Some thoughts on the importance of open source and open access for emerging digital scholarship

Stefan Gradmann, Berlin

1 Introduction

Both the open source and the open access movements have their roots in the 'hard' sciences rather than in the social Sciences and Humanities (SSH). They have been concerned, traditionally, with open access to source code for computational data processing and with open access to scientific information published as journal articles.

Still, the basic assumption of the present contribution is that there is a specific open source and open access agenda within the SSH and that this may affect these disciplines—once such an agenda is fully in place—in a way hardly conceivable in the 'hard' sciences.

However, understanding the full impact and potential of such approaches in the SSH requires reflection upon broader methodological issues. Two vectors or primary oppositions are of specific interest in this respect:

- the scholarly information continuum as a whole and its evolution from print based to electronic working paradigms and the revolutionary changes that can be foreseen as a consequence
- the specific difference of the SSH as opposed to the Science-Technology-Medicine (STM) culture of relating signifiers to significates and the specific impact of the digital revolution resulting from this specific difference.

Exploring these two vectors this contribution will try to indicate constituent elements of an 'open' agenda for the digital humanities.

2 Evolution of the scholarly information continuum from print to XML

As W. McCarty has put it, „Academic publishing is one part of a system of highly interdependent components. Change one component [...] and system-wide effects follow. Hence if we want to be practical we have to consider how to deal with the whole system.“[1] Thus, in order to understand the coming paradigm shifts it is useful to first consider the evolution of the print based scholarly information continuum which has been stable and basically unchanged for centuries. This continuum can be conceived as a circular work flow centered around basically monolithic and static printed information objects and is sketched in Figure 1 below:

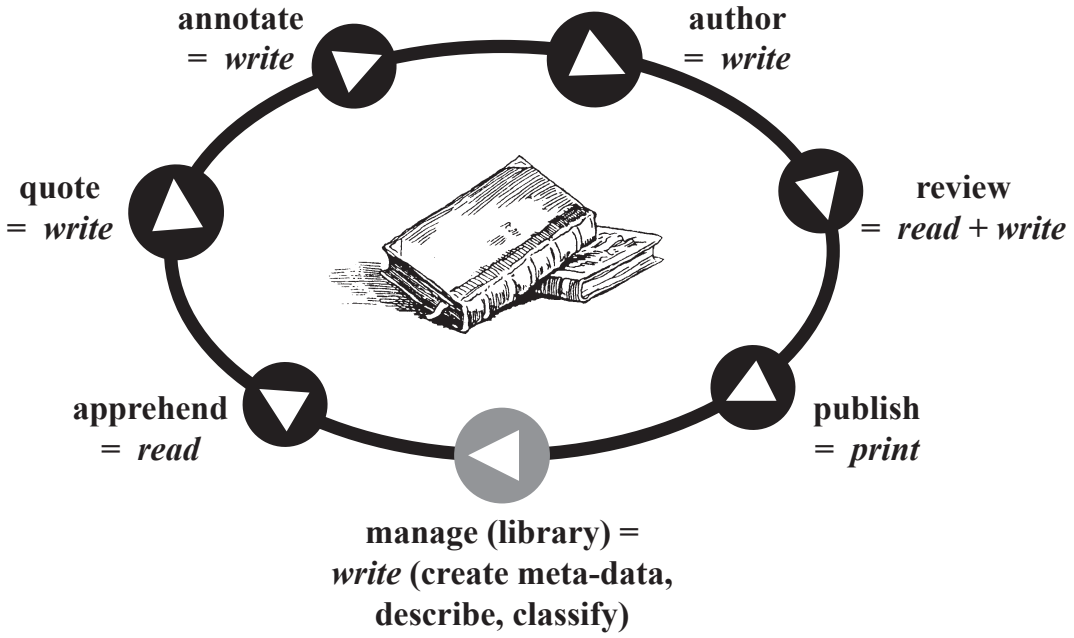

Figure 1: The traditional scholarly information continuum

In this traditional view of the scholarly information chain typical stages such as ‚authoring', ‚reviewing', ‚publishing', 'managing', ‚apprehension', ‚quotation' and ‚annotation' of scholarly information objects were implemented using very few and very stable cultural techniques (basically reading and writing). Furthermore, these stages were organized in linear, circular workflows with no or at most marginal modifications in sequence and centered around well understood, monolithic entities (documents).

