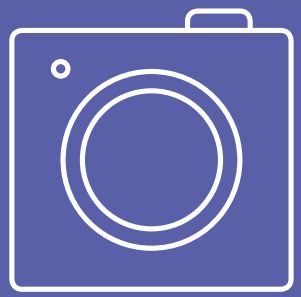
KMI RESEARCH SEMINAR
31 MAY, OPEN UNIVERSITY



Data provenance and trustworthiness assessment in photo archives

Marilena Daquino | marilena.daquino2@unibo.it

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

OUTLINE

about me

insights into Digital Humanities / Photo archives

research questions

the Zeri & LODE project

approach and methodology

data provenance and trustworthiness in photo archives

partial findings and issues

WYSIWHY demo

About me

— Research visiting student from the University of Bologna, Italy

2013 graduated at the Department of Histories and Culture

About me

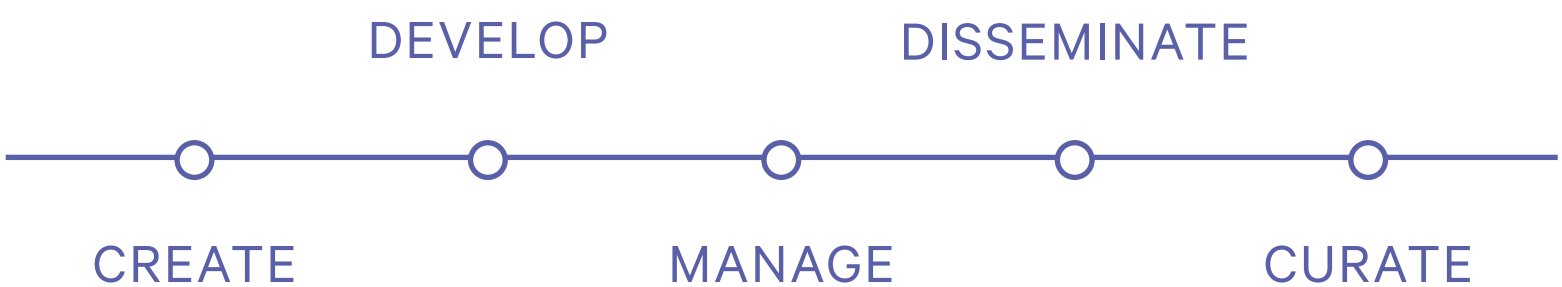
— Research fellow at the CRR-MM, University of Bologna

MULTIMEDIA CENTER
University of Bologna

2013 graduated at the Department of Histories and Culture

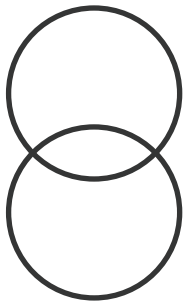
2014 — working at the Multimedia Centre (CRR-MM)

under the Library Division (AlmaDL - Digital Library)
for supporting scholars to deal with their research data



About me

— 2nd year PhD student, University of Bologna, Italy



2013 graduated at the Department of Histories and Culture

2014 — working at the Multimedia Centre (CRR-MM)

2015 — PhD student at the *Department of Classical Philology and Italian Studies* in Digital Humanities *

GENERAL TOPIC *semantic web applications and archival studies with a focus on Art historical photo archives* **

Digital Humanities

— IMHO

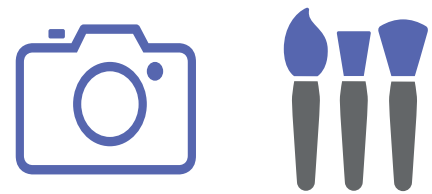
1 consuming technologies so as to **innovate/change/discover** methodologies for the research in humanities

- — new research questions in humanities
- — new methods to answer a research question
- — new research fields

2 contribute to research in Computer Science providing domain-(in)dependent solutions/applications

