Digital Text : Conceptual and Methodological Frontiers

Maria Clara PAIXÃO DE SOUSA
Universidade Estadual de Campinas (UNICAMP), Brazil
mariaclara.ps@gmail.com

0. Introduction

The upsurge of text circulation in the electronic environment has been attracting the attention of various fields of investigation in recent years. Historians have observed the role of digital texts in transforming the cultural practice of writing and reading; the cognitive sciences have investigated digital formats as knowledge representation models; literary studies have interrogated digital writing as a genre[1]. A less populated debate on the "digital revolution" is their place in the history of the technology of language processing in written form, a sphere of investigation traditionally in charge of philology, linguistics, and textual studies. It could be said that in this regard, '*Digital Text*' is an object in wait for conceptual and methodological exploration. This paper investigates some routes in this direction.

We shall examine digital texts from a strictly material perspective, examining their characteristics as *artifacts* and the functionings of their *diffusion* (that is: of the transmission chain from producing to receiving them), and endeavor to understand their significance in the material transformation of written language production. We shall argue that digital text, rather than an incremental point in the evolution of text-production techniques, is a watershed in the history of text diffusion. In digital text production, an unprecedented stage is introduced into the chain of information processing: the mediation of codification by mathematical programming. This singles out digital text as an entirely novel form of written language. As such, it claims a central place as an object of study; and at the same time, it constitutes a break-

[1] See CHARTIER, 2001; LESK, 1997, among others.

through instrument for the varied fields to which the written word, in any form, is an object of investigation – an instrument which can only be fully explored if the material singularities of digital texts in terms of information codification are taken into account.

1. “Digital Text” as a concept

1.1 Material singularity of “digital texts”

“*Texts*”, in a very strict, material sense, are spatial-temporal ‘bridges’– language registers produced in one point of space-time which may be received in a different point of space-time. In order to build those bridges, human cultures have invented means of representing information and means of materially holding this representation, producing different artifacts that can it carry away.

The history of the techniques for ‘writing’, in this material sense[2], is the history of the transformations in the technology of carrying away linguistic units through space/time by means of a symbolic system. The fundamental means of representing (or codifying) information invented by human cultures consists of a system of correspondences between graphic symbols and linguistic information (concepts or sounds) – i.e., writing. This symbolic correspondence is achieved visually; and the traditional means to materially ‘hold’ the representation are strictly linked to its visual essence – quite simply, somehow the graphic symbols have to be made apparent.

Different cultures have invented different techniques with this purpose; in a general way, techniques whereby an instrument traces the symbols into or onto some support matter. For example, the instrument ‘chisel’ can trace symbols into hard support matters such as ‘stones’, and the instrument ‘chalk’ can trace symbols onto those hard support matters. With

[2] Or the perspective of the “*history of the forms or techniques*”, as contrasted to the “*history of the cultural practices*” and “*the history of reading*” (CHARTIER 2001).

time the techniques were perfected, and the aid of other materials was included – so that the instruments would deposit apparent matter onto bland support matters (as in: a pen depositing pigments onto animal skins, and later paper; or a stamp pressing the pigments onto paper, with the mechanical press). Some technical transformations are considered crucial points into the development of this technology – the substitution of bland support matters (skin, paper) for hard support matters (stone); the invention of mechanical instruments (the press)[3].

The techniques for 'holding' the symbolic information have changed quite a lot, therefore; on the other hand, the methods for codifying language have not varied much in spirit. They have all consisted of a system of immediate correspondences between graphic symbols and linguistic information (concepts or sounds). It is true that the systems can take different forms and different symbols; that is not our point here, but rather, that their functioning is always based purely on immediate visual correspondence.

Where then, in this history, lays the place of digital texts? Are they a new technique for holding the symbolic system, in the same line of development that took us from clay tablets to the printed books? We shall argue that there is more to it; in our perspective, digital texts are not an 'evolved' format of the same technology historically involved in producing texts – they are artifacts produced by a different technology altogether, in which the linguistic correspondence in the symbolic system is established by means of mathematical representation.

That which makes a digital text *digital*, then, is not simply in the technique for *holding the symbolic system*, but also in the technology for *building the correspondences between the symbols and the linguistic information*. This is unprecedented in the history of text production; in the previously

[3] This historical development provides room for complex debates, such as the relation between such technical transformations and the history of the cultural practices of writing and reading. Particularly, as CHARTIER 2001 points out, the transformation of cultural practices cannot be simply derived from the transformations in the formal or technical transformations; material and cultural transformations are two distinct lines in the history of reading – here we are strictly focusing some aspects of the material transformations.

known transformations in text production, new techniques were involved mainly to perfect the ways of *carrying about* codified information. With traditional texts (be them constructed as engravings on a piece of clay or on a stone, or paintings on a piece of paper – i.e., whatever the technique is for holding the symbols), the command of a particular writing system (and a particular language, of course) sufficed for the writers to write and for the readers to read the text – i.e., for the information thereby contained to be processed. With digital texts, something else is needed for that beyond the receivers' and the producers' command of a writing system for the information to be processed. Human writers and readers do not process (codify and de-codify) the information immediately: they need an artificial logic programming to mediate this job.