With the advent of digital media and working instruments this functional sequence remained practically unchanged in a first phase, during which

the individual steps were simply electrified using digital means to emulate what had been done using traditional cultural techniques before as indicated in Figure 2 below:

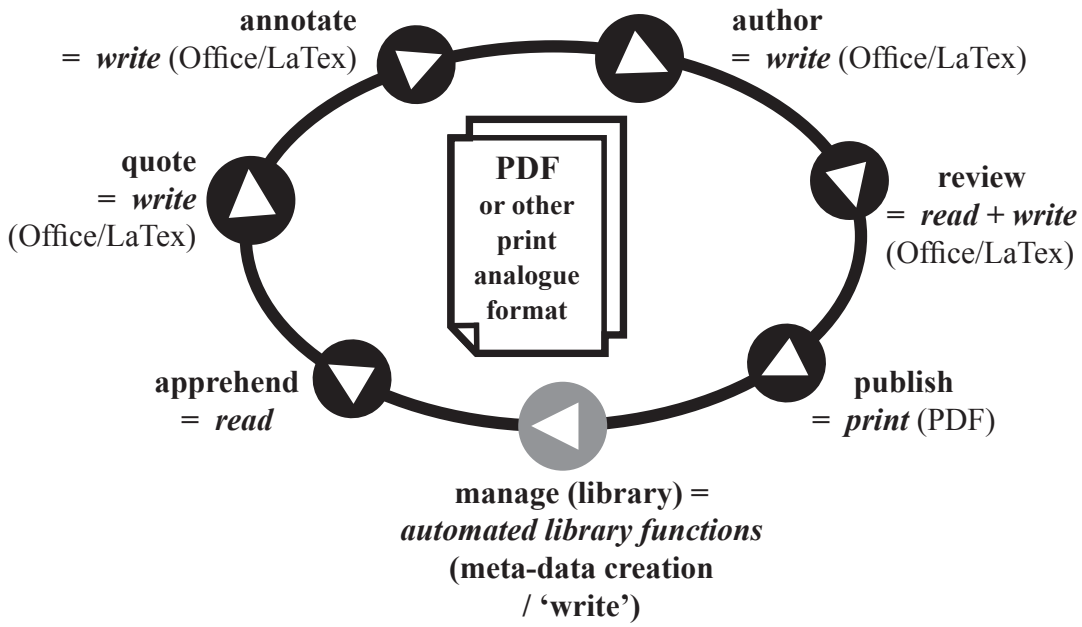

Figure 2: The traditional continuum in emulation mode

This scholarly value chain in emulation mode is somewhat similar to incunabulae in early print age: just as the latter have been preserving major characteristics of medieval folios the former kept (and partially still keeps) typical characteristics of the traditional value chain. Not only is the circular sequence preserved, but also its individual stages remain functionally unchanged and the use of well known cultural techniques remains constitutive. The same is true for the information object at the center of the circle which uses print-analogue formats such as PDF to emulate basic characteristics of the 'bookish' information support.

The first real qualitative change within this functional continuum happens with its transition to a third phase which is illustrated in Figure 3 below including some of the questions related to this process. In this third phase individual stages in the still basically unchanged linear function paradigm are now remodeled digitally and thus undergo substantial changes. Transition to this phase is currently under way and more or less advanced depending on the different scientific cultures.