Art historical photo archives

— all you need if you are an art historian



the most complex scenario in the Cultural Heritage domain

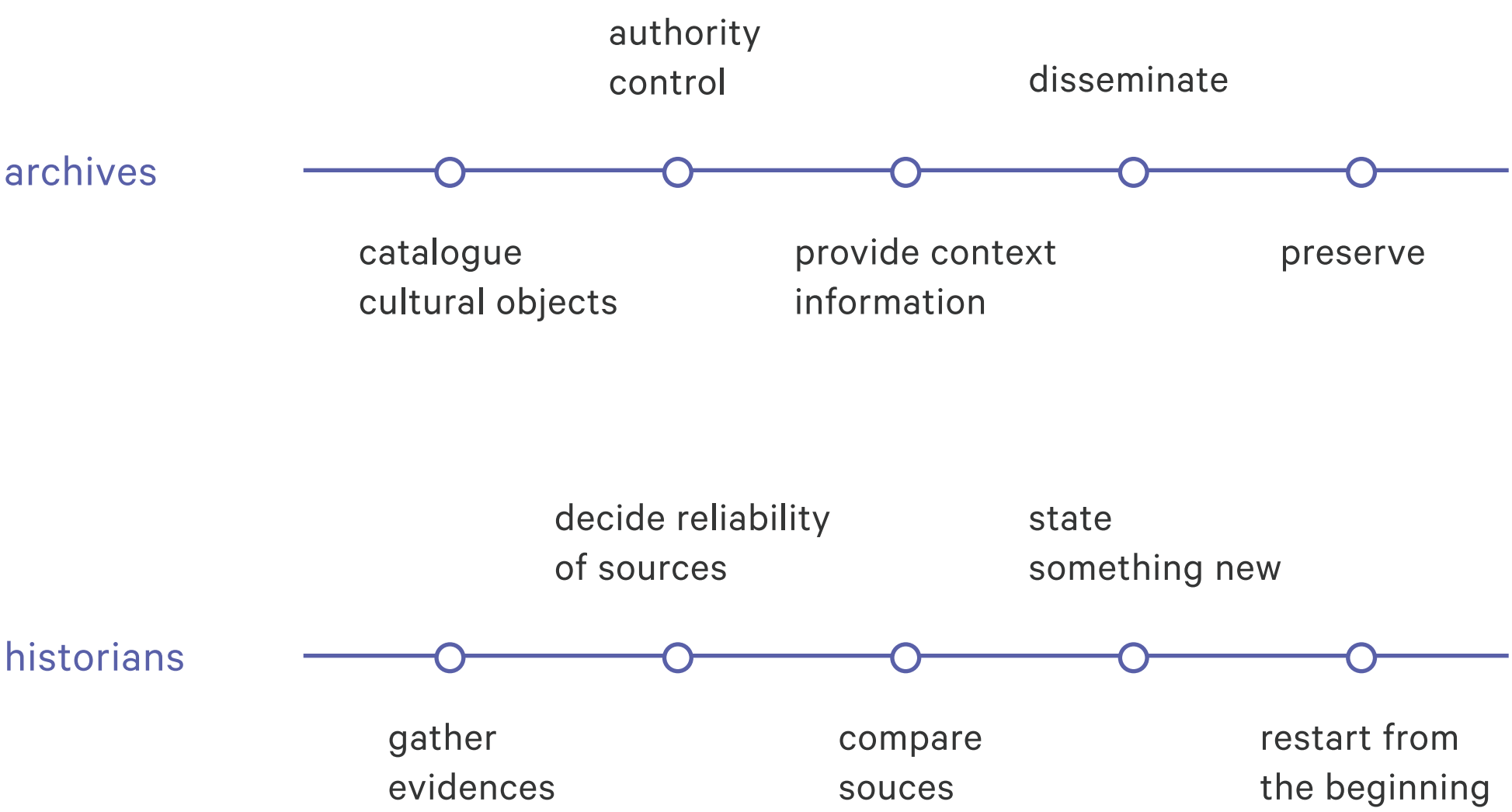
G L A M **H**

*contain pieces of information belonging to domains such as galleries, libraries, archives, museums... and **history** of course*

(will) provide structured, curated and trusted linked open data that will be reused (among the others) by **scholars**

