What is a database?

Monday Morning Session 1

MJC-BSANA-Data-Workshop-2024

Paula Granados García, The British Museum

Paula Loreto Granados García is a Digital Humanities expert focusing on digital curation, data governance, and interoperability. She is the Digital and Operations Manager of the Endangered Material Knowledge Programme (EMKP) at The British Museum, where she oversees the digital aspect of the Programme, ensuring best practice in rights and ethics management, and developing innovative approaches to digital curation.

Dr. Granados received her Ph.D. in archaeological research and Linked Open Data from the Open University in 2020. Her research focuses on the intersection between Digital Humanities and Cultural Heritage, by looking at how digital technologies can help individuals access, share, and build upon archaeological and ethnographic knowledge in different interdisciplinary streams: data and knowledge modelling via Linked Open Data technologies, the creation and mediation of knowledge in museums, and collaborative curation of cultural heritage initiatives through digital media and experimental technologies.

She currently holds a Visiting Research Fellowship at the Digital Humanities Research Hub, School of Advanced Study, University of London. She is Digital Repository Advisor for the GCRF Network+ grant "Imagining Futures through Un/archived Pasts" at the University of Exeter, and she is a member of the Steering Committee of the Linked Pasts Symposium and the Advisory Board for the InDIGenius Languages and Digital Technology Initiative.
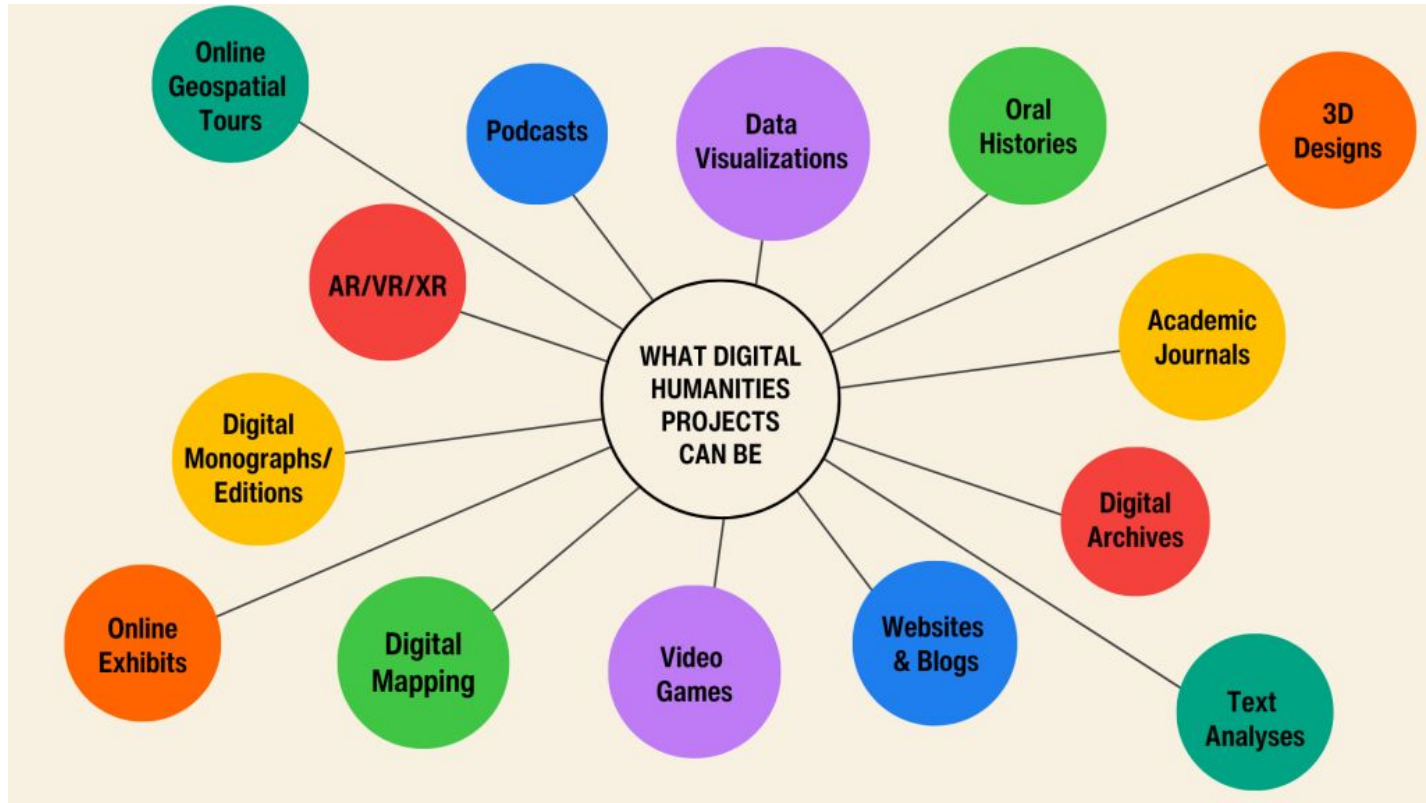
BANEA
Banjercha

KADALLU
Babshankur

ABU RAMLA
Amizibil

3
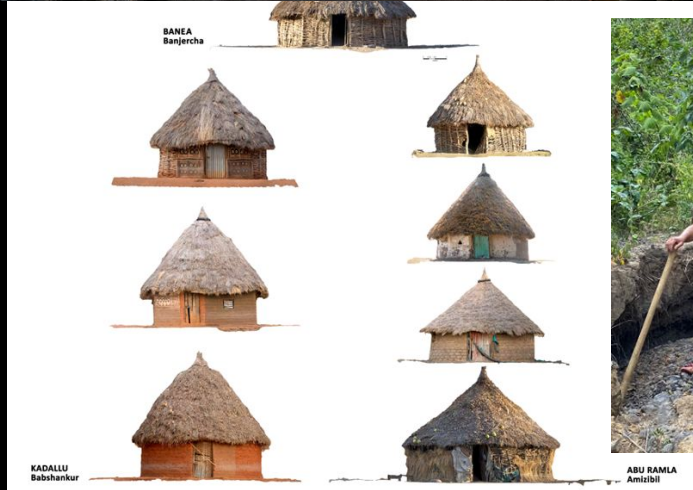
Session outline:

1. The Data Lifecycle
2. Licenses
3. Structured data
4. What is a database?
5. Types of Databases
6. Schemas
7. Designing your database
8. Data recording
9. Virtual Research Environments

