSservice/adcd121ea5a2496287701b6a	无著和尚塔

⇒ Title of Each POI

IIIF

{scheme}://{server}/{prefix}/{identifier}/{region}/{size}/{rotation}/{quality}.{format}

It supports users to annotate POI manually according to their metadata framework or ontology model

Marcia L. Zeng. HELDIG DH Summit 2020

Source: Wang, et. al. 2020.



# Dunhuang Mural Thesaurus

## Case: Digital Dunhuang

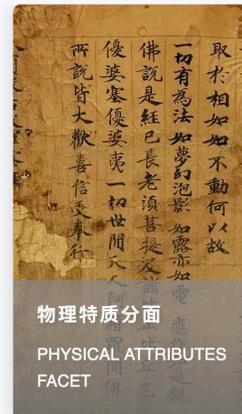
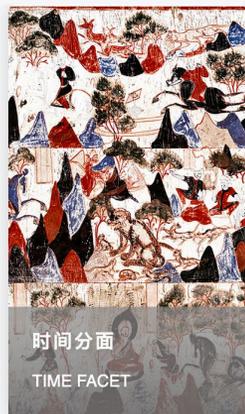
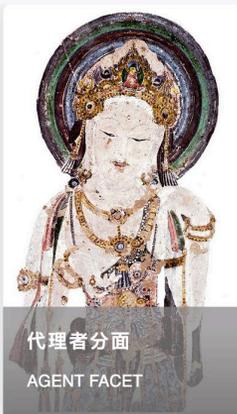
### Covering:

- Mural protection and restoration
- religion
- iconography
- cave archaeology
- humanity, culture, and other research perspectives.

Type	Number
Facets	5
Hierarchy terms	83
Concepts	3199
Instances	989
Total Terms	4276