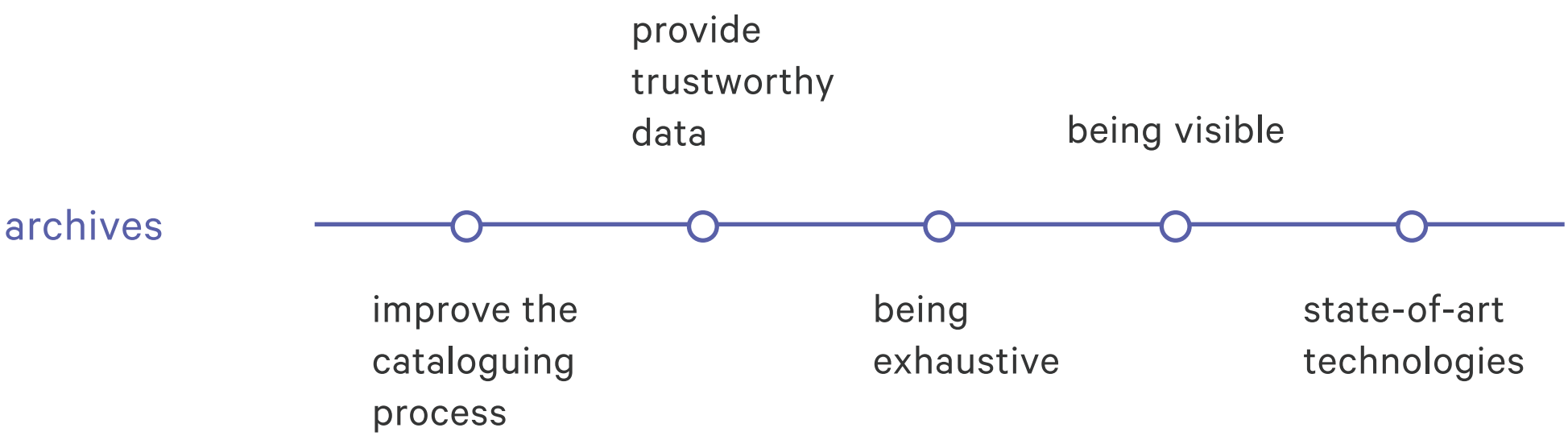
Archives and historians

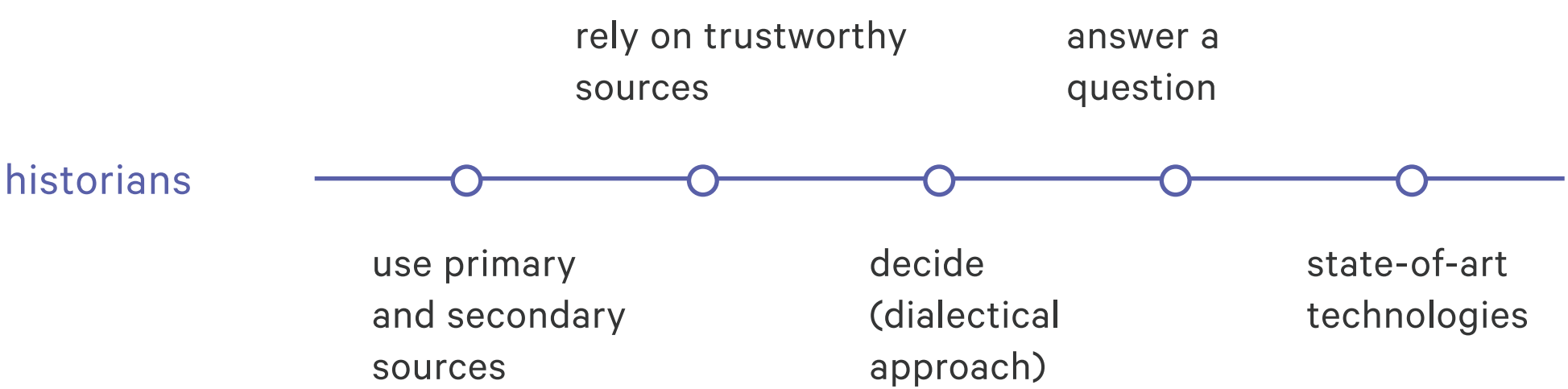
— what do they do?