Let us illustrate this with the level of character strings. In writing a text by hand, the producer ultimately traces characters of a writing system onto the support matter (e.g. paper) with an instrument (e.g. pen) and apparent matter (e.g. ink); they codify the character immediately; and the character is de-codified immediately into conceptual information by the reader, through visual contact:

{traced graphic symbol A = viewed graphic symbol A}.

This process can be aided by more sophisticated instruments, such as typewriters. In this case, instead of being *traced* by the hand of the producer, the characters are *stamped* on the paper by handles in a mechanic apparatus (the typewriter). Notice however that *information* is still codified directly by the producer, and de-codified directly by the receiver, just as with hand-drawn characters. So here we'd have the sequence:

{stamped graphic symbol A = viewed graphic symbol A}.

How does this compare to “writing” a text on a computer? At first sight, the process appears to be similar to that of writing texts with a typewriter: the producer presses keys with characters on them, and then the characters appear on a screen. But the similarity is of course illusory. With typewriters, between the action of pressing the key and the appearance of a correspondent character on paper, there is a *mechanical process* (i.e., the key activates a handle that raises a stamp; the stamp prints the character on the page thanks to ink stored in a tape). With computers, on the way between our pressing of the key and the appearance of the character on the screen, there lies not a mechanical process, but a mathematical process, in which information is codified and de-codified, so that mathematical relations stored in the form of electronic pulses are transformed-retransformed into graphic (human readable) symbols. So in this case, instead of **{traced/stamped graphic symbol A = viewed graphic symbol A}*,*** we’d have a sequence somewhat like:

{command x activated by keys [shift +‘A’] > code &#0065 > viewed graphic symbol A}.

Notice that now, between the ultimate action by the part of the producer (*pressing a key*) and the ‘final product’ to be processed by the reader (*viewed graphic symbol*) there is an intermediate stage (*code &#0065*), in which commands activated by the computer keyboard are dealt with by mathematical programming which needs to follow conventions for character codification.

This brief comparison of computers and typewriters as “writing instruments” reveals that the difference between hand-writing and typewriting is incremental: the basic technology is the same, ie, *graphic symbols are being deposited onto support matter*; in typewriting, essentially, the technique of tracing is substituted by the technique of stamping. So the evolution between hand-writing and typewriting may be termed as an incremental stage in a gradual evolution of one technology, the technology of imprinting graphic symbols onto some material. In contrast, the difference between this technology and digital text processing cannot be termed as an incremental stage in a technical evolution. It is not the simply the case that we have the aid of a new instrument that improves the existing technology for imprinting graphic symbols, but rather, a new technology for text production and diffusion. The technology of text production is the technology of turning a system of symbols readable. In the case of non-digital texts, this has meant simply, turning the symbols *visible*; the different techniques for 'carrying' information in a non-digital text (engraving, painting, or pressing symbols onto some material) are derived from the visual essence of the codification system. In the case of digital texts, the means of codifying the information is not simply visual (and conducted inside human minds), but includes digital representation (conducted artificially, in electronic media); the technology, here, consists in making those artificial representations humanly readable. It is then not a technical difference – it is another technology, i.e., another combination of codifying and carrying away the code.

It is therefore in the combination of information codification and means of diffusion that we may observe the material singularity of digital versus other forms of text.

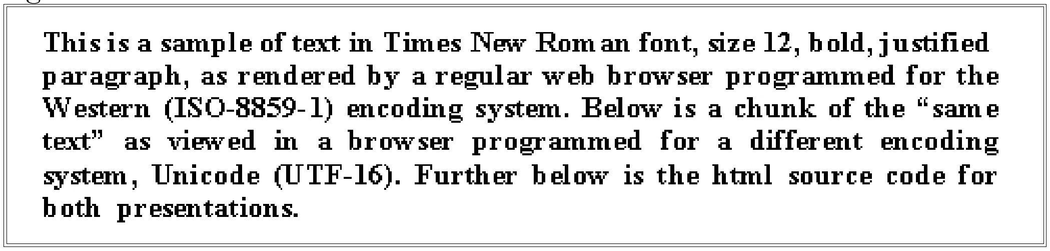
1.2 The artifact and the processes

At this point we have to face the somewhat staggering task of defining what, after all, a digital text actually *is*. We could start by vaguely determining that they are texts in which the use of digital technologies (mathematic representation) is, at some level, involved. At the process level, we have seen a little of what mathematic representation does – it intermediates the correspondence between symbols that humans may produce and receive via visual contact.