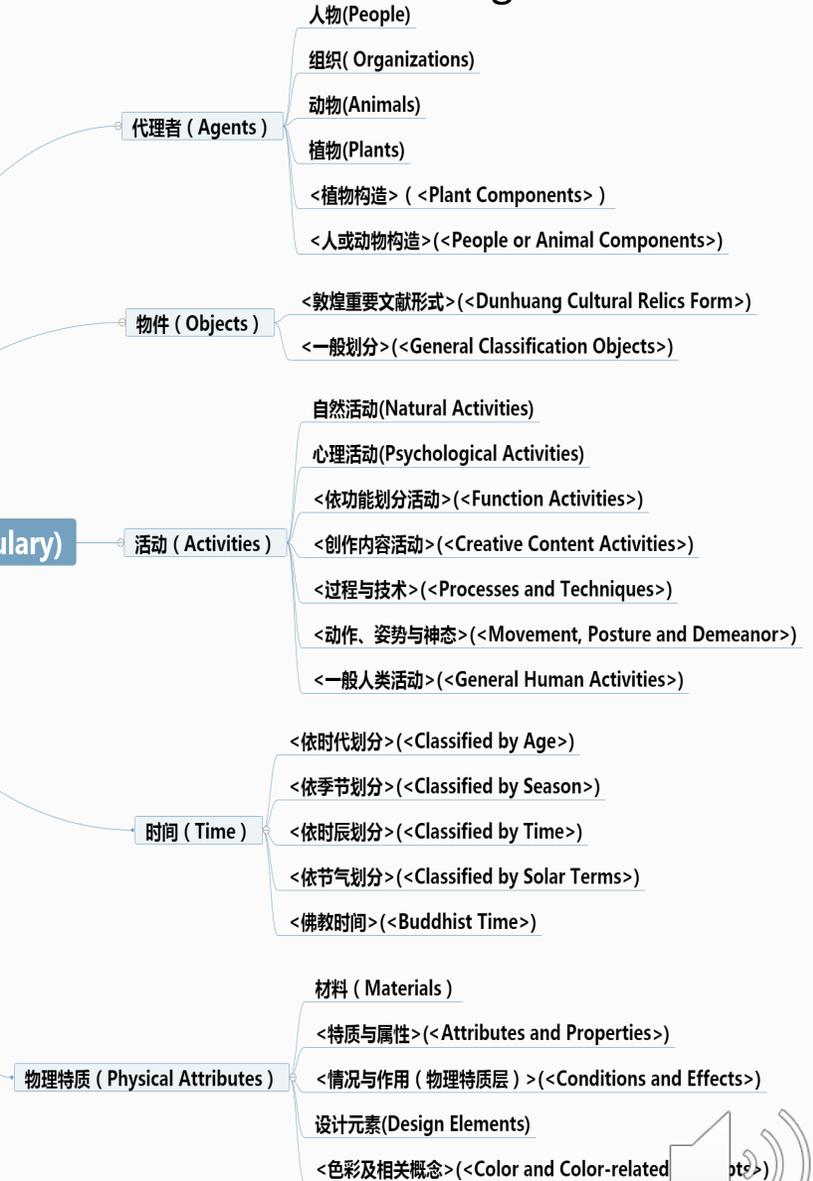
### FIVE FACETS

### 敦煌壁画主题词表五大分面



### Facets and second-level categories

### 敦煌词表(Dunhuang Vocabulary)



<http://dh.whu.edu.cn/dhvocab/home>

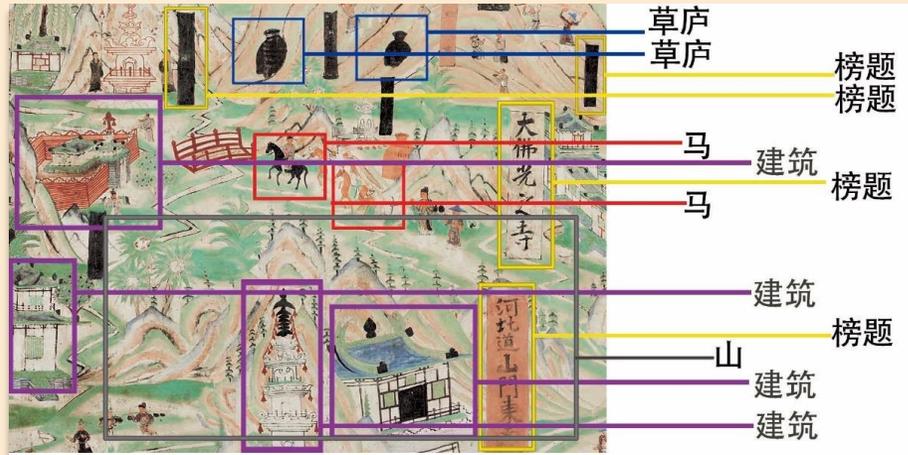
Image captured 2020-11-25

# Dunhuang Mural Thesaurus used in Semantic annotation

## Case: Digital Dunhuang

### Highlights

- Semantic annotation (deeper than normal tagging)
- SKOSified annotated figures
- Datifying after digitizing
- Contextualization supported by thesaurus
- Re-story-telling



Source: Wang, et. al. 2020.

Marcia L. Zeng. HELDIG DH Summit 2020



## Online Coins of the Roman Empire (OCRE)

<http://numismatics.org/ocre/>

A revolutionary new tool designed to help in the identification, cataloging, and research of the rich and varied coinage of the Roman Empire.

All coin types from Augustus in 31 BC to Zeno in AD 491 (representing five centuries of Roman imperial numismatics) have been published.

OCRE incorporated 107,000+ physical coins related to these coin types from 21 different datasets, originated from

- large collections a
- smaller civic or university museums,
- archaeological databases, and
- the Domuztepe excavations published through OpenContext which publishes research data on the web (Gruber 2017).



Online Coins of the Roman Empire (OCRE), a joint project of the [American Numismatic Society](#) and the [Institute for the Study of the Ancient World](#) at New York University, is a revolutionary new tool designed to help in the identification, cataloging, and research of the rich and varied coinage of the Roman Empire. The project records every published type of Roman Imperial Coinage from Augustus in 31 BC, until the death of Zeno in AD 491. This is an easy to use digital corpus, with downloadable catalog entries, incorporating over 43,000 types of coins.

As of April 2017, OCRE provides links to examples present in nearly 20 American and European databases (both archaeological and museum in context), including the [ANS collection](#), the [Münzkabinett of the State Museum of Berlin](#), and the [British Museum](#), now totalling over 100,000 physical specimens. Between these collections, OCRE is now able to illustrate 50% of the imperial coin types that it contains. Moving forward, as more collections join the project, it will eventually incorporate and display almost all recorded Roman Imperial coin types. Furthermore, it draws findspot information from another ANS-developed resource, [Coin Hoards of the Roman Republic](#), enabling the mapping of the

Language ▾

- Arabic
- Bulgarian
- Danish
- German
- Greek
- English
- Spanish
- French
- Hungarian
- Italian
- Dutch
- Polish
- Romanian
- Russian
- Swedish
- Turkish
- Ukrainian





# Visualize your queries on-the-fly

<http://numismatics.org/ocre/>

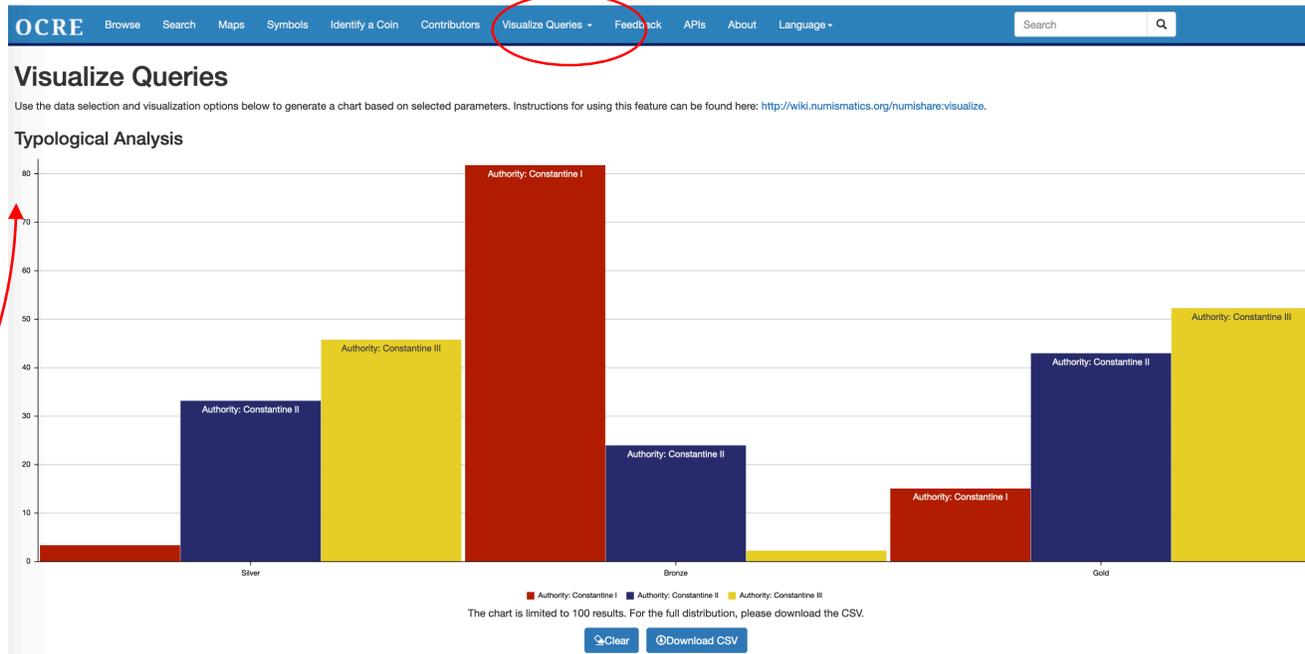
Images captured 2020-12-03

Case: OCRE

## Highlights

A pioneering, revolutionary new tool

- Ontology /knowledge base design
- Extremely user-friendly accesses
- Innovative research-oriented functions
- Smart data in DH



## SMART data

- Adequately represent a sufficient number of relevant features of humanistic objects of inquiry to enable the necessary level of precision and nuance required by humanities scholars
- Provide users with a sufficient amount of data to enable quantitative methods of inquiry, helping researchers to surpass the limitations inherent in methods based on close reading strategies (Schöch, 2013).