Archives and historians

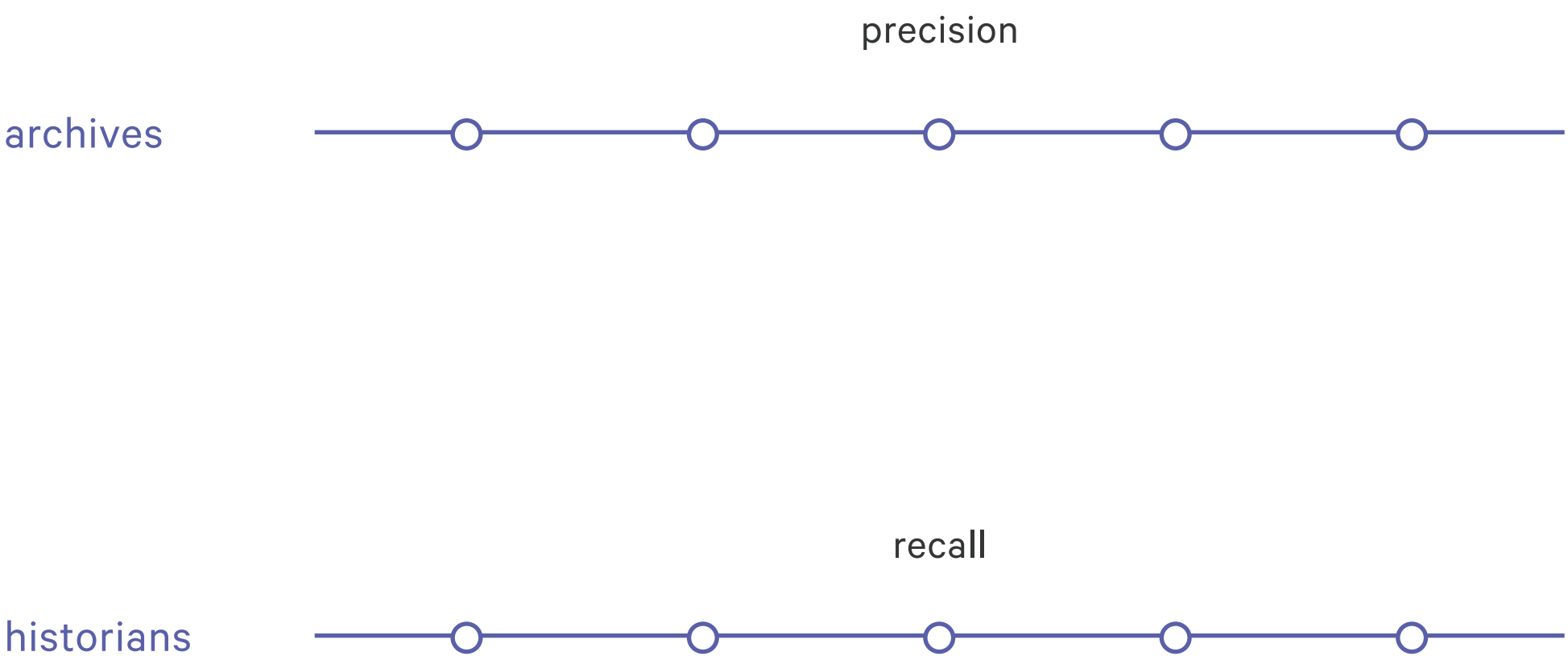
— what do they need?

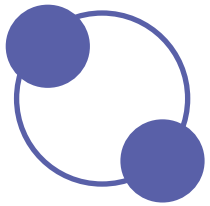




Archives and historians

— what do they want?





Archives and historians

— what they need, but they don't know it...

