DHDK 2018/2019 SEMINARS - OCTOBER 31ST, BOLOGNA

Getting started in the Digital Humanities

An how-to guide

Marilena Daquino
FICLIT, University of Bologna

outline

Humanities Data

data life cycle

tools, services, methodologies

Research in Digital Humanities

research fields, journals, communities

Research Project Management

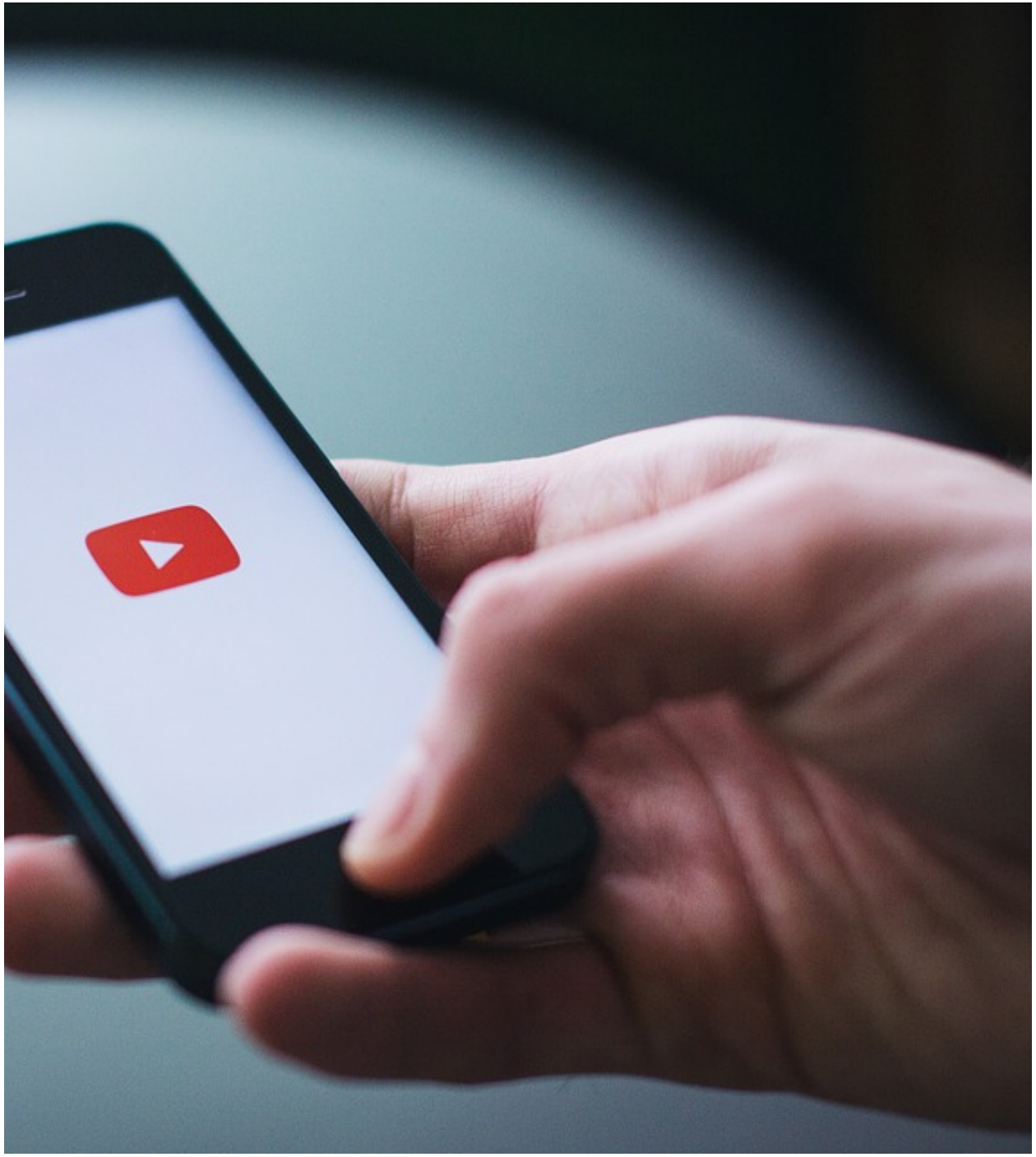
create your project

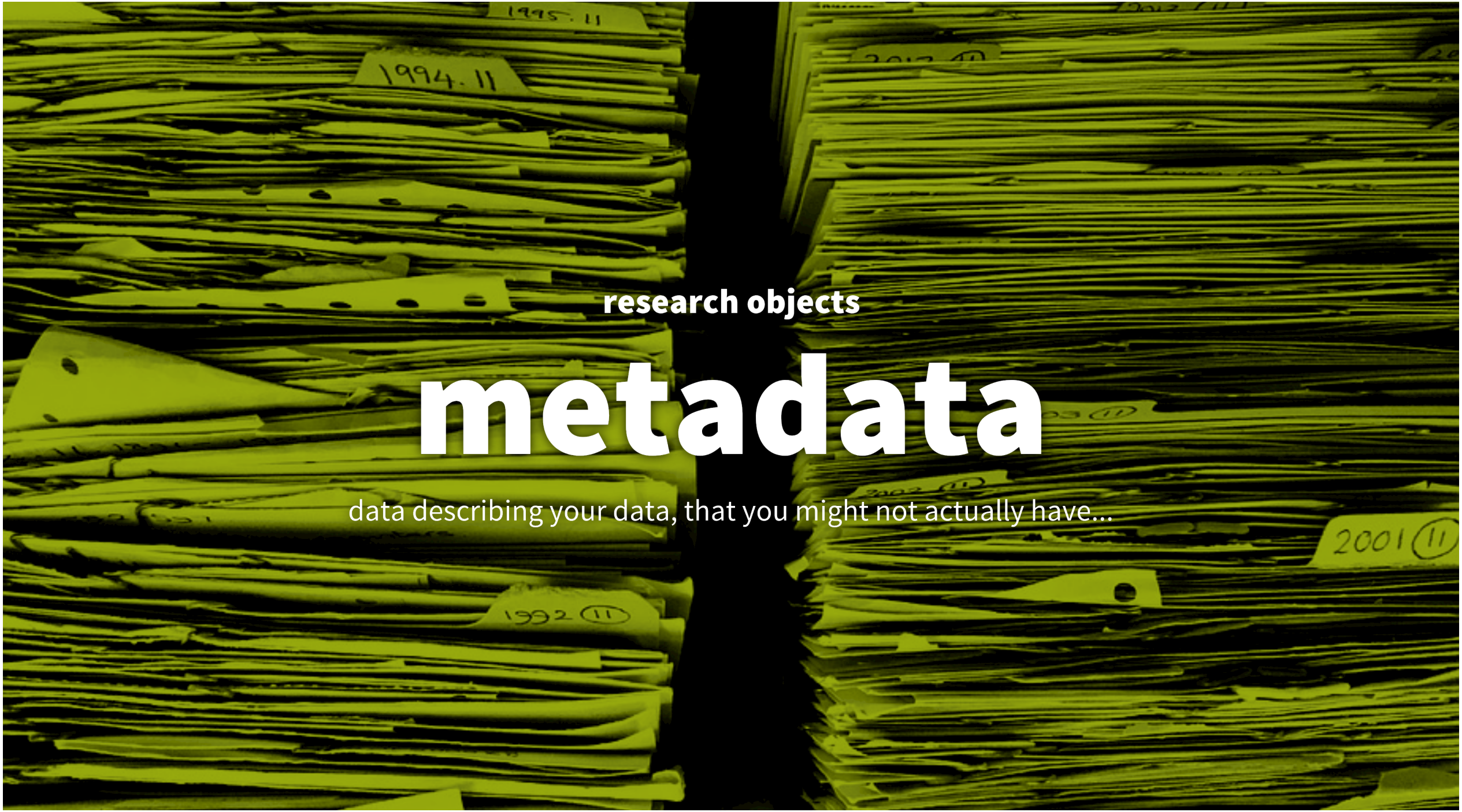
anatomy of a project presentation

Humanities data

Humanities data in a DH centre

https://www.youtube.com/watch?v=zdlOC0sFo5k





research objects

metadata

data describing your data, that you might not actually have...

What are Humanities data?

1 **Research data**

must comply with requirements for being shared and reused

2 **Cultural Heritage data**

must comply with community standards and users' requirements

Research data

unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results

A classification of research data

observational

e.g. survey data

experimental

e.g. stemmatics

simulation

e.g. virtual exhibition

derived

e.g. data mining, text analysis

reference or canonical

e.g. text corpora, databases

What type of data

e.g. literary texts

raw/initially processed data

e.g. plain text

‘research ready’ processed data, cleaned and annotated

e.g. a critical apparatus

published output dataset

e.g. a scholarly edition

cataloguing data

e.g. bibliographic information

What data formats

Text

e.g. plain text, RTF, Word, PDF, XML

Numerical

e.g. Excel, csv

Multimedia

e.g. jpeg, tiff, IIIF, mp3

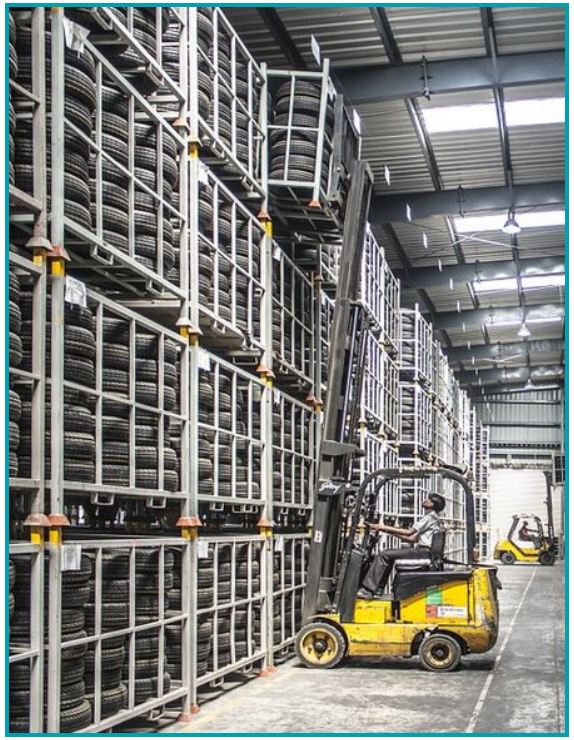
Software

e.g. Java, C, python

Research output

Documents

questionaires, transcriptions, codebooks, images, movies, slides, articles, books