Examples of data in the Humanities

5 Reasons Why All Graduate Education Should Include the Digital Humanities. Edinburgh University Press. October 2023. Infographic created by Bailey Betik using Canva

Digital Humanities and Cultural Heritage

2024 Digital Humanities Trends

Artificial Intelligence
Storytelling
3D Printing
Immersive/Digital Exhibitions
Mass digitisation
Digitisation standards
Community Engagement
FAIR/CARE

Data sustainability
Environment sustainability
National Databases
Data Access and accessibility
Information Modelling
Crowdsourcing
Interoperability
Intangible Heritage

The Data Lifecycle
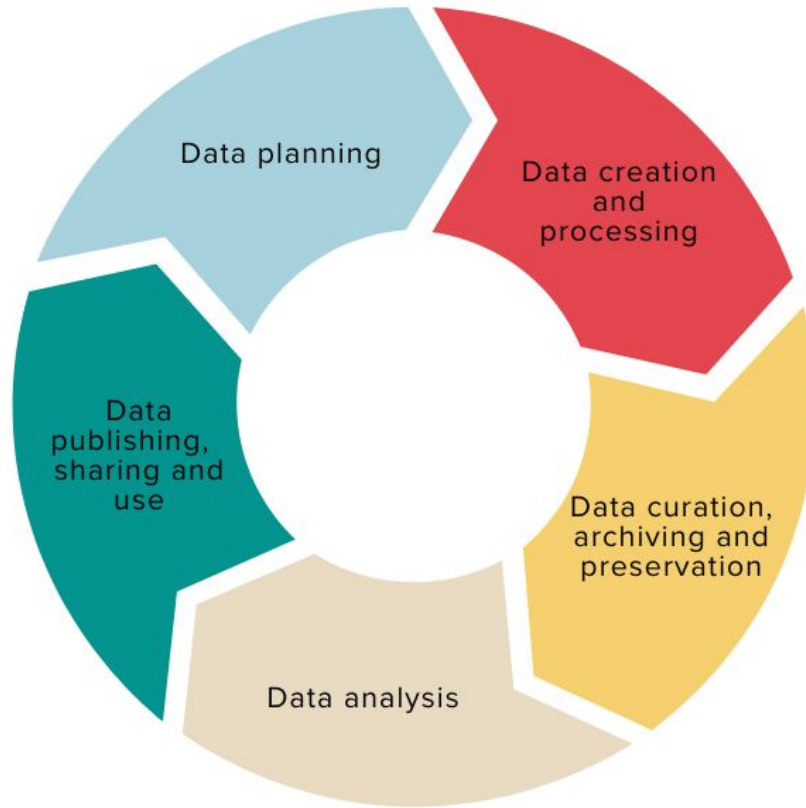
- Represents all of the stages of data from its creation to its distribution and reuse.
- Begins with a researcher developing a concept for a study
- Once a study concept is developed, data is then collected/ created and processed.
- Data is then stored in a server where it can queried and analysed.
- Finally data is published in a location (i.e. repository, registry) where it can be discovered by other researchers.

The data lifecycle © by the 3D Service Suite, University of Brighton under CC BY 4.0

Data management

- All research projects generate data in one way or another.
- Good data management practice involves the organization, preservation, and reuse of data by documenting the steps throughout the process.
- Storing data in the right formats and conditions helps ensure that future generations of researchers can see and take advantage of it.
- There are several legal and ethical issues to consider before sharing research data. In addition, it is important to get credit for your data by quoting it correctly.

Licensing and Open Access

Attribution
BY
Others can copy, distribute, display, perform and remix your work if they credit your name as requested by you

No Derivative Works
ND
Others can only copy, distribute, display or perform verbatim copies of your work

Share Alike
SA
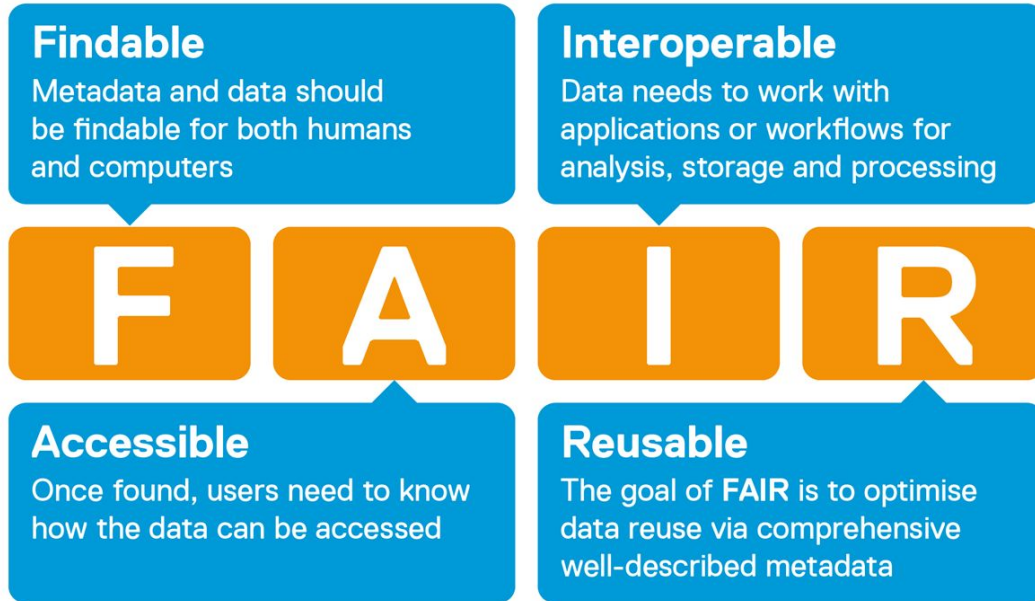Others can distribute your work only under a license identical to the one you have chosen for your work

Non-Commercial
NC
Others can copy, distribute, display, perform or remix your work but for non-commercial purposes only.

- Creative Commons licenses provide a legal framework to allow the right to reuse a copyrighted publication.
- The fewer restrictions a license implies, the greater the chances of using and distributing content.

9

FAIR principles

Findable
Metadata and data should be findable for both humans and computers

Interoperable
Data needs to work with applications or workflows for analysis, storage and processing

F A I R

Accessible
Once found, users need to know how the data can be accessed

Reusable
The goal of **FAIR** is to optimise data reuse via comprehensive well-described metadata

- The FAIR principles guide was published as a scientific paper in 2016.
- These principles were developed primarily with interoperability, linking and re-use of information in mind.