DATA RECONCILIATION	authority files to identify univocally resources
DATA CITATION	authoritative data sources to be cited
DATA AGGREGATION	ontology matching and data mashup
DATA EXPLORATION	serendipity is not enough...
DATA COMPLEXITY	completeness so as to really reuse data for research aims

INTERPRETATION AND AUTHORITATIVENESS | **DATA PROVENANCE AND TRUSTWORTHINESS**

The **Zeri & LODE** project

— the Zeri Photo Archive in Linked Open Data

INITIAL AIM *to publish Zeri's LOD in order to share data with PHAROS partners (International Consortium of Photo Archives)*

The first photo archive sharing its complete metadata element set in Linked Open Data

Daquino, Marilena, et al. (2016). "Enhancing semantic expressivity in the cultural heritage domain: exposing the Zeri Photo Archive as Linked Open Data." arXiv preprint arXiv:1605.01188 - forthcoming JOCCH

The **Zeri & LODE** project

— what they needed?

MODELING mapping the two national cataloguing standards to ontologies

RECONCILING data to trusted authority files and to other datasets

PUBLISHING data for developers' reuse, and to enrich the current database

The **Zeri & LODE** project

— what they wanted?

an increased **impact** on their stakeholders
(archive users, researchers, cataloguers, funders...)

authoritativeness by means of their tools for research
(comprehensive information, multiple sources)

significative increase of **quality** of data
(authority control, data cleaning, complete and structured data)

facilitate the **creation** of new datasets/archives
(reuse of their authority files, completion of partial information)

Tobias and the angel

— my problem statement by example

an artwork	Tobias and the angel
3 archives	**ZERI** Zeri Photo Archive **FRICK** Frick Art Reference Library **TATTI** Villa I Tatti - Harvard University
photographs (evidences)	

Tobias and the angel

— my problem statement by example

	authorship attributions	sources/criteria of attributions
ZERI	Apollonio di Giovanni	F. Zeri's archival classification (post 1965)
FRICK	Anonymous florentine, 15th c.	Davanzati catalogue (1916); R. Offner's attribution (1925); F. Zeri's attribution (1965)
TATTI	Bicci di Lorenzo	Inscription on verso of photo n. 710295; Berenson's classification (1967)

Tobias and the angel

— my problem statement by example

MERGE contradictory or partial information
by ensuring consistency of data

EVALUATE methodologies, sources (and common believes)
to rate the quality of an attribution

RECOMMEND the most authoritative attributions to users
and refine/update the data source

Research questions

— for humanists

authority

How to be the **most authoritative** source of information?
So other archives will use Zeri's catalogue as an authority file

innovation

How to **discover something new** by comparing different sources?
So that I don't have to do it manually

enrichment

How to be the **most complete** source of information?
So users will come to visit our catalogue rather than Google..

dissemination

How provide a **better experience** to the target user?
So historians will be more engaged

Research questions

— for digital humanists

modeling

How can other cataloguers **take advantage of Zeri's data**?
Which data they can reuse to improve their job?

methodology

Given a set of rules/believes shared among archives, how to exploit LOD to highlight the **most authoritative attribution**?

enhancement

How can Zeri's cataloguers **refine their data** by taking into account other data sources?

research

How can a historian use such knowledge/technologies to **compare sources**?

Research questions

— for computer scientists

modeling	Which **data models** satisfy requirements for data reuse?
evaluation	Which qualitative/quantitative methods can be applied to evaluate the **trustworthiness of a triple pattern**?
querying	How to **fetch data** from the web to assess the validity of a statement and update a source graph?
development	How to realize a **mash up application** to exploit all the previous findings?

Approach and methodology

— a ring to bring them all...

mapping content standards to RDF
ontology development
training dataset creation

Start from a complex use case
to represent the domain
— the Zeri & LODE project

conceptual framework
provenance annotations
trustworthiness policy

Define quantitative/qualitative methods
to evaluate the validity of a statement
— provenance and trustworthiness

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

Combine methods to fetch/merge/refine
and enrich data sources
— WYSIWHY

The **Zeri & LODE** project

— two mapping documents

mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

F ENTRY AND OA ENTRY

two national cataloguing standards issued by the Ministry of Cultural Heritage (MIBACT)

~200 fields

Mapping F Entry to RDF | https://doi.org/10.6084/m9.figshare.3175273.v1
Mapping OA Entry to RDF | https://doi.org/10.6084/m9.figshare.3175057.v1