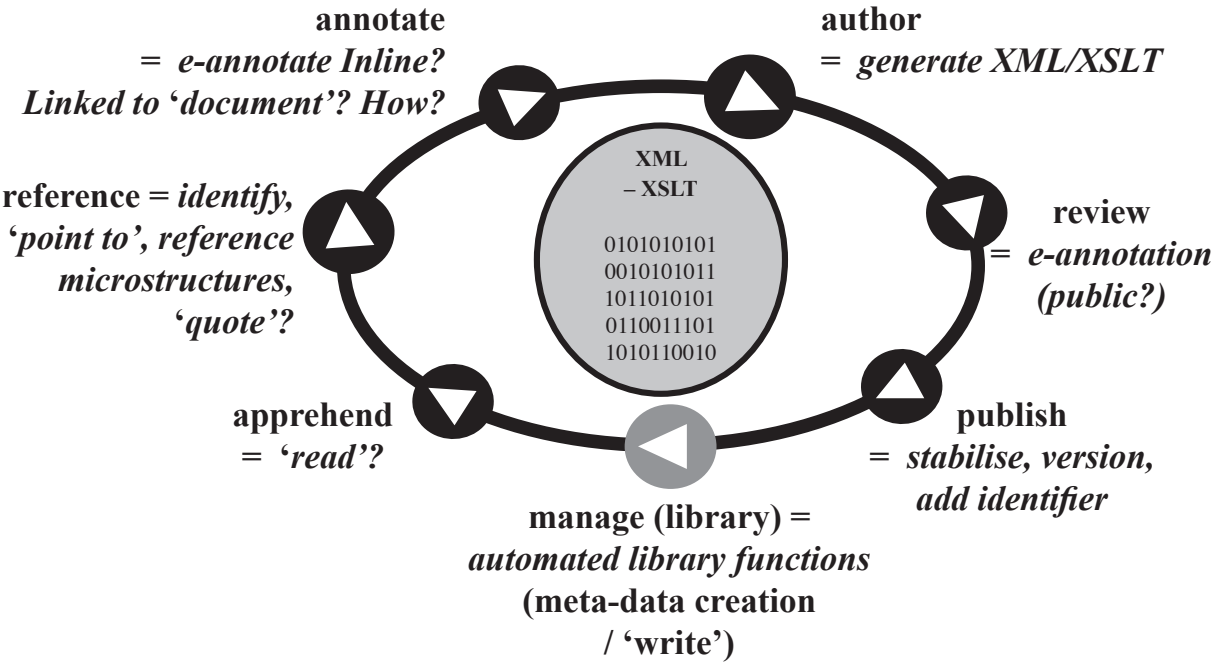

Figure 3: Scholarly information continuum ... going digital!

Authoring of scholarly documents, for instance, turns into generating of content using some XML syntax and appropriate presentation modes using XSLT or similar processing techniques. The reviewing stage turns into a more or less public and open procedure of digital annotation. ‚Publishing' in this context may be equivalent of stabilizing document content, applying version information and a unique identifier. ‚Quotation' instead of replicating parts of external documents more and more turns into identifying external information objects and referencing to its internal micro structure. It remains unclear, to which extent the term 'reading' can still be applied to the related acts of apprehension. And it becomes more and more evident that the 'library' metaphor is increasingly inappropriate for the fundamentally changed management methods for digital information objects.

Even if the formative power of traditional cultural techniques rapidly decreases within the individual stages as part of the transition from analogue to digital representation modes at different stages of the scholarly communication continuum, other basic characteristics of the traditional continuum remain unchanged in this stage: the scholarly value chain remains linear-circular and is focused around a seemingly still well understood monolithic information object, the 'document'.

However, these two remaining characteristics in turn may be subject to de-construction in a next phase that is already casting its shadows and which is likely to influence the continuum as indicated below in Figure 4:

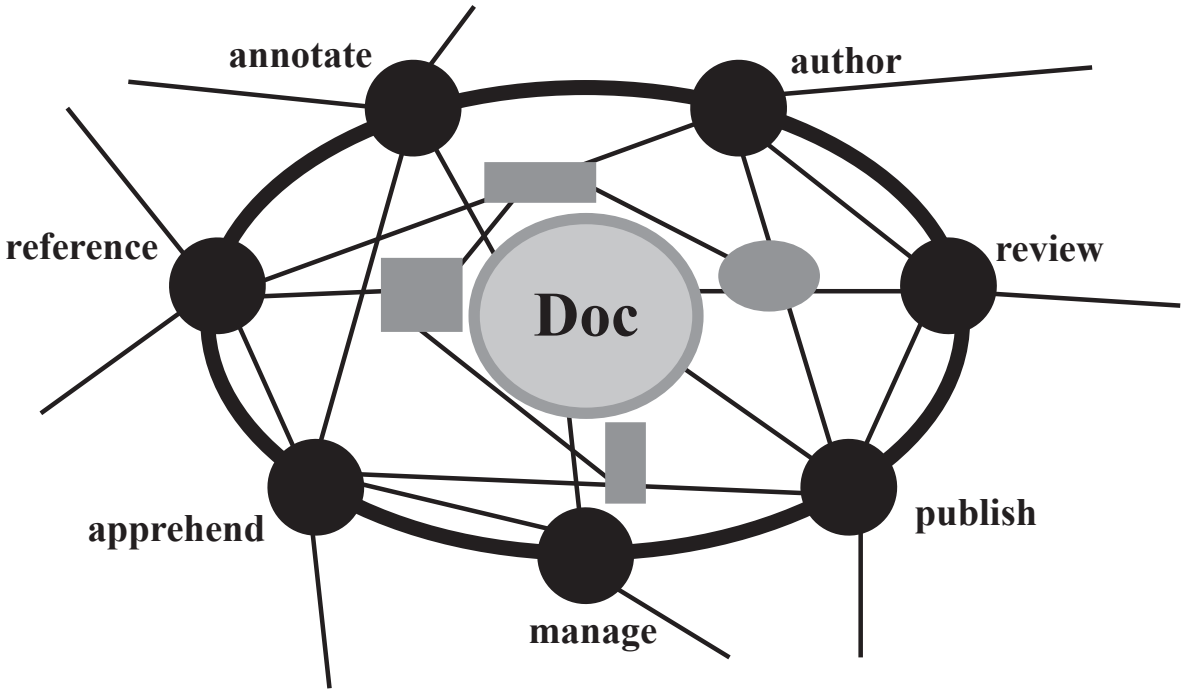

Figure 4: A de-construction of the scholarly information continuum

Two tendencies can already be outlined regarding this future phase: the stages that used to be organized in a sequential-circular will increasingly relate to each other in almost any networked order and the central information object, the 'document' looses its monolithic character, itself becomes a networked cluster of information entities with increasingly dynamic and diffuse borders.

We will thus be facing a triple paradigm shift but which has specific consequences with respect to the different scholarly / scientific cultures.

If one accepts—at least as a working hypothesis—the distinction established by C. P. Snow in his Rede lecture on "The Two Cultures" and considers the respective consequences of the triple paradigm shift for the sciences (henceforth STM) and the humanities (SSH) striking differences are almost evident.

In such a perspective, the *erosion of the linear/circular function paradigm* only marginally affects the way 'publication' is conceived in the SSH because of the prevalent 'monolithic' publication practice in this culture:

- journal articles and related peer reviewing procedures are still rather marginal
- authors still tend to work in 'splendid isolation' in the SSH with collaborative authoring still being an exception (such as the present contribution!).

The *declining formative power of traditional cultural techniques* certainly

affects the SSH (and probably much more than the sciences), but this does not specifically affect the publishing function.

However, the *de-construction of the 'document' notion in digital, networked settings* vitally affects the SSH in a very specific way. This process fundamentally changes the conditions of production and publication as well as the conditions of apprehension and reuse of scholarly documents. The consequences touch the very core of scholarly work which in both of its main strands of work is fundamentally concerned with documents both as objects and as instruments of scholarly activity. As shown in Figure 5 below, both the 'aggregation' (arrows pointing down) and the 'modeling' strands have their point of origin in digital corpora (and thus most of the time in document clusters) and produce new documents in turn!

