# Editing a 15th century political treatise using the computer: a back and forth between meaning and information

Matthias Gille Levenson

# Editing a 15th century political treatise using the computer: a back and forth between meaning and information*

Matthias Gille Levenson

## Introduction: towards a comparative edition

I am working on the edition of part of the *Regimiento de los prínçipes*, a political treatise of the XIV^th^ century, that is itself the translation of the famous *De Regimine Principum* of Gilles of Rome. I am focusing my study on the discourse of chivalry and nobility in this text. I am influenced by the so called "New Philology"[1] and by the works of Roger Chartier on the materiality of the texts: we cannot put aside this materiality when studying a text. As Roger Chartier points out,

> *La sociologie des textes telle que l'a définie D.F McKenzie conduit à tenir chaque état d'une œuvre comme l'un de ses états historiques,* qu'il faut comprendre, respecter et, possiblement, éditer. *Le concept d'un* "ideal copy text", *existant en deçà ou au-delà des différentes formes imprimées (ou manuscrites) d'une œuvre, est une illusion que la critique textuelle doit abandonner au profit de l'analyse des effets produits sur le texte, ses lecteurs et, éventuellement, son auteur, par chacune de ses existences matérielles*[2].

My aim is to study the discourses that I have in the different manuscripts of my textual tradition, as well as its homogeneity: in other words, I want to evaluate how much of a text I have in the *Regimiento*. What can we say about the use of the different manuscripts that circulated in these 150 years ? What do they say, individually ? How were the texts read, received, or transformed by the actors that decided to re-produce them ? From an ecdotic point of view, I am not interested in the ideal text. I need to produce *as many texts as I have witnesses*. For this reason, I plan to do not a critical edition, but rather a comparative edition: I will not choose between "good" and "bad" variants in order to access and produce the original text. This presentation will focus on the methodology that is being used to produce this edition.

*I would like to thank Jesús Velasco for giving me the oportunity to talk, as well as to Karina López and Marjorie Burghart for the time they took to read this text, and to help me correcting it.

[1]See for example the works of Bernard Cerquiglini, and the volume 65 of *Speculum*, published in 1990.

[2]Roger Chartier, "Les chemins de l'écrit, ou le retour à Monte Verità", in: *Scripta volant, verba manent. Schriftkulturen in. Europa zwischen 1500 und 1900/Les cultures de lécrit en Europe entre 1500 et 1900*, Zurich, Schwabe, 2008, pp. 483–493, p. 490.

## Data, information, meaning

Let's compare versions of an extract of the *Regimiento*. G, J, Z are the sigla of three witnesses, two manuscripts and an incunable. We are in the gloss of the chapter 21 of the last part of the text, that speaks of the funeral honours given by Alexander the Great to Darius' wife.

> *E llamó ý un muy grant sabio de los iudíos que dezían Apelles, que era de Jherusalém* (J)
> *E llamó ý un grand sabio de los iudíos que dezían Apelles, que era de Gerusalén* (G)
> *E llamó ý un muy grand sabio de los indios que dezían Apelles que era grand pintor* (Z)

One possible representation of the difference between these three sentences would be:

**8** muy *J Z* | *om.* G **9** iudíos *J G* | indios *Z* **9** Jherusalém *J G* | grand pintor *Z*

The second apparatus entry means that in the witness Z, "*iudío*", was replaced by "*indio*". That is meaning, a translation made by the human mind of some information: a sequence of characters, a pipe "|" and another sequence of characters; this information is meaningful for those who know the conventions of the critical apparatus.

This is one difference between us and the computer. The computer 'sees' the information, and may process it. From this, we get or we create a meaning by reading the information. In an article published in 2007, Chaims Zins analyses how searchers define data, information and meaning more than one hundred papers presented at an conference on Information Science. highlights that there are as many definitions of these concepts as people who try to define them[3]. I decided to choose one definition from the 130 that the article presents.

> **Data** are the basic individual items of numeric or other information, garnered through observation; but in themselves, without context, they are devoid of information. **Information** is that which is conveyed, and possibly amenable to analysis and interpretation, through data and the context in which the data are assembled[4].

As for the term "meaning", the *Cambridge Dictionary* tells us:

> The meaning[5] of something is what it expresses or represents[6].

