Tutorial/Workshop Session 2: Claire Lemercier: From Sources to Databases: Data extraction out of Humanistic Sources

Complicated relations are nothing more than rows and columns

- please do not become slaves of programs; the basic structure of data is in quantitative columns

- will not be discussing software, but beginning with some sort of material that is accessible

- underdiscussed part of research: inputting?

- put data in rows and columns

- transform data as literally(?) as possible

- "code data" not in a sense of programming but in terms of signifying

- separate "inputing data" and "coding data" (become prisoners of own coding)

- flexible ways of inputting data that keeps open the possibilites of dealing with data but still adding structure

- call for people who have data and how we would input or code it according to the (negative) principles she'd like to show us (principles of what we shouldn't do)

-- this notetaker is confused about what she means by material... :(

Case Study 1: Caravanserais

Criteria:

- - name
- - geographic coord.
- - cited/discovered by (first name, surname, source title)
- - material (stone, brick...)
- - date of construction
- - water installation
- - roads/paths
- - archeological data (+/- details)
- - who built it
- - near to village?

(ca. 72 caravanserais, Syria)

General aims of building the dataset:

1) She is trying to see the development of the building techniques, so her interest is in the material and archaeological results, reconstructing patterns of patronage (based on who commissioned the road)

2) reconstruct the road networks on the basis of caravanserais and stopovers cited by different travellers

3) see the caravanserais distribution patterns

Problem in humanities: incertitude!

Question: if you have several sources, how do you deal with it in data extraction?

L'entité sur laquelle je travaille est-elle si évidente que ça? It is an option to focus on the source and not on the unit (here a caravanserail)

It's not because it is a table that you have to be 100% sure about your data -- shouldn't you always be 100% sure about your data?? you can also express your doubt (for example, datation: year 1455, 15th cent., uncertain, unknown, ...)

[clarification: you should be sure about your data (what you've recorded) but not about its exactitude and relevance/usefulness to the research project as yet (?)]

- should use software that is flexible but allows for retrospective notes/columns/changes, etc.

- if your data is more structured, you should always be able to transform the data into another table

Comment séparer les informations qui sont miennes ou qui viennent d'autres collègues/études? Il est possible de faire des "doubles colonnes" par ex :

| DATE |

|1455 | because I found it in this reference... |

|1472-73| depends on the day of birth of X |

Case Study 2 - Italian politician (?):

Heterougenous documents written by one person :

- type

- date

- opinions -> concepts (closed or open list) How? Most frequent lemmas

General aims of building the dataset:

- Have a better understanding of the chronology of ideas by the studied politician

Case Study 3 - Arabic versions of the New Testament:

How to deal with uncertainty ?

- It's better not to have empty cells.... Specify if unknown, or if inexistant, and why it is unknown or inexistant. (always good to explain why you have no data - not just for statistical purposes)

- Il vaut mieux avoir des ; que des espaces (reconnus par excel)

Recommended software:

- pajek

- "clustering" is a family of techniques used to create typologies according to similarities

- - R is the software that helps map this (see below)
- - beginner's "how-to" tutorial offered by Lemercier in another forum (she will send us PDF later) That would be good too for those of us who couldn't attend this workshop.

- family of algorithms called 'clustering" not with excel or calc but with statistical software called R (http://www.r-project.org/)
- TraMineR (see notes from Lemercier's lecture)

- package that provides you with something to click: RCommander (the most used statistical collations)

- additional package for correspondance analysis "FactormineR" -- allows you to export your data back to where you're more comfortable (Fehler! Linkverweis ungÜltig.

- alternatives to R - SAS, Stata, SPSS [don't follow that recommandation, please] <-- it wasn't mine! I'm just recording!! it's all acronyms to me..! ;)

[that's why I'm putting it in-between brackets : it's my opinion, I'm not even present at this workshop]

(hi, it's Claire - I said you should definitely use R (several times) but if it happens that in your University you have commercial software and nice colleagues who want to work with you on it, well, why not)