FAIR and CARE

FAIR and CARE

Practice 'CARE' in data collection	Engage 'CARE' in data stewardship	Implement 'CARE' in data community	Use 'FAIR' with 'CARE' in data applications
Define cultural metadata Record provenance in metadata	Use appropriate governance models Make data 'FAIR'	Indigenous ethics inform access Use tools for transparency, integrity and provenance	Fairness, Accountability, Transparency Assess equity

Carroll, S.R., Herczog, E., Hudson, M. *et al.* Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Sci Data* 8, 108 (2021). https://doi.org/10.1038/s41597-021-00892-
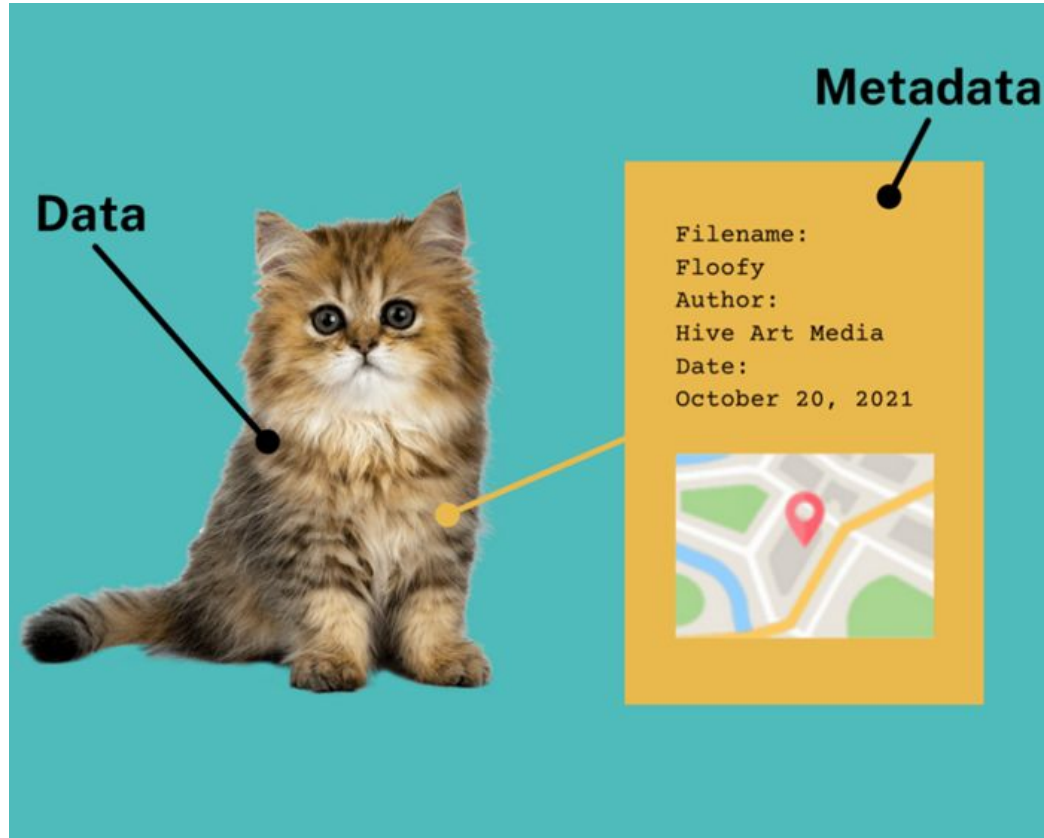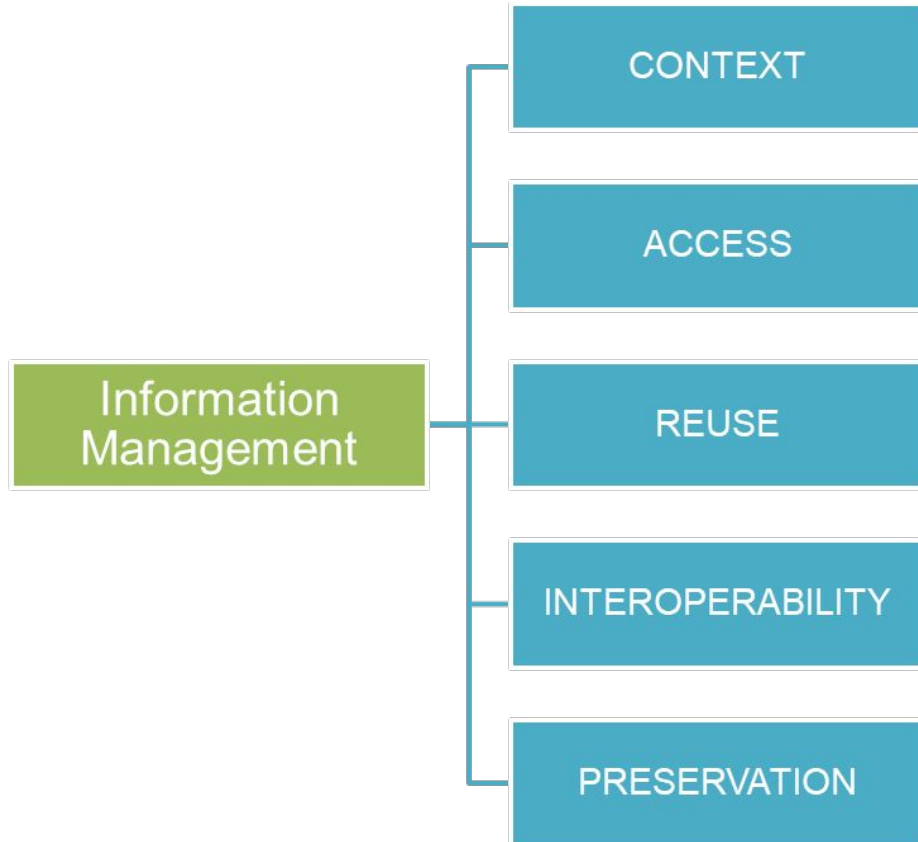
Data Vs. Metadata

What is Metadata?

'a set of data that describes and gives information about other data'.
Literally "data about data". Metadata helps users make data findable, accessible, interoperable and reusable "FAIR".

The term is today widely used but frequently unspecified, thus understood in different ways by the diverse professional communities that use metadata in their records.

Why is metadata important?