The **Zeri & LODE** project

— two models for describing photo archives

mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

F ENTRY AND OA ENTRY ONTOLOGIES

developed by using SAMOD methodology

reuse of domain and task ontologies

- CIDOC-CRM (artworks)
- SPAR Ontologies (photos, bibliography)
- PROV Ontology (attributions, influence between artworks)

F Entry Ontology | http://www.essepuntato.it/2014/03/fentry
OA Entry Ontology | http://purl.org/emmedi/oaentry

The **Zeri & LODE** project

— a Linked Dataset

mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

ZERI PHOTO ARCHIVE RDF DATASET

scope: photos of Modern Art artworks

links to: DBpedia, Wikidata, geoNames, VIAF, Getty AAT, Getty ULAN, ICONCLASS

~11M RDF triples

Project page | https://w3id.org/zericatalog/

The **Zeri & LODE** project

— research questions and humanists' needs

modeling

How can other cataloguers **take advantage of Zeri's data**?
Which data they can reuse to improve their job?

DATA RECONCILIATION	authority files to identify univocally resources
DATA COMPLEXITY	completeness so as to really reuse data for research aims

Daquino, Marilena (2016): Presentation ICCD Seminar - Photographs in the Semantic Web. figshare.
https://doi.org/10.6084/m9.figshare.4479488.v1

The **Zeri & LODE** project

— an example of data reuse

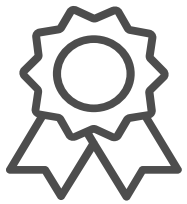
dating a new photograph by means of Zeri's data

a photographer may change name several times during his activity.
Recording these changes helps to date a new photograph

time-indexed value in context (TVC) ontology pattern

```
<organization/2087/anderson>
     tvc:hasValue <name/stab-d-anderson/1877-1938> .

<name/stab-d-anderson/1877-1938> a tvc:ValueInTime ;
        tvc:atTime <date/1877-1938> ;
        tvc:withValue <name/stab-d-anderson> .
```



provenance/trustworthiness

— a model to represent interpretations (attributions)

mapping content standards to RDF
ontology development
training dataset creation

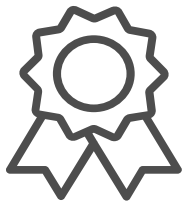
conceptual framework
provenance annotations
trustworthiness policy

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

HICO ONTOLOGY (EXTENSION OF PROV-O)
to describe the provenance of a RDF triple which represents an attribution, including: criteria, sources used to support a questionable information, and RDF creator

A CONTROLLED VOCABULARY FOR CRITERIA
e.g. bibliography, technical analysis of photographs

RATING OF CRITERIA
to assess the trustworthiness of an attribution
e.g. last recorded attribution



Provenance

— in data integration

mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

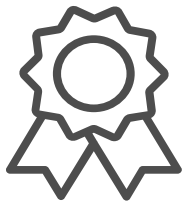
catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

PROVENANCE REPRESENTATION

Annotation approach: store data
Inversion approach: examine the source

Agent-oriented: who created data
Object-oriented: the history of the entity
Process-oriented: activities/observations needed to generate an entity

Named graphs
Federated approach and merge according to a common vocabulary



Provenance

— in data integration

mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

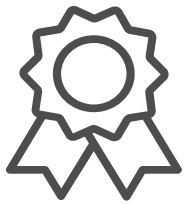
PROVENANCE GRANULARITY

Workflow provenance: software and processes

Data provenance: history of entities

Instance-level mapping

Schema-level mapping



Trustworthiness

— in data sources

mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

DOMAIN-DEPENDENT

First hand information

Trust ratings provided by third parties

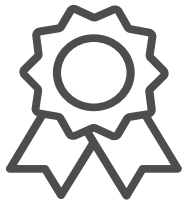
DOMAIN-INDEPENDENT

Information explicitly annotated

+ transitivity of attributes

Number of sources in agreement

Retrieval date



provenance/trustworthiness

— research questions and humanists' needs

methodology

Given a set of rules/believes shared among archives, how to exploit LOD to highlight the **most authoritative attribution**?

