Beyond Embedded Markup

Buzzetti, Dino
dino.buzzetti@gmail.com
formerly University of Bologna, Italy

Thaller, Manfred
manfred.thaller@uni-koeln.de
University at Cologne, Germany

1. Introduction (Manfred Thaller)

Dino Buzzetti and

The unquestionable success of embedded markup methods and practice for the encoding of texts in the humanities is indisputably confirmed by the several projects and the authoritative collections of encoded texts now made available to the scholarly community.

The reasons of this success are many and diverse. An important one, however, consists in the influence that technological innovations have had on the accepted methodological principles of humanities computing. As John Unsworth has shown, we have been witnessing different phases of prevailing concerns in humanities computing projects and research: the chief orientation of interests has shifted from 'tools,' in the '50s, '60s, and '70s, to 'sources,' in the '80s and '90s, and now seems to be turning back from sources to tools (Unsworth 2004). From a computational point of view, what this change in orientation entailed was a shift of the attention focus from *processing* to *representation*, from developing algorithms applicable to information contents, to producing digital surrogates to replicate and visualise primary surce materials. The introduction of graphic interfaces, the development of WYSIWYG word processing systems on PCs, and the astounding explosion of the World Wide Web have undoubtedly favoured the expansion of this process.

The original purpose of the Web was actually to allow remote access to documents and to visualise them, and the languages developed to produce Web resources, HTML and now increasingly XML, are data representation languages and not data processing languages. Processing Web resources is heavily dependent on the structure these languages assign them, i.e. on hierarchical tree structures. XSLT, the language used to process XML data, 'takes a tree structure as its input, and generates another tree structure as its output' (Kay 2005). The point of view of the so-called 'document community' as opposed to that of the 'data processing' or 'database community' – i.e. 'to standardize the representation of data' *vs* 'to standardize the semantics of data' – was heavily influential on the decisions of the scholarly community, where 'attempts to define semantics [...] met with resistance,' and 'most notably' so in the Text Encoding Initiative. Thus, 'the route proposed by SGML,' and later XML, was to them 'a reasonable one' and embedded markup established itself as a standard for the encoding of texts in the humanities (Raymond 1996: 27-28).

However, from the original surmise that what text 'really is,' is nothing but an 'ordered hierarchy of content objects,' – the so-called OHCO thesis (De Rose et al. 1990) – the inadequacies of embedded markup have soon come to the fore, just for the sake of *representing* textual variation, not to mention *processing* textual information content. The need to overcome these difficulties has prompted several attempts to propose different approaches to text encoding. Among them we may mention the Layered Markup and Annotation Language (LMNL), (Piez s.d.), the eCommentary Machine web application (eComma), (Brown s.d.), the 'extended string' model, (Thaller 2006), and the Computer Aided Textual Markup and Analysis (CATMA) desktop application (CATMA s.d.).

One of the most efficient implementations of similar attempts to meet the difficulties of embedded markup is the proposal of 'standoff properties' as introduced by Desmond Schmidt. Properties can be assigned to given ranges of text that may nest and overlap, and can be stored separately from the text and in different formats suitable to specific needs. Standoff properties can be used to represent different textual features, for either interpretation or rendition purposes, and they can be organised in different sets, that can be easily merged, thus allowing for different encoding perspectives. The same technique is applicable to semantic annotation needs and stands as a viable and efficient alternative to it. But most importantly, as Desmond Schmidt has pointed out, whereas 'it is virtually impossible to freely exchange and interoperate with TEI-encoded texts,' with standoff properties 'interoperability and interchange are enhanced because the components are small and can be freely recombined and reused' (Schmidt, forthcoming). Moreover, the afforded flexibility of the standoff representation of property sets allows its 'textualisation': it can be expressed as a simple string of characters and its variations can be represented in the Multi-Version Document (MVD) format, i.e. as an oriented graph with a start-node and an end-node. Possible interpretative variants and encodings can then be easily compared and analysed.