Datasets

spreadsheets, corpora, data files

Methodologies, workflows

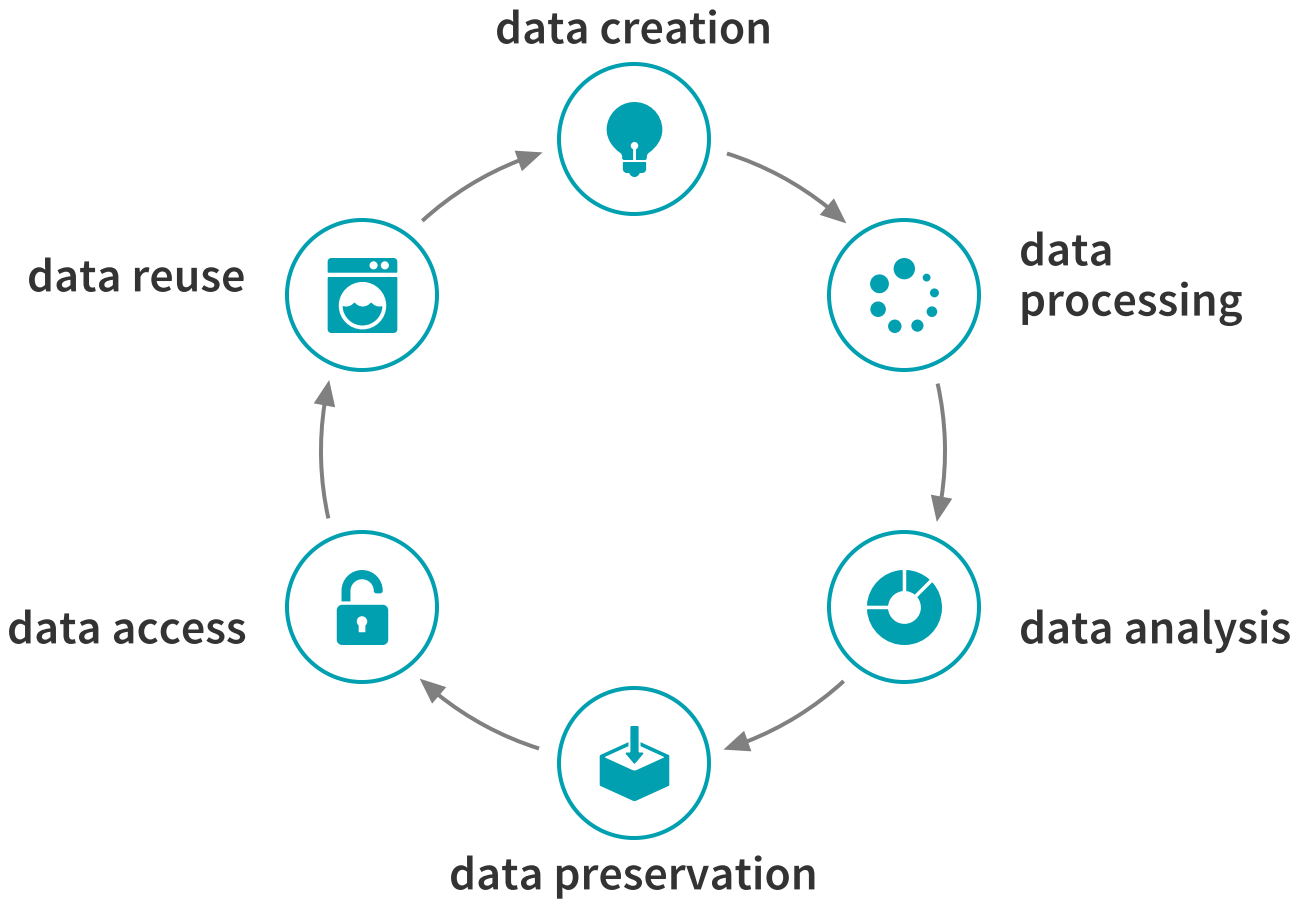
best practises, models, procedures

Applications

web applications, algorithms, scripts

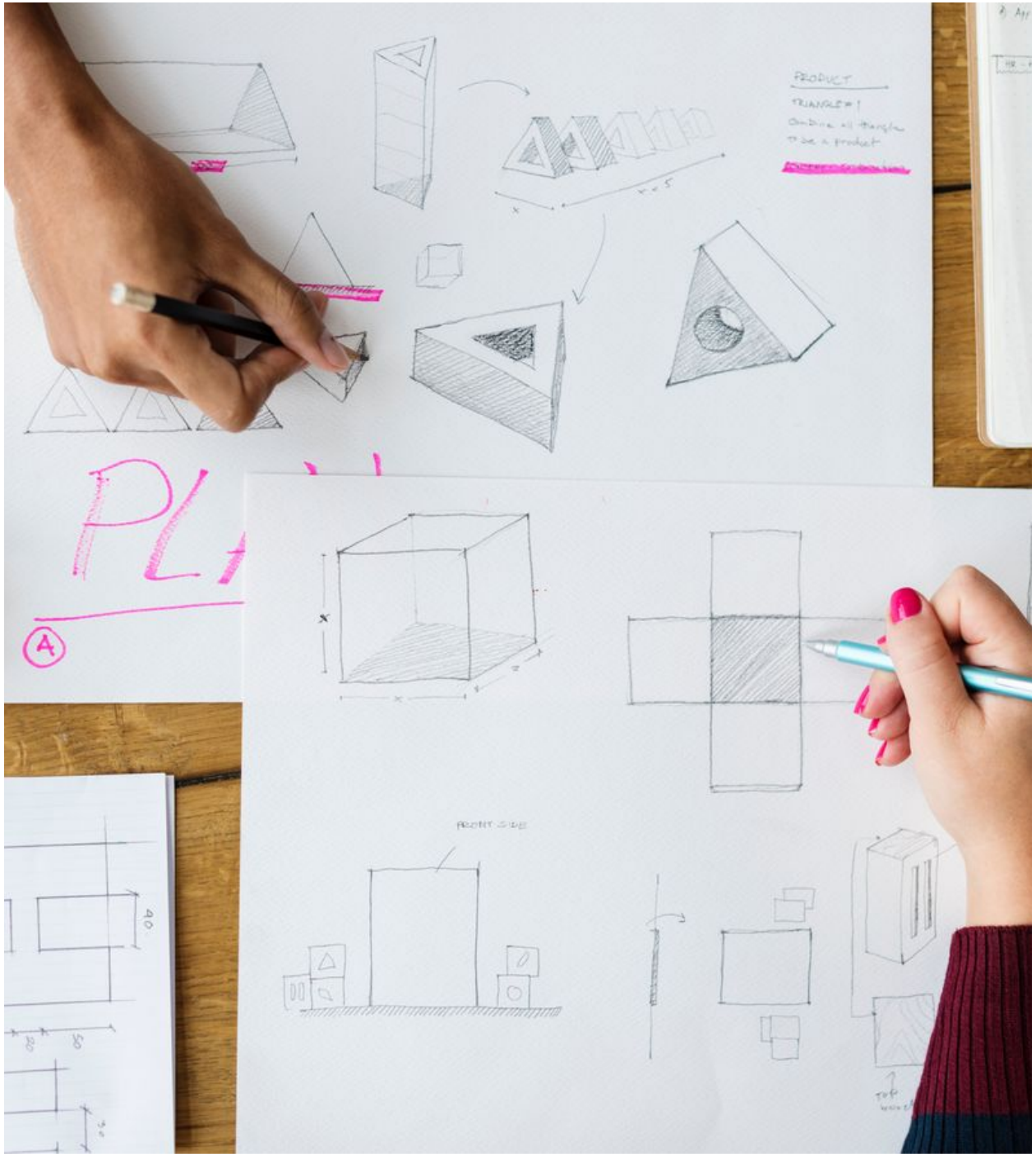
Research data life cycle

in six steps



data creation

or data collection



What is Humanities data about?

cultural objects

artworks, photographs, manuscripts, books, articles...

held by galleries, libraries, archives, and museums (GLAM)

intangible heritage

historical events, traditions, people relationships, creative industries, iconography

Planning data creation/collection

Manage expectations

- data are created or collected according to an intended purpose and to answer precise research questions
- formulate competency questions to understand what questions your data should be able to answer

Identify requirements

- identify data quality metrics
- identify (technical and community) standards
- determine intellectual property rights and licensing options for data reuse

Appraisal and selection

- define minimum requirements (content, volume) according to the use case
- identify valuable data among the ones collected for research purposes

An example: planning a scholarly edition

Manage expectations

- collect witnesses of a text and secondary literature are gathered to reconstruct the story of a text
- manuscripts provide insights on author's variants, creative process, changes over time that allow to define a timeline

Identify requirements

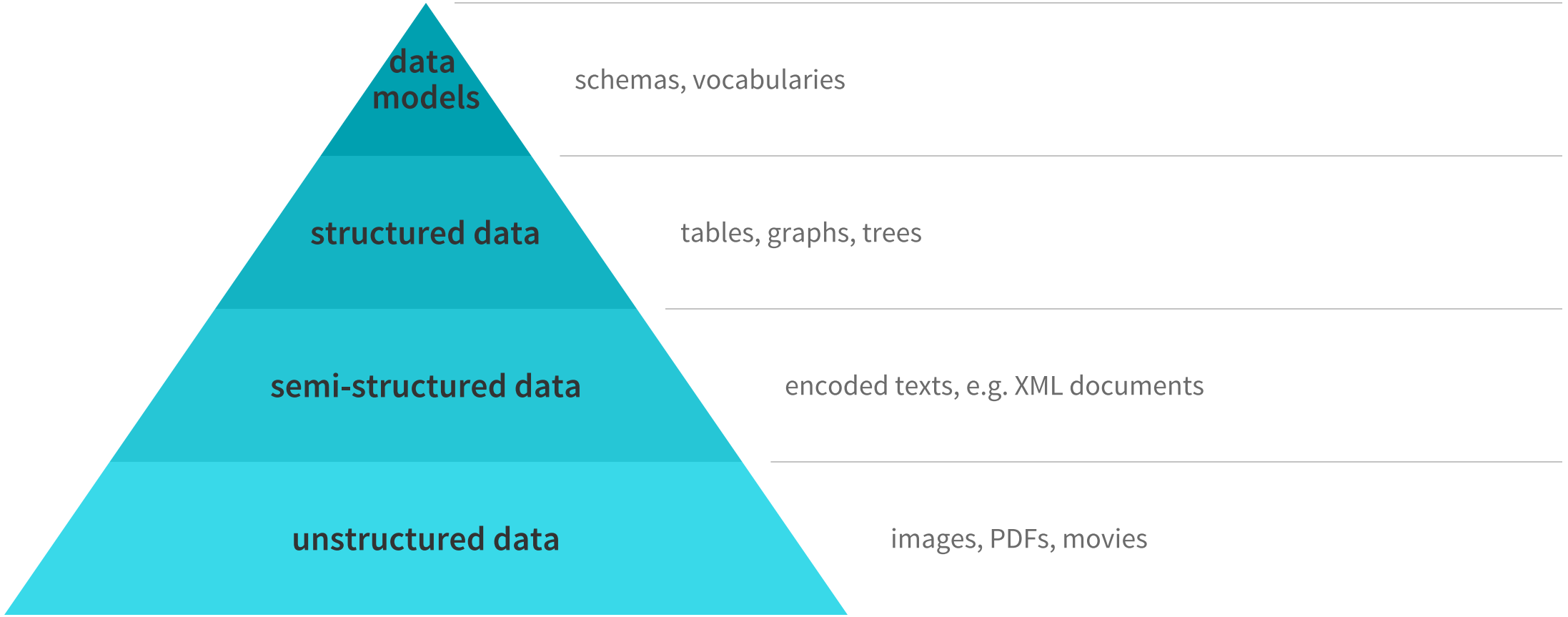
- transcription rules should be homogeneous in all the witnesses
- encoded transcription using XML syntax and TEI vocabulary
- texts are available in Public Domain and the resulting edition is released in CC-BY

Appraisal and selection

- not all the texts deserve to be transcribed, only the ones that include original contents (e.g. discard secondary literature)
- address relevant authority files for identifying people, organisations, places; select finding aids, catalogues and relevant secondary sources

Data creation (or collection)

which data you will create or collect



Unstructured data

e.g. facsimiles of a text to be transcribed

- **licenses**

  double-check what you can do with third-party images, videos, audio contents

- **quality**

  quality is a fit-for-purpose value; consider feasible storage solutions

- **format**

  for long-term preservation purposes (e.g. TIFF, AVI, WAV)

- **describe**

  add metadata to your multimedia contents

Where to start from

licenses

https://creativecommons.org/licenses/

formats

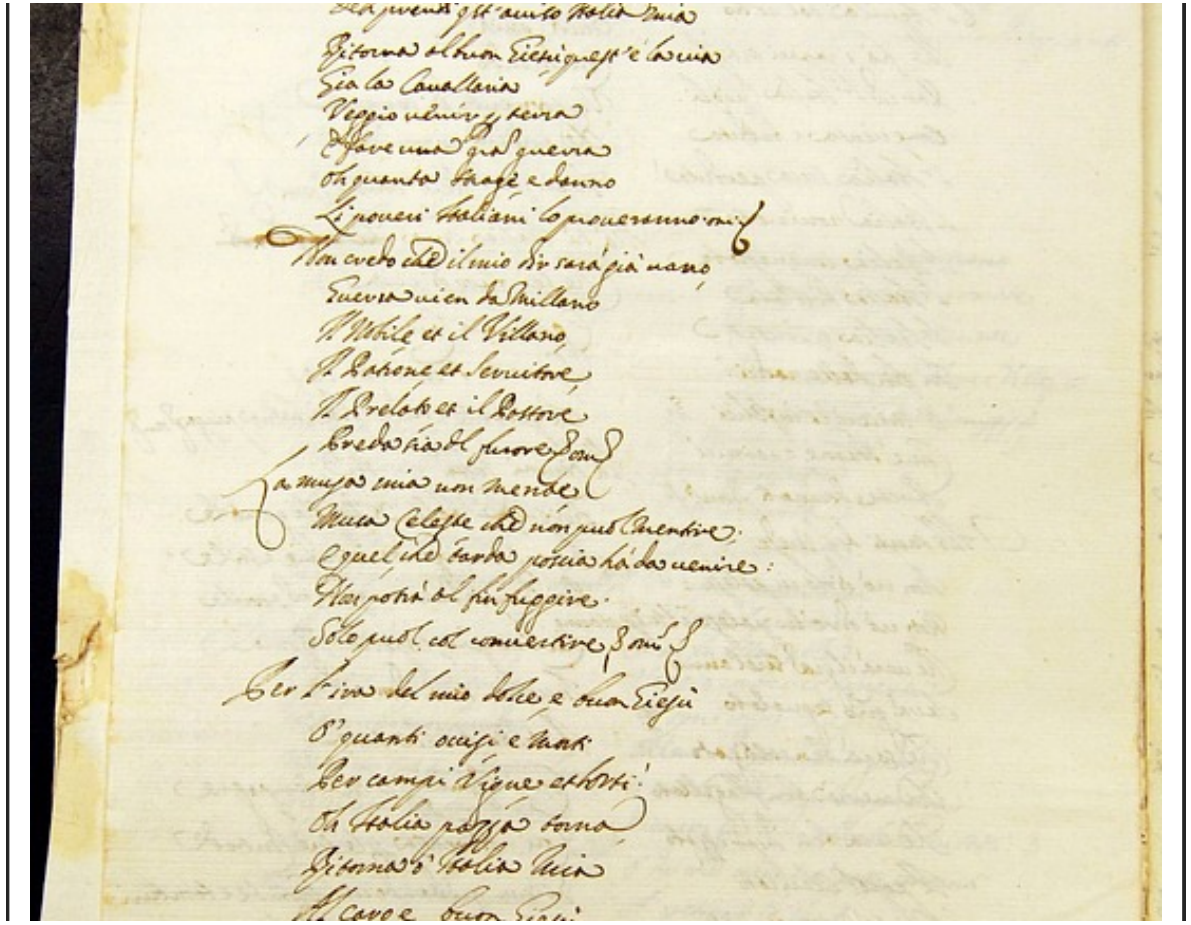
http://www.loc.gov/preservation/resources/rfs/

metadata element sets

http://dublincore.org/documents/dces/

an example

a collection of digitizations of a manuscript that will be part of a digital edition



- **describe your collection of images and single items**

  describe the digital medium (i.e. images) first

  (author of the picture, format of the image, date of shot, etc.)

