

D
a
t
a
b
a
s
e
s

for young historians

Databases
*for young
historians:
theory
&
practice*

NO PREVIOUS
KNOWLEDGE
REQUIRED

Course:
**25–27 January and
19 February 2016**
Huygens-ING building
(The Hague)



HUIZINGA
INSTITUUT



DAY 1 + DAY 2
Pim van Bree & Geert Kessels
@LAB1100

An introduction to
"database development"

Day 1

09.30 Welcome & Coffee

10.00 Introduction of programme & participants

10.30 Presentation on databases in the humanities, examples and good practices

From Roberto Busa to WIKIDATA.

11.00 Exercise: Text to Database (groups)

Brief explanation on how to conceptualise a data model. One short text on a research question is read by each group. Each group conceptualises a data model based on this text. Each group will present their data model to give insight on the choices they have made.

12.30 Lunch

13.30 Conceptualise a data model (individual/groups)

Participants may form groups or work individually to conceptualise a data model based on their own research question.

14.15 Presentations/discussion of models

Each data model will be presented to discuss its strengths, weaknesses and consequences (in terms of feasibility/workload).

15.00 Break

15.30 LibreOffice Base: learn how to create a database

Hands on tutorial on how to create a database in LibreOffice Base.

16.15 LibreOffice Base: create your own database (individual/groups)

Participants may form groups or work individually to create a database in LibreOffice Base, based on the data model they have conceptualised.

17.00 End of day 1

Day 2

09.30 Welcome & Coffee

10.00 Presentations/discussions of databases

Each database created during day 1 will be presented and discussed.

11.00 Introduction to nodegoat

We will briefly go through an exemplary project: Mapping Notes & Nodes.

11.30 Learn how to enter data into nodegoat

Hands on tutorial on how to enter data into a data model in nodegoat.

12.30 Lunch

13.30 Learn how to build a data model in nodegoat (individual/groups)

Hands on tutorial on how to create a data model in nodegoat. Next, participants may form groups or work individually to create a project in nodegoat, based on the data model they have conceptualised.

14.30 Enter data into your own data model (individual/groups)

Once the data model is ready, data can be entered into nodegoat to produce geographic and social network visualisations.

15.00 Break

15.30 Presentation of results

Each project will be presented to discuss its strengths, weaknesses and consequences (in terms of feasibility/workload).

16.30 Linked Data

A brief introduction on the principles behind Linked Data and how this can be used to make datasets interoperable.

17.00 End of day 2

Humanities & Databases: some points in time

- First computational data storage and handling in the humanities: [Index Thomisticus](#) by Roberto Busa (1949-1980, in cooperation with IBM).
- [Manfred Thaller](#)'s 'The Historical Workstation Project': [Kleio](#), since 1978.
- [Arachne](#) (central Object database of the German Archaeological Institute (DAI) and the Archaeological Institute of the University of Cologne).
- [ECARTICO](#) (UvA, people involved in the 'cultural industries' of the Low Countries in the sixteenth and seventeenth centuries).
- Wikipedia → DBPedia, WIKIDATA.
- other interesting examples?

Humanities & Databases: opportunities

- Store / structure your own research data (card catalogue)
- Count, filter, query (diachronically)
- Geographical mappings (patterns/information highways)
- Network analysis (hubs/paths/clusters)
- Object-oriented approach (actor-network theory)
- Connect your data to shared resources (linked data)
- Publish datasets (figshare/github)
- ...

Humanities & Databases: challenges

- Conceptual Challenge (data modeling)
- Epistemological Challenge (data completeness/uncertainty)
- Technological Challenges (interfaces)

Huge time investment



Humanities & Databases: common issues

- How to determine the scope of your research?
- How to deal with unknown/uncertain primary source material?
- How to use/import 'structured' data?
- How to reference entries in a dataset and how to deal with conflicting sources?
- How to deal with unique/specific objects in a table/type? (pt. 1)
- How to deal with unique/specific objects in a table/type? (pt. 2)

Humanities & Databases: common issues (1/6)

How to determine the scope of your research?

Level of detail and range of subject matter depends on

...what your current research questions are but also on possible future hypotheses. (i.e. are you open to formulate new research questions while working on your dataset?)

...the purpose of use: individual / project-based / commons (e.g. linked open data)

...the period in time you research. You might have to deal with conceptual changes through time (e.g. 'capacity' vs 'occupation') or changing characteristics of your objects (e.g. Deutsches Kaiserreich/Weimarer Republik/Deutsches Reich/Großdeutsches Reich/BDR/DDR).