Information Management

- CONTEXT
- ACCESS
- REUSE
- INTEROPERABILITY
- PRESERVATION

Structured Data

A	B	C	D	E	F
Settlement	Nomsma ID	Coordinates	Coordinates	Province 1	Province 2
Abdera	nomisma.org/id/abdera_hispania	36.733.333	-3.016.667	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
Abra	http://nomisma.org/id/abra	37.765.419	-3.959.262	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
Acci	http://nomisma.org/id/acci	37.300.309	-3.134.636	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
Acinippo	http://nomisma.org/id/acinippo	36.832.412	-5.240.746	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
Aipora	http://nomisma.org/id/aipora	36.779.174	-6.354.245	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
Alba	http://nomisma.org/id/alba			http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
Arsa/Arse	http://nomisma.org/id/arsa			http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
Asido	http://nomisma.org/id/asido	36.467.791	-5.927.867	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior

Table:
- Graphical layout for presenting data in a structured way.
- Used since ancient times.
- It arranges data in a regular grid.

Structured Data

	A	B	C	D	E	F
	Settlement	Nomsma ID	Coordinates	Coordinates	Province 1	Province 2
	Abdera	nomisma.org/id/abdera_hispania	36.733.333	-3.016.667	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
	Abra	http://nomisma.org/id/abra	37.765.419	-3.959.262	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
	Acci	http://nomisma.org/id/acci	37.300.309	-3.134.636	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
	Acinippo	http://nomisma.org/id/acinippo	36.832.412	-5.240.746	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
	Aipora	http://nomisma.org/id/aipora	36.779.174	-6.354.245	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
	Alba	http://nomisma.org/id/alba			http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
	Arsa/Arse	http://nomisma.org/id/arsa			http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior
	Asido	http://nomisma.org/id/asido	36.467.791	-5.927.867	http://nomisma.org/id/hispania#this	http://nomisma.org/id/hispania_ulterior

Table:
- We can refer to columns, rows or an individual cell as the intersection between columns and rows.
- We can also refer to groups of cells, called ranges.
- This becomes easier when we name or number the columns and rows.

Structured Data

1, Abdera, nomisma.org/id/abdera_hispania, 36733333, -3016667, http://nomisma.org/id/hispania_ulterior,
2, Abra, nomisma.org/id/abra, 37765419, -3959262, http://nomisma.org/id/hispania_ulterior

Database

"Organised collection of information that enables efficient storage, retrieval and manipulation of data."

- dataset: a collection of data, primarily used for analysis
- database: a system designed for storing and managing data efficiently.

Is this a database?

| Town	Series	BM	G&B	URI Item	Modern name of the mint	Region	Description	Hoard	Similarities	FromDate L	ToDate L	FromDate	ToDate	Material	Denomination
Abdera	1=1	CGR253141	17		Cerro de Montecristo/ desembocadura Ab	Almeria	Phoenician colonia from 8th ct BC	Obulco and Cas		2nd ct./ BCE	1st ct./ BCE	-(199-100)	-(99-1)	Ae	Unit
Abdera	1=2	CGR252651	17		Cerro de Montecristo/ desembocadura Ab	Almeria	Phoenician colonia from 8th ct BC	Obulco and Cas		2nd ct./ BCE	1st ct./ BCE	-(199-100)	-(99-1)	Ae	Half
Abdera	1=2A		17		Cerro de Montecristo/ desembocadura Ab	Almeria	Phoenician colonia from 8th ct BC	Obulco and Cas		2nd ct./ BCE	1st ct./ BCE	-(199-100)	-(99-1)	Ae	Fourth
Abdera	2=3	CGR252586	17	http://ceres.mcu.	Cerro de Montecristo/ desembocadura Ab	Almeria	Phoenician colonia from 8th ct BC	Obulco and Cas		1st ct./ BCE	1st ct./ BCE	-99	-1	Ae	Unit
Abdera	2=4		17	https://www.num	Cerro de Montecristo/ desembocadura Ab	Almeria	Phoenician colonia from 8th ct BC	Obulco and Cas		1st ct./ BCE	1st ct./ BCE	-99	-1	Ae	Half
Abdera	2=5	CGR277100	17		Cerro de Montecristo/ desembocadura Ab	Almeria	Phoenician colonia from 8th ct BC	Obulco and Cas		1st ct./ BCE	1st ct./ BCE	-99	-1	Ae	Fourth
Abdera	2=6		17		Cerro de Montecristo/ desembocadura Ab	Almeria	Phoenician colonia from 8th ct BC	Obulco and Cas		1st ct./ BCE	1st ct./ BCE	-99	-1	Ae	Eight
Abdera	3=7		18		Cerro de Montecristo/ desembocadura Ab	Almeria	Phoenician colonia from 8th ct BC	Obulco and Cas		1st ct./ BCE	1st ct./ BCE	-99	-1	Ae	As
Abdera	3=8		18		Cerro de Montecristo/ desembocadura Ab	Almeria	Phoenician colonia from 8th ct BC	Obulco and Cas		14 CE	37 CE	14	37	Ae	As
Abra	1=1		18	http://www.coinp	Torredonjimeno?	Jaén	Uses meridional writing for Iberian	Obulco		2nd ct./ BCE	2nd ct./ BCE	-199	-100	Ae	Duplex
Abra	2=2	CGR39744	18	http://www.junta	Torredonjimeno?	Jaén	Uses meridional writing for Iberian	Obulco		2nd ct./ BCE	2nd ct./ BCE	-199	-100	Ae	Duplex
Abra	3=3		19	http://www.junta	Torredonjimeno?	Jaén	Hybrid Abra and Obulco	Obulco		2nd ct./ BCE	2nd ct./ BCE	-199	-100	Ae	Triplex
Acci	1=1	CGR299208	19		Guadix	Granada	Control over Sierra Moreno paths, later colony use			27 BCE	23 BCE	-27	23	Ae	As
Acci	1=2		19		Guadix	Granada	Control over Sierra Moreno paths, later colony use			27 BCE	23 BCE	-27	23	Ae	Semis
Acci	2=3	CGR299209	20		Guadix	Granada	Control over Sierra Moreno paths, later colony use			12 BCE	12 BCE	-12	-12	Ae	As
Acci	2=4	CGR299211	20		Guadix	Granada	Control over Sierra Moreno paths, later colony use			12 BCE	12 BCE	-12	-12	Ae	Semis
Acci	3=5	CGR299213	20		Guadix	Granada	Control over Sierra Moreno paths, later colony use			14 BCE	19 CE	-14	19	Ae	Dupondius
Acci	4=6		20		Guadix	Granada	Control over Sierra Moreno paths, later colony use			14 BCE	19 CE	-14	19	Ae	Dupondius
Acci	4=7		20		Guadix	Granada	Control over Sierra Moreno paths, later colony use			19 CE	19 CE	-19	19	Ae	As
Acci	4=8		20		Guadix	Granada	Control over Sierra Moreno paths, later colony use			19 CE	19 CE	-19	19	Ae	Semis
Acci	5=10		21		Guadix	Granada	Control over Sierra Moreno paths, later colony use			37 CE	41 CE	-37	41	Ae	As
Acci	5=11		21		Guadix	Granada	Control over Sierra Moreno paths, later colony use			37 CE	41 CE	-37	41	Ae	Semis
Acci	5=9		21		Guadix	Granada	Control over Sierra Moreno paths, later colony use			37 CE	41 CE	-37	41	Ae	Dupondius