- **describe the content of the images**

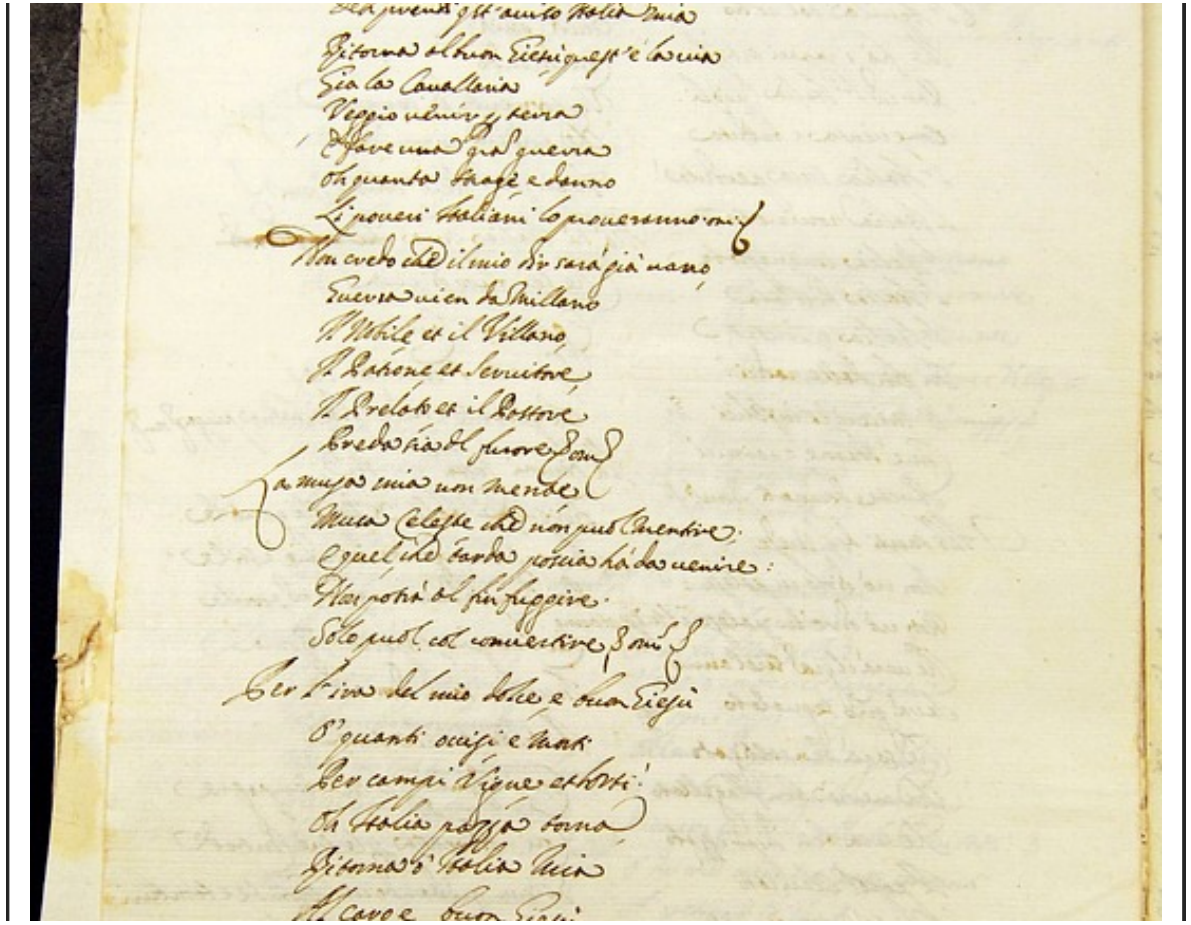
  describe the physical medium, the provenance, its content, etc.

- **describe your intellectual contribution**

  authors, contributors, dates of the digital edition

an example

a digitized page of a manuscript



Dublin core (image)

title = 'MS5, p.6'

creator = 'marilena daquino'

subject = 'manuscript'

date = '2018'

type = 'image'

format = 'JPG'

source="RC607.A26W574 1996"

rights = 'Public Domain'

Some solutions for collecting multimedia contents

see more at http://dirtdirectory.org/tadirah/sharing

- **Omeka.net**

  a web-publishing platform to curate collections and create exhibitions of digitized content. Does not require the user to provide hosting or to maintain their installation in any way. Hosting more than 500MB of content requires a paid account.
  http://www.omeka.net/

- **Recogito 2**

  an online platform for collaborative document annotation. Recogito provides a personal workspace where you can upload, collect and organize your source materials - texts and images - and collaborate in their annotation and interpretation.
  http://recogito.pelagios.org/

- **StoryMapJS**

  a project which aims to help journalists and historians tell stories by using maps, highlighting the locations of a series of events. https://storymap.knightlab.com/

- **FigShare**

  a repository where users can make all of their research outputs available in a citable, shareable and discoverable manner. All file formats can be published, including videos and datasets that are often demoted to the supplemental materials section in current publishing models. https://figshare.com/

Semi-structured data

e.g. encoded transcription of a text

- **community standards and best practices**

  do not reinvent the wheel. Use what already exists and what other people in your community use for managing semi-structured data

- **define dimensions and a model**

  structure as much as possible your contents in a tree fashion; manage exceptions; decide what is in and what is out; conceptualize a model

Where to start from

Markup languages

https://www.w3.org/XML/

but also HTML

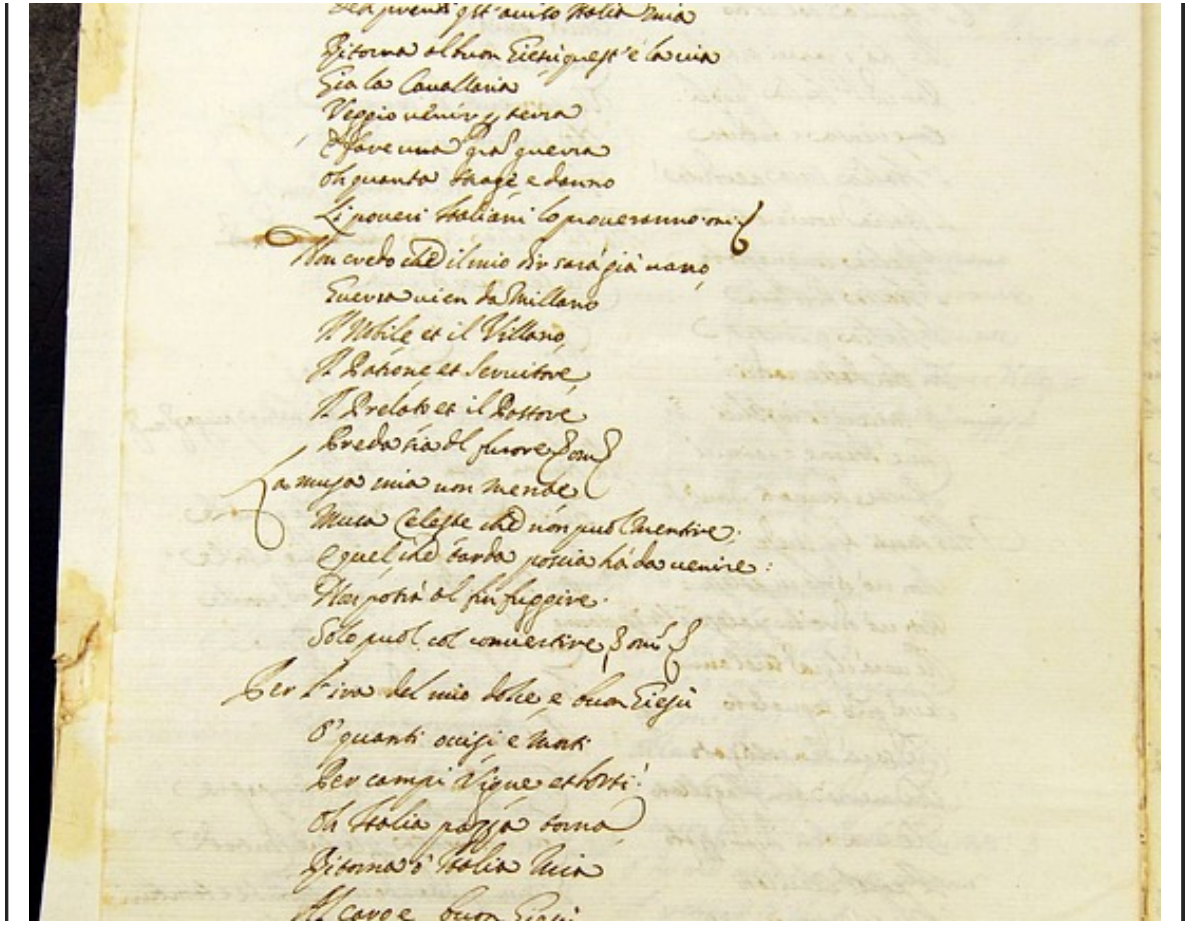
https://www.w3.org/standards/webdesign/htmlcss