OCRE Browse Search Maps Symbols Identify a Coin Contributors Visualize Queries Feedback APIs About Language

## Visualize Queries

Use the data selection and visualization options below to generate a chart based on selected parameters. Instructions for using this feature can be found here: <http://wiki.numismatics.org/numishare:visualize>.

### Typological Analysis

#### Select Category for Analysis

Select a category below to generate a graph showing the quantitative distribution of the selected category.

Material

#### Numeric response type

- Percentage
- Count

#### Compare Queries

You can compare multiple queries to generate a more complex chart. In a single category, the drop-down menu will include only those mints that produce results for that category.

Group **+** Add Query Field

Authority (Person) Constantine I

Group **+** Add Query Field

Authority (Person) Constantine II

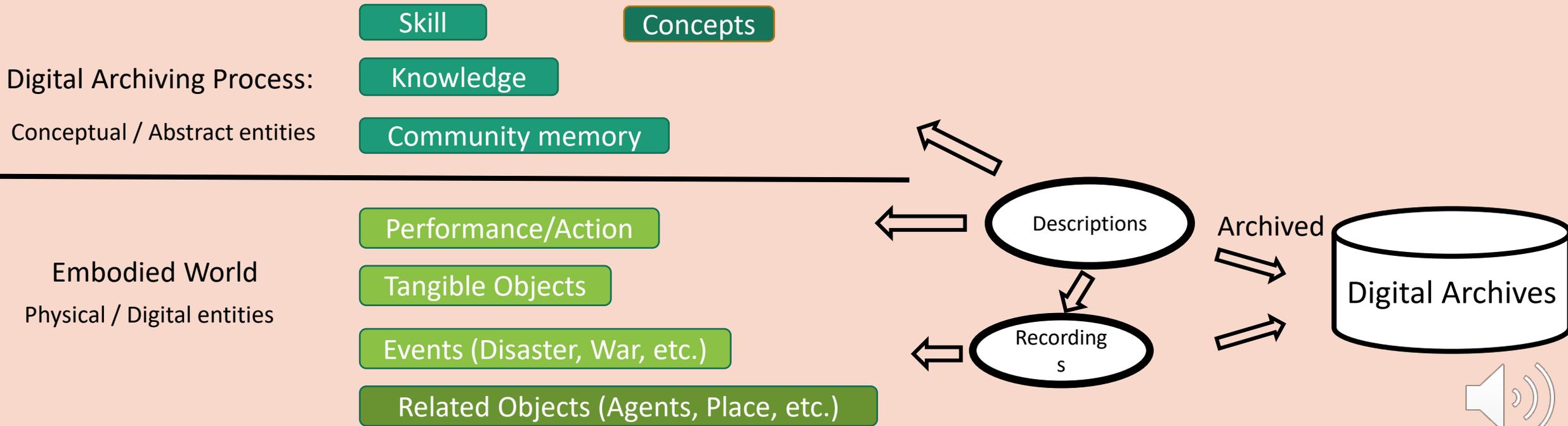
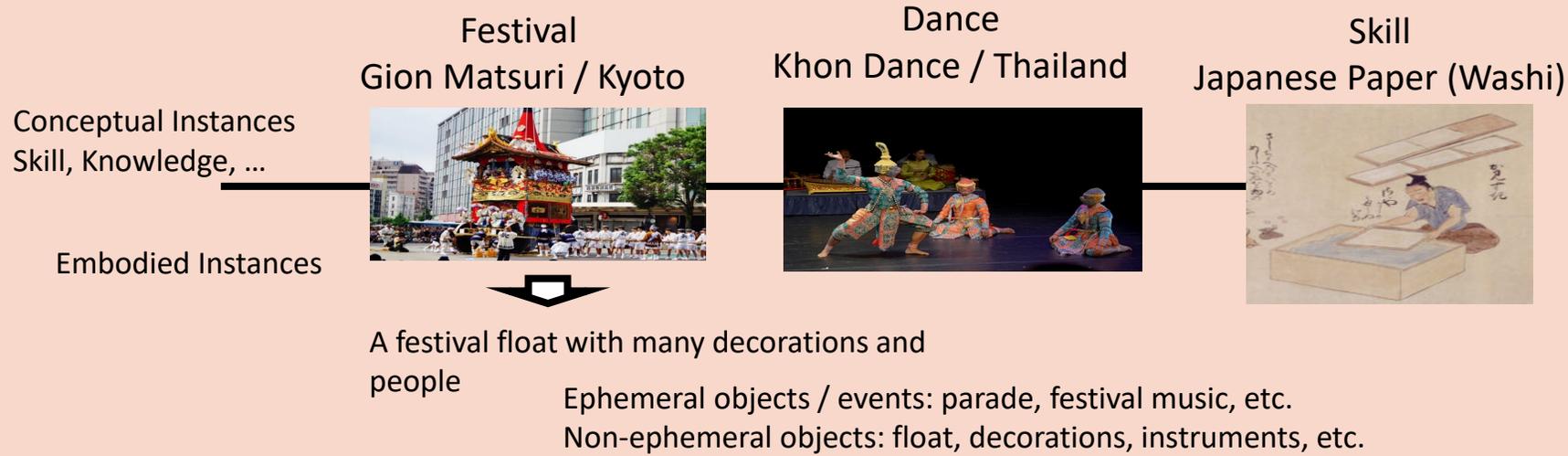
Group **+** Add Query Field

Authority (Person) Constantine III

Generate Chart

Select...

