Communing Texts

A talk given on the second day of the conference Off the Press *held at WORM, Rotterdam, on May 23, 2014.*

I am going to talk about publishing in the humanities, including scanning culture, and its unrealised potentials online. For this I will treat the internet not only as a platform for storage and distribution but also as a medium with its own specific means for reading and writing, and consider the relevance of plain text and its various rendering formats, such as HTML, XML, markdown, wikitext and TeX.

One of the main reasons why books today are downloaded and bookmarked but hardly read is the fact that they may contain something relevant but they begin at the beginning and end at the end; or at least we are used to treat them in this way. E-book readers and browsers are equipped with fulltext search functionality but the search for "how does the internet change the way we read" doesn't yield anything interesting but the diversion of attention. Whilst there are dozens books written on this issue. When being insistent, one easily ends up with a folder with dozens other books, stucked with how to read them. There is a plethora of books online, yet there are indeed mostly machines reading them.

It is surely tempting to celebrate or to despise the age of artificial intelligence, flat ontology and narrowing down the differences between humans and machines, and to write books as if only for machines or return to the analogue, but we may as well look back and reconsider the beauty of simple linear reading of the age of print, not for nostalgia but for what we can learn from it.

This perspective implies treating texts in their context, and particularly in the way they commute, how they are brought in relations with one another, into a community, by the mere act of writing, through a technique that have developed over time into what we have came to call *referencing*. While in the early days referring to texts was practised simply as verbal description of a referred writing, over millenia it evolved into a technique with standardised practices and styles, and accordingly: it gained *precision*. This precision is however nothing machinic, since referring to particular passages in other texts instead of texts as wholes is an act of comradeship because it spares the reader time when locating the passage. It also makes apparent that it is through contexts that the web of printed books has been woven. But even though referencing in its precision has been meant to be very concrete, particularly the advent of the web made apparent that it is instead *virtual*. And for the reader, laborous to follow. The web has shown and taught us that a reference from one document to another can be plastic. To follow a reference from a printed book the reader has to stand up, walk down the street to a library, pick up the referred volume, flip through its pages until the referred one

is found and then follow the text until the passage most probably implied in the text is identified, while on the web the reader, *ideally*, merely moves her finger a few milimeters. To click or tap; the difference between the long way and the short way is obviously the hyperlink. Of course, in the absence of the short way, even scholars are used to follow the reference the long way only as an exception: there was established an unwritten rule to write for readers who are familiar with literature in the respective field (what in turn reproduces disciplinarity of the reader and writer), while in the case of unfamiliarity with referred passage the reader inducts its content by interpreting its interpretation of the writer. The beauty of reading across references was never fully realised. But now our question is, can we be so certain that this practice is still necessary today?

The web silently brought about a way to *implement* the plasticity of this pointing although it has not been realised as the legacy of referencing as we know it from print. Today, when linking a text and having a particular passage in mind, and even describing it in detail, the majority of links physically point merely to the beginning of the text. Hyperlinks are linking documents as wholes by default and the use of anchors in texts has been hardly thought of as a *requirement* to enable precise linking.

If we look at popular online journalism and its use of hyperlinks within the text body we may claim that rarely someone can afford to read all those linked articles, not even talking about hundreds of pages long reports and the like and if something is wrong, it would get corrected via comments anyway. On the internet, the writer is meant to be in more immediate feedback with the reader. But not always readers are keen to comment and not always they are allowed to. We may be easily driven to forget that quoting half of the sentence is never quoting a full sentence, and if there ought to be the entire quote, its source text in its whole length would need to be quoted. Think of the quote "information wants to be free," which is rarely quoted with its wider context taken into account. Even factoids, numbers, can be carbon-quoted but if taken out of the context their meaning can be shaped significantly. The reason for aversion to follow a reference may well be that we are usually pointed to begin reading another text from its beginning.

While this is exactly where the practices of linking as on the web and referencing as in scholarly work may benefit from one another. The question is *how* to bring them closer together.

An approach I am going to propose requires a conceptual leap to something we have not been taught.

For centuries, the primary format of the text has been the page, a vessel, a medium, a frame containing text embedded between straight, less or more explicit, horizontal and vertical borders. Even before the material of the page such as papyrus and paper appeared, the text was already contained in lines and columns, a structure

which we have learnt to perceive as a grid. The idea of the grid allows us to view text as being structured in lines and pages, that are in turn in hand if something is to be referred to. Pages are counted as the distance from the beginning of the book, and lines as the distance from the beginning of the page. It is not surprising because it is in accord with inherent quality of its material medium – a sheet of paper has a shape which in turn shapes a body of a text. This tradition goes as far as to the Ancient times and the bookroll in which we indeed find textual grids.

A crucial difference between print and digital is that text files such as HTML documents nor markdown documents nor database-driven texts did inherit this quality. Their containers are simply not structured into pages, precisely because of the nature of their materiality as media. Files are written on memory drives in scattered chunks, beginning at point A and ending at point B of a drive, continuing from C until D, and so on. Where does each of these chunks start is ultimately independent from what it contains.