Is this a database?

Main characteristics of a good database

1. **Data Integrity**: maintain the accuracy, consistency, and reliability of data. Enforce rules to prevent invalid or inconsistent data.
2. **Flexibility**: support various data models and data types to accommodate diverse requirements and use cases.
3. **Ease of Use**: have a user-friendly interface and intuitive tools for database administration, development, and querying. It should provide comprehensive **documentation,** and support resources to assist users in effectively utilizing the database.
4. **Interoperability**: integrate with other systems, applications, and tools through standard protocols and APIs.
5. **Backup and Recovery**: offer robust backup and recovery capabilities to protect against data loss or corruption. Plean for regular backups, point-in-time recovery, and disaster recovery strategies to restore data.

Types of databases

Flat Files	Non-Relational (NOSQL)	Relational (SQL)
- e.g. excel files - commonly used in DH - can be easily integrated with analytical tools	- more flexible and schema-less approach - well-suited for managing diverse data types and accommodating rapidly changing data structures - graph databases	- Connect information in separate tables. - Able to pull information from across different tables.
- not efficient - difficult to update	- minimally structured	- may encounter challenges for non-structured data

Relational Databases
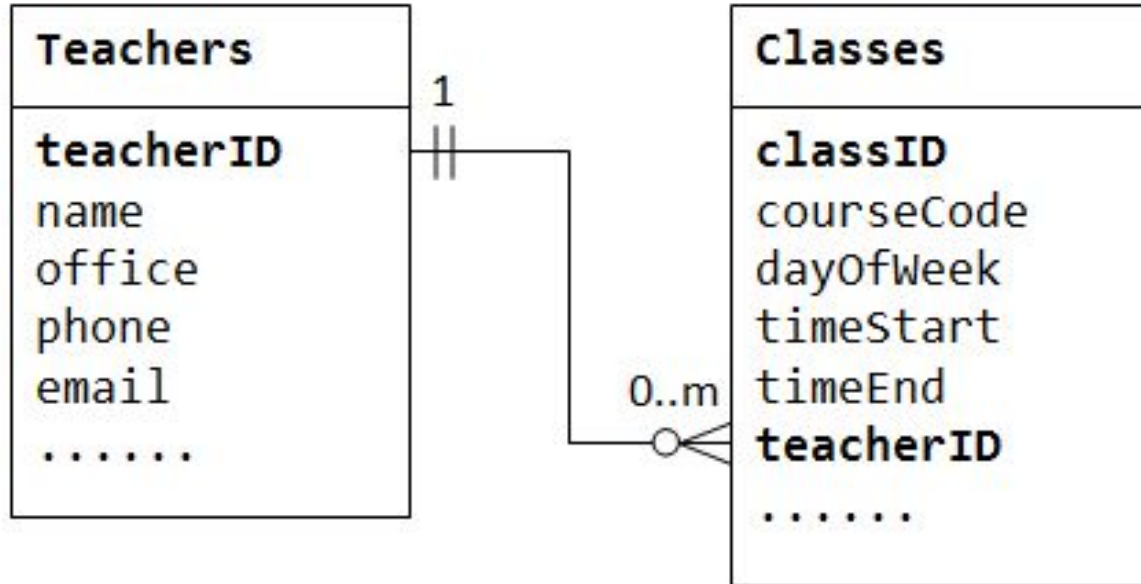
SQL: Structured Query Language

A standard language for storing, manipulating and retrieving data in databases.