But a text is not only a process; it is also an artifact – what kind of an artifact is a digital text? As artifacts, digital texts are nothing but *mathematically encoded information*. This encoded information is presented in the form of humanly readable “writing”; but it is not actually *writing*, only sets of codes programmed to appear as writing.

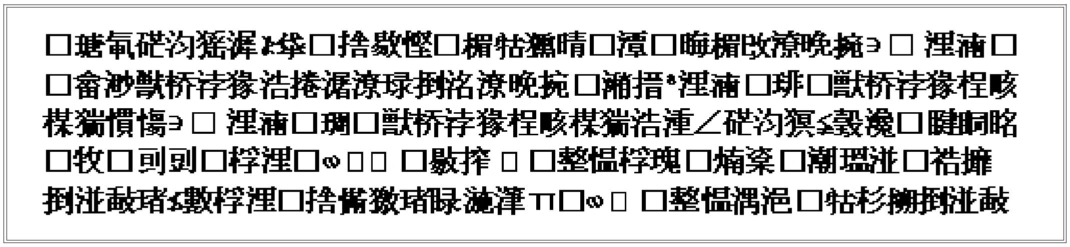
This can be illustrated by comparing the figures below; *Figure 1* shows a sample text produced digitally, as it would be read in a browser with Western (ISO-8859-1) encoding system; *Figure 2* shows the “same text” as it would be viewed in a browser programmed for a different encoding system, Unicode (UTF-16); *Figure 3* further below shows the source code that corresponds both to Figures 1 and 2.

Figure 1



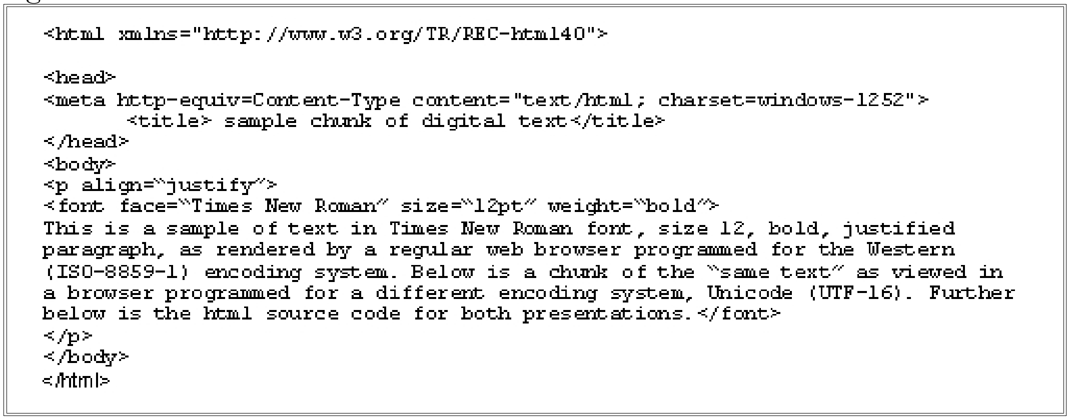
This is a sample of text in Times New Rom an font, size 12, bold, justified paragraph, as rendered by a regular web browser programmed for the Western (ISO-8859-1) encoding system. Below is a chunk of the “sam e text” as viewed in a browser programmed for a different encoding system, Unicode (UTF-16). Further below is the html source code for both presentations.

Figure 2



□塘氧砣沟猺湃ⅹ绦□捨敹慳□楣牯瀱晴□潭□畤楣攺潦晚捥∍ □ 浬湔□
□畬渺敹栣渺捨捲潦湣潳爾捨晚捥□湡摨*浬湔□琲□敹栣渺敹程畢
楳獨慣慳∍ □ 浬湔□瑒□敹栣渺敹程畢楳獨湡湶∠砣沟猺ⅹ穀湶□湭桪瑮
□敹□刵刾□桴浬□∞□□ □敹捾 □ □整慻桴瑌□湧渼□湮瑵湴□湮攮
敹湴敷畭ⅹ獸桴浬□捨湩敹桴畭湳湧ㅠ□∞□ □整慻湤瑭□牯杉湨敹湴湴