Meaning is information that makes sense. Meaning is the information when read by the human mind: this is why the same document can be either meaning or information, depending on what one does with it. This conceptual distinction is important to me, because it helps me to understand why and how I do the things I do to produce my texts. In this presentation I will talk of the dialectic between meaning and information, that is: what are the steps I have to take before the production of meaning, the publication of my edition ? How the conceptual distinction between 'information' and 'meaning' can help us (can help me) understand what I am doing ? How to describe my work in terms of back and forth between meaning and information ?

My use of the machine is based on the following assumption: **computers can manipulate information better than we can**, and they can help us in producing more homogeneous or consistent documents. But ultimately what is important for us as humans is to deal with meaning. How do we get from information to meaning? We will see that we have to complete successive tasks of translation between meaning and information to get the desired result, that is: **a comparative edition for each of the transcribed witnesses, with a meaningful and readable apparatus**[7].

[3] Chaim Zins, "Conceptual approaches for defining data, information, and knowledge", *Journal of the American Society for Information Science and Technology* 58.4 (2007), pp. 479–493.

[4] Ibid., p. 481.

[5] I've decided to choose the term "meaning" because the conceptual opposition *information vs meaning* is far more productive that *information vs knowledge*: as far as we are concerned, the computer doesn't get the *meaning* of a text.

[6] https://dictionary.cambridge.org/dictionary/english/meaning

[7] The graphical variant will have to be excluded from the apparatus.

## The workflow

The traditional way to edit a text is as follows: transcription of a witness, collation with the other witnesses, creation of the apparatus[8]. My own approach adds two steps (which are absolutely not original): the TEI encoding and a phase of addition of some grammatical information: lemmas[9] and Part of Speech[10] (PoS). If we try to think in terms of manipulation of the information, the steps I take are:

1. Acquisition of the information: the transcription of the texts
2. Structuring of the information: the TEI encoding
3. Enrichment of the information: lemmatisation and PoStagging
4. Production of more meaningful information: collation and creation of the apparatus
5. Translation of the information: production of readable documents: pdfs (and/or web-based interface)

The presentation will follow this workflow, step by step; I will try to show the links between all of them.

# 1 Acquiring the information: the transcription

If realized with the help of the computer, the transcription is an acquisition of the information, or a first translation between data and information: after that, we only have more or less structured information.

We use programs that are called OCR programms: optical character recognition, based on neural networks. I use Ocropy for recognizing the incunable. How does it work ? We train a model, and use this model to recognize new text. The model is trained by feeding it with veridic data, in our case text and image. This data is called ground truth. With this ground truth the programm will start creating a recognition model. With this first model, we will predict new text, and correct it to transform it into new ground truth, re-train the model, etc. Once a given error rate is reached, we can stop an use the model for transcribing the whole text. As surprising as it can seem, the ideal rate has not to be the lowest possible, because we don't want the model to be overfitting on the training data: we want it to be accurate on new data.

I will insist on the importance of the correction: a prediction using neural networks is based on probabilities, and thus is prone to errors[11].

[8] Depending on the methodology one chooses, we may add to this list the production of a *stemma*, and the selection of the "good" variants. I will not talk about these two steps in my presentation: I am not interested in re-creating a virtual original text; I am not advanced enough in my work to talk about the *stemma codicum*.

[9] https://en.wikipedia.org/wiki/Lemma_(morphology).

[10] https://en.wikipedia.org/wiki/Part_of_speech.

[11] I have used the tool called Ocropus to predict the text of the incunable. The error rate would be too high for using it to predict the text of manuscripts, but more recent tools are being developped and are reaching an acceptable error rate with manuscripts: see Kraken (Benjamin Kiessling, "Kraken - an Universal Text Recognizer for the Humanities", in: , DH2019 : Complexity, Utrecht, 2019) (open-source) or Transkribus (Philip Kahle et al., "Transkribus-a service platform for transcription, recognition and retrieval of historical documents", in: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4, IEEE, 2017, pp. 19–24) (not fully open-source).