Select all records from the Customers table:

```
SELECT * FROM Customers;
```

| CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

https://www.w3schools.com/sql/default.asp

Basic database concepts

- **Entity**: the thing that you want to collect data about (rows)
- **Attribute**: the information about the entity that you want to collect and store (columns)
- **Domain**: a set of values from which an attribute can take a value in each row. Usually, a data type is used to specify domain for an attribute.
- **Database Management System (DBMS):** software that enables users to create, manage, and interact with databases. It provides tools and utilities for tasks such as data storage, retrieval, manipulation, and security. Examples of popular DBMS include MySQL, PostgreSQL, Microsoft SQL Server, Oracle Database, and MongoDB.
- **Queries**: commands or statements used to retrieve, manipulate, or analyze data in a database.
- **Schema**: The schema of a database defines its structure, including the tables, columns, data types, and relationships between tables.
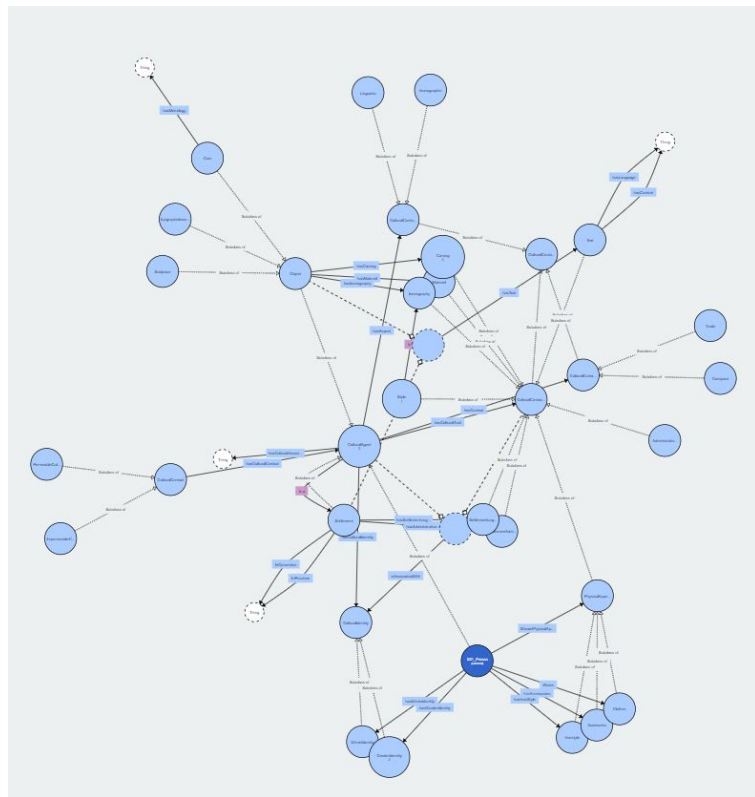
Metadata schemas
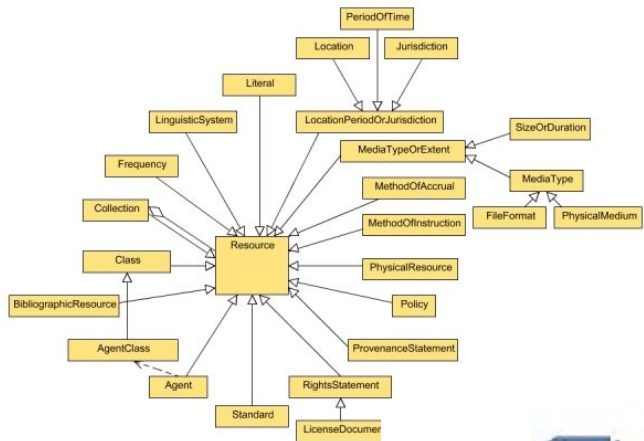
Dublin Core
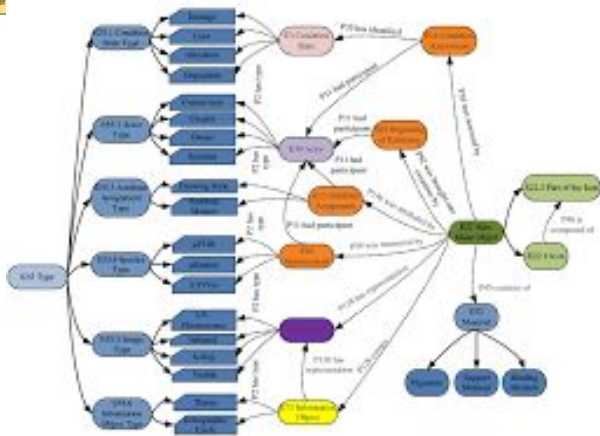
Cidoc-CRM

Cultural Contact Ontology

Dublin Core

https://www.dublincore.org/

- Metadata standard used to describe digital resources such as documents, images, videos, and web pages.
- It provides a set of elements and guidelines for creating simple and interoperable metadata to facilitate the discovery and management of digital resources.
- The Dublin Core Metadata Initiative (DCMI) maintains and develops the Dublin Core standard
- Includes elements such as title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. These elements help to provide basic information about digital resources, making them easier to find and manage across different systems and platforms.

Dublin Core

```xml
<metadata>
    <dc:title>Golden Gate Bridge</dc:title>
    <dc:creator>Anne Smith</dc:creator>
    <dc:subject>Architecture</dc:subject>
    <dc:description>A stunning view of the Golden Gate Bridge on a clear day.</dc:des
    <dc:publisher>San Francisco Tourism Board</dc:publisher>
    <dc:date>2023-04-15</dc:date>
    <dc:type>Image</dc:type>
    <dc:format>JPEG</dc:format>
    <dc:identifier>https://example.com/photo123</dc:identifier>
    <dc:source>Personal collection</dc:source>
    <dc:language>English</dc:language>
    <dc:rights>Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC
</metadata>
```

CIDOC - CRM

- "CIDOC Conceptual Reference Model."
- An ontology, or a formal model, used to describe knowledge about cultural heritage and museum collections.
- Developed by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM).
- Based on the principles of object-oriented modeling and uses concepts such as classes, properties, and relationships to represent the structure and semantics of cultural heritage information.
- One of the key features of CIDOC CRM is its ability to support interoperability and data exchange between different cultural heritage institutions and systems. By providing a common vocabulary and structure for describing cultural heritage information, CIDOC CRM enables institutions to share and integrate their data more effectively, facilitating collaboration, research, and access to cultural heritage collections.
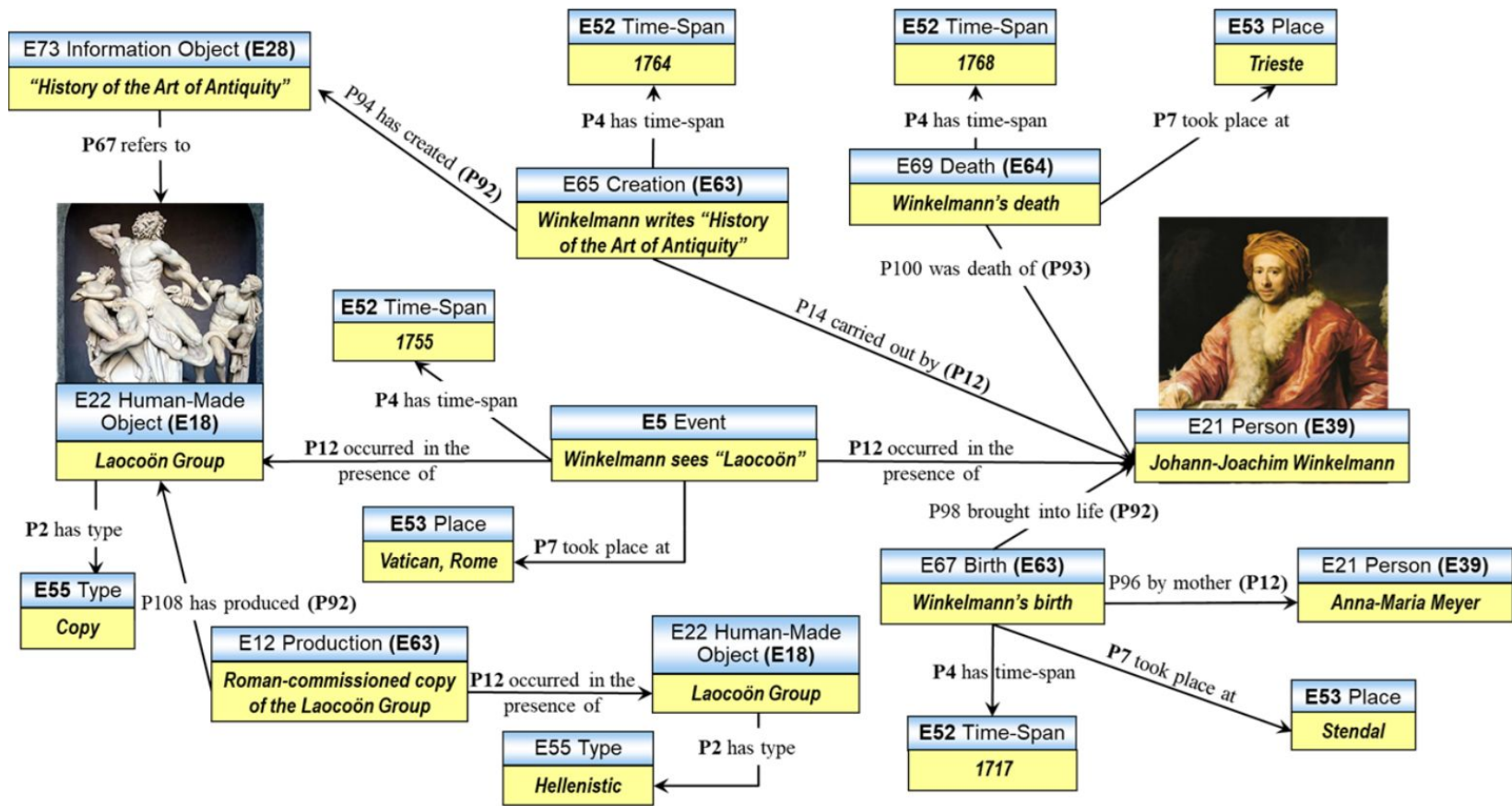
Material Culture Ethnography Metadata Schema