In sum, approaches of this kind seem to be affording viable solutions to the challenge of putting to good use the invaluable wealth of data now made accessible and encoded, by 'building tools' that would enable us to proceed 'beyond

representation' (Unsworth 2004) and to process their information content. Standoff solutions can provide suitable means to deal with the different kinds of information conveyed by textual data structures and to assign them adequate data models for purposeful processing.

As it happens, however, the basic distinction between *data* and *information* is often overlooked, just as the clear severing of the two basic components of a text, its *expression* and *content*. This lack of distinction often leads to technical and conceptual shortcomings, sometimes intrinsic to the use of embedded markup. In this respect, standoff solutions can usefully complete and supplement embedded markup techniques with additional contrivances to distinguish between rendition and content features and to treat them appropriately. Specific case studies (see Buzzetti 2012, and Thaller 2012) can better exemplify the kind of problems that would require solutions that go beyond the mere representational scope of embedded markup and heed basic conceptual distinctions, such as those between data and information, or interpretation and rendition. To what extent these solution may also converge with the new technologies developed in the context of the Semantic Web would deserve a careful and more documented enquiry.

1.1. References

Brown, T. (s.d.). eComma: A commentary machine. http://ecomma.cwrl.utexas.edu/e392k/ (accessed 23 March 2012).

Buzzetti, D. (2012). Bringing Together Markup and Semantic Annotation. In this volume.

CATMA (s.d.). CATMA – Computer Aided Textual Markup & Analysis. http://www.catma.de/ (accessed 23 March 2012).

DeRose, St., et al. (1990). What Is Text, Really? *Journal of Computing in Higher Education* 1(2): 3-26.

Kay, M. (2005). What Kind of Language is XSLT? An analysis and overview. http://www.ibm.com/developerworks/library/x-xslt (accessed 23 March 2012).

Piez, W. (s.d.). LMNL Activity. http://www.piez.org/wendell/LMNL/lmnl-page.html (accessed 23 March 2012).

Schmidt, D. (forthcoming). Standoff Properties in HRIT.

Thaller, M. (2006). Strings, Texts and Meaning. In *Digital Humanities 2006: Conference Abstracts*. Paris: CATI – Université Paris-Sorbonne, pp. 212-214.

Thaller, M. (2012). What Is a Text within the Digital Humanities, or Some of Them, at Least? In this volume.

Unsworth, J. (2004). Forms of Attention: Digital humanities beyond representation. Paper delivered at *The Face of Text: Computer-Assisted Text Analysis in the Humanities*, the third conference of the Canadian Symposium on Text Analysis (CaSTA), McMaster University, November 19-21, 2004. http://people.lis.illinois.edu/~unsworth/FOA/ (accessed 13 March 2012).

2. What is a text within the Digital Humanities, or some of them, at least? (Manfred Thaller)

(i) The Humanities are a very broad field. The following ideas relate to those Humanities disciplines, which are dealing with 'historical texts' – or at least they started from them. 'Historical' in this context defines any text, which has been created by actors, which we cannot consult any more. This creates a complication when we understand an existing text as a message from a sender to a recipient – an understanding which is absolutely fundamental to modern information technology, as it is the model which has been used within Shannon's article of 1948, one of the corner stones of modern information theory and for most computer scientist, *the* corner stone of Computer Science upon which the later has been built. All of the measures Shannon proposes require an understanding, what the message that has been transmitted by the sender contained before transmission. Another important restriction Shannon acknowledges himself:

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.*

(Shannon 1948: 379)

The fact that information processing systems start with a model which ignores semantics from page one. is ultimately the reason, why meaning has to be added to the signal stream in ways, which allow the transmission (or processing) of that information as an integral part of the signal stream – today usually as embedded markup. Embedded into a signal stream, which has been created by a sender; so embedding anything into it would, according to the model of Shannon, require the markup being part of the transmitted message. This is indeed, what SGML has been created for: To enter the intentions of the

producer of a document about the formatting (and, up to a degree the meaning) of a data stream in such a way, that they would be independent of the requirements of specific devices.

When we are not able to check the meaning of a message with the sender we have to distinguish between the message, even if we do not understand it, and our assumptions about interpreting them. As we do not know the intent of the sender, the result of the 'transmission' of a historical text across time cannot be determined conclusively.