## 2 Structuring the information. Digital encoding and TEI

If not the most important, the XML-TEI encoding is one of the most important steps of my workflow, mainly because of the standard characteristic of the TEI[12]. I will show in this part that the difference between information and meaning is not as obvious as it can seem first: a TEI conformant file, as the ideal description of a given text, is processable by the machine, but also readable and can be understood by the human being, thanks to its semantic nature. Why ? Because the TEI is a "conceptual model of textuality", as says Fabio Ciotti[13]. It proposes an ontology on the structure of the text, not informatically speaking (the TEI is not what informaticians call an ontology), but rather in its philosophical meaning. The TEI data model doesn't explain or states on what *is* a text, but it does say what *is inside* a text, what can be the different components of a text: the different elements, but also their interelation. TEI gives us a semantism. It is human-readable, but also machine-readable: in other words, it is meaning, but it is also information that can be processed by the computer, to create multiple outputs, or indices, to extract some particular information, etc. Last but not least, it is a community driven standard: it evolves with the needs and the reflections of the community of users.

The paradox here is that the TEI encoding, for my project at least, is the step that can be the less automatized, precisely because of this semantic nature: we are here in the first phase of construction of the meaning, in the most important transition from information to meaning, a phase that has to be realized by the human being, by the editor.

## 3 Enrichment of the information: Lemmatisation and PoStagging

For my project, I need to be able to make efficient searchs within my text. For medieval languages, that can be tricky because the spelling of the words is not fixed yet: I need to add some information that will help me go over these issues. Grammatical annotations can be helpful here: I am talking about lemmatisation and PoStagging, that I mentionned before. The lemmatisation is the act of indicating for each word the canonical word it refers to; the PoStagging is about indicating the morphology of the word, with a code that indicate its nature, its gender, the mode if applied, the person, etc. This code, named *tagset*, is a convention: in my case I am using the EAGLES tagset[14].

I will give an example: the legal aspects of my text have to be studied, as Jesús Velasco helped me understand. In Spanish, one particular mode and time is very often used in legal contexts: the future subjunctive. With a non PoStagged text, it would be really difficult to extract all of the occurrences. With a PoStagged corpus, if we want to find the verbs that are in future subjunctive, we will try to find all of the words whose PoS will match the code for future subjunctive.

This is the step I call enrichment of the structured information: the interesting thing is that I can add this grammatical information to the TEI encoding: it will be really helpful in order to process the text, in particular to collate it, as I will show.

## 4 Comparison of the information. The *collatio*.

How do we get from *n* individual transcription to a compared text ? With this step I will be able to get to some meaning, and to answer to some my questions: how much of a text do I

[12] For Text Encoding Initiative, see here.

[13] Fabio Ciotti, "A Formal Ontology for the Text Encoding Initiative", *Umanistica Digitale* 2.3 (Nov. 18, 2018).

[14] See here. I have been using a dictionnary of forms, lemmas and PoS created for the project Freeling (Lluís Padró, "Analizadores Multilingües en FreeLing", *Linguamática* 3.2 (2011), pp. 13–20), that decided to follow this tagset.

have ? What is the importance of the variation ? The collation is a way to add relevance to the information: from several witnesses, or, in my case, several transcription, I will produce a single document – per witness – that gathers all of the information about the variation. The process of collating several witnesses is the following: we want to compare the texts, so we need to find the equivalent texts (to align the texts) and to compare the content of each passage. These two steps are virtual: a human mind cannot dissociate them, but I need this distinction to simulate the process of collation that I will explain in detail.

The important idea is that here the words are represented by their attributes, form (the sequence of characters of the word), lemma and PoS. This deconstruction, as we will see, is a translation too, this time from meaning (the concept of "word") to information (the attributes that can identify a word). Another point is that we use here the computer to create more meaningful content, to clean up the apparatus from unuseful information: the computer deals with information to make the meaning emerge more easily. However, the created documents will not be critical editions, something that only the human mind can achieve; however this method could be used to create critical editions. It has issues: the main one here is related to what is called "overlapping", a situation that is not allowed by the specifications of the XML format, and that might be generated here.

Once the collation phase is done, the program will transform the result into XML, and inject this XML information to the individual TEI structured transcriptions. The interests of this method are multiple: it makes me gain a lot of time, and it allows me to have each particular manuscript as a base, with the apparatus as the result of the comparison between them.

# 5 Transformation of the information: the creation of the output

The final creation of a meaningful object: the edition. I use the typesetting language/program LaTeX for the pdf output, and I will show samples of the result. An illustration of what can be the web-based output of the edition[15], on the passage I've been talking about, can also be found here.