Humanities & Databases: common issues (2/6)

How to deal with unknown/uncertain primary source material?

Even though it might be impossible to establish an exact date/location, there are multiple ways that will allow you to make statements on temporality/locality.

- Use a period instead of a fixed date (Statue 1 was made between 1855 and 1861).
- Use a 'Before ...' statement (Artist X was born before creation of first artwork).
- Use a 'After ...' statement (Event Z took place after coronation of Ruler Y).
- Include these options explicitly in your date model (Date inferred: true/false, Date uncertain: true/false). This allows you to filter on these values (you could also use scales for certainty).
- The same works for locations (and for any other statement): explicitly set whether a location/statement is uncertain/inferred.
- Related question: do I really need to research the date of birth of this obscure figure just for the sake of completeness of my dataset?

Humanities & Databases: common issues (3/6)

How to use/import 'structured' data?

You might have (semi-)structured primary sources that can be used to populate your database. For example:

- Textual data found at similar/recurring places in document sources
- Indexes / registers
- Member / participation lists

However, you probably have to deal with:

- Issues inherent to textual sources; its freedom & inconsistency (previous slide)
- Disambiguation of the data

So, structured data \neq actionable data!

Humanities & Databases: common issues (4/6)

How to reference entries in a dataset and how to deal with conflicting sources?

- Databases can reflect historiographical debates (D.O.B. based on source I = xxxx, D.O.B. based on source II = xxxx). It is important to incorporate this flexibility in your data model.
- Connect statements/entries directly to your bibliography. This can be done per row/object or per field in a row. (You could for example use [Zotero](#) or [Mendeley](#) for this.)
- Version management can reflect your own research process/decisions/development. (cf. 'digital forensics', i.e.: <http://dhbenelux.org/wp-content/uploads/2015/04/30.pdf>)

Humanities & Databases: common issues (5/6)

How to deal with unique/specific objects in a table/type? (pt. 1)

Every row/object does **not** need unique columns/attributes!

	d.o.b.	d.o.d.	capacity	catholic	gave lectures	traveled to Rome	friends with Pope
person 1	1565?	2/2/1645	painter	no	-	yes	-
person 2	may?	1-2-??	sculptor	-	yes	-	yes