(ii) That data – as transmitted in signal streams – and information, as handled by humans, are not identical is a truism. They have long been seen as separate strata in information theory. (For a recent overview of the discussion see Rowley 2007.) A main difference between Shannon and the 'data – information – knowledge – wisdom' hierarchy has always been, that the former leads directly to an intuitive understanding of systems which can be realized by software engineering, while the later cannot. This is also true of attempts to use a similar scheme to understand information systems, notably Langefors (1995) *infological equation*.

(1) $I = i (D, S, t)$

Information (I) is understood here as the result of a process of interpretation (i) that is applied to data (D), applying previous knowledge (S) within the time available (t). The great attraction of this model is that – unlike Shannon's – it explicitly promises to model the meaning of messages, which are explicitly excluded from consideration by Shannon. To emphasize the difference between the models, we could say that Shannon assumes information to exist *statically*, therefore it can be broken into discrete units, independent of any process dealing with it, while Langefors understands information to be the result of a *dynamic* process, which, having a relationship to time, goes through different stages: So the amount of information existing at tn is not – or not necessarily – equal to the amount of information at tn-1, the ongoing process i having had the chance to produce more of it in the meantime.

The previous knowledge – S – can of course be easily seen as embodied in the interpreting scholar, who looks at the data. For the creation of systems of information processing Thaller (2009a: 228) has shown that Langefors original equation can be developed further. When we assume that knowledge is transformed from a static entity into a dynamic process, as Langefors has proposed for information, we can – via a few steps omitted in this abstract – reach

(2) $Ix = i (I_{x-\alpha}, s(I_{x-\beta}, t), t)$

Roughly: Information at point *x* is the result of the interpretation of an earlier level of information, in the light of knowledge generated from earlier knowledge, at a point of time *t*. As this allows the interpretation of data – e.g. a 'transmission' of a sender not living any more – as a process, which does not have to terminate, it is a better model for the handling of Humanities' texts as Shannon's.

(iii) This abstract model can be turned into an architecture for a representation of information, which can be processed by software. Thaller (2009b) has lead a project team within the digital preservation project PLANETS (cf. http://www.planets-project.eu/), which used this abstract model for the development of tools, which work on the comparison of the information contained within two different representations of an item according to two different technical formats. (Roughly: Does a PDF document contain exactly the same 'text' as a Word document.) For this purpose it is assumed, that all information represented in persistent form on a computer consists of a set of tokens carrying information, which exists within an *n*-dimensional interpretative space, each dimension of that space describing one 'meaning' to be assigned to it. Such a meaning can be a request directed at the rendering system processing the data to render a byte sequence in a specific way, or a connection to a semantic label empowering an information retrieval system. As such a representation is fully recursive, the requirements of formalism (2) above are fulfilled. For texts this can be simplified to an introductory example, where a text is seen as a chain of characters, each of which can be described by arbitrarily many *orthogonal* properties. (Whether the string *Biggin* within a text describes a person or an airfield is independent of whether that string is represented as italics or not; whether the string 'To be or not to be' is assigned to the speaker *Hamlet* is independent of whether it appears on page 13 or 367 of a book.)

(iv) Returning to the argument of section (i) we can see, that there is a direct correspondence between the two arguments. On the one hand the necessity to keep (a) the symbols transmitted within a 'message' from a sender who is irrevocably in the past and (b) our intellectual interpretations of them cleanly and unmistakably separate. On the other hand the necessity to distinguish clearly between (a) the tokens which transmit the data contained within a byte stream and (b) the technical information necessary to interpret that byte stream within a rendering system. If it is useful to transfer information transported within files with different formats into a representation, where the transmitted data are kept completely separate from the technical data needed to interpret them on a technical level, it is highly plausible, that that is even more the case,

when we are discussing interpretations of texts left to us by authors we can not consult any more.