Rich text editors

https://www.sublimetext.com/2

an example

an HTML document including information on a digitized page of a manuscript



Dublin core HTML

<meta name='DC.Title' content='MS5, p.6'>

<meta name='DC.Creator' content= 'marilena daquino'>

<meta name='DC.Subject' content = 'manuscript'>

<meta name='DC.Date' content= '2018'>

<meta name='DC.Type' content= 'image'>

<meta name='DC.Format' content= 'JPG'>

<meta name='DC.Source' content='RC607.A26W574 1996'>

<meta name='DC.rights' content= 'Public Domain'>

Some solutions for creating semi-structured data

- **oXygen XML editor**

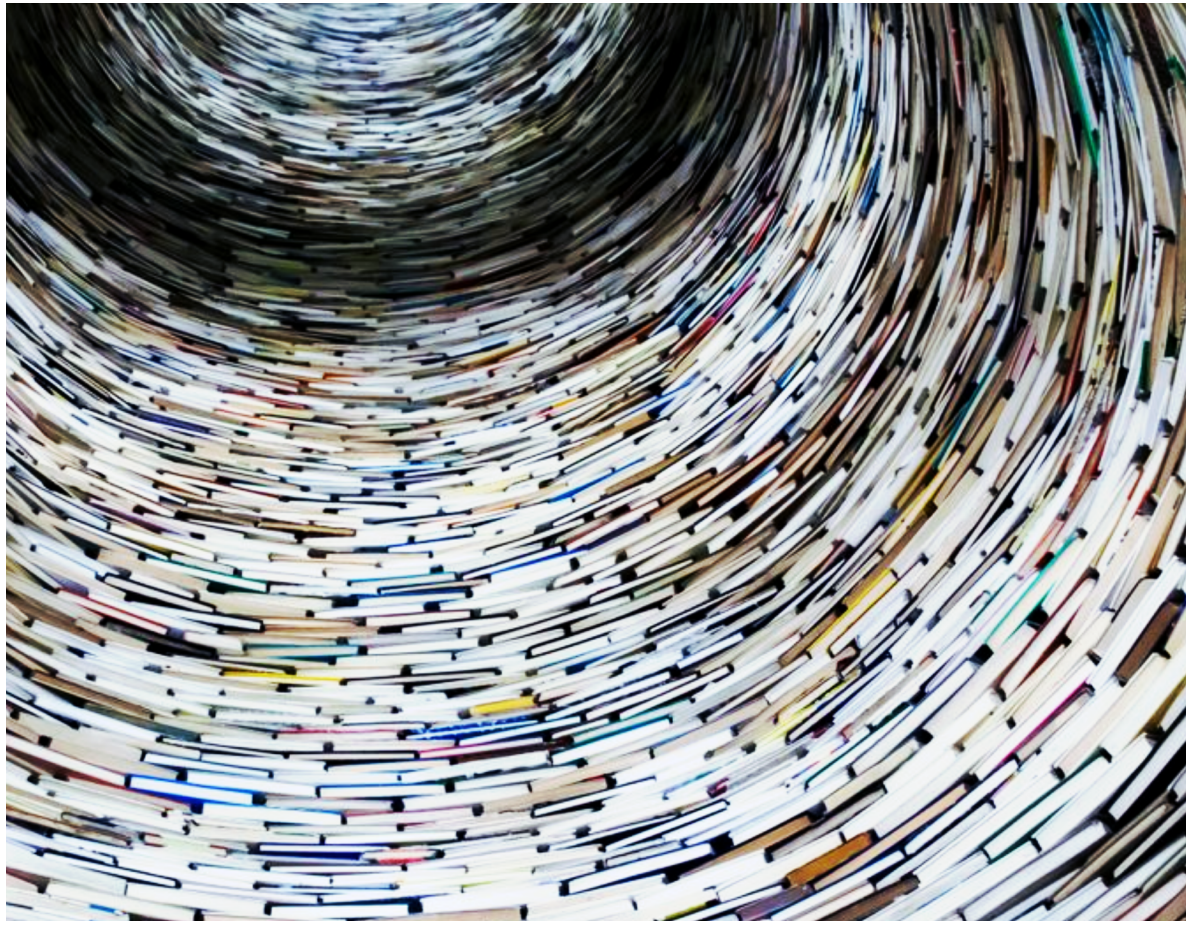
  a cross-platform XML editor that may be used to create and validate XML documents and associated schema. OXygen XML Editor works with all XML-based technologies, including XML databases and web services and comes with ready-to-use DITA, DocBook, TEI, and XHTML support. Academic license available.
  http://www.oxygenxml.com

- **eXist-db**

  an open source NO SQL database management system that stores XML data according to the XML data model and features efficient, index-based XQuery processing.
  http://exist-db.org

Structured data

e.g. bibliographic records



- **put everything in tables**

  organize contents in tables and link tables with each others

- **define and document your variables**

  column names, data types (e.g. strings, dates, URLs)

- **model your database**

  describe how data variables interact with each others

Where to start from

Entity-Relationship model

https://en.wikipedia.org/wikiEntity%E2%80%93relationship_model/

Data types

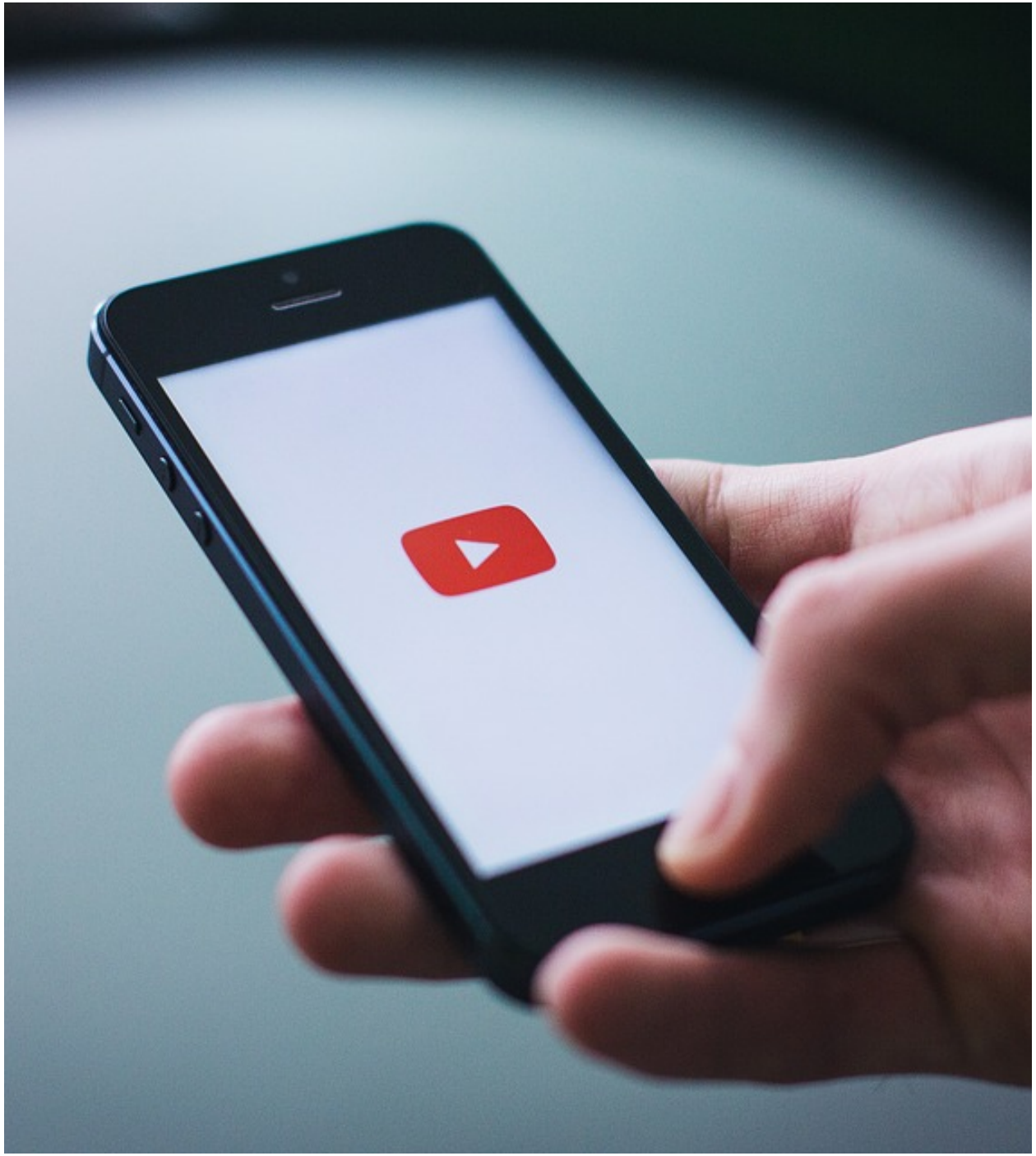
https://www.w3.org/TRxmlschema11-2/#built-in-datatypes/

SQL query language

https://en.wikipedia.org/wiki/SQL

ER diagram

https://www.youtube.com/watch?v=eJWpP2JeSGU



Some solutions for creating structured data

relational databases and beyond

- **MySQL**

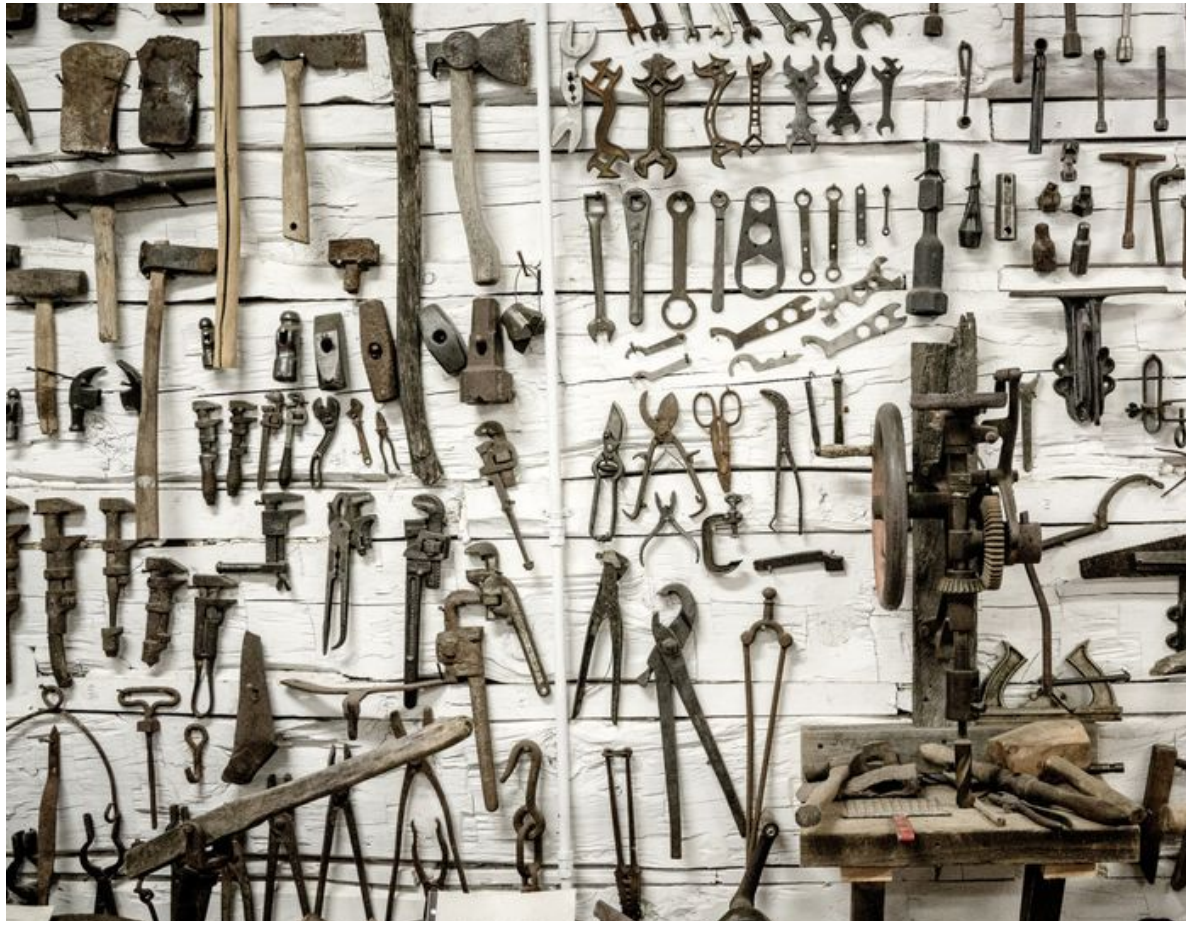
  an open-source relational database management system (RDBMS) including a number of features for creating, managing and querying structured data. http://mysql.com

- **any table!**

  but not Excel! use non-proprietary formats (such as CSV, JSON) as much as possible to allow compatibility with other software and enable people to reuse your data without legacy tools

Schemas, vocabularies, data models

describe pieces of data according to existing guidelines



- **thesauri and taxonomies**

  controlled lists of terms (concepts) + hierarchical/associative relations

- **vocabularies and ontologies**

  more complex relations between concepts

- **application profiles**

  define how to model your data according to a given vocabulary

Where to start from

Text Encoding Initiative

http://www.tei-c.org/

EAD/EAC archival standards

https://www.loc.gov/ead/

https://eac.staatsbibliothek-berlin.de/

LIDO museum standard

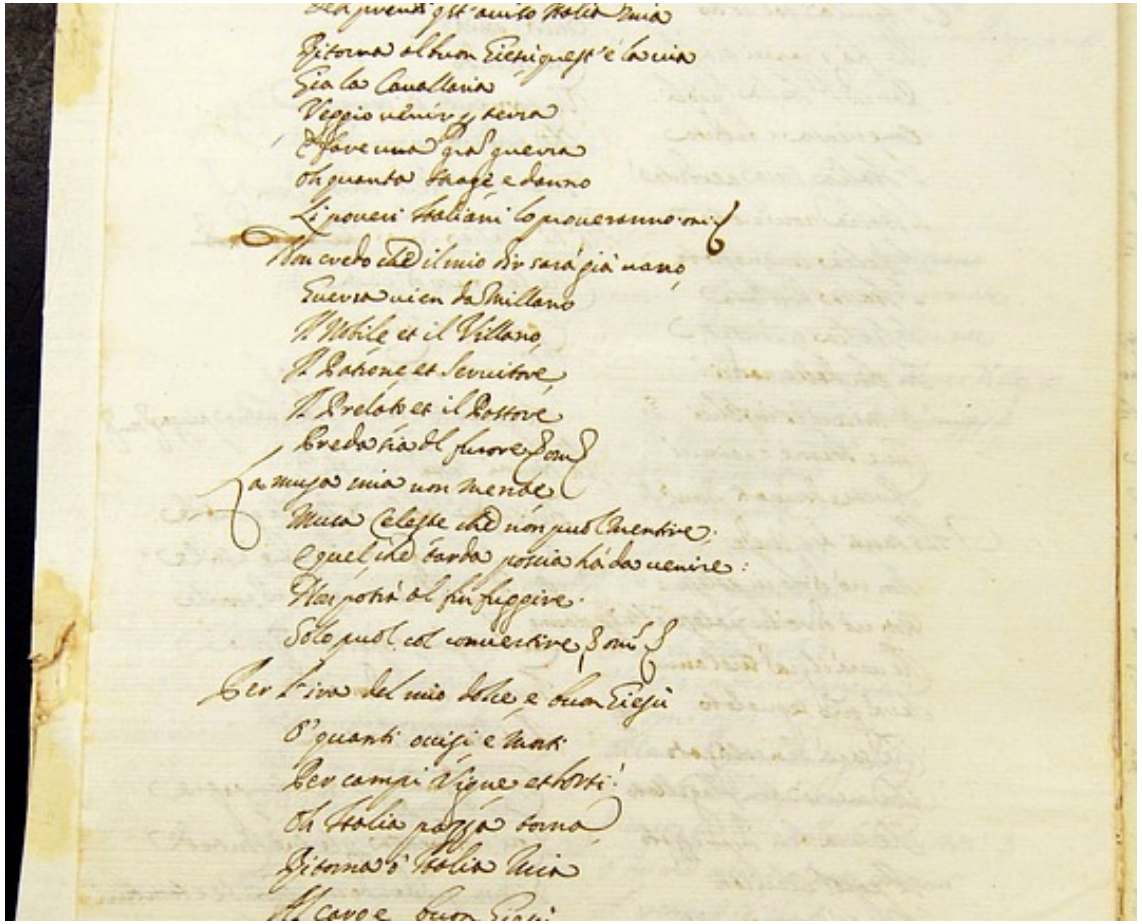
http://network.icom.museumcidoc/working-groups/lido/what-is-lido/