- Authority (Person)
- Authority (State)
- Coin Type
- From Date
- To Date
- Denomination
- Deity
- Issuer
- Manufacture
- Material
- Mint
- Object Type
- Portrait
- Region
- Stated Authority



## Metadata Models for Organizing Digital Archives on the Web

### Metadata-Centric Projects at Tsukuba

Digital Archive = a collection of digital resources, mainly of cultural and historical resources

- **Metadata-Centric Projects at Tsukuba**

1. Great East Japan - Enhancing digital archives' usability
2. **Japan Earthquake Archive**  
- Aggregating metadata within and across archives after the Great East Japan Earthquake and Tsunami (2011-03-11)
3. Manga - Aggregating resources on the web
4. Cultural heritage objects for digital archives – tangible and intangible cultural heritage

- **Common Research Goals**

- **Enrich values of digital archives by metadata aggregation** within and across digital archives and linking institutional digital archives and web resources

#### Great East Japan Earthquake Archive Challenges

- Quality of metadata (especially photographs and videos)
- Item-based metadata – relationships among the items during one event
- What do users want to find?
  - A specific photo?
  - A group of photos?
  - Around a place or an event?
  - From one community or across all archives and web?

#### Aggregation by:

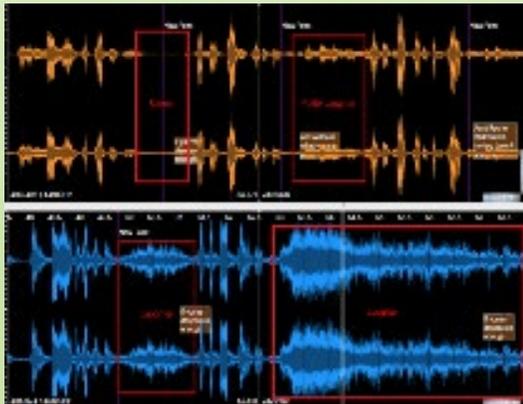
- Temporal
- Special
- Agent
- Subject information

# Other notable LAM projects for cultural heritage and historical resources

## Audio

**Linked Reading** project, U. Cincinnati

- Elliston Poetry Archive: 700+ recordings of poetry or poetry-related content
- a linked data infrastructure and sound analysis platform
- using semantic analysis of the printed text, coupled with sonic analysis of the audio archives

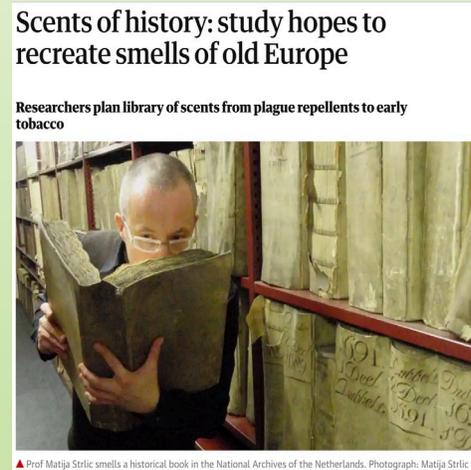


<http://dsc.uc.edu/projects/elliston-poetry-archive-sonification-experiments>

## Scents

**“Odeuropa”** Project

- 1<sup>st</sup> step: develop AI to screen historical texts in seven languages for descriptions of odours – and their context – as well as to spot aromatic items within images, such as paintings.



Scents of history: study hopes to recreate smells of old Europe

Researchers plan library of scents from plague repellents to early tobacco

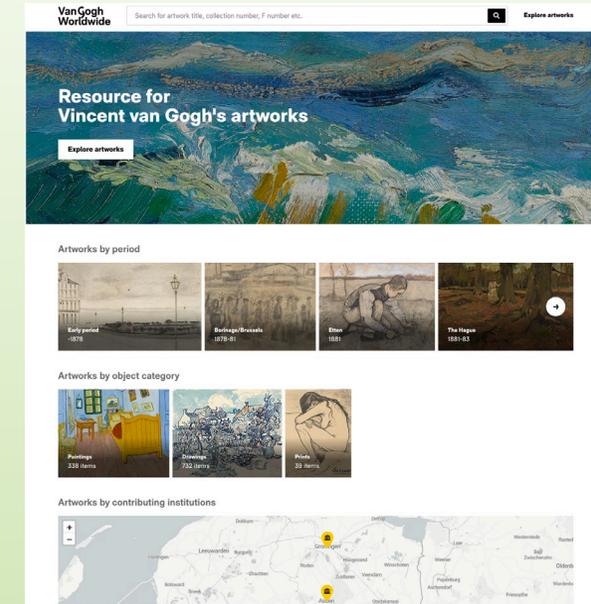
▲ Prof Matija Strlic smells a historical book in the National Archives of the Netherlands. Photograph: Matija Strlic.

<https://www.theguardian.com/science/2020/nov/17/scents-of-history-study-hopes-to-recreate-smells-of-old-europe>

## Art-historical + technical info

**Van Gogh Worldwide**

- Launched 2020-11-05
- Uses *Linked Art Model*
- Provides all published scholarly research results about the artworks, inter-linked by means of thesauri.
- Art-historical + technical information *about* the work of Vincent van Gogh (1853-1890).



<https://vangoghworldwide.org/>

Vincent van Gogh  
The Letters

by period  
by correspondent  
by place  
with sketches

Search  
keyword or number(s) >>  
Advanced search  
Search results

Van Gogh as a letter-writer  
Correspondents  
Biographical & historical context  
Publication history

About this edition  
Chronology  
Concordance, lists, bibliography  
Book edition

316 216 160 << 315 | 317 >>

To Theo van Gogh. The Hague, on or about Thursday, 15 February 1883.