Figure 3



```
<html xmlns="http://www.w3.org/TR/REC-html40">

<head>
<meta http-equiv=Content-Type content="text/html; charset=windows-1252">
       <title> sample chunk of digital text</title>
</head>
<body>
<p align=“justify”>
<font face=“Times New Roman” size=“12pt” weight=“bold”>
This is a sample of text in Times New Roman font, size 12, bold, justified
paragraph, as rendered by a regular web browser programmed for the Western
(ISO-8859-1) encoding system. Below is a chunk of the “same text” as viewed in
a browser programmed for a different encoding system, Unicode (UTF-16). Further
below is the html source code for both presentations.</font>
</p>
</body>
</html>
```

If we regard the source code as the core material defining "*text*", then Figure 1 and Figure 2 have to be called "*the same text*"; yet how could they be "*the same text*", if one same reader with a command of the western alphabet might perfectly read the writing represented in Figure 1, but not the writing represented in Figure 2? If we consider that the text is the presentation (i.e., the humanly readable form rendered by a program working on the source code), then clearly Figures 1 and 2 are two "*different texts*"; yet how could they be "*different texts*", if materially they have been rendered by the same source code? From the material point of view, what should be conceived as the artifact *text* – the source code (Figure 3) or the potential presentations rendered by a browser (Figures 1 and 2)?

Because the source code is what materially constitutes the text, it cannot be dissociated from the concept "*text*"; and because the presentation is what we as humans can process, it cannot be dissociated from it either. Actually, then, in their final form (i.e., as a product), digital texts are *a layered combination* of mathematically encoded information and humanly readable presentation; it is this combination that we perceive as *the text.*.

The definition for digital text, then, includes the double dimension of process and product; this is important in trying to define a typology of digital texts. If we limit the definition to the dimension "product", "*digital texts*" would be texts whose <u>final form</u> is digital, i.e., actuated in an electronic environment (which relies on numerical representations). In this sense a text which is not received in the electronic environment would not be considered a "*digital text*", even in those cases when it might have been produced within the electronic environment (such as a printed form of a text that has been processed in a computer)[4]. Conversely, in this case, the term could, ultimately, be used to refer to a text that is received in the electronic environment, even though it has not been originally produced in the electronic

[4] There is a parallel duplicity of terms in other forms of text production: for example, "Printed Text" is either used to describe the object's final form, or to its production process. Evidently, many texts that we call "printed" have been produced in handwriting process and subsequently printed.

environment (such as a digital photograph of a hand-written text). If we expand to include the dimension "process", however, the term "*digital*" would be applied to texts that have been *processed* in an electronic device. This includes partial processing, as in texts that have been produced digitally, but received in other forms; and global processing, as in texts for which the whole transmission chain – production (writing) through reception (reading) – is actualized in an electronic environment[5]. For our purposes here this seems to be a more adequate approach; we shall consider then as *digital* those texts in whose construction language has been *processed* in the electronic environment (whether or not the final form of the text is received electronically).

It is important to realize, then, that even texts whose "final form" are not digital, but that have been processed digitally, may be included in our description of "mathematically encoded information presented in the form of humanly readable writing". Take a printed issue of a digitally produced document; it is presented in the printed form, but it has been processed digitally up to the moment in which a printer (via logical programming) managed to deposit the ink onto the paper. It is now an artifact "printed text", but some of the issues that pertain digital text diffusion will apply to its analyzes as well.

We have then, tentatively, defined digital texts as artificially encoded linguistic material rendered humanly readable with the intermediation of logical programming. This allows us to briefly explore the functionings of digital text diffusion, which consists, essencially, in multi-layered copying of a source code.

1.2 Digital texts diffusion

In the process of reaching out in time and space, texts may be altered and transformed – a central fact for textual studies and textual criticism, the art and science of retrieving back

[5] Less clearly so in the reverse case – a digital copy of a text produced in other medium. In a processual sense, digital copies produced by transcribing the original text would be classified as "digital text" (because the transcription is the processing); however, digital copies produced as images would not (because there is no processing of text in the electronic medium, only the production of an image of a text).

alterations and understanding how the elements of written culture are transmitted. Different factors are involved in this transformation: factors pertaining material deterioration, and factors pertaining the process of transmission itself. Traditionally, the potential factors that can impair text integrity are classified into endogenous (internal, i.e., material decay of support matter and apparent matter) and exogenous (external, i.e., caused by interferences in the chain of transmission).

The potential factors of endogenous modification of digital texts could be located on hardware decay and software decay[6] – while for the studies of manuscript or printed texts diffusion, paper and ink decay are the main endogenous factors impairing text integrity. This rather obvious difference already indicates that digital texts deserve a singular approach in studies of diffusion.

However, it is when we consider the factors pertaining the transmission process (i.e., exogenous) that the singularity of digital texts for such studies become more interesting. For traditional textual studies, *copy errors* are the major source of exogenous interference along the diffusion chain (cf. BLECUA, 1983 among others). For digital texts, the process of "*copy*" acquires an expanded significance; digital texts are always diffused as copies, but not copies produced by humans – rather, copies produced by programmed machines.