FRBR family of library standards

https://www.loc.gov/cds/downloads/FRBR.PDF

an example

a XML/TEI document including the transcription of a digitized page of a manuscript



XML/TEI document

```
<TEIheader>
  <titleStmt><title>MS5</title></titleStmt>
  <sourceDesc>RC607.A26W574 1996 </sourceDesc>
</TEIheader>
<text>
 <body>
    <pb n='6'/>

Lorem ipsum dolor sit amet, consectetur adipiscing...

 </body>
</text>
```

Some resources you may find useful

- **AAT vocabularies**

  materials, types of object
  http://www.getty.edu/research/tools/vocabularies/aat/

- **Library of Congress Subject Headings**

  subjects - http://id.loc.gov/authorities/subjects

- **ICONCLASS**

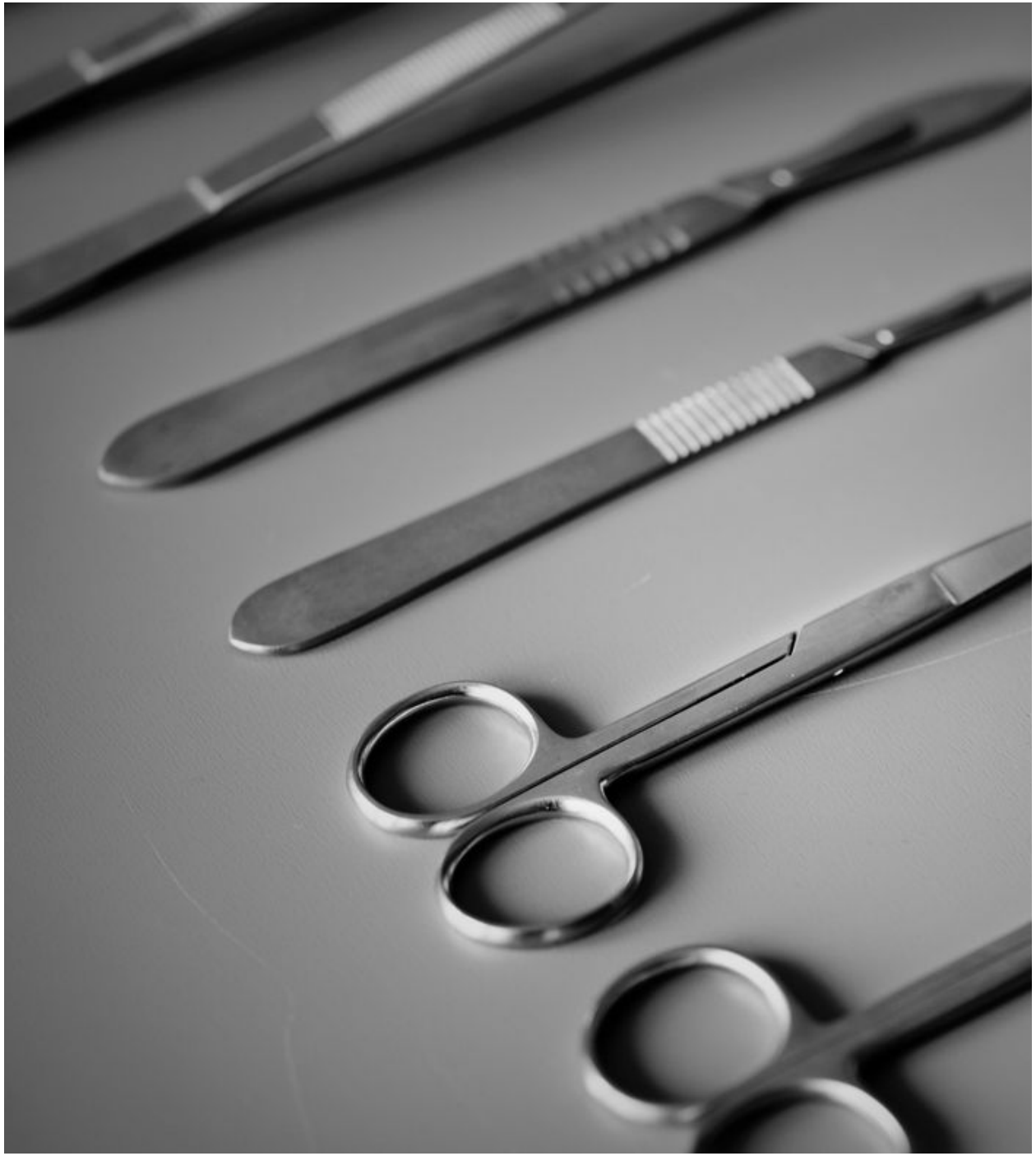
  iconographic themes - http://www.iconclass.nl/

- **VIAF**

  people and organisations - http://viaf.org

data processing

the collection and manipulation of data
to produce meaningful information



Process your data before analyzing it

e.g. get a raw transcription of your text by performing OCR on digitized facsimiles

convert

to another format or language

e.g. JPEG to TIFF

validate

data cleansing, check correctness

e.g. double-check of OCR

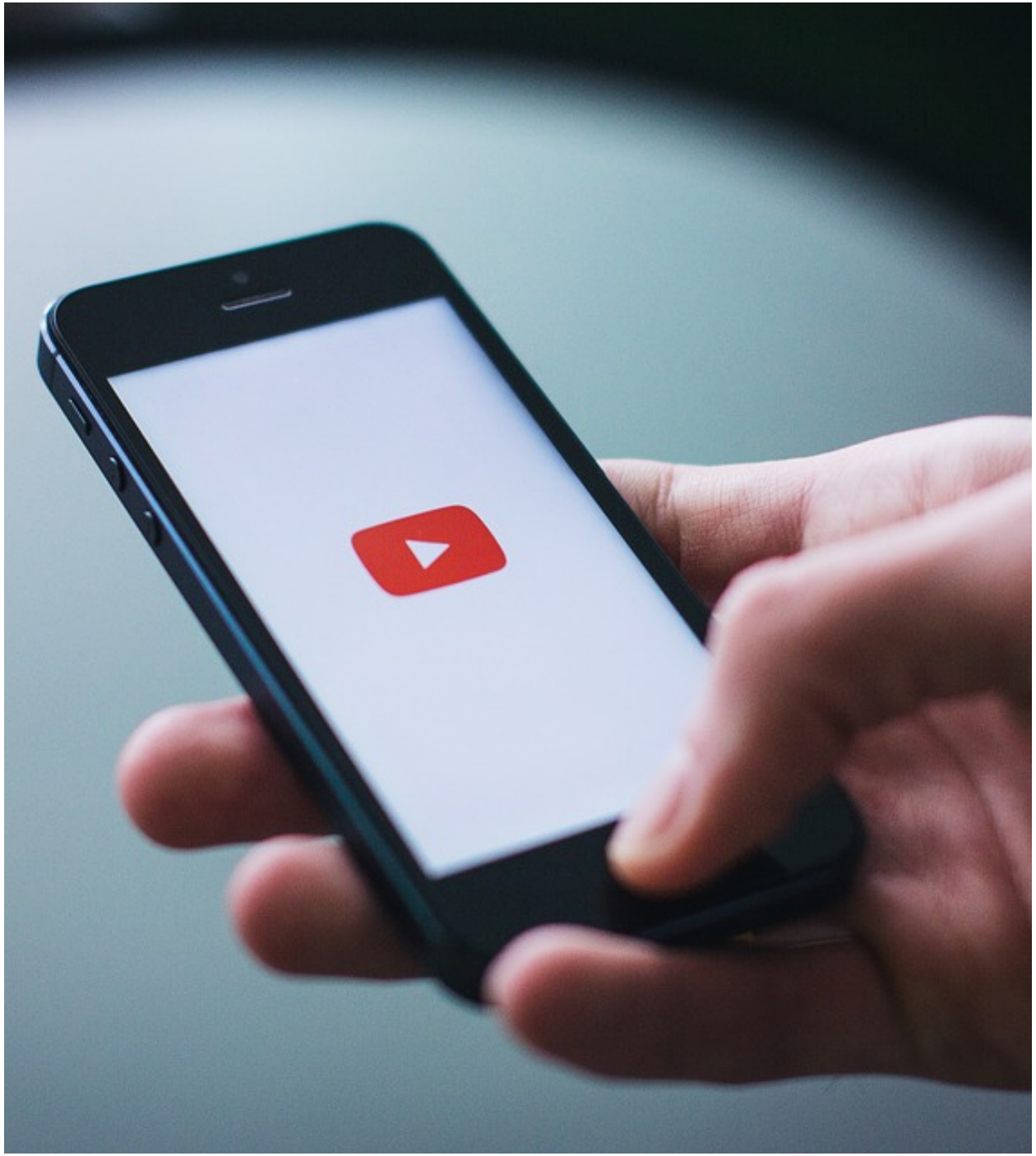
arrange or aggregate

sort and group items

e.g. group pages by witness or collate pages by page number

OCR

https://www.youtube.com/watch?v=cAkklvGE5io



Data quality

humanities data, like other research data, must be valid, accurate, complete, and consistent

- **validity**

  define datatype and range constraints (e.g. dates)

- **accuracy**

  fit for purpose; homogeneity and richness of data

- **completeness (or representativeness)**

  fit-for-purpose; whatever allows you to perform your analysis

- **consistency**

  across systems, e.g. usage of identifiers, non-contradictory information

Some solutions for cleaning your data

- **OpenRefine**

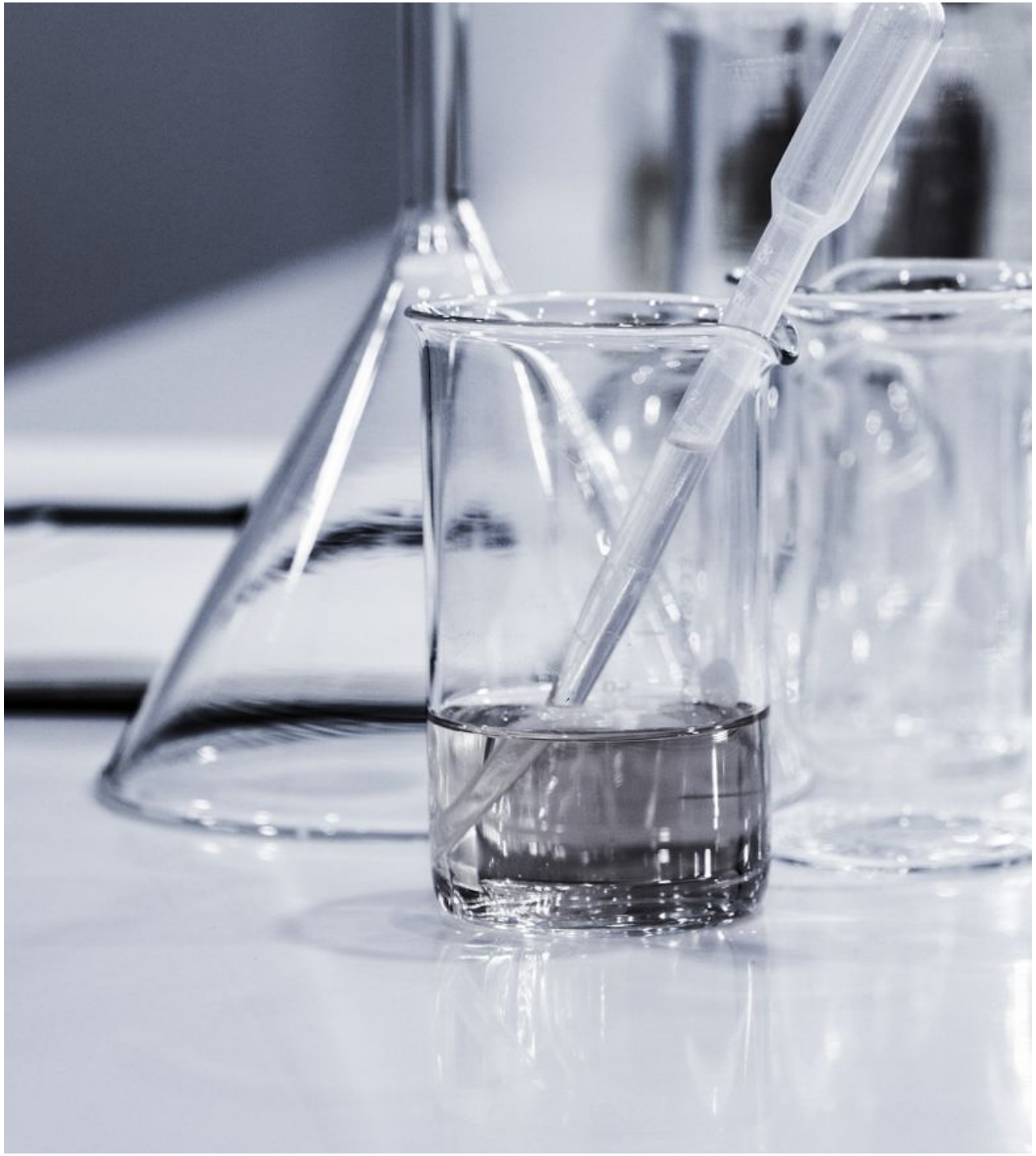
  an open source desktop application for data cleanup and transformation to other formats