This in turn is highly compatible to an architecture for virtual research environments for manuscript related work, where Humanities' work on historical texts is understood to consist of adding layers of changing and potentially conflicting interpretation unto a set of images of the manuscript to be interpreted. Ebner et al. (2011) have recently described an architecture for a virtual research environment for medieval manuscripts which implements this overall architecture, though using embedded markup for some of the layers for the time being.

To summarize the argument: (1) All texts, for which we cannot consult the producer, should be understood as a sequence of tokens, where we should keep the representation of the tokens and the representation of our interpretation thereof completely separate. (2) Such representations can be grounded in information theory. (3) These representations are useful as blueprints for software on highly divergent levels of abstraction.

2.1. References

Ebner, D., J. Graf, and M. Thaller (2011). A Virtual Research Environment for the Handling of Medieval Charters. Paper presented at the conference *Supporting Digital Humanities: Answering the unaskable*, Copenhagen, November 17-18, 2011 (forthcoming).

Langefors, B. (1995). *Essays on Infology*. Lund: Studentliteratur.

Rowley, J. (2007). The Wisdom Hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science* 33(2): 163-180.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* 27(3)**:** 379-423 and 27(4): 623-656.

Thaller, M. (2009a). The Cologne Information Model: Representing information persistently. In M. Thaller (ed.), *The eXtensible Characterisation Languages – XCL*. Hamburg: Kovač, pp. 223-240.

Thaller, M., ed. (2009b). *The eXtensible Characterisation Languages – XCL*. Hamburg: Kovač.

3. Bringing together markup and semantic annotation (Dino Buzzetti)

Far from having been convincingly clarified, the relation between *markup* and *semantics* still appears to be a perplexing one. The BECHAMEL project, a consistent and systematic attempt to provide a semantics for document markup (Renear et al. 2002), aimed at introducing mechanisms and rules for mapping syntactic markup structures into semantic domains of objects, properties and relations. In a convincing article, however, Dubin and Birnbaum (2004) acknowledge that 'all the distinctions that we're able to explicate using BECHAMEL' could either 'guide the re-tagging of documents with richer markup' or 'be serialized in the form of RDF or a topic map.' (p. 8) But, in the first case, is the prospect of expressing all semantic information through the markup a viable solution? As it has been pointed out, semantic and interpretative encoding prevents interoperability, and since any 'attempt to make a document interoperational' is 'likely to result in decreased expressiveness,' markup scholars are ready to admit that interoperability is 'the wrong goal for scholarly humanities text encoding.' (Bauman 2011) On the other hand, a purely semantic description is clearly incomplete, for it might disregard equally possible and semantically equivalent textual variants.

Keeping inline markup and semantic information distinctly severed proves to be a more appropriate approach. In the case of scholarly digital editions, the sole concern with markup has left us with 'a problem that still exists,' for 'we need (we still need) to demonstrate the usefulness of all the stuff we have digitized over the last decade and more – and usefulness not just in the form of increased access, but specifically, in what we can do with the stuff once we get it' (Unsworth 2003). How can we proceed 'beyond representation' and build tools that shall 'put us into new relationships with our texts' and enable us to process their information content? (Unsworth 2004) Embedded markup can best serve as a comprehensive information carrier for textual information content, but we need further solutions to process content in an efficient and functional way. For embedded markup provides a data structure, but it does not beget a suitable data model. (Raymond 1992, 1996) It defines a format, not a semantics to process its information content. Whether Semantic Web technologies do provide satisfactory data models for humanities research is still an open question, but the problem yet remains how markup and semantic description techniques can be suitably related, by heeding carefully the basic distinction between data and information content. TEI extensions on the one side and the RDFa syntax on the other, do not seem to provide an adequate approach, failing as they do to keep format and content concerns duly severed. The apparent markup overload they produce carries with it a dubious Ptolemaic flavour.