Take the "texts" that we read on the web; for each sample hypertext we may access, there is, naturally, a corresponding source code. In a regular situation of remote digital text diffusion, what the writer "writes", ultimately, is the source code (even if they are not aware of this fact, as we shall comment further on); but what the reader "*reads*" is the presentation – the code as translated by the programming. Consider, further, that a text which is received in the

[6] We consider informatic decay as an endogenous (i.e., internal, non-human factor), because it may take effect independently from any external action – it is not necessarily a transmission failure; bugs in text processing softwares, software obsolescence, etc., may affect a text that has been quietly stored in a machine without anyone ever handling it.

electronic environment such as the world computer web may be "*read*" by several people at the same time. However: that which is stored in the source computer is a document containing the encoded information (the source code); and that which is received by each of the simultaneous readers are multiple renderings of this source code by multiple browsers (i.e., processors) – actually, multiple *copies* of this rendering. What the receiver receives is a presentation of the source code as rendered by a translating machine; also in local access this is so, be it on a screen, or via a printing program.

In any case, remote or local, digital text diffusion functions by means of copying of multiple layers of source code and presentation. This brings us back to to the previous statement that for digital texts, the process of "*copy*" acquires an expanded significance for the study of text diffusion.

Traditionally, interference in the transmission chain is mainly studied as regards manuscript cultures – reproduced by means of human copy, rather than mechanic duplication. In copying, humans may alter the text (among others) by logical influence – copying is an active process to which mobility is inherent. In the diffusion of digital texts, copy is also a stage which involves logical processes. It is important to stress this out, as one could argue that printed texts are also diffused as copies – as in multiple reproductions of an original document. But this is an entirely different process altogether: a mechanical press will mechanically reproduce a source document with no logical process involved; a computer will reproduce a source document by means of logical processing. Text reproduction in the digital environment differ from human copying in that the logical processes involved are artificial; but they differ from to mechanic reproduction in that there is a logic process involved at all.

This makes digital diffusion more akin to human copying than to mechanical reproduction – an interesting state of affairs for the study of copy errors and '*mobility*', a term used, traditionally, to describe the tendency towards transformation by copying in manuscript writing cultures. “*Copy errors*” in the logical stage of digital text processing are a major source of problems for the integrity of digital texts. Such ‘errors’ may be better defined as interferences in the logical programming at some point into the transmission chain; problems pertaining the encoding, or problems pertaining the rendering of the presentations by subsequent programming.

This could again be exemplified with character processing: a typical diffusion problem, at the producer’s end, would be those situations in which the producer cannot find the proper key combination for diacritics (such as ´, ^, ~, **ç** etc.) in a given computer keyboard; and at the receivers’ end, the difficulty of reading web documents with character encoding problems, in which a sequence such as ‘*diacrítico*’ reads ‘*diacr*茅*tico*’. Mediation of artificial processing occurs in other levels of digital text processing, such as spatial organization. One rather ironical example is the separation of digital texts into “*pages*”. A “*page*” is, naturally, a spatial unit closely connected to paper as support material – there are obviously no such things as ‘pages’ on a computer screen. Nevertheless, most available text processing applications make up a visually recognizable space similar to a “*page*”, so that we can write comfortably and so that we can preview the results of printing a document. Obviously, though, the “*page*” that appears on the screen is nothing but a visual representation fabricated by codes of which the text producers and receivers are usually oblivious – unless, thanks to some programming inadequacy, this representation fails to work, and at the receiving end, the copy of a digital document appears with the wrong ‘page-breaking’ (a situation only too familiar to anyone who has tried to share the edition of a document between different machines or softwares). This goes to show that the intermediate stage of information codification via digital programming represents fertile ground for potential loss of information by diffusion.

For textual studies to be able to bring digital texts into their horizon as objects, the stage of mathematical codification and de-codification must be included as an area of investigation. At the same time, this expanding of horizons towards the inclusion of logical programming bears on those disciplines to which texts and their diffusion are an instrument of investigation; digital processing constitute breakthrough tools, in the form of scholarly editions that take full advantage of controlled encoding.

2. Digital text as a method

2.1 Transparency and control in information codification

We have argued that the mediation of artificial language in the codification of information singles out digital texts as watershed in the history of text diffusion. Interestingly, this intermediate stage– mediation of artificial processing – is one to which the producers and the readers are usually oblivious. It is mostly when the transmission chain is truncated that the average producer/receiver may become aware of the intermediation process. For instance, the brute fact of character encoding is normally only perceived by the receivers and producers when a link in the transmission chain presents malfunction (i.e., encoding inadequacies or presentation inadequacies); the same is true for other information codified in digital documents, such as spatial organization.