- Standard developed by the EMKP to describe ethnographic materials, such as artifacts, photographs, audio recordings, and field notes, within the context of cultural anthropology and ethnography.
- It provides a structured framework for documenting various aspects of these materials, including provenance, contextual information, date, author, access rights etc.
- Developed to be compatible with other widely used standards, mapped to the Dublin Core Metadata Element Set (DCMES). Additionally some elements and their definitions in the Group 'Persons' were integrated and adapted from the Isle Metadata Initiative (IMDI) and OLAC standards, while some elements and their definitions in the Group 'Assets' were integrated and adapted from SPECTRUM and The CIDOC Conceptual Reference Model (CRM).

Material Culture Ethnography Metadata Schema
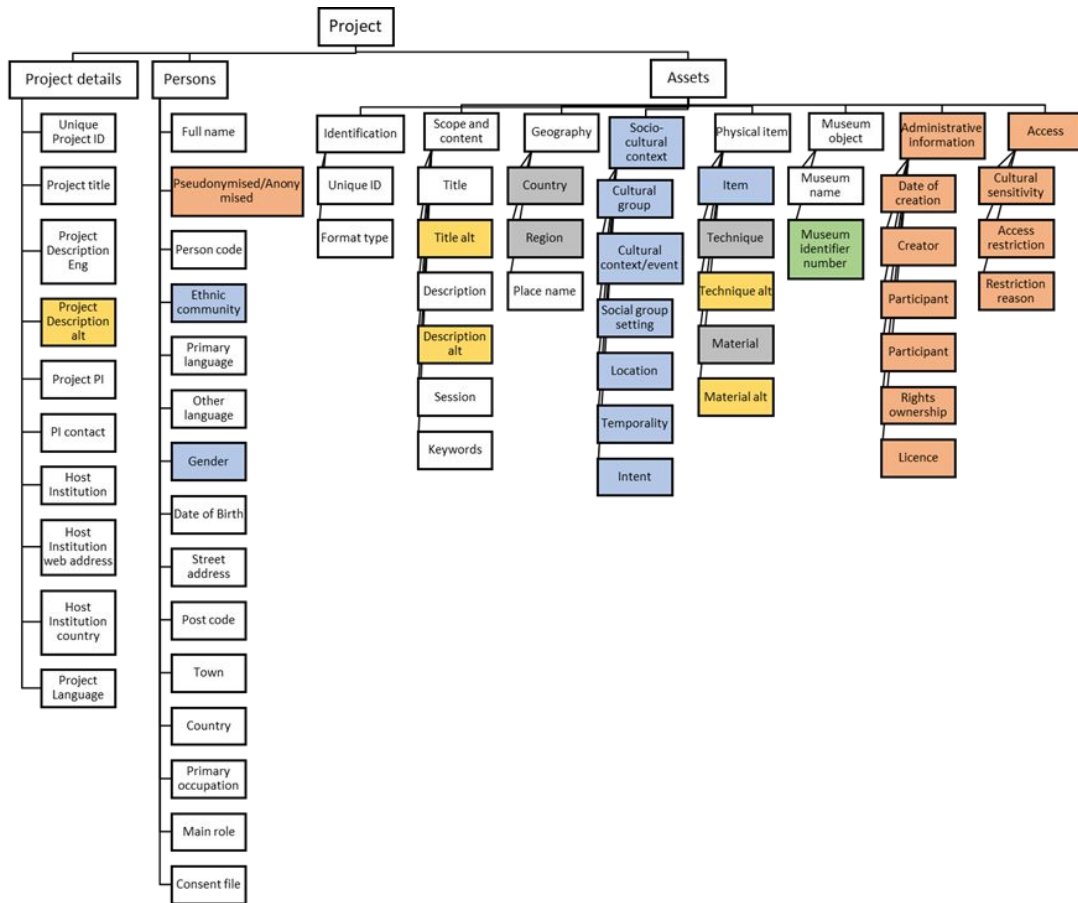
Designing your database

1. Identifcation and surveying of relevant data types and sources.

2. Available data consumption and edition.

3. New data collection and integration.

4. Data Modelling

5. Alignment

Designing your database

Remember: every database is a subjective interpretation of reality.

- Identify the research goal - define purpose and scope.
- Understand your data - Know what you are recording
- Keep it as simple as you can - concise but descriptive
- Plan/record your schema
- Ensure data integrity mechanisms - e.g. vocabularies
- Find/create user-friendly interfaces for data entry, querying and visualization.
- Document your process

Recording your data: Best practices

- Information has to be simple, complete and coherent
- Provide appropriate descriptions
- Explain limitations and possibilities to support reuse
- No more than one value per cell (unless otherwise stated)
- Avoid special characters (e.g. @, !, <>)
- Provide persistent identifiers and permanent links when possible
- When copying information, use a text editor to avoid hidden characters

Data Validation!