The relation between embedded markup and semantic description languages is an indetermination relationship. Dubin and Birnbaum (2004) fully recognise its very nature: 'the *same markup* can convey different meanings in different contexts,' and 'markup can communicate the *same meaning* in different ways using very different syntax.' It is, on both sides, a one-to-many relation. If you fix the syntax, the semantics may vary in various contexts, and vice versa, if you fix the semantics, you can use a different syntax to express the same content. Contrary to the tenets of hard artificial intelligence – 'if you take care of the syntax, the semantics will take care of itself' (Haugeland 1985: 106) – and of current analytic philosophy of language – 'to give the logical form of a sentence' is to 'bring it within the scope of a semantic theory' (Davidson 1980: 144) – there is no one-to-one correspondence between the logical form of a phrase and the structure of its semantic content. We should not take for granted that by processing a string of characters representing a text, we process its information content, for we can, and often do, process a string without processing the content. And far from being a drawback, this circumstance is actually an advantage, for by dealing with indetermination we can effectively chart variation.

Both the *expression* and the *content* (Hjelmslev 1961) of the text are open to variation. Dealing with textual variants is the task of textual criticism, just as dealing with interpretative variants is that of the literary critic. But we are not at loss in tackling these problems with computational means. We can exploit the ambivalent status of markup to represent the dynamics of variation. (Buzzetti 2009) As a diacritical mark, the markup can be construed either as belonging to the text or as providing an external description of its structure. We may therefore attribute to the markup both a descriptive and a performative function. (Renear 2000) Assumed in its performative capacity, the markup can be seen as an instruction, or as a rule, to determine the semantic structure of the text, whereas taken declaratively it can be equated to a variant of the string of characters that constitutes the text. Referring to a stand-off metadata representation of the textual information content, or to a stand-off markup of sorts with semantic import, we can all the same assume their structural marks in both acceptations, declarative and performative, and get an overall dynamic model of textual and interpretative variation, as shown in Figure 1:

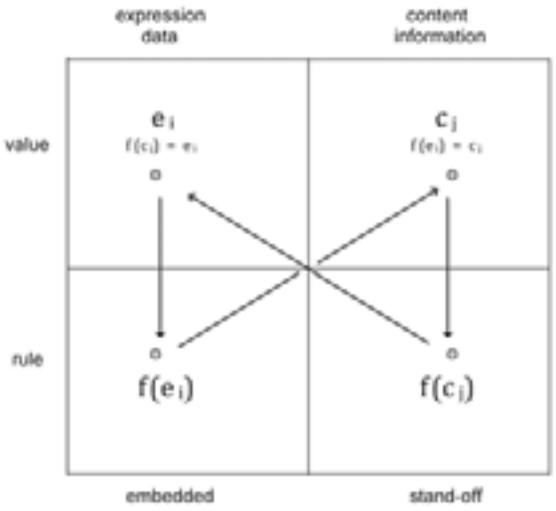

Figure 1

In this diagram, *ei* represents a specific element or construct of the *expression* of the text, conceived of as the set of all tokens that compose it, or E = { *e1* , *e2* , … , *en* }. In its performative capacity that element assumes a different logical status, and can be construed as a function *f (ei) = ci* mapping into the set of all tokens of a given content representation C = { *c1* , *c2* , … , *cn* }, whose specific elements *ci* act in a similar way as a function *f (ci) = ei* mapping into the set E of all the elements of the expression of the text.

Both kinds of variants, textual and interpretative, can be collectively represented, as a kind of 'logical sum' (Thaller 1993: 64), by means of an MVD (Multi-Version Document) graph, as shown by Schmidt and Colomb (2009). Each path of an MVD graph – a directed graph with a start-node and an end-node – represents a different version of the text. A totally isomorphic graph can be obtained also for interpretative variants. In the case of a Topic Maps representation of textual content, an MVD graph was obtained by collating textualized XTM representations of different maps referring to the same text (Isolani et al. 2009).

A comprehensive representation of this kind, of both textual and interpretative variants through MVD graphs, aims at finding efficient ways to determine which paths of the one graph, or which versions of the text, are compatible with specific paths of the other, or with different interpretations of its information content. Both graphs can be used to process the information they represent: the textual variants graph in order to visualise and display different views and versions of the text of a digital edition; the interpretative variants graph in order to process its information content. Promising and different approaches to that end have been proposed by Schmidt (forthcoming), in the context of the HRIT (Humanities Resources, Infrastructure and Tools) project, and by Thaller

(2009), through the development of the XCL (eXtensible Characterisation Language) language. The two methods can offer different implementations of the model here described for specific tasks of the editorial practice, and the pursuit of interoperability between them sets the goal for further development and research.