Intermediation in the codification of information, a fundamental stage in the production chain of any kind of digital text, is not always evident – or 'transparent', as we shall term it; rather, current forms of digital texts processing can vary quite a lot in how 'transparent' the intermediation of information codification is. We shall now see that the level up to which human writers and human readers can understand and control the processes of text codification makes all the difference for the potentialities of digital text as an object of study, and as a methodological instrument.

The least transparent processes seem to be associated with *partially* digitally processed formats (documents that are *produced* in digital environments, but not necessarily meant to be *received* in digital environments; typically, to be stored in a computer and eventually printed). This includes most of today's texts formats that can be produced in text processing applications[7]. These formats are typically processor-bound, that is, the formats depend on the text processing application used to write/read them[8]. They also share the contingency of proximity to paper formats, as most current text processors go a long way towards adapting digital processing to non-digital processing, for comfort in production and reception.

The wealth of sophisticated encoding possibilities embedded in such programs makes digital processing quite intuitive and straightforward for any literate person, which is certainly a convenient state of affairs. But as an accessory consequence, this sophistication of programming embedded in "intuitive" applications results in increasingly complex intermediation processes, upon which the text producers and receivers have little or no control. This has impacts in studies of text diffusion and mobility, and in other fields for which text processing and text edition is a crucial scientific tool.

Take, for example, linguistic studies to which the exact form of text structure can be a crucial factor in the investigation – such as historical linguistics studies based on texts. It is quite unusual nowadays for such studies to form paper databases; rather, they make use of digitally processed editions. Moreover, they need texts edited in a specialized format – philological editions, in which for example original orthography and text organization is kept

[7] For example: DOC files ("*Document*", texts processed by Microsoft Word), ODT files ("*Open Document Text*", processed by Sun Open Office), PDF files ("*Portable Document Files*", processed by Adobe Writer/Reader).

[8] The code used in such formats can be open (Sun .ODT) or closed (Microsoft .DOC, Adobe .PDF); open codes mean that the code can be studied and manipulated by programmers. However, our main point remains: even in open code cases, in such formats the codification is not meant to be manipulated by the regular users.

faithfully close to the originals. When such editions are produced in regular text processors, the editors are not in a position to fully control the codification of crucial information such as correct graphic symbols for characters, faithful reproduction of spatial organization, etc.; in regular text processors, as mentioned, those elements are codified by increasingly complex and opaque intermediation processes. Specialized editions cannot afford to let formatting and information organization elements to be modified by programming stages in the course of diffusion; and in order to ensure some control over the encoding of those elements, they need to turn to more transparent text processing.

Relatively more transparent processes of information codification are associated with *globally* digitally processed formats (ie., produced in digital environments and meant to be received in digital environments). This group is formed mainly by the Hypertext family, including formats such as HTML ("*HyperText Markup Language*"), XML ("*eXtensible Markup Language*") and XHTML ("*eXtensible HyperText Markup Language*"). Notice, crucially, that in this case the formats are not processor-bound. The classification of the formats and the differences between them refer to the *language* used for marking up the texts for subsequent processing; these languages are not processor-dependent[9], rather they are regulated by an international consortium, the W3C – cf. (W3C 1997a) for the regulation on HTML, and (W3C 1997b) for the regulation of XML. It makes sense for Hypertext to be a processor-independent format; it has been conceived for global digital processing, that is, to be produced and received in the electronic environment of the World Wide Web; it must be remotely readable by (any) machines via a browser (the de-codifier).

[9] There are, of course, a number of available applications, commercial or otherwise, that can handle hypertext formats; the important point here is that hypertext can be processed independently of a particular processor application.

For our purposes here this means, essentially, that codification in hypertext can be made transparent to the producer, and to the receiver. More than that: in hypertext construction the codification of information can be manipulated by the producer within a scheme that is controlled and normatively regulated. This is in fact the very spirit of hypertext markup language: it permits a human text constructor to control the levels of information to be handled by the intermediate stage of mathematical programming. Text markup languages within the spirit of hypertext can also be quite flexible– especially so with the extensible languages, such as *Extensible Markup Language* or XML (W3C, 1997a), which predicts a properly structured syntax, but an open semantics. This means that in preparing a text to be processed, the editor can very much markup any level of information that they regard as relevant. As we shall see below, this possibility has been turned to the advantage of several initiatives in specialized text editing.