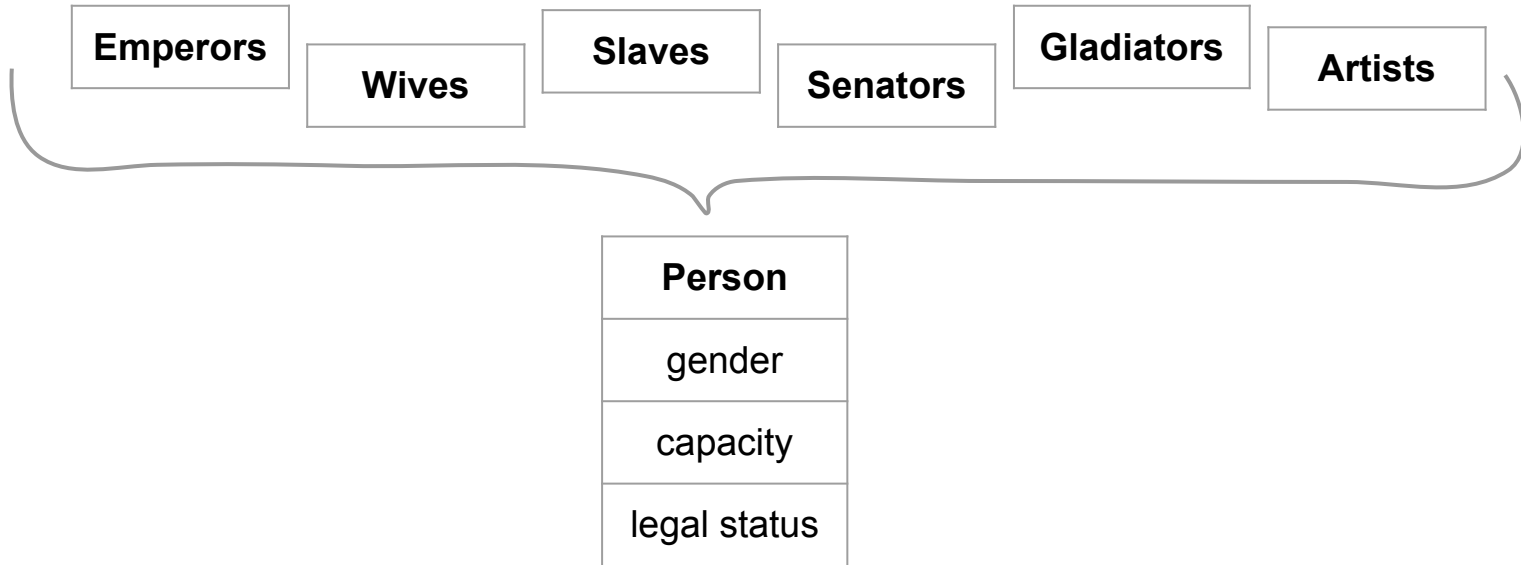


	d.o.b.	d.o.d.	capacity	religion	notes
person 1	1565?	2/2/1645	painter	catholic	Known to have traveled to Rome.
person 2	may?	1-2-??	sculptor lecturer	unknown	Known to be friends with the Pope.

Humanities & Databases: common issues (6/6)

How to deal with unique/specific objects in a table/type? (pt. 2)

Design your tables/types as broad as possible. This ensures flexibility and helps you to avoid overlap between tables/types.



Databases: terminology

"A database is an organized collection of data. It is the collection of schemas, tables, queries, reports, views and other objects."

"A database management system (DBMS) is a computer software application that interacts with the user, other applications, and the database itself to capture and analyze data. (...) Sometimes a DBMS is loosely referred to as a 'database'."

- <https://en.wikipedia.org/wiki/Database>

Excel \neq a database

Databases: terminology

(A selection of) different kind of database models:

- Relational model (e.g. [Zuccaro](#))
- Hierarchical model (e.g. [ICONCLASS](#))
- Document model (e.g. [Wikidata](#), [Talk of Europe](#))
- Graph. Computing & traversal.

Databases: terminology

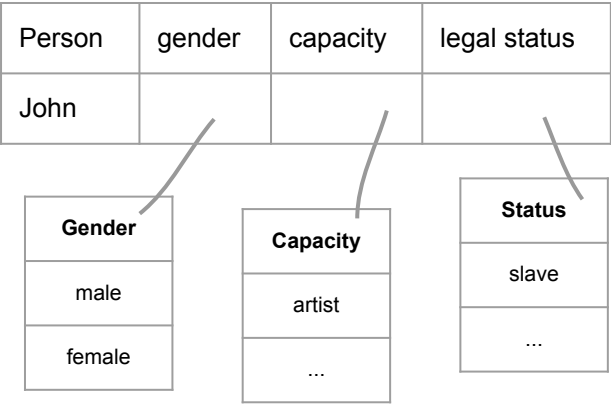
Your own understanding of your data model and conceptual comprehension matters most.

Where and to what end do you create relations between entities. Ontology vs convention.

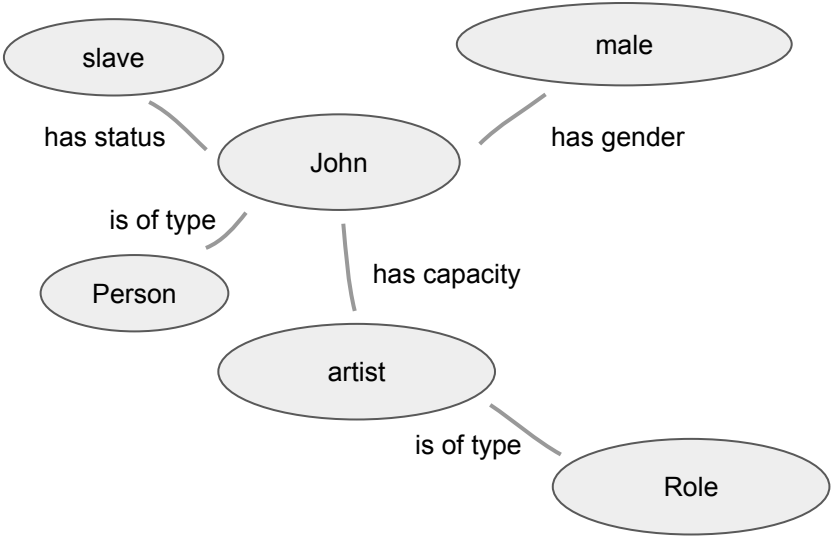
The database technology used is only the means of accessing that data model.