  http://openrefine.org/

- **your beautiful hands :)**

  despite many solutions are available for automatically cleaning data, human supervision is often required to ensure accuracy

data analysis



What types of analysis

according to the type of data at hand, you can perform several types of analysis

text analysis

Natural Language Processing, linguistic and stylistic analysis

data mining

quantitative and qualitative analysis for extracting patterns and latent knowledge

computer vision

mathematical algorithms for pattern recognitions in images and videos

statistics

qualitative/quantitative analysis on large scale datasets for describing some phenomenon

Some solutions for analyzing your data

- **R studio**

  examine data through visualizations and broad summary statistics.

  https://www.rstudio.com/

- **Python libraries**

  Program by your own, reuse existing libraries dedicated to data analysis

  e.g. Pandas (statistics), NLTK (NLP)

data preservation

store and curate, fight obsolescence

Store, share, preserve

save more copies of your work, it will be easier for people to find it in the future

places for your data

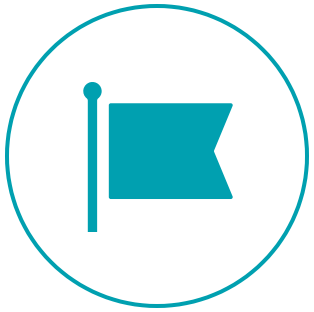
1. store your data in a repository for development
2. disseminate your data where people can find it - use your community channels
3. save a copy for long-term preservation in bespoke repositories

prepare your data

avoid proprietary formats, document your work, use naming conventions, save backup versions

Data preservation

ensure access to data over time

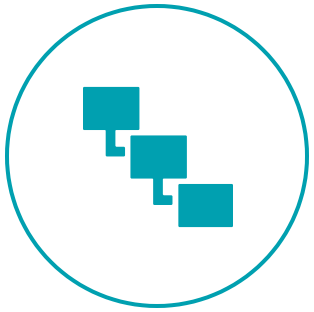


identify

attribute a permanent identifier to your data, so that it will be easier to cite it and retrieve it (e.g. DOI)

curate

provide information for future reuse, e.g. provenance, documentation, links to other resources (e.g. publications)



versioning

define what has to be stored for long-term preservation, and save a reasonable number of versions (e.g. incremental versions of same data, raw and processed)

An example: your digital edition

create your data
plain text
images

process your data
OCR
data cleansing

analyze your data
annotated images
encoded transcription (critical edition)

store data
work repo

document
your data and your work

disseminate
publish a scholarly digital edition

save a copy
provide a DOI
store plain text, code used, and the final edition

Some storage solutions

- **github**

  public git repository for versioning of working copies of your data/code. It is not for long-term preservation

  http://github.com/

- **institutional repositories**

  publication and data long-term preservation repositories

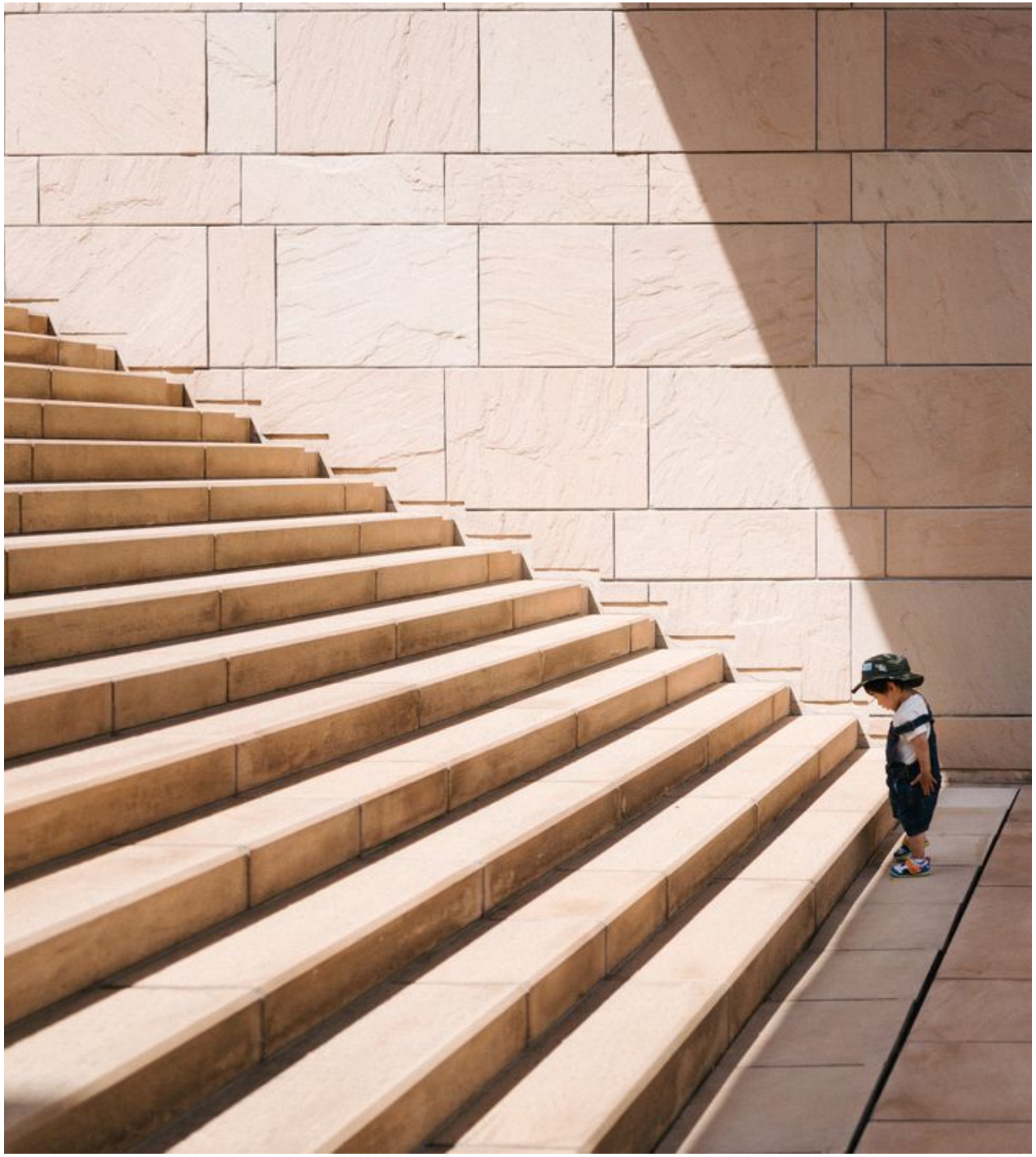
  @unibo https://cris.unibo.it/ and https://amsacta.unibo.it/

- **figshare (again)**

  https://figshare.com/

data access and reuse

publishing, licensing and services
make data discoverable and reusable



FAIR data principles

https://www.force11.org/group/fairgroup/fairprinciples

Findable

- uniquely identified
- described with metadata
- indexed in a searchable resource

Accessible

- metadata are accessible by means of the identifier
- metadata are retrieved even if data do not longer exist

Interoperable

- metadata conforms to shared vocabularies
- metadata include references to existing vocabularies

Reusable

- metadata are rich enough for being understood
- metadata include provenance information
- metadata include information on licenses

Access data over time and in different contexts

- **define licenses for reuse**

  whether an embargo period is necessary make sure it is respected. Prefer PD od CC licenses

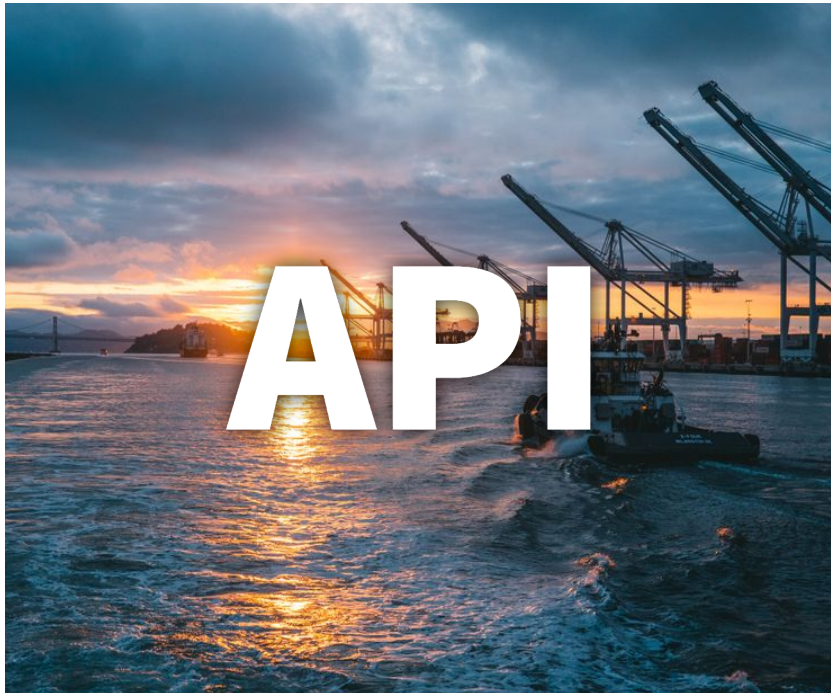
- **provide context**

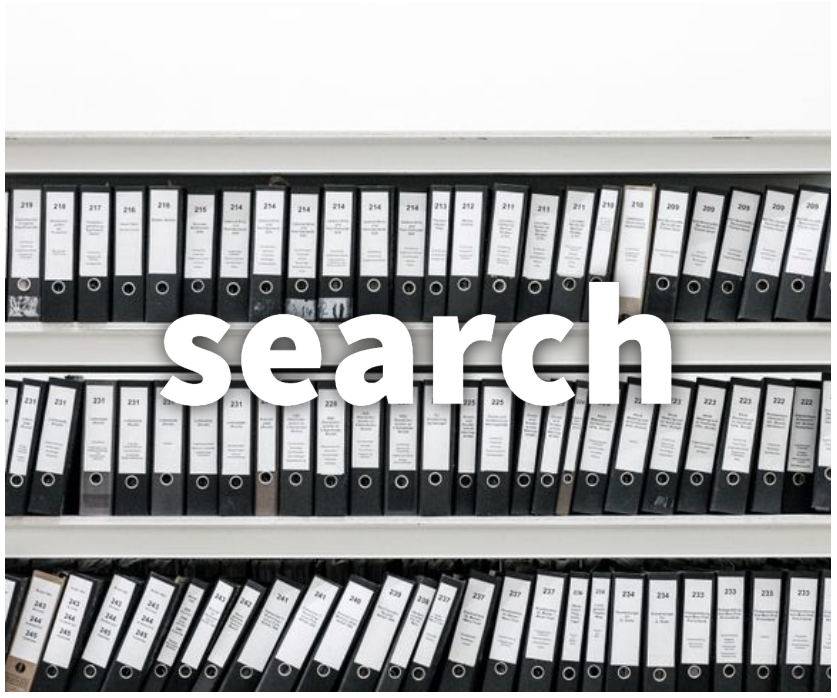
  give all the information required for reuse and reproducibility, e.g. software, code, tools used (eventually, store them as well)