SEARCH THIS LETTER

original text + line endings facsimile translation notes artworks

316 | Show metadata

My dear Theo,

Sincerest thanks for your letter, and the enclosure was most welcome; it really does help me. I begin by telling you that it's a great relief to me that the past of the woman you write about is entirely different from what I first instinctively thought. Namely that she has known not only misery and straitened circumstances, but also other things, so I believe that she'll appreciate you fully, also as regards civilization and broader views, much more so than a woman who has been hurt by misery from an early age and knows no better than to

1. This phrasing indicates that Theo sent extra money; this is confirmed by a remark in letter 323, ll. 66-67. See also Date.

2. A year before Van Gogh had noted these words Michelet's *La femme* (in a slightly different form) drawing *Sorrow* - see letter 216, n. 3.

3. For this pronunciation by Theo, see letter 291

original text + line endings facsimile translation no

1. This phrasing indicates that Theo sent extra money confirmed by a remark in letter 323, ll. 66-67. See

2. Means: 'om haar karakter te vormen' (to shape character).

3. A year before Van Gogh had noted these words Michelet's *La femme* (in a slightly different form) drawing *Sorrow* - see letter 216, n. 3.

3. For this pronunciation by Theo, see letter 291

SEARCH THIS LETTER

Vincent van Gogh  
Series of 1883-1890, 1882, 1883, 1884, 1885, 1886, 1887, 1888, 1889, 1890, 1891, 1892, 1893, 1894, 1895, 1896, 1897, 1898, 1899, 1900, 1901, 1902, 1903, 1904, 1905, 1906, 1907, 1908, 1909, 1910, 1911, 1912, 1913, 1914, 1915, 1916, 1917, 1918, 1919, 1920, 1921, 1922, 1923, 1924, 1925, 1926, 1927, 1928, 1929, 1930, 1931, 1932, 1933, 1934, 1935, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 2680, 2681, 2682, 2683, 2684, 2685, 2686, 2687, 2688, 2689, 2690, 2691, 2692, 2693, 2694, 2695, 2696, 2697, 2698, 2699, 2700, 2701, 2702, 2703, 2704, 2705, 2706, 2707, 2708, 2709, 2710, 2711, 2712, 2713, 2714, 2715, 2716, 2717, 2718, 2719, 2720, 2721, 2722, 2723, 2724, 2725, 2726, 2727, 2728, 2729, 2730, 2731, 2732, 2733, 2734, 2735, 2736, 2737, 2738, 2739, 2740, 2741, 2742, 2743, 2744, 2745, 2746, 2747, 2748, 2749, 2750, 2751, 2752, 2753, 2754, 2755, 2756, 2757, 2758, 2759, 2760, 2761, 2762, 2763, 2764, 2765, 2766, 2767, 2768, 2769, 2770, 2771, 2772, 2773, 2774, 2775, 2776, 2777, 2778, 2779, 2780, 2781, 2782, 2783, 2784, 2785, 2786, 2787, 2788, 2789, 2790, 2791, 2792, 2793, 2794, 2795, 2796, 2797, 2798, 2799, 2800, 2801, 2802, 2803, 2804, 2805, 2806, 2807, 2808, 2809, 2810, 2811, 2812, 2813, 2814, 2815, 2816, 2817, 2818, 2819, 2820, 2821, 2822, 2823, 2824, 2825, 2826, 2827, 2828, 2829, 2830, 2831, 2832, 2833, 2834, 2835, 2836, 2837, 2838, 2839, 2840, 2841, 2842, 2843, 2844, 2845, 2846, 2847, 2848, 2849, 2850, 2851, 2852, 2853, 2854, 2855, 2856, 2857, 2858, 2859, 2860, 2861, 2862, 2863, 2864, 2865, 2866, 2867, 2868, 2869, 2870, 2871, 2872, 2873, 2874, 2875, 2876, 2877, 2878, 2879, 2880, 2881, 2882, 2883, 2884, 2885, 2886, 2887, 2888, 2889, 2890, 2891, 2892, 2893, 2894, 2895, 2896, 2897, 2898, 2899, 2900, 2901, 2902, 2903, 2904, 2905, 2906, 2907, 2908, 2909, 2910, 2911, 2912, 2913, 2914, 2915, 2916, 2917, 2918, 2919, 2920, 2921, 2922, 2923, 2924, 2925, 2926, 2927, 2928, 2929, 2930, 2931, 2932, 2933, 2934, 2935, 2936, 2937, 2938, 2939, 2940, 2941, 2942, 2943, 2944, 2945, 2946, 2947, 2948, 2949, 2950, 2951, 2952, 2953, 2954, 2955, 2956, 2957, 2958, 2959, 2960, 2961, 2962, 2963, 2964, 2965, 2966, 2967, 2968, 2969, 2970, 2971, 2972, 2973, 2974, 2975, 2976, 2977, 2978, 2979, 2980, 2981, 2982, 2983, 2984, 2985, 2986, 2987, 2988, 2989, 2990, 2991, 2992, 2993, 2994, 2995, 2996, 2997, 2998, 2999, 3000, 3001, 3002, 3003, 3004, 3005, 3006, 3007, 3008, 3009, 3010, 3011, 3012, 3013, 3014, 3015, 3016, 3017, 3018, 3019, 3020, 3021, 3022, 3023, 3024, 3025, 3026, 3027, 3028, 3029, 3030, 3031, 3032, 3033, 3034, 3035, 3036, 3037, 3038, 3039, 3040, 3041, 3042, 3043, 3044, 3045, 3046, 3047, 