DATA CITATION authoritative data sources to be cited

WYSIWHY

— what you see is why

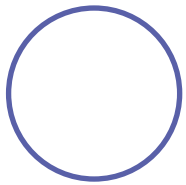
a tool for

HISTORIANS / USERS

knowledge discovery, comparison of sources

ARCHIVES / DATA PUBLISHERS

trustworthiness evaluating, data refining and enrichment by means of crowdsourcing



WYSIWHY

— what you see is why

mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

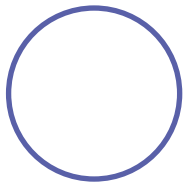
catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

given a simple triple pattern representing
a questionable information

INSTANCE-LEVEL MAPPING
look into the source graph (i.e. Zeri's data)
for explicit equivalences (subject) and look
iteratively into graphs of other equivalent URIs

CATALOG OF TRUSTED DATASETS
look into the catalog whether aforementioned
equivalent URI bases are found

SCHEMA-LEVEL MAPPING
look into a mapping document the BTP
needed to rewrite a SPARQL query



WYSIWHY

— what you see is why

mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

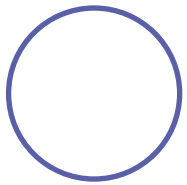
catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

QUERY
look into a settings document how to query the graphs

- content negotiation
- SPARQL endpoint
- Linked Data Fragments

time for retrieving results may vary
LDF improve significantly the speed

MERGE RESULTS
create a new graph and group observations by found objects



WYSIWHY

— what you see is why

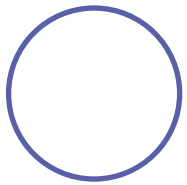
mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

INSTANCE-LEVEL MAPPING
compare found objects with the one proposed in the originary triple pattern

STATISTICS
sources in agreement and in disagreement



WYSI**WHY**

— what you see is why

mapping content standards to RDF
ontology development
training dataset creation

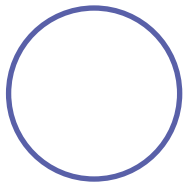
conceptual framework
provenance annotations
trustworthiness policy

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

DO MORE!
provide final users of all the information
they need to assess the validity of a statement

- cited primary sources
- bibliography
- images/photographs
- criteria/motivations
- date of attribution

...



WYSIWHY

— what you see is why

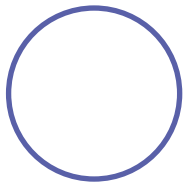
mapping content standards to RDF
ontology development
training dataset creation

conceptual framework
provenance annotations
trustworthiness policy

catalog of trusted datasets
dereferenced URIs / SPARQL / LDF
mashup web application

REFINE DATA
users provide crowdsourced data linking when they find mistakes in other datasets

refined data are gathered in linksets



WYSI**WHY**

— research questions and humanists' needs

enhancement How can Zeri's cataloguers **refine their data** by taking into account other data sources?

research How can a historian use such knowledge/technologies to **compare sources**?

DATA AGGREGATION ontology matching and data mashup

DATA EXPLORATION serendipity is not enough...

Conclusions

— issues and future works

Data reconciliation of artworks is a hard task

- similar data are not yet easy to be found (waiting for PHAROS)
- names of artworks are misleading
- automatic tools for reconciling lead to mistakes

IMPROVEMENTS

- linksets can be created and used instead of fetching data
- test an authority file driven approach - look into datasets for entities equivalent to a shared URI (e.g. VIAF)
- image recognition to find similarities

- make assumptions on the basis of sources/motivations provided in data sources

Conclusions

— issues and future works

Fetching data from the web might be slooooow

- SPARQL endpoints bottle necks

IMPROVEMENTS

- linksets will improve the speed
- Linked Data Fragments halve time of query

Crowdsourcing has to be managed

- provision of questionable statements

IMPROVEMENTS

- define a strategy for data management and trustworthiness policy
- update the triple store with approved information

KMI RESEARCH SEMINAR
31 MAY, OPEN UNIVERSITY

Thank you!

Marilena Daquino | marilena.daquino2@unibo.it