References

Bauman, S. (2011). Interchange vs. Interoperability. In *Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies*, vol. 7, doi:10.4242/BalisageVol7.Bauman01 (accessed 13 March 2012).

Buzzetti, D. (2009). Digital Editions and Text Processing. In M. Deegan and K. Sutherland (eds.), *Text Editing, Print, and the Digital World*. Aldershot: Ashgate, pp. 45-62.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Dubin, D., and D. Birnbaum (2004). Interpretation Beyond Markup. In B. T. Usdin (ed.), *Proceedings of the Extreme Markup Languages 2004 Conference*, Montreal, Quebec, August 2-6, 2004 http://www.ideals.illinois.edu/bitstream/handle/2142/11838/EML2004Dubin01.pdf (accessed 13 March 2012).

Haugeland, J. (1985). *Artificial Intelligence: The very idea*. Cambridge, MA: MIT Press.

Hjelmslev, L. (1961). *Prolegomena to a Theory of Language*. Madison, WI: U of Wisconsin P.

Isolani, A., C. Lorito, Ch. Genovesi, D. Marotta, M. Matteoli, and C. Tozzini (2009). Topic Maps and MVD for the Representations of Interpretative Variants. In *Digital Humanities 2009: Conference Abstracts*, Proceedings of the 2nd ALLC, ACH and SDH-SEMI Joint International Conference (University of Maryland, College Park, June 22-25, 2009), College Park, The Maryland Institute for Technology in the Humanities (MITH), pp. 8-11.

Raymond, D. R., et al. (1992). Markup Reconsidered. Paper presented at the *First International Workshop on Principles of Document Processing*, Washington, DC, 22-23 October 1992. http://www.cs.uwaterloo.ca/~fwtompa/.papers/markup.ps (accessed 13 March 2012).

Raymond, D. R., F. Tompa, and D. Wood (1996). From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML. *Computer Standards and Interfaces* 18(1): 25-36.

Renear, A. (2000). The descriptive/procedural distinction is flawed. *Markup Languages: Theory & Practice* 2(4): 411-20.

Renear, A., M. Sperberg-McQueen, and C. Huitfeldt (2002). Towards a semantics for XML markup. In R. Furuta, J. I. Maletic, and E. Munson (eds.), *DocEng'02: Proceedings of the 2002 ACM Symposium on Document Engineering*, McLean, VA, November 8-9, 2002, New York, NY: ACM Press, pp. 119-126.

Schmidt, D., and R. Colomb (2009). A Data Structure for Representing Multi-version Texts Online. *International Journal of Human Computer Studies* 67(6): 497-514.

Thaller, M. (1993). Historical Information Science: Is There Such a Thing? New Comments on an Old Idea. In T. Orlandi (ed), *Discipline umanistiche e informatica: Il problema dell'integrazione*. Roma: Accademia Nazionale dei Lincei, pp. 51-86.

Thaller, M., ed. (2009). *The eXtensible Characterisation Language: XCL*. Hamburg: Kovač.

Unsworth, J. (2003). Tool-Time, or 'Haven't We Been Here Already?' Ten Years in Humanities Computing. Paper presented at the conference *Transforming Disciplines: The Humanities and Computer Science*, Washington, DC, 17-18 Jnuary 2003. http://people.lis.illinois.edu/~unsworth/carnegie-ninch.03.html (accessed 13 March 2012).

Unsworth, J. (2004). Forms of Attention: Digital humanities beyond representation. Paper delivered at *The Face of Text: Computer-Assisted Text Analysis in the Humanities*, the third conference of the Canadian Symposium on Text Analysis (CaSTA), McMaster University, November 19-21, 2004. http://people.lis.illinois.edu/~unsworth/FOA/ (accessed 13 March 2012).