3048, 3049, 3050, 3051, 3052, 3053, 3054, 3055, 3056, 3057, 3058, 3059, 3060, 3061, 3062, 3063, 3064, 3065, 3066, 3067, 3068, 3069, 3070, 3071, 3072, 3073, 3074, 3075, 3076, 3077, 3078, 3079, 3080, 3081, 3082, 3083, 3084, 3085, 3086, 3087, 3088, 3089, 3090, 3091, 3092, 3093, 3094, 3095, 3096, 3097, 3098, 3099, 3100, 3101, 3102, 3103, 3104, 3105, 3106, 3107, 3108, 3109, 3110, 3111, 3112, 3113, 3114, 3115, 3116, 3117, 3118, 3119, 3120, 3121, 3122, 3123, 3124, 3125, 3126, 3127, 3128, 3129, 3130, 3131, 3132, 3133, 3134, 3135, 3136, 3137, 3138, 3139, 3140, 3141, 3142, 3143, 3144, 3145, 3146, 3147, 3148, 3149, 3150, 3151, 3152, 3153, 3154, 3155, 3156, 3157, 3158, 3159, 3160, 3161, 3162, 3163, 3164, 3165, 3166, 3167, 3168, 3169, 3170, 3171, 3172, 3173, 3174, 3175, 3176, 3177, 3178, 3179, 3180, 3181, 3182, 3183, 3184, 3185, 3186, 3187, 3188, 3189, 3190, 3191, 3192, 3193, 3194, 3195, 3196, 3197, 3198, 3199, 3200, 3201, 3202, 3203, 3204, 3205, 3206, 3207, 3208, 3209, 3210, 3211, 3212, 3213, 3214, 3215, 3216, 3217, 3218, 3219, 3220, 3221, 3222, 3223, 3224, 3225, 3226, 3227, 3228, 3229, 3230, 3231, 3232, 3233, 3234, 3235, 3236, 3237, 3238, 3239, 3240, 3241, 3242, 3243, 3244, 3245, 3246, 3247, 3248, 3249, 3250, 3251, 3252, 3253, 3254, 3255, 3256, 3257, 3258, 3259, 3260, 3261, 3262, 3263, 3264, 3265, 3266, 3267, 3268, 3269, 3270, 3271, 3272, 3273, 3274, 3275, 3276, 3277, 3278, 3279, 3280, 3281, 3282, 3283, 3284, 3285, 3286, 3287, 3288, 3289, 3290, 3291, 3292, 3293, 3294, 3295, 3296, 3297, 3298, 3299, 3300, 3301, 3302, 3303, 3304, 3305, 3306, 3307, 3308, 3309, 3310, 3311, 3312, 3313, 3314, 3315, 3316, 3317, 3318, 3319, 3320, 3321, 3322, 3323, 3324, 3325, 3326, 3327, 3328, 3329, 3330, 3331, 3332, 3333, 3334, 3335, 3336, 3337, 3338, 3339, 3340, 3341, 3342, 3343, 3344, 3345, 3346, 3347, 3348, 3349, 3350, 3351, 3352, 3353, 3354, 3355, 3356, 3357, 3358, 3359, 3360, 3361, 3362, 3363, 3364, 3365, 3366, 3367, 3368, 3369, 3370, 3371, 3372, 3373, 3374, 3375, 3376, 3377, 3378, 3379, 3380, 3381, 3382, 3383, 3384, 3385, 3386, 3387, 3388, 3389, 3390, 3391, 3392, 3393, 3394, 3395, 3396, 3397, 3398, 3399, 3400, 3401, 3402, 3403, 3404, 3405, 3406, 3407, 3408, 3409, 3410, 3411, 3412, 3413, 3414, 3415, 3416, 3417, 3418, 3419, 3420, 3421, 3422, 3423, 3424, 3425, 3426, 3427, 3428, 3429, 3430, 3431, 3432, 3433, 3434, 3435, 3436, 3437, 3438, 3439, 3440, 3441, 3442, 3443, 3444, 3445, 3446, 3447, 3448, 3449, 3450, 3451, 3452, 3453, 3454, 3455, 3456, 3457, 3458, 3459, 3460, 3461, 3462, 3463, 3464, 3465, 3466, 3467, 3468, 3469, 3470, 3471, 3472, 3473, 3474, 3475, 3476, 3477, 3478, 3479, 3480, 3481, 3482, 3483, 3484, 3485, 3486, 3487, 3488, 3489, 3490, 3491, 3492, 3493, 3494, 3495, 3496, 3497, 3498, 3499, 3500, 3501, 3502, 3503, 3504, 3505, 3506, 3507, 3508, 3509, 3510, 3511, 3512, 3513, 3514, 3515, 3516, 3517, 3518, 3519, 3520, 3521, 3522, 3523, 3524, 3525, 3526, 3527, 3528, 3529, 3530, 3531, 3532, 3533, 3534, 3535, 3536, 3537, 3538, 3539, 3540, 3541, 3542, 3543, 3544, 3545, 3546, 3547, 3548, 3549, 3550, 3551, 3552, 3553, 3554, 3555, 3556, 3557, 3558, 3559, 3560, 3561, 3562, 3563, 3564, 3565, 3566, 3567, 3568, 3569, 3570, 3571, 3572, 3573, 3574, 3575, 3576, 3577, 3578, 3579, 3580, 3581, 3582, 3583, 3584, 3585, 3586, 3587, 3588, 3589, 3590, 3591, 3592, 3593, 3594, 3595, 3596, 3597, 3598, 3599, 3600, 3601, 3602, 3603, 3604, 3605, 3606, 3607, 3608, 36

# Outline

Introduction

Semantic Enrichment Approaches

- Structured Data
- Semi-structured Data
- Unstructured Data

Summary and Conclusions





# Enhancing Historical and Cultural Heritage Data to Support Digital Humanities Research

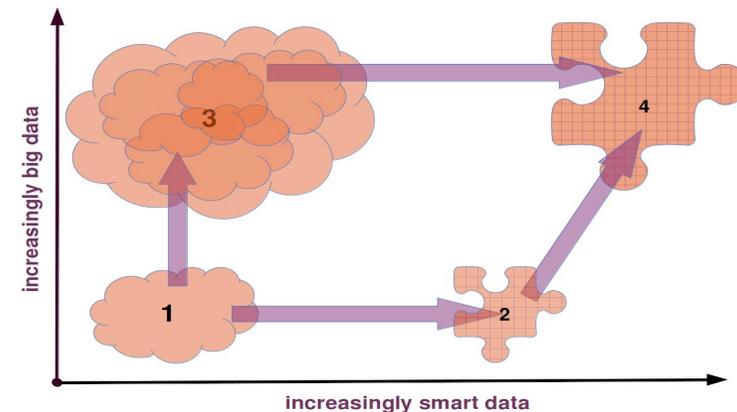
- Enhancing subject access through controlled vocabularies that have embraced Linked Data
- Transforming the semi-structured and unstructured data into structured data through semantic analysis and aligning with standardized controlled vocabularies
- Linking and contextualizing existing structured data across data silos
- Enabling one-to-many usages of LAM data in supporting digital humanities