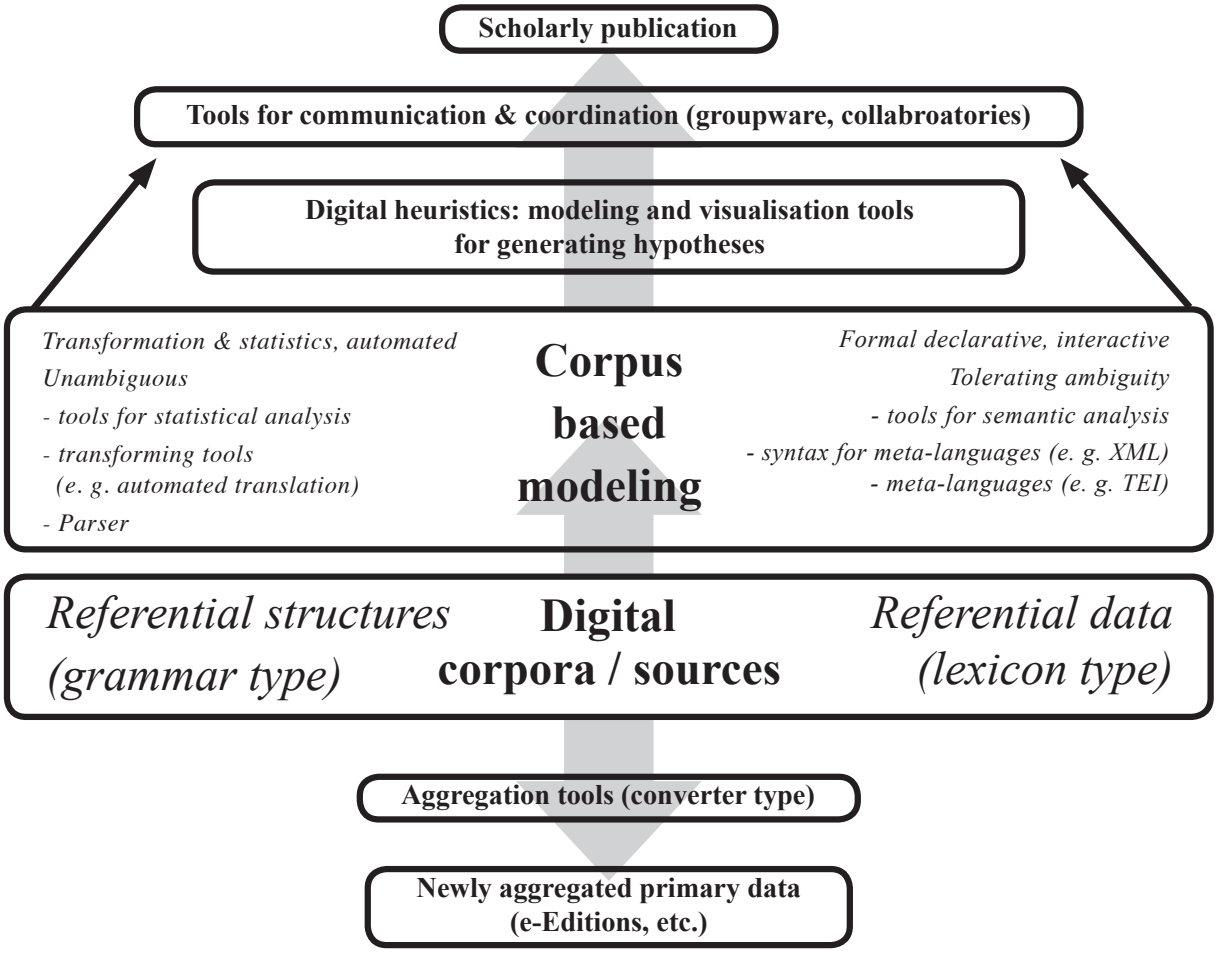

Figure 5: The two strands of scholarly work based on document corpora

And this observation organically leads to a closer investigation of the specific relation between the SSH (especially the hermeneutical rooted disciplines) and the constituent representation modes of documents as complex signs.

3 The pandora box of semiotics …

When considering this issue in more detail it becomes clear that signification and document modeling in all discussion related to electronic publishing up to now have basically been coined on the information model prevailing in the empirical sciences. In this model, scientific research as the core activity is completely dissociated from the publication process. Only once 'research' has yielded 'results' these in turn are 'packaged' in discourse and published (typically as a journal article): in this extremely robust and not very complex 'container' model of scientific publishing it is perfectly sufficient to remain on 'emulation level' as outlined above, since the publishing stage is not at the core of scientific work, anyway.

However, scholarly publishing in the SSH takes place in a substantially different information model: scholarly research and discoursive 'packaging' cannot be separated in this perspective and the published results of the core scholarly activity are again documents. This accordingly results in complex document models and publishing formats heavily intertwined with core research operations. In such a view, the 'container' models used in 'hard sciences' publishing are over-reductionist and inappropriate and complex relations between signifiers and significates are constitutive.