2.2 Digital text in scholarly editions

The instrumentation of text encoding for the production of scholarly editions has turned out as one of the most interesting frontiers of digital text production in recent years. Digital processing, and automatic, controlled codification, has been turned to the advantage of specialized edition processes with various aims; for general textual and literary studies, the wealth of electronic editions available nowadays is, in itself, an evidence of the role of digital texts as instances in the preservation of written tradition[10]. In the specific field of linguistic analysis, digital annotation has been applied to encode different levels into the texts, from graphic organization to morphology or syntax, providing research with large databases of

[10] We have in mind initiatives such as the Oxford Text Archive (http://ota.ahds.ac.uk/), the Oxford Digital Library (www.odl.ox.ac.uk), Biblioteca Virtual Miguel de Cervantes (http://www.cervantesvirtual.com/index.jsp), Biblioteca Nacional de Lisboa Digital (http://bnd.bn.pt/), Projeto Vercial (http://alfarrabio.di.uminho.pt/vercial/index.html), Victorian Web (http://www.victorianweb.org), to cite only a few of the most influential current text archives.

unprecedented volumes of linguistic information, in the form of annotated corpora[11], most of whom follow standards regulated by consortia such as the XML Corpus Encoding Standard (http://www.xml-ces.org/).

Great part of the current annotation standards apply to marking up elements such as cataloguing information or 'metadata', text organization (paragraphs; sections), and linguistic levels (morpheme; word; sintagm); however, information regarding stages in previous text diffusion and editorial interference can also be marked into texts, with potentially interesting results for electronic scholarly editions. Recently, we have conducted an experiment towards the annotation of diffusion layers, within the Tycho Brahe Annotated Corpus of Historical Portuguese (TBACHP, 2006). The corpus is formed by 15th-19th century Portuguese texts, which have been marked up with XML for regular text-organization information and for the annotation of modernization of orthography[12]; the basic idea in this *Controlled Edition Technique* (TRIPPEL and PAIXÃO DE SOUSA, 2006) is to allow our own interferences as editors to be marked up and recovered in a text. Editorial interference is annotated within layers in one document, so that each layer may be produced separately as different presentations (for human readers and subsequent automatic tools), on demand[13]; in the presentation documents thus produced, each item that has been interfered with can be linked to its counterpart in the equivalent version; a glossary of editiors' interferences can also be produced[14].

[11] The several text markup projects in this spirit have produced annotated corpora such as the Corpus Diacrónico del Español (http://www.rae.es), the British National Corpus (http://www.natcorp.ox.ac.uk/), the American National Corpus (http://www.americannationalcorpus.org/), the Lácio-Web (http://www.nilc.icmc.usp.br/lacioweb/), and the several corpora united under web-based resource centers such as the Open Language Archive Community (http://www.language-archives.org/).

[12] Modernization of orthography is a necessary edition stage for those texts, as the final aim is for them to be processed by automatic programming which annotates morphology, and which cannot satisfactorily handle the variation in orthography found in Classical Portuguese writing.

[13] Via XSLT (Extended Stylesheet Transformation Language, cf. www.w3.org/TR/xslt) transformations.

[14] Cf. http://www.ime.usp.br/~tycho/participants/psousa/memorias/sample_1.html for an example of the encoding of a text with one edition layer in this system.

More recently, we have been exploring the annotation of some other specific text particularities, such as deterioration from the original support material of transcribed documents, such as ink blotting, corrosion, bug-holes, etc. If it is the case that there are other editions available for a transcribed text, the interferences of the previous editors are annotated as well, in overlapping layers into the interferences of the present editor[15]. The same technique can be applied to compare multiple previous editions of one text, by annotating each subsequent stage[16]; and ultimately, it could be applied for the comparison of subsequent stages of text elaboration by one author, making this an interesting tool for genetic critique. The fundamental spirit of the "*Critical Hyper-editions*" elaborated in this recent project (PAIXÃO DE SOUSA, 2006) is to capture, by digital annotation, the signs of text mobility in subsequent edition processes.

This brief account of some recent experiments in text annotation has intended to illustrate the potential role of digital processing as a tool for textual studies. Electronic editions with scholarly purposes – linguistic analysis, literary studies, textual studies – are "specialized" in two senses: they involve specialized knowledge of text editing and specialized knowledge of electronic processing of language.

This is not to say that specialized text editors who make use of electronic media will suddenly turn into computer scientists; it only means that artificial intelligence has to be included as a related area in textual studies, much in the same way that paleography or codicology have traditionally been so. Scholars have traditionally relied on their knowledge of texts as cultural objects *and as material artifacts* to produce specialized editions – a conjunction of horizons which includes cultural, literary, linguistic *and material* dimensions (such as a typology

[15] Cf. http://www.ime.usp.br/~tycho/participants/psousa/memorias/sample_2.html for an example of the encoding of a text with two edition layers in this system.
[16] Cf. http://www.ime.usp.br/~tycho/participants/psousa/memorias/sample_3.html for an example of the encoding of two different editions on text in this system.

of support matters, a typology of alphabets, a typology of transmission failures, etc.). The same conjunction between cultural and material knowledge is needed for editions carried out digitally – only here, the material dimension must include digital programming.