- **create several access points**

  provide diverse services for accessing your data


browse


API


search


data
dump


repositories
and
archives

Learn more

Research data life cycle models

DCC Curation Lifecycle Model

http://www.dcc.ac.uk/resources/curation-lifecycle-model

DataONE data management tool

https://www.dataone.org/data-management-planning

UK Data Archive lifecycle

https://www.ukdataservice.ac.uk/manage-data/lifecycle

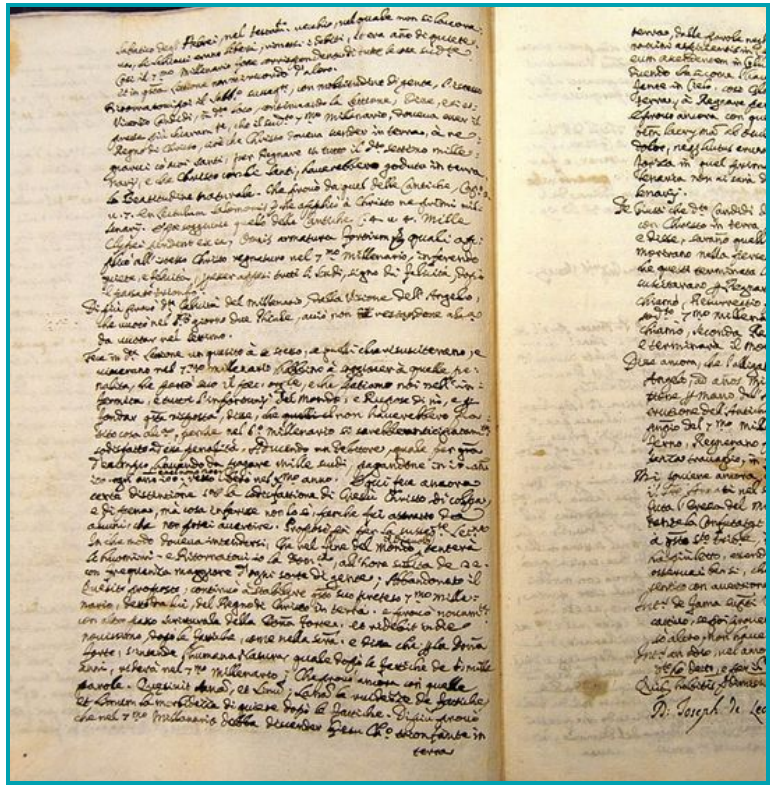
Corti et al.

http://www.sagepub.com/sites/default/files/upm-binaries/61019_Corti__Managing_and_sharing_research_data.pdf

Cultural Heritage data

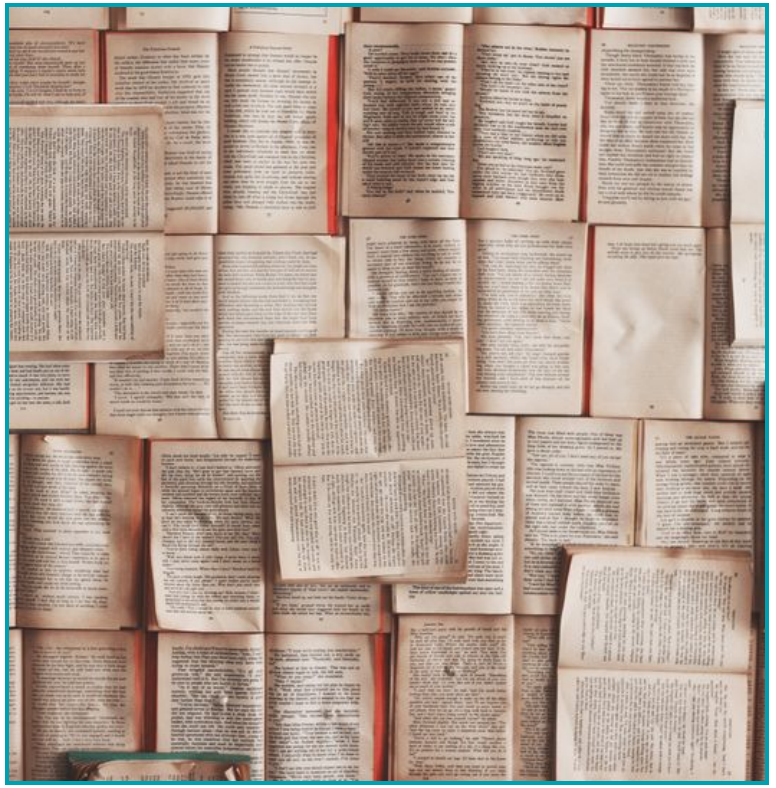
cataloguing, archival, literary,
historical, art historical metadata

What are cultural heritage data?



digitizations

or artefacts preserved in GLAMs



metadata

highly curated information on cultural objects provided by GLAMs

annotations and links

crowdsourced or automatically extracted

Are CH data research data?

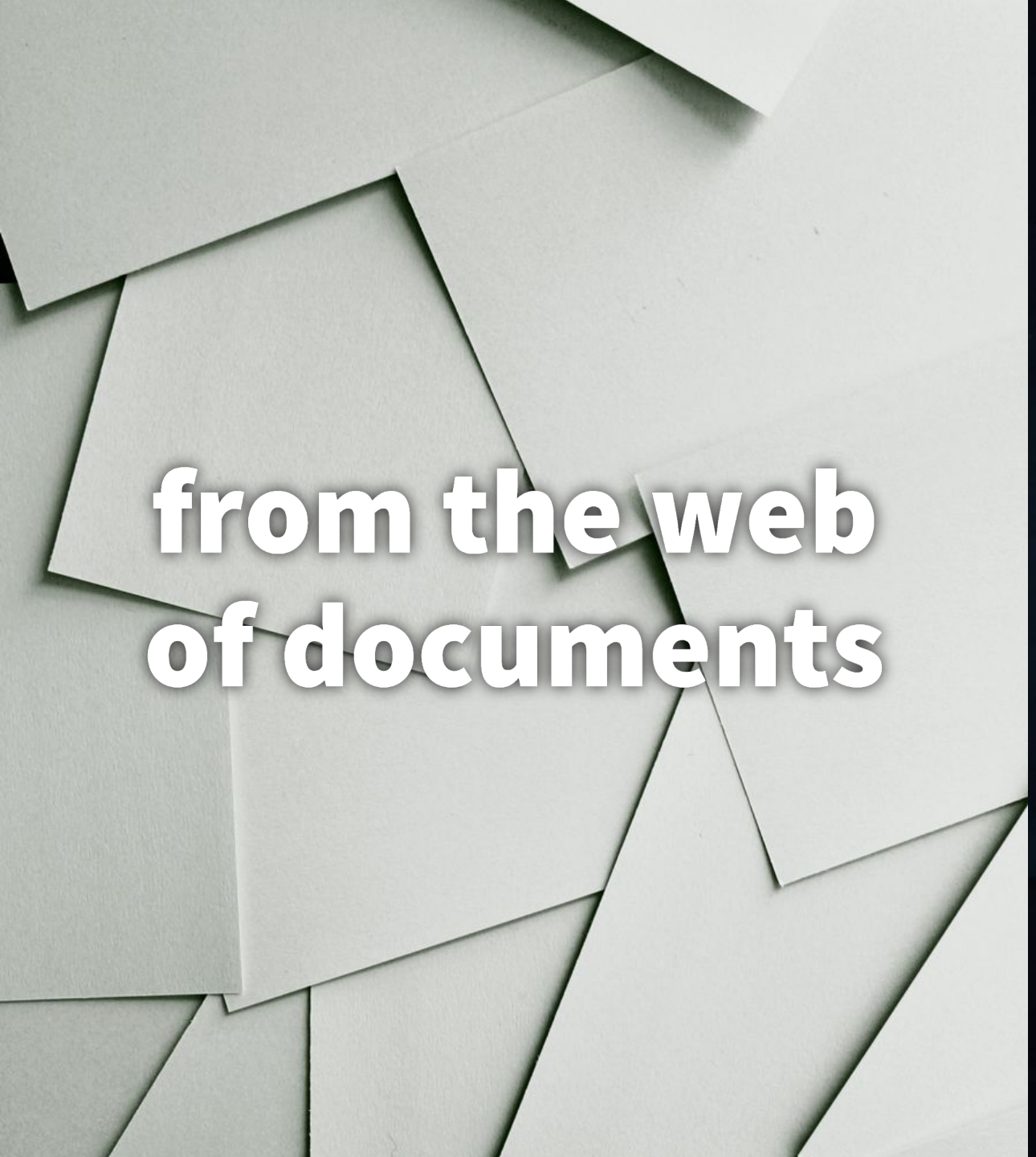
1. created by trusted providers
2. include accurate information
3. adopt community standards
4. mainly open for reuse

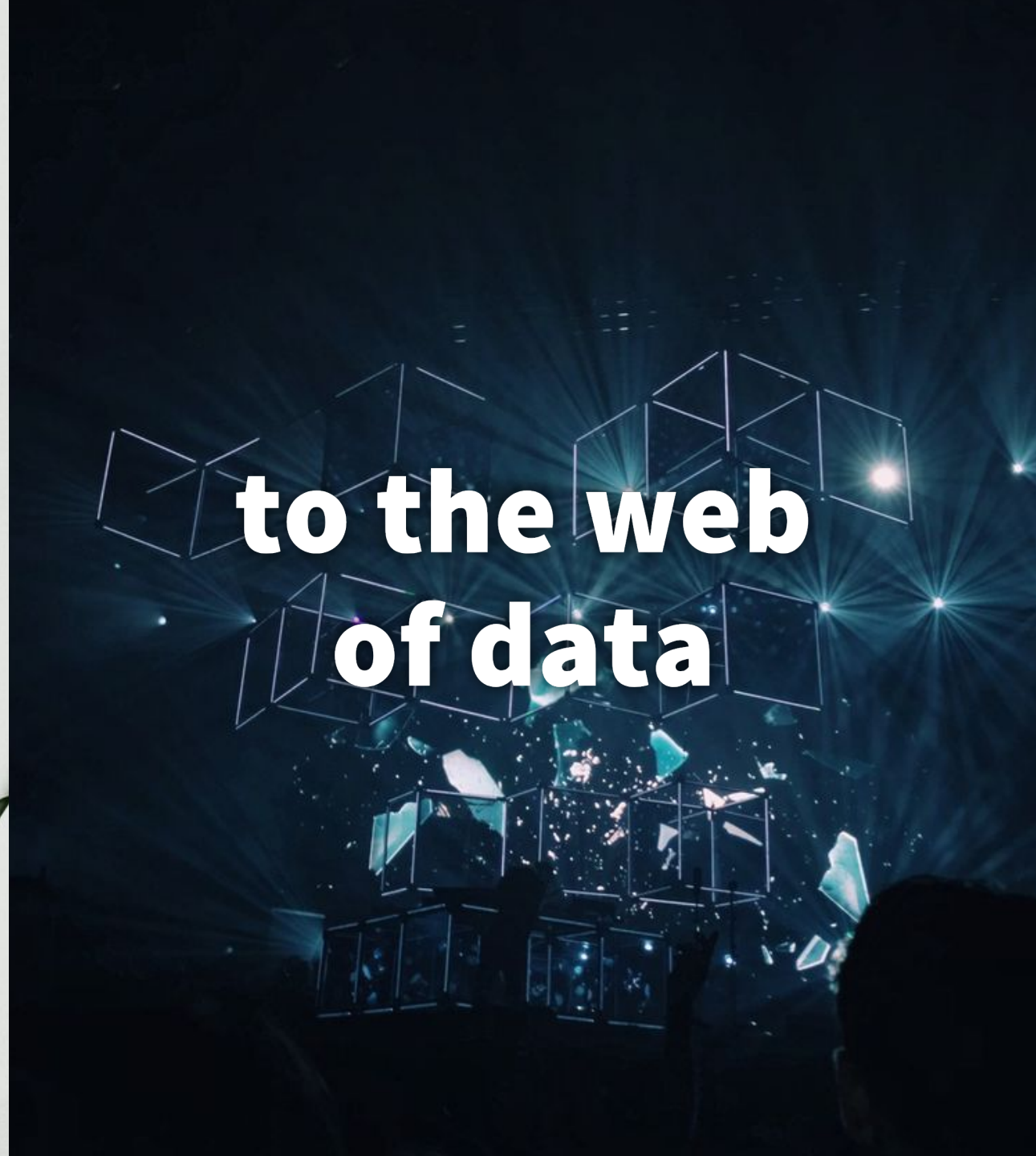
but... interoperability is still a big issue