Clearly, behind the different information models underlying the respective publication cultures of the STM and the SSH another, even more fundamental semiological difference is hidden. In fact, dominant discourse in electronic STM publishing communities (mostly emanating from computer science) uses terms such as 'document', 'sign' or 'name' quite naively and without referring to their inherent semiological complexity. This results in a (technically) high level nominalist regression: the 'Pointer -> Object'-Model, in which 'words' are supposed to point to 'real' things as in Figure 6 below:

Figure 6: Words pointing to 'things'

The perfect incarnation of such a thinking are the 'ontologies' of the semantic web![2]

As opposed to this very simple mode of conceiving the relation between

words and things it is useful to consider the linguistic model of significance that has developed in the 20th century starting from DeSaussures theory of the sign and considerably refined by Hjelmslev, Eco and others[3] as indicated in the (much simplified) Figure 7 hereafter:

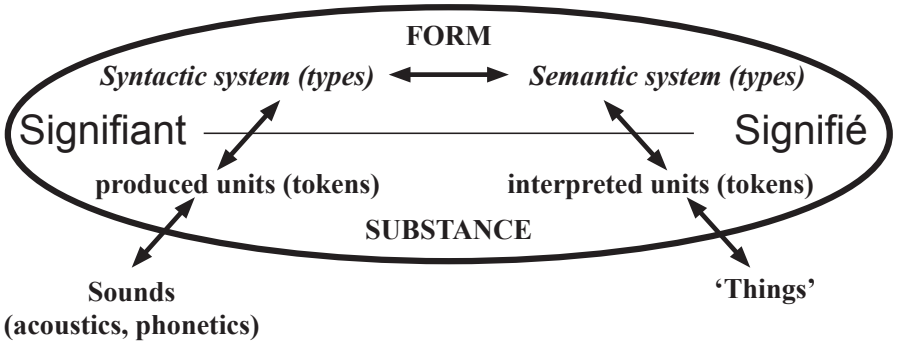

Figure 7: A simplified model of the semiological space

Signifiers and significates cannot be dissociated in this vision as it is impossible to consider form and substance of constituents independently: produced and interpreted individual units always have to be seen as part of they respective systemic context. And both sounds and real 'things' are *not* part of the representational space in such a view.

Such thinking has once been declared by a senior computer scientist as "opening the pandora box of semiotics"—but the fact is that exactly such thinking is required to understand the way the SSH relate to documents, which in turn must be conceived as complex significant units and themselves are part of a system made up of such units (vulgo 'litarature').

It then becomes clear that (electronic) text is not just a transcription of speech acts (parole) and at same time it must be noted that the notion of 'text' basically remains a blank spot in linguistics and still is subject to fundamental research as a complex, semiological digital object. In such an approach the model used above might tentatively translate to electronic documents as in Figure 8 below:

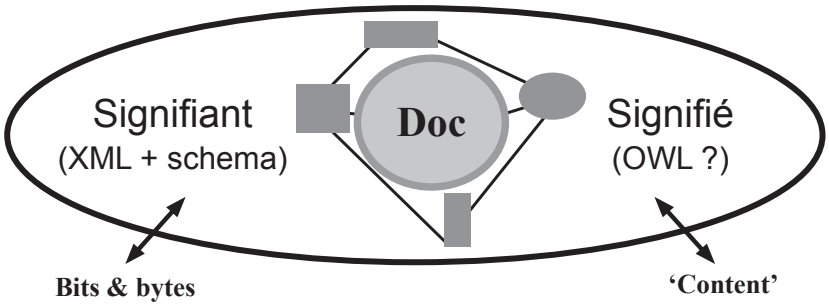

Figure 8: A tentative representational model for electronic documents

4 … and a way to re-think the ‘document’ notion

The heart of the issue thus seems to better understand the metamorphosis of the ‘document’ notion in the digital context—and a very competent attempt in this sense has been made by the French research group RTP-DOC (CNRS) that has used the pseudonym Roger T. Pédauque to publish fundamental work relating to the de-construction of the ‘document’ notion currently under way in the digital, networked context.[4]

RTP-DOC presents the evolution of the ‘document’ notion in the passage from printed to digital documents along three paradigms:

- Form (vu=’Look at’, morphosyntax), as material or non-material structured object, the corresponding chapter is *forme, signe et médium, les re-formulations du numérique*;
- Sign (lu=’read’, semantics), as meaningful instance and thus both intentional and part of a sign system, the corresponding chapter is *Le texte en jeu: permanence et transformations du document*;
- Medium (su=’Knowledge, Interpretation, Apprehension’, Pragmatics) as a vector of communication, part of a social reality with constituting temporal and spatial processes of mediation, the corresponding chapter is *Document et modernités*.