# Conclusions

In order to use the computer, there is a series of translations to be made between meaning and information. One of them can be considered as a deconstruction of the aforementioned objects: the word, for example, will be represented by its attributes (string and grammatical information, lemma and PoS). The same can be said of philological methods: the deconstruction of the cognitive process of the *collatio*, for example, is what we have to look after if we want to simulate this ecdotic method. As we can see, the computer is here to accomplish the majority of the mechanisable tasks, and these tasks alone. The result is always to be controled by the editor: as I've said before, the correction phase is essential.

The very last translation operation would be, of course, the close reading of the text, in order to create the final meaning: I am talking about the comments on the edited sources, the construction of the discourse upon all of the texts. The apparatus shows a difference between "*iudío*" (jew) and "*indio*" (indian). How can I explain it ? The history of the incunable is important here: it was printed in 1494 by people that were somehow close to the *Reyes Católicos*, as I have demonstrated in my article; the edition of 1494 could be part of the cultural and political project of the

[15] The apparatus are still to be added.

Catholic Kings[16]. The change made by the incunable is not likely to be a correction; it seems to be a conscious and wanted modification, not an error or a misreading[17]: if there is no mention anywhere of the Greek painter Apelles being Jewish, neither is there any other mention of him being related to the Indians! This is why I think this modification can be interpreted as an act of censorship, the silencing of a minority that had been expulsed from the Peninsula two years before.

## After the production: what comes next ?

I would like to highlight three concepts that are important to me, and to ask a question, to start the discussion.

## Three ideas

Concerning the production of a scientific document as a digital edition can be considered, the most important points are, in my view, the:

**Citability of the information.** A published and accessible web-based interface, as Elena Pierazzo calls them[18] is very often a work in progress: if we take this into account, that means there is a risk of inconsistency of the citations: how can one be sure that the text one cites will be the same in two months, or in a year ? This leads to the need to have consistent and fixed documents when they are published, or at least to record the different versions of the document. One solution is to use a version control software like git: each version is identified with the git hash, the "unique" identifier for the version of a project.

**Perennity of the information.** The lifespan of the web-based interfaces depends on funding a project can acquire. After a certain time, the funding ceases and we cannot be assured that the produced text will be accessible for the scientific community, either because the server doesn't exist anymore, or because the code is obsolete and cannot be read by modern browsers[19]. This is a real problem for the consistency of the scienfic works, a problem that takes its source in the separation between the information and its support: what is the value of an article based on a text that is not available anymore ? This is why the source (XML, in this case), as the bearer of the semantic value, should be available and stored to be kept for the long term, and why I think, at the very least, a pdf version of the edited text should be made available.

**Accessibility of the information.** The standardization of the information is an important step towards its processability; but if you don't know how to access this information, it can be considered as a lost/unuseful information. The International Image Interoperability Framework (IIIF[20])

[16]Matthias Gille Levenson, "L'évolution du Regimiento de los prínçipes (1345-1494), conditionnée par le pouvoir politique ?", in: *Écritures du Pouvoir*, ed. by Véronique Lamazou-Duplan, Ausonius Edition, vol. 2, Scripta Medievalia, 2019, pp. 137–148.

[17]Regardless of the importance of the variant, we have here a good example of what the computer can and cannot do: when transcribing the text, I couldn't tell if the word was "*iudío*" or "*indio*". The context, and the indication that Apelles was from Jerusalem (information that doesn't appear in the incunable), helped me decide what was written in each witness.

[18]Elena Pierazzo, *Digital Scholarly Editing: Theories, Models and Methods*, Ashgate Publishing, 2015.

[19]I don't think it is a good idea to rely on external initiatives like archive.org to make sure a scientific document is accessible.

[20]https://iiif.io/

proposes standardized ways to retrieve an image and its metadata. The requests to access the images from a server are documented and are the same for each implementation. Distributed Text Services (DTS[21]) does the same, but with the text instead of the image: it provides guidelines to create a standard API (application programming interface) with a standard grammar to access a text. Hopefully, the use of this new standard will increase the interoperability of the texts, and will ease their reusability.

## A question

**Question:** Where is the edition ? Is it in the TEI encoding, or in the output ?

[21] https://w3id.org/dts