Databases: terminology

Relational model vs. graph oriented models



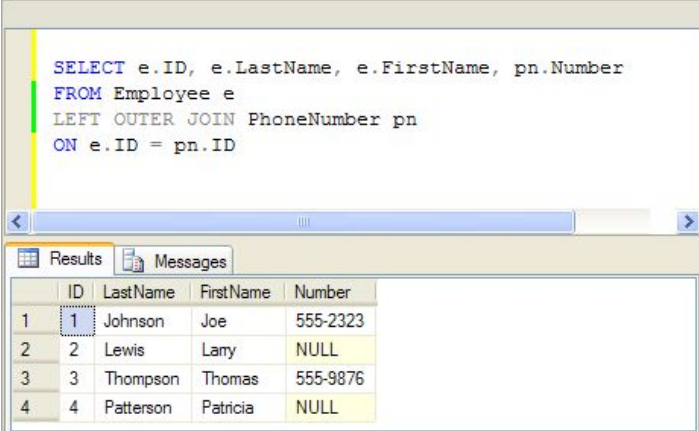
= & ≠



Databases: terminology

Today and tomorrow we will work with the relational model.

Most relational databases use SQL, see: <https://en.wikipedia.org/wiki/SQL>



The screenshot shows a SQL query window with the following text:

```
SELECT e.ID, e.LastName, e.FirstName, pn.Number
FROM Employee e
LEFT OUTER JOIN PhoneNumber pn
ON e.ID = pn.ID
```

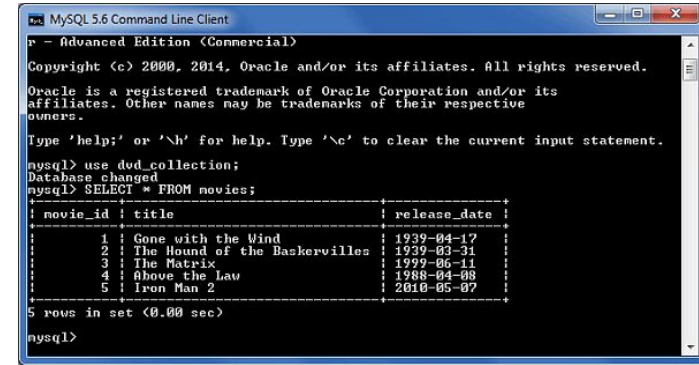
Below the query window, there is a 'Results' tab showing a table with the following data:

	ID	LastName	FirstName	Number
1	1	Johnson	Joe	555-2323
2	2	Lewis	Lary	NULL
3	3	Thompson	Thomas	555-9876
4	4	Patterson	Patricia	NULL

Databases: terminology

Relational DBMS without a graphic user interface:

- MySQL
- PostgreSQL

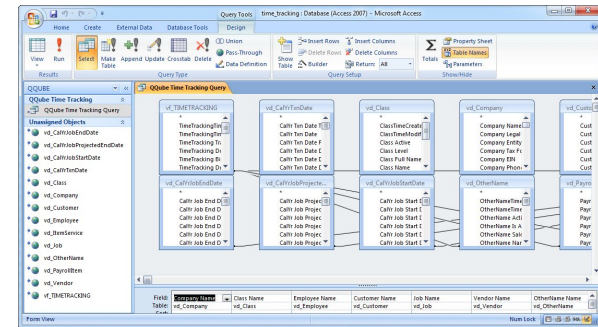


```
mysql> use dvd_collection;
Database changed
mysql> SELECT * FROM movies;
+-----+-----+-----+
| movie_id | title                | release_date |
+-----+-----+-----+
| 1         | Gone with the Wind   | 1939-04-17   |
| 2         | The Hound of the     | 1939-03-31   |
| 3         | The Matrix           | 1999-06-11   |
| 4         | Above the Law        | 1988-04-08   |
| 5         | Iron Man 2           | 2010-05-07   |
+-----+-----+-----+
5 rows in set (0.00 sec)

mysql>
```

Relational DBMS with a graphic user interface:

- MS Access
- Filemaker
- dBase
- HSQLDB



Databases: terminology

We will be using LibreOffice Base today which is only a graphic user interface and uses HSQLDB as its database engine.

LibreOffice Base can also be used with other databases, for example: <http://www.linuxuser.co.uk/tutorials/make-a-small-business-database-with-libreoffice>

Why LibreOffice Base in this course?

- Cross platform availability (for any java issues [click here](#))
- Free

Databases: terminology

Table 1: Person

id	first name	last name	d.o.b.	gender	bio
1	Ludovit	Stur	15-1-1815	m	Slovak hero.
2	Jacob	Grimm	1-2-1789	m	German philologist.