In DH, emphasis has been on:

➤ transforming unstructured data to → structured data

## Structured Data Trending:

- Machine ~~readable~~-understandable data
- Machine ~~readable~~ actionable data
- Accurate (no error) data in the processes of interlinking, citing, transferring, rights-permission, use and reuse
- One → to → Many uses and high efficiency processing data



## 5-star Linked Data

The baseline of our work is the [5-star Linked Data model](#), proposed [originally](#) by Tim Berners-Lee.

- ★ Make data available on the Web in whatever format.
- ★★ Make data available as structured data (e.g., Excel instead of an image scan of a table).
- ★★★ Use non-proprietary formats (e.g., CSV instead of Excel format).
- ★★★★ Use URIs to denote things, so that people can point at your data.
- ★★★★★ Link your data to other data to provide context.

## 7-star Linked Data Service

However, in our opinion, providing 5-star Linked Data is just the beginning. To actually make use of the datasets, consumers need more support in getting to know and access them, as well as a better grasp of their quality and provenance. To this end, we extend the model with two additional stars:

- ★★★★★★ Provide your data with a schema and documentation so that people can *understand and re-use* your data easily.
- ★★★★★★ Validate your data and denote its provenance so that people can *trust the quality* of your data.

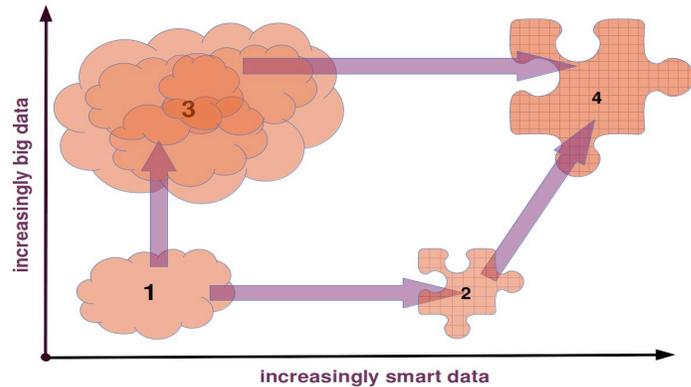
"7-star" [Linked Data Finland portal](http://www.ldf.fi/) <http://www.ldf.fi/>

Marcia L. Zeng. HELDIG DH Summit 2020



# Results of Today – Visions for Tomorrow

## Enhancing Historical and Cultural Heritage Data to Support Digital Humanities Research



Schöch, C. 2013. Big? Smart? Clean? Messy? Data in the humanities. *Journal for Digital Humanities*. 2(3): pp.2-13.