3. Final Remarks

Digital text is yet to reveal our historical time its full potential, both as an artifact to be technically developed and as an object to be conceptually explored. The conceptual exploration of digital texts requires some important challenges to be faced by some of the fields that have traditionally been dedicated to the study of text diffusion and its correlated issues. Theoretical approaches to digital texts in the material perspective must take into account mathematical programming as a stage in the chain of production – more evidently so, the study of 'mobility' in the diffusion process cannot be carried out without considering this fact. The challenge is then posed for philology and textual studies: their scope of action must expand to include this cycle of procedures. Quite simply, those fields must start to regard artificial intelligence as one of their related areas.

On the side of technical developments, it could be said that digital texts, as artifacts, are still constructed in a general way within much dependency on typically "*paper-bound*" concepts. We witness today a transition stage, similar maybe to that which took place with the advent of the mechanical press. As CHARTIER (2001) and EISENSTEIN (1998) have pointed out, the first books out of the first presses were simply printed transposition of texts as they would appear in manuscripts; it took some time until the technical paradigm of manuscript production gave way, and the full technical potentials of printing were explored. With time, digital text production may become relatively more independent as well, and the full technical potentials of digital text production may be explored in different directions.

One area in which this exploration is already in motion is the use of digital processing for scholarly editions, in experiments towards the exploration of the digital environment as an instrument for text studies in general in a way that could never be achieved with other text processing technologies. Controlled text annotation, in which any number of layers of information can be merged into one document, allows the several different aspects in a text to be captured and analyzed – graphic organization, content, linguistic structure (lexical, morphological, syntactic, semantic), and marks of diffusion stages – both vertically and horizontally, in parallel to other texts. This is very much what textual studies have been pursuing over the centuries (in their toil of notating slight differences of terms in the margins of manuscripts, tracing back footprints of previous editors, and drawing intricate interpretation signs for editors to come): the weaving of delicate tangled webs of correspondences between different versions of different documents. This interweaving of different dimensions or layers of information acquires enhanced technical possibilities as we start to work with digital processing, in the intersections between code and presentation. We can push the frontiers of *text* as "*a system of roots that can be excavated into itself*" (CARVALHO, 2003)[17] – and from this spiral continuum of information layers, renewed perspectives may come to flourish.

REFERENCES

BLECUA, Alberto. *Manual de crítica textual.* Madrid: Castalia, 1983 [1987].

CARVALHO, Rosa Borges Santos (2003). "A Filologia e seu Objeto: Diferentes perspectivas de estudo". *Philologus - Revista do Círculo Fluminense de Estudos Filológicos e Lingüísticos*, ano 9, n. 26, Rio de Janeiro. *http://www.filologia.org.br/revista/artigo/9(26)03.htm*

[17] In the original: "*Todas essas disciplinas [...] têm tomado o texto como um sistema de raízes que pode ser escavado nele próprio, ou seja, constrói-se teoria a partir da análise de seus componentes, daí afloram as diversas abordagens conforme os modelos teóricos e métodos adotados*" (CARVALHO, 2003).

CHARTIER, R (2001). *Cultura Escrita, Literatura e História*. Porto Alegre: Artmed.

EISENSTEIN, Elizabeth (1998). *A Revolução da cultura impressa*. São Paulo: Ática.

IDE, Nancy and ROMARY, Laurent (2000): "XML Support for Annotated Language Resources". *Linguistic Exploration: Workshop on Web-Based Language Documentation and Description. Dec 12 - Dec 15, 2000, University of Pennsylvania.*

LESK, Michael (1997). "Hypertext". *Practical Digital Libraries*. New York: Morgan Kaufmann.

PAIXÃO DE SOUSA, Maria Clara (2006). "Edições Críticas Eletrônicas: Fundamentos e Diretrizes". *www.ime.usp.br/~tycho/participants/psousa/memorias/critical_hyper/ece.html*

TRIPPEL, Thorsten and PAIXÃO DE SOUSA, Maria Clara (2006). "Building a historical corpus for Classical Portuguese: some technological aspects". *Papers from the V International Conference on Language Resources and Evaluation,* Genoa: LREC.

TBACHP (2006). "Tycho Brahe Annotated Corpus of Historical Portuguese". *<http://www.ime.usp.br/~tycho/corpus>*

W3C (1997a). "Extensible Markup Language". *http://www.w3.org/XML*.

W3C (1997b). "HyperText Markup Language".*http://www.w3.org/MarkUp*.