Primary Key



Table 2: Letter

id	sender	receiver	date	send location	receive location
1	1	2	8-6-1851	Modra	Berlin
2	2	1	13-9-1852	Berlin	Pressburg

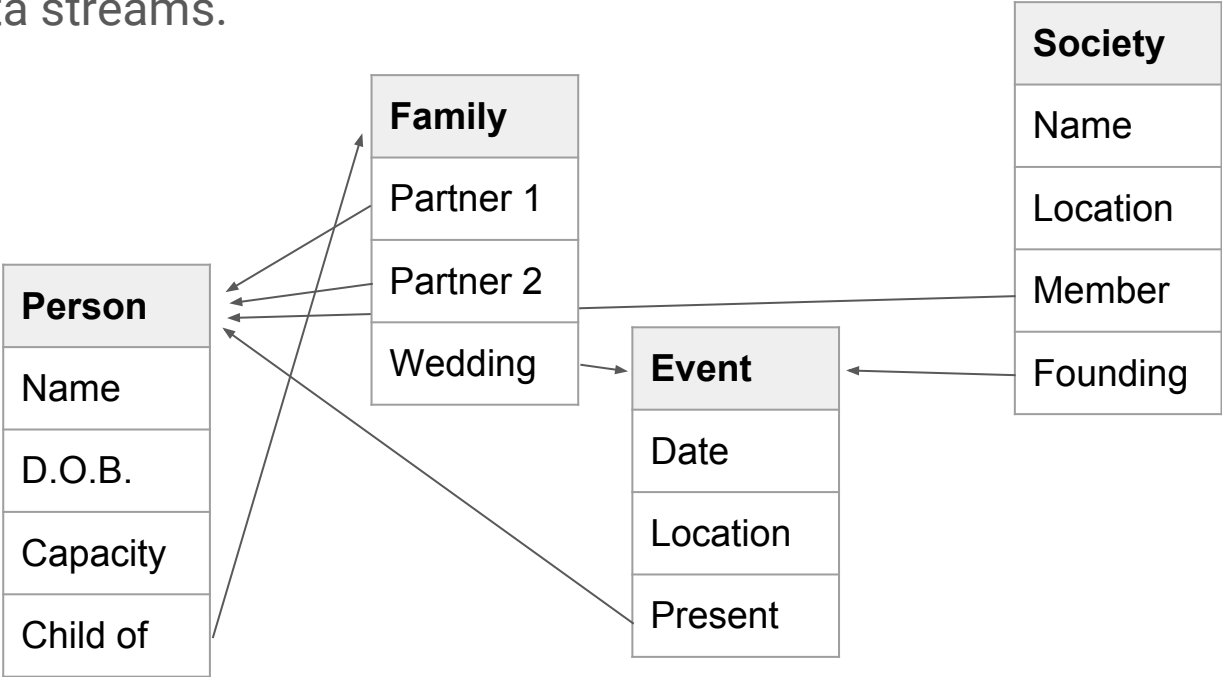
Foreign keys



Data modeling

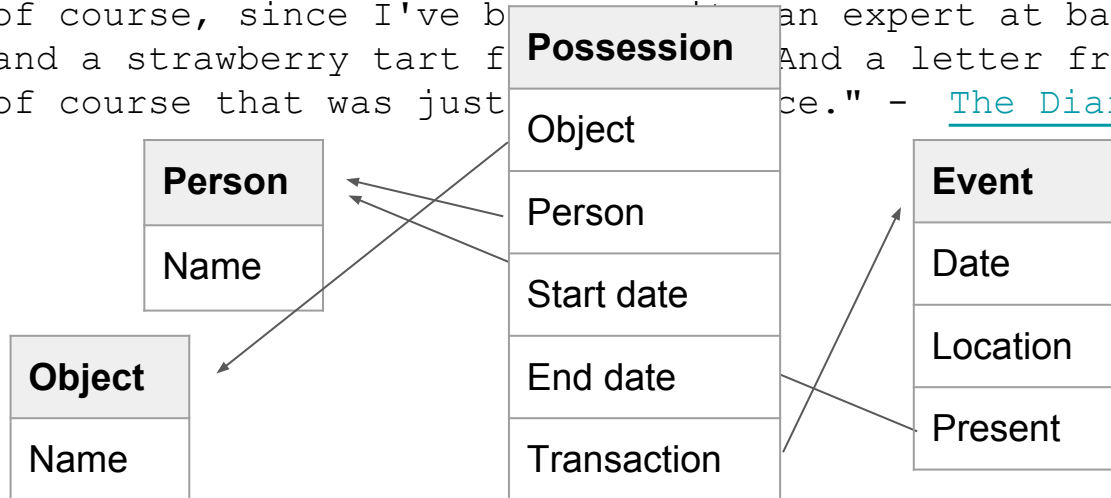
Start by imagining the research scope as wide as possible. Then proceed by narrowing down by means of structured data streams. Next, make connections between these data streams.