In each of the three conception paradigms one of the aspects is used as a dominant, yet non-exclusive vector for developing equations that distinguish traditional, electronic and future web-based document notions with each of these equation triples resulting in a definition of the respective nature of the ‘electronic document’.

Thus, the ‘form’ vector, in which object nature is constitutive, can be summed up in these three equations:

1. Traditional document = medium + inscription
2. Electronic document = structures + data
3. XML-document = structured data + stylesheet

And these in turn result in a first definition: “An electronic document is a data set organized in a stable structure associated with formatting rules to allow it to be read both by its designer and its readers”.

Likewise. The ‘sign’ focused on the meaningful nature of documents yields the following three equations:

1. Traditional document = inscription + meaning
2. Electronic document = informed text + knowledge
3. Semantic Web document = informed text + ontologies

And the resulting definition reads: “An electronic document is a text whose elements can potentially be analyzed by a knowledge system in view of its exploitation by a competent reader”.

Finally, the ‘medium’ vector organized around the ‘document’ as social phenomenon hast these three equations:

1. Traditional document = inscription + legitimacy
2. Electronic document = text + procedure
3. Web-Document = publication + measured usage/access

With the following definition associated: “An electronic document is a trace of social relations reconstructed by computer systems.”

Without referring more in detail the rich discussions within RTP-DOC it should be evident that the conceptual framework proposed by this group could serve as an excellent fundament for re-building consensus regarding the ‘document’ notion and for a better understanding of the nature of digital, networked document resources. Such an understanding in turn is required in order to better understand the specific impact of digital publication techniques in the SSH, as the ‘document’ notion is at the semiological heart of hermeneutical based scholarly work.

6 Our Cultural Commonwealth: the need of a triple 'open' agenda in the SSH

All the observations made in the preceding chapters converge in what one could call a triple 'open' agenda for the Social Sciences and the Humanities as it is already partly expressed in the report on cyberinfrastructure for the social sciences and humanities prepared for the American Council of Learned Societies Commission under the title "Our Cultural Commonwealth"[5].

6.1 Open and standardized document models

First of all and as should be clear from the above the effectiveness of the digital paradigm shift in the humanities vitally depends on open and non-proprietary techniques for document modelling and authoring. This is even more evident if one considers not just isolated documents, but webs of interrelated documents pointing and referring to each other. And this evidence gets particularly striking if one considers the need to maintain coherent webs of documents over time for decades or even centuries. In such a perspective introducing document protection technology such as for DRM in the audiovisual industry would create ridiculous and nightmarish functional scenarios!

6.2 Open sources …

Second, for innovative processing of digital sources to work at all a very specific understanding of the term 'open source' needs to be consequently and systematically applied: such scholarly work requires the free availability of all source material!

Hence the primary characteristic of cyberinfrastructure according to a statement made on source material made by the ACLS report: "It will be accessible as a public good".

6.3 … and open source processing instruments

Finally, the heuristics used for novel corpus modelling and aggregation work as well as their technical implementations and foundations need to be open source in the more traditional sense of the term as well as based on open standards to enable future digital hermeneutical heuristics. This is emphasised in recommendation 7 of the ACLS report stating: "Develop and maintain open standards and robust tools".

The author of this contribution is convinced that at least substantial progress in all three areas of this triple agenda of openness is required for genuine digital scholarship to happen at all!

Endnotes

[1] http://lists.village.virginia.edu/lists_archive/Humanist/v17/0336.html

[2] The following paper gives a very valuable discussion of the profound inappropriateness of positivist ontology based approaches in the SSH: Benel, Aurélien et al.: Truth in the Digital Library: From Ontological to Hermeneutical Systems. Proceedings of the fifth European Conference on Research and Advanced Technology for Digital Libraries (Lecture Notes in Computer Science, 2163). Heidelberg 2001, pp.366-377.

[3] Probably the best introduction to this semiological approach still are Eco, Umberto: La struttura assente. Milano 1968 and Eco, Umberto: A Theory of Semiotics. Bloomingrton 1976.

[4] Two publications are of interest here: Pédauque, Roger T.: Le document à la lumière du numérique. Toulouse 2006 and Pédauque, Roger T.: La redocumentarisation du monde. Toulouse 2007 as well as the web presence of the group at http://rtp-doc.enssib.fr.

[5] http://www.acls.org/cyberinfrastructure/