5. heterogeneous content standards for describing CH objects
6. legacy formats and tools make data hard to be integrated
7. different degrees of data quality hinder possible reuse

Some solutions and research trends

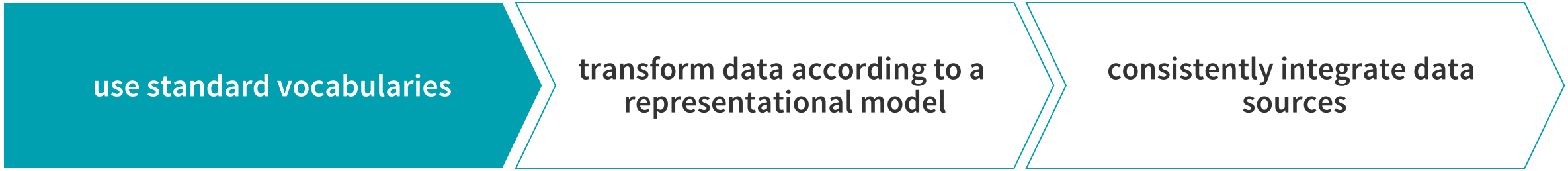
Semantic web technologies


from the web
of documents

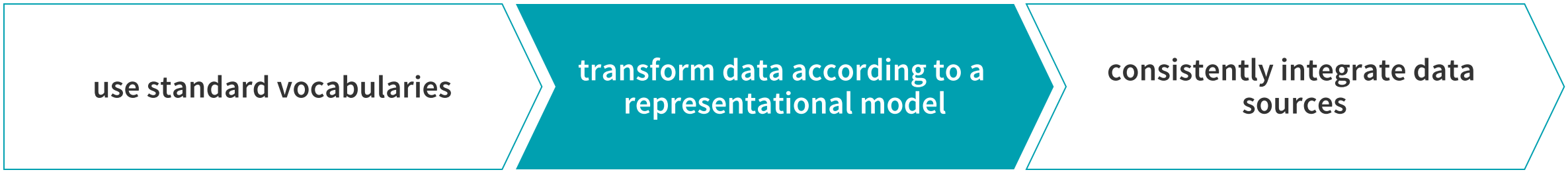

to the web
of data

How it works

ontologies and thesauri to overcome semantic heterogeneity



How it works



overcome technological barriers and represent data at the same way

How it works

use identifiers to refer to the same entities and link information included in different sources

What are the benefits?

enrichment

integrate similar data sources to enhance your data collection and offer a comprehensive view on a topic (e.g. transcription + bibliography + books available in library)

mashup

integrate information from different domains to create new services (e.g. hotel booking + maps + flight info + exhibition in town)

new knowledge

inferential rules allow to extract latent knowledge from data and discover new information (e.g. associative relations)

Research in DH

Some research fields

Digital Philology

- texts
- text analysis, style analysis, data mining

Digital Art History

- artworks, photographs
- data mining, computer vision

Digital History

- texts, photographs, artworks
- data mining, multimodal linking

Computational linguistics

- texts
- NLP, data mining

Who is involved?

GLAMs

creators, keepers

Researchers

adopters, contributors

Research Centres

aggregators, funders, researchers

Communities

journals, conferences, workshops

Some (EU) research centres and departments

- **Oxford e-centre**
  http://digital.humanities.ox.ac.uk/
- **Cologne CCeH**
  http://cceh.uni-koeln.de/
- **King's College London - DH Dept.**
  https://www.kcl.ac.uk/artshums/depts/ddh/index.aspx
- **Göttingen Centre for Digital Humanities**
  http://www.gcdh.de/en
- **Utrecht Digital Humanities Lab**
  https://dig.hum.uu.nl/
- **DHARC @unibo**
  (forthcoming)

Some DH (EU) infrastructures and associations

- **DARIAH**

  European network for DH research and education

  https://www.dariah.eu/

- **CLARIN**

  infrastructure for making available language resources

  https://www.clarin.eu/

- **EADH**

  European association focused on literature and language

  https://eadh.org/

- **AIUCD**

  Italian association of DH

  http://www.aiucd.it/

Project outputs in the Digital Humanities

Digital Libraries

aggregators of cultural objects metadata, multimedia and services providers

Web Applications

dissemination, browsing and searching of projects focused on narrow research problems

Tools

solutions to perform tasks related to data creation, analysis, and annotation

Methodologies

hypotheses, research problems, approaches, guidelines, solutions

Scholarship

articles, chapters, books, blog posts

Some relevant projects in the DH

- **Europeana**

  aggregator of cultural objects belonging to European collections
  https://www.europeana.eu/portal/en

- **PHAROSresearch**

  aggregator of pictures of artworks
  http://pharosartresearch.org

- **Pleiades**

  gazetteer of ancient places
  https://pleiades.stoa.org/

- **Perseus**

  collection of ancient texts
  http://catalog.perseus.org/

DH Journals and conferences

- **Journal of Open Humanities Data**

  https://openhumanitiesdata.metajnl.com/

- **Umanistica Digitale**

  http://umanisticadigitale.unibo.it

- **Digital Humanities Quarterly (DHQ)**

  http://www.digitalhumanities.org/dhq

- **DSH**

  http://llc.oxfordjournals.org/

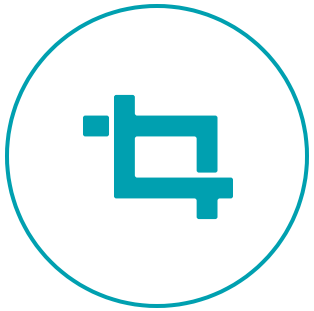
- **Journal of TEI**

  http://journal.tei-c.org/

Research Project management

what is a DH project and how to set up a project?

A DH project requires you to...



frame your research questions

while Computer Science focus on cutting-edge IT solutions, DH aim at boosting Humanities research

clarify the contribution to your field

it may be theoretical, procedural, or yet another tool...

present results and make them reusable

disseminating research outputs includes both sharing data and documents

Start a project

working with the Federico Zeri Photo archive @unibo - a true life story

what data is in art historical photo archives

map the heritage at your disposal

photographs

artworks

archival documents

bibliography

what to do with data in art historical photo archives

address several possible use cases, research fields that would benefit of your work

art history

study artists, styles, genres, impact of art in society

history of photography

study photographers, networks and creative industries

history of collecting

reconstruct the history of collections and the art market

connoisseurship

ascribe an artwork to an artist

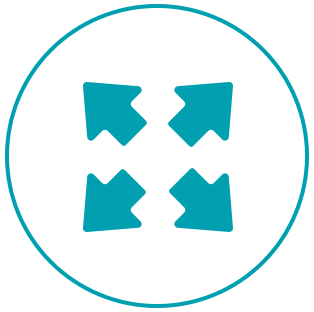
Connoisseurship

define a scenario and stick on that one

how connoisseurs work

- gather as many sources of information as possible
- compare sources and evaluate their trustworthiness
- rely on authoritative sources providers
- judge on the basis of their expertise in fine arts

Understand your research problems

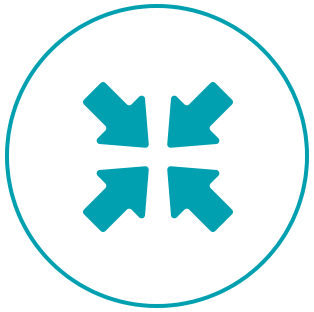


heterogeneity of cataloguing standards among providers

for describing different objects (artworks, books, etc.) and recording information about questionable information

no guidance for defining authoritativeness of contradictory statements

authoritativeness of sources (textual) and authoritativeness of people (cognitive) that provide contradictory information



no means for supporting connoisseurs' work

meaning ranking attributions on the basis of authoritative providers, bibliography, sources in agreement, etc.

Define your research questions

what defines a research question

- **topic**

  e.g. choosing the most authoritative attribution

- **nature of research endeavor**

  create, discover, explain, describe, compare

- **questions**

  what, who, where, how, when, why

- **variables**

  data providers' accuracy, authoritativeness, subjectivity of statements

- **relations between variables**

  impact, correlation, causes

Define your research questions

an example

how to make easier the integration of heterogeneous art historical data sources?

how to define authoritative sources of information in art history?

how to support scholars and cataloguers' decision-making process?

For whom you are doing it?

identify stakeholders and potential beneficiaries of your work

scholars

facilitate their daily work

cataloguers

avoid them time consuming activities

art dealers

provide sources for their statements

Define assumptions, hypotheses, restrictions

- **assumptions**

  something you give for granted and that you won't demonstrate, e.g. data providers record sufficient information for validating veracity of their statements

- **hypotheses**

  something you think may work and solve your problem, but you have to prove it! e.g. Linked Data may support scholars' tasks such as gathering and comparing sources

- **restrictions**

  we cannot solve all the problems of the world... narrow your scenario to something achievable. e.g. focus on art historical photo archives because of the richness of their art historical data

Survey the state of the art

Schemas, vocabularies and ontologies for the cultural heritage

Information Quality measures and metrics that fit-for-purpose

Ranking models, semantic crawlers, web development

Plan your work

what are the expected research outputs?

ontologies, exemplar datasets, golden standards for publishing art historical data

conceptual framework of measures and metrics for assessing authoritativeness of painting attributions

semantic crawler for harvesting and ranking contradictory painting attributions

Evaluate your results

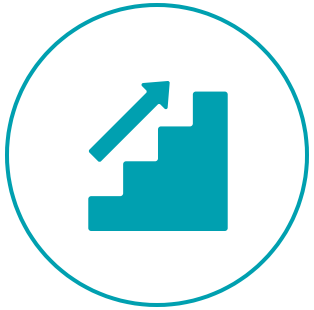
according to the nature of results

comparison with golden standards

for results that leverage existing standards and provide an improvement, e.g. datasets, ontologies

user-centered evaluation

for results that are hard to be assessed automatically and that depend on users' perception, e.g. ranking model



benchmarking

compare application performance with existing services, e.g. speed of crawler

Publish your (documented) results

Website

website including explanations and references

http://data.fondazionezeri.unibo.it

Web application (if applicable)

web application for evaluating results

http://purl.og/emmedi/mauth/search

Journal articles

plenty, never enough..

Presentations

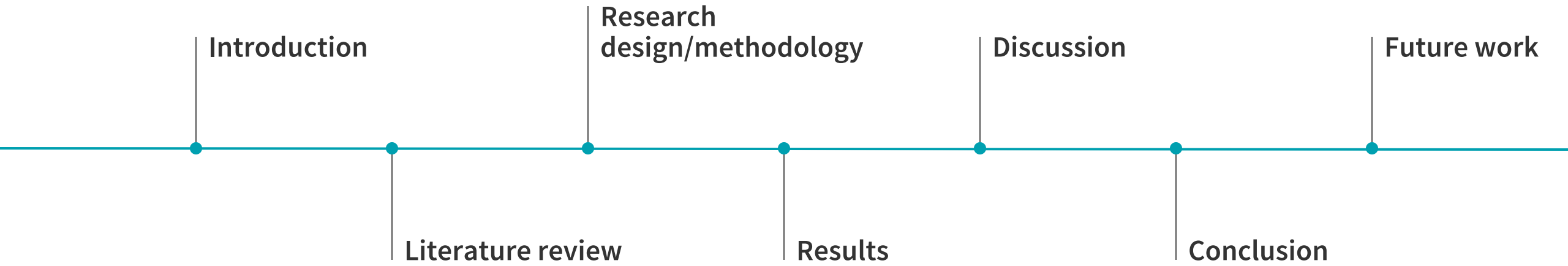
presentation slides for talks

Present your results

what you need to know to prepare
a good DH presentation

A research presentation should include...

seven things you will never forget



Introduction

what is your research about

- background, e.g. art historical photo archives and art history enquiry
- problem, e.g. questionability of statements
- motivating scenario, e.g. connoisseurship
- challenge, e.g. formal definition of authoritativeness
- approach, e.g. use Semantic Web technologies to gather and compare statements
- evaluation, e.g. user-centered evaluation of a mashup application

Literature review

what is already done

- topic/issue related literature
- highlight gaps
- rationale or justification of your research

Design/Methodology

your research questions

hypotheses, assumptions and restrictions

how you went about your research

use an existing methodology, e.g. design-science methodology

approach to the research

all the steps you went through to get your results

Present results

what you found out

overview of your findings, e.g. vocabularies, measurements, applications

evaluation

explain thoroughly how you demonstrated the validity of your results

Discuss results

what your results mean

- make new statements on the basis of results
- explain limits of your contribution

Conclusion

So what? summarize your contribution

- recall hypotheses and research questions and make statements about the contribution of your work

Future work

what happens now

moving from your restrictions, describe future research lines

Learn more

Write that PhD

https://twitter.com/writethatphd?lang=en

Write a structured abstract

https://pdfs.semanticscholar.org194f/
91e45c1784f379c91788a748459157e57304.pdf/

That's all folks!

Thanks