Data Silos → Data Lakes → Data Planets → Data Universe

## “Sampo” Model and Sampo-UI Framework

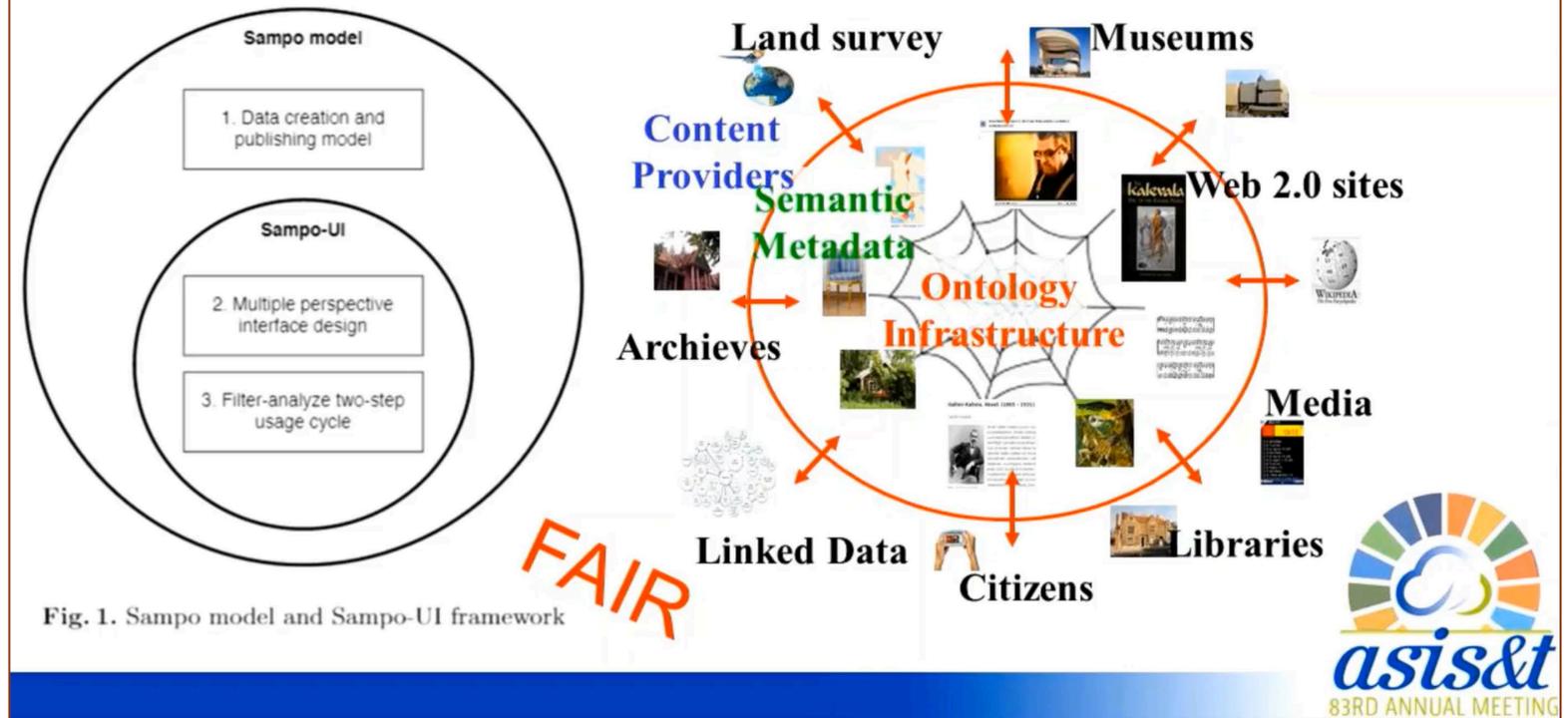


Fig. 1. Sampo model and Sampo-UI framework

Hyvönen E. 2020. Building a National Level Linked Open Data Infrastructure for Digital Humanities in Finland. 83rd Annual Meeting of the Association for Information Science and Technology, Oct. 22- Nov.1, 2020. <https://vimeo.com/460086143>



# References (1)

Borgman, C.L., 2015. *Big data, little data, no data: Scholarship in the networked world*. MIT press.

Chen, S. & Lu, L. 2020, Linked Data as Method for Supporting DH-Research on the Cultural Resources of Chinese Wooden Slips and the Interpretation of Ancient Chinese Characters. DCMi Virtual 2020. <http://shorturl.at/otzX4>

Daquino, M., Mambelli, F., Peroni, S., Tomasi, F. and Vitali, F., 2017. Enhancing semantic expressivity in the cultural heritage domain: exposing the Zeri Photo Archive as Linked Open Data. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(4), pp.1-21.

Gracy, K. and Zeng, M. 2015. Creating Linked Data within Archival Description: Tools for Extracting, Validating, and Encoding Access Points for Finding Aids. *DH 2015*, June 29–July 3, 2015, Sydney, Australia.

Europeana Task Force on Enrichment and Evaluation. 2015. Report on Enrichment and Evaluation. 29/10/2015. [http://pro.europeana.eu/files/Europeana\\_Professional/EuropeanaTech/EuropeanaTech\\_taskforces/Enrichment\\_Evaluation/FinalReport\\_EnrichmentEvaluation\\_102015.pdf](http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf)

Europeana Semantic Enrichment Framework, Documentation. 17th November 2016 (updated 2017, 2018, 2020). Available from <https://pro.europeana.eu/page/europeana-semantic-enrichment>

Hyvönen, E., Tuominen, J., Alonen, M. and Mäkelä, E., 2014, May. Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In *European Semantic Web Conference* (pp. 226-230). Springer, Cham.

Hyvönen, E. 2020. Linked Open Data infrastructure for Digital Humanities in Finland. In: *Proc. Of the Digital Humanities in the Nordic Countries* (DHN 2020) (pp. 254-259). <https://seco.cs.aalto.fi/publications/2020/hyvonen-lodi4dh-dhn-2020.pdf>

Hyvönen, E. 2020. “Sampo” model and semantic portals for Digital Humanities on the Semantic Web. In: *Digital Humanities in Nordic Countries* (DHN 2020) (pp. 373-378). <https://seco.cs.aalto.fi/publications/2020/hyvonen-sampos-dhn-2020.pdf>

Hyvönen, E. 2020. Using the Semantic Web in Digital Humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semantic Web* 11(1) pp. 187–193. <http://semantic-web-journal.net/system/files/swj2310.pdf>



# References (2)

O'Neill, Ed, and Jeff Mixter. 2013. "Maximizing the Usage of Value Vocabularies in the Linked Data Ecosystem." *76th Annual Meeting of the American Society for Information Science and Technology (ASIS&T), Montreal, Canada, Nov. 2-6, 2013.*

<http://nkos.slis.kent.edu/ASIST2013/ONeill-Mixter.pptx>

Schöch, C. 2013. Big? Smart? Clean? Messy? Data in the humanities. *Journal for Digital Humanities*. 2(3): pp.2-13.

<http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/>

Sugimoto, S., Kiryakos, S., Wijesundara, C., Monika, W., Mihara, T. and Nagamori, M., 2018. Metadata models for organizing digital archives on the Web: Metadata-centric projects at Tsukuba and lessons learned. In *International Conference on Dublin Core and Metadata Applications, 2018* (pp. 95-105).

Sugimoto, S. 2019. Modeling Culture – a perspective from digital archives and metadata. In *International Conference on Dublin Core and Metadata Applications, 2019*.

Svensson, P. 2010. The Landscape of Digital Humanities. *Digital Humanities Quarterly*. 4(1).

<http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html>

Weitz, J., Toves, J., Vizine-Goetz, D., Naught, N. and Bremer, R., 2016. Mining MARC's hidden treasures: initial investigations into how notes of the past might shape our future. *Journal of library metadata*, 16(3-4), pp.166-180..

Wang, X., et al. 2020. Representation and Display of Digital Images of Cultural Heritage: A Semantic Enrichment Approach. *Knowledge Organization* (forthcoming).

Zeng, M.L., Gracy, K.F. and Žumer, M., 2014. Using a semantic analysis tool to generate subject access points: A study using Panofsky's theory and two research samples. *Knowledge Organization*. 41(6), pp.440-451.

Zeng, M.L., 2019. Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article. *El profesional de la información*, 28(1). <http://www.elprofesionaldeinformacion.com/contenidos/2019/ene/03.html> [open access]