For example:



Data modeling

"A little after seven I went to Daddy and Mama and then to the living room to open my presents, and you were the first thing I saw, maybe one of my nicest presents. Then a bouquet of roses, some peonies and a potted plant. From Daddy and Mama I got a blue blouse, a game, a bottle of grape juice, which to my mind tastes a bit like wine (after all, wine is made from grapes), a puzzle, a jar of cold cream, 2.50 guilders and a gift certificate for two books. I got another book as well, Camera Obscura (but Margot already has it, so I exchanged mine for something else), a platter of homemade cookies (which I made myself, of course, since I've been an expert at baking cookies), lots of candy and a strawberry tart for Daddy. And a letter from Grammy, right on time, but of course that was just a letter." - [The Diary Of Anne Frank](#)



Data modeling - Task 1

Read the first six pages of Christopher Clark's *The Sleepwalkers* (page 19 to 25).

Use this text as a starting point for a data model. The goal is not to make the most comprehensive/correct data model, but to make a data model that makes sense and is useful for a research process. This can be from a cultural/gender/Serbian/comparative/transnational etc. perspective.

Formulate any number of tables with any number of columns and any number of relations.

Think about the common issues we've discussed and how this model can be used to answer research questions.

Share your results

[Open this document](#) [link removed] and enter a link to your shared document

Example: [https://docs.google.](https://docs.google.com/presentation/d/16YrlgQflyZK02QIKTd3kd6r6qAhn6O87UeRysWLixRA/presentation)

[com/presentation/d/16YrlgQflyZK02QIKTd3kd6r6qAhn6O87UeRysWLixRA/presentation](https://docs.google.com/presentation/d/16YrlgQflyZK02QIKTd3kd6r6qAhn6O87UeRysWLixRA/presentation)

Data modeling - Task 2

Conceptualise a data model based on your own research question.

Formulate any number of tables with any number of columns and any number of relations.

Think about the common issues we've discussed and how this model can be used to answer research questions.

Share your results

[Open this document](#) [link removed] and enter a link to your shared document

LibreOffice Base

Together we will create a database with two tables ('Person', 'Letter') and one form to enter letters.

(For later) walkthrough: <https://docs.google.com/document/d/137QRdSLEiOBW98NghtliGzWxCHX0fMBuqjOyHLOLNws/edit?usp=sharing>

See this video for a database with 'many to many' relationships: https://www.youtube.com/watch?v=GYawY08u3_s

nodegoat

- Web-based research environment
- Create and manage any number of datasets
- Collaborative data entry / data ingestion processes
- Analyse and visualise complex datasets relationally, diachronically and spatially; trailblazing

Go to <http://nodegoat.net> and log in with the username 'demo_mnn' and password 'demo' to explore the '[Mapping Nodes & Notes](#)' project.

nodegoat

You will first enter data in a shared environment with a pre-defined data design.

Secondly, you will learn how to create your own data design and be able to set up your own project with a custom data design in nodegoat.

(For later) see the nodegoat video tutorials: <https://www.youtube.com/watch?v=eLDRNiJrRUc&list=PLXc6y7I7xxxIwd64QppyAA0G2ECsNGJCx>

Linked Data for historians

It is good practice to (manually) store persistent identifiers or URIs in your dataset. For example: <http://dbpedia.org/resource/Rembrandt>, <https://www.wikidata.org/wiki/Q5598>, <http://viaf.org/viaf/64013650/>.

Benefits:

- Data disambiguation (you mean [Francis Bacon](#) and not [Francis Bacon](#))
- Enhance interoperability
 - For other researchers to reuse your data
 - For aggregators to harvest your data (like <http://correspsearch.bbaw.de/>)

See further: <http://nodegoat.net/blog.s/12/linked-data-vs-curation-island>

Questions or feedback:

info@LAB1100.com / [@LAB1100](#)

For nodegoat:

[FAQ](#) / [Forum](#) / [@nodegoat](#)