Forensic archaeologists would confirm that when a portion of a text survives, in the case of ASCII documents it is not a page here and page there, or the first half of the book, but textual blocks from completely arbitrary places of the document.

This may sound unrelated to how we, humans, structure our writing in HTML documents, emails, Office documents, even computer code, but it is a reminder that we structure them for habitual (interfaces are rectangular) and cultural (human-readability) reasons rather then for a technical necessity that would stem from material properties of the medium. This distinction is apparent for example in HTML, XML, wikitext and TeX documents with their content being both stored

on the physical drive and treated when rendered for reading interfaces as single flow of text, and the same goes for other texts when treated with automatic line-break setting turned off. Because line-breaks and spaces and everything else is merely a number corresponding to a symbol in character set.

So how to address a section in this kind of document? An option offers itself – how computers do, or rather how we made them do it – as a position of the beginning of the section in the array, in one long line. It would mean to treat the text document not in its grid-like format but as line, which merely adapts to properties of its display when rendered. As it is nicely implied in the animated logo of this event and as we know it from EPUBs for example.

The general format of bibliographic record is:

Author. Title. Publisher. [Place.] Date. [Page.] URL.

In the case of 'reference-linking' we can refer to a passage by including the information about its beginning and length determined by the character position within the text (in analogy to *pp.* operator used for printed publications) as well as the text version information (in printed texts served by edition and date of publication). So what is common in printed text as the page information is here replaced by the character position range and version. Such a reference-link is more precise while addressing particular section of a particular version of a document regardless of how it is rendered on an interface.

It is a relatively simple idea and its implementation does not be seem to be very hard, although I wonder why it has not been implemented already. I discussed it with several people yesterday to find out there were indeed already attempts in this direction. Adam Hyde pointed me to a proposal for *fuzzy anchors* presented on the blog of the Hypothes.is initiative last year, which in order to overcome the need for versioning employs diff algorithms to locate the referred section, although it is too complicated to be explained in this setting.[1] Aaaarg has recently implemented in its PDF reader an option to generate URLs for a particular point in the scanned document which itself is a great improvement although it treats texts as images, thus being specific to a particular scan of a book, and generated links are not public URLs.

[1]Proposals for paragraph-based hyperlinking can be traced back to the work of Douglas Engelbart, and today there is a number of related ideas, some of which were implemented on a small scale: fuzzy anchoring, http://hypothes.is/blog/fuzzy-anchoring/; purple numbers, http://project.cim3.net/wiki/PMWX_White_Paper_2008; robust anchors, http://github.com/hypothesis/h/wiki/robust-anchors; *Emphasis*, http://open.blogs.nytimes.com/2011/01/11/emphasis-update-and-source; and others http://en.wikipedia.org/wiki/Fragment_identifier#Proposals. The dependence on structural elements such as paragraphs is one of their shortcoming making them not suitable for texts with longer paragraphs (e.g. Adorno's *Aesthetic Theory*), visual poetry or computer code; another is the requirement to store anchors along the text.

Using the character position in references requires an agreement on how to count. There are at least two options. One is to include all source code in positioning, which means measuring the distance from the anchor such as the beginning of the text, the beginning of the chapter, or the beginning of the paragraph. The second option is to make a distinction between operators and operands, and count only in operands. Here there are further options where to make the line between them. We can consider as operands only characters with phonetic properties – letters, numbers and symbols, stripping the text from operators that are there to shape sonic and visual rendering of the text such as whitespaces, commas, periods, HTML and markdown and other tags so that we are left with the body of the text to count in. This would mean to render operators unreferrable and count as in *scriptio continua*.

Scriptio continua is a very old example of the linear onedimensional treatment of the text. Let's look again at the bookroll with Plato's writing. Even though it is 'designed' into grids on a closer look it reveals the lack of any other structural elements – there are no spaces, commas, periods or line-breaks, the text is merely one flow, one long line.

Phaedrus was written in the fourth century BC (this copy comes from the second century AD). Word and paragraph separators were reintroduced much later, between the second and sixth century AD when rolls were gradually transcribed into codices that were bound as pages and numbered (a dramatic change in publishing comparable to digital changes today).[2]

'Reference-linking' has not been prominent in discussions about sharing books online and I only came to realise its significance during my preparations for this event. There is a tremendous amount of very old, recent and new texts online but we haven't done much in opening them up to contextual reading. In this there are publishers of all 'grounds' together.

We are equipped to treat the internet not only as repository and library but to take into account its potentials of reading that has been hiding in front of our very eyes. To expand the notion of hyperlink by taking into account techniques of referencing and to expand the notion of referencing by realising its plasticity which has always been imagined as if it is there. To mesh texts with public URLs to enable entaglement of referencing and hyperlinks. Here, open access gains its further relevance and importance.

Dušan Barok

Written May 21-23, 2014, in Vienna and Rotterdam. Revised May 28, 2014.

[2]Works which happened not to be of interest at the time ceased to be copied and mostly disappeared. On the book roll and its gradual replacement by the codex see William A. Johnson, "The Ancient Book", in *The Oxford Handbook of Papyrology*, ed. Roger S. Bagnall, Oxford, 2009, pp 256-281. http://google.com/books?id=6GRcLuc124oC&pg=PA256