- Data validation refers to the process of ensuring the accuracy and quality of data.
- Data validation is Key to ensure consistency and searchability
- Excel and google spreadsheets offer different features for data validation.
- Use standard and controlled vocabularies for consistency and to avoid ambiguity

Controlled Vocabularies

To facilitate the quality of documentation, findability, access and long-term preservation of data. Provide standardised terms and language to describe artefacts, objects, or concepts within a specific domain. Some of the most well-known vocabularies in the cultural heritage domain are:

- The Getty Art & Architecture Thesaurus (AAT)
- The Getty Thesaurus of Geographic Names (TGN)
- The Union List of Artist Names (ULAN)
- The Library of Congress Subject Headings (LCSH)

The enriched records include URLs pointing to the Getty vocabularies that provide:

- preferred concept labels in multiple languages
- alternative labels in multiple languages
- broader and narrower concepts

Controlled Vocabularies

Tools to map vocabularies also exist, such as the Vocabulary Matching Tool by the University of South Wales and the Vocabulary Matching Tool by the EU Horizon 2020-funded AriadnePlus project.

Data identifiers

Define and consistently use a suitable file name to keep track of your entities.

1. Unique (global uniqueness)
2. Consistent
3. Simple
4. Descriptive - Informative
5. Immutable
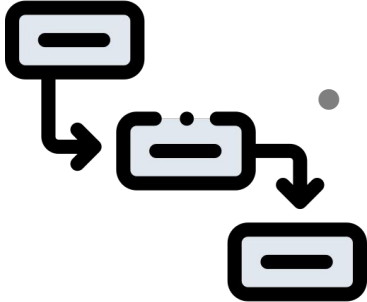6. Validated

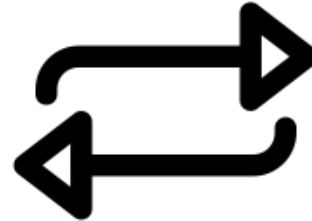# Limitations of databases

- Difficulty in access

- Data isolation

- Data inconsistency

- Data dependence

- Data redundancy

Examples of databases:

- [Thesaurus Linguae Graecae](#) (TLG): extensive collection of ancient Greek texts spanning from Homer to the fall of Byzantium in AD 1453.

- [Prosopography of the Byzantine World](#) (PBW): biographical database of individuals who lived in the Byzantine Empire and surrounding regions, specially persons of significance, including emperors, bishops, scholars, and other historical figures.

- [Byzantine Seals Collection](#): Rich source of information about the Byzantine empire. Ranging from the fourth to fifteenth centuries and covering a wide array of Byzantine society, seals provide a unique lens through which to view social, institutional, religious, and artistic developments in the empire.

Exercise:

Are you familiar with the databases listed in the previous slide?

Based on the limitations of databases that we have mentioned before and the best practices for databases that we have also looked at:

- Review these resources
- Can you say why they are good or bad?
- Can you rate them from 0 to 5 stars?
- What is the criteria that you are going to follow to rate/review them?
- What is the best/worst example?

Virtual Research Environments

- Collaboration support (wen forums and wikis)
- Shared access
- Document hosting
- Discipline Specific tools (data analysis, visualisation or simulation management)
- Backup and version history
- Support the design and execution of research workflows
- Access control and Security measures to protect sensitive data
- Cloud-based
- browser- based and access to directory
- Friendly user interfaces that allow document management, storage system and collaboration.

Microsoft Sharepoint

- Web-based collaborative platform that integrates natively with Microsoft 365.
- Recommended if you use Microsoft Office and Teams.
- Primarily sold as a document management and storage system.
- Compatible with other Microsoft applications.
- Includes team collaboration groupware capabilities: project scheduling, special collaboration, shared mailboxes, project related document storage and collaboration.

Google Drive

- File storage and synchronization service
- Encompasses Google Docs, Sheets, Slides.
- 15GB of free storage shared with Gmail and Google Photos.
- Google drive's privacy policy has been very critizided.
- A number of external applications are available from Web Store.
- It incorporates a File viewing feature that allows several formats including:Autodesk AutoCad (.DXF)
  - Scalable Vector Graphics (.SVG)
  - PostScript (.EPS, .PS)
  - Python (.PY)
  - Fonts (.TTF)
  - XML Paper Specification (.XPS)
  - Archive file types (.ZIP, .RAR, tar, gzip)
 - .MTS files
  - Raw Image formats (.DNG)

Google Drive

D4Science

- Promotes Open Science through a data infrastructure service.
- community-driven virtual research environment.
- Hosted by the Instituto di Scienza e Technologie dell'Informazione of National Research Council (Italy).
- Developed and supported by several European-funded projects.
- It also provides a Jupiter-based notebook environment.

GitHub

- Developer platform that allows developers to create, store, manage and share their code.
- Commonly used to host open source software
- World largest source code host in 2023.
- It uses Git software (distributed version control system that tracks changes in computer files).
- Headquartered in California, subsidiary of Microsoft since 2018.
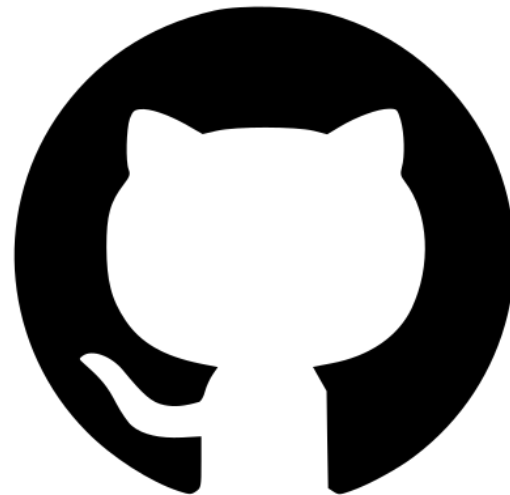
Exercise:

Design your own database:

- Identify the research goal - define purpose and scope.
- Understand your data - Know what you are recording
- Plan/record your schema
- Ensure data integrity mechanisms - e.g. vocabularies
- Find/create user-friendly interfaces for data entry, querying and visualization.
- Document your process