# Where Do Humanities Computing and Digital Libraries Meet?

Dino Buzzetti

University of Bologna, Italy
buzzetti@philo.unibo.it

## 1 Introduction

It is in libraries that humanists have always found their basic and essential instrumentation. Libraries can be described as the humanist's lab. Obviously this applies also to digital humanists, who deal with digital objects for research purposes, and to digital libraries that store collections in digital form. But digital objects produced for research purposes are not just inactive artefacts and 'digital library objects are more than collections of bits,' for 'the content of even the most basic digital object has some structure' and to enable access and transactions additional information or 'metadata' is required. [1] So 'if, unlike print,' digital editions 'are also open-ended and collaborative work-sites rather than static closed electronic objects' (p. 77), [2] it can be legitimately asked how a digital repository for objects of this kind can enable effective access to the interactive functionalities they provide. In a digital research context, the issue of how the architecture of a digital library could meet the needs of the working practices increasingly adopted by digital humanists seems therefore of primary importance.

But how can we define humanities computing and what are its requirements? A plausible answer can be found in the final report of a European Thematic Network on Advanced Computing in the Humanities (ACO*HUM):

> […] we will attempt to define the core in terms of the traditional combination of data structures and algorithms, applied to the requirements of a discipline: (a) the methods needed to represent the information within a specific domain of knowledge in such a way that this information can be processed by computational systems result in the data structures required by a specific discipline; (b) the methods needed to formulate the research questions and specific procedures of a given domain of knowledge in such a way as to benefit from the application of computational processing result in the algorithms applicable to a given discipline. [3]

In this understanding, digital objects representing primary source materials, should be endowed with specific functionalities capable of answering specific research questions. Accessing this kind of resources should not prevent the applicability of such functionalities and that is precisely the point where digital humanities and digital libraries can actually meet.

M. Agosti et al. (Eds.): IRCDL 2012, CCIS 354, pp. 4–10, 2012.

## 2 The Creation of Digital Resources

The creation of digital resources for the humanities, however, has not remained unaffected by major technological developments. Humanities computing research practices have remarkably changed along with the availability of different computational means. Whereas with the use of mainframes the emphasis was placed chiefly on content processing, with the advent of personal computers and even more so with the introduction of the WorldWideWeb, the interest shifted to the representation of the original source materials. As John Unsworth has timely observed,

> we are, I think, on the verge of what seems to me the third major phase in humanities computing, which has moved from tools in the 50s, 60s, and 70s, to primary sources in the 80s and 90s, and now seems to be moving back to tools […]. I think we are arriving at a moment when the form of the attention that we pay to primary source materials is shifting from digitizing to analyzing, from artifacts to aggregates, and from representation to abstraction. [4]

And again, clearly, the now emerging 'third phase' in humanities computing is substantially enhanced by the development of the new Semantic Web technologies. Nowadays research practices in humanities computing are actually moving back from representing sources in digital form to designing tools to process their information content:

> We've spent a generation furiously building digital libraries, and I'm sure that we'll now be building tools to use in those libraries, equally furiously, for at least another generation. [4]

But do the needs of advanced digital humanities practice and research find satisfactory support in current digital library environments and architecture? Can digital libraries designers and digital humanists join their efforts to set up a common research agenda?

## 3 The Case of Digital Editions

To better trace these developments, we may consider, by way of example, the case of digital editions. In the late 80's and early 90's a digital edition was thought of as a way of representing a text and its entire textual tradition as a database, [5] because at that time, in order to bind passages of text to selections of their manuscript images it was necessary to integrate textual and visual elements in a single DBMS capable of handling both kinds of structured information. Accordingly, and more to the point, the transcription of the original documents was not meant, like a diplomatic transcription, as a means to convey to the reader 'a closer idea of the nature of the source' (p. 145), but 'as data to be processed'; and so, in this understanding, it was assumed that the transcription of a document

> becomes an activity of data modelling and encoding in order to elicit as much information as possible from the manuscript and to infer new analytical results. From this point of view, both the image and the transcript are not regarded as physical reproductions referring back to the original document but rather as analytical data pointing toward a new logical representation of the source (p. 148). [6]

The emphasis was still on processing, even after the introduction of graphic user interfaces. A digital text representation was still conceived of as data for further processing rather than as a means to visualise a physical document. But things gradually changed as the emphasis shifted more and more towards visualisation on graphic Web browsers and computer screens. The Web was chiefly meant for remote access and visual display, whereas WYSIWYG systems and page description languages promoted an ever increasing tendency towards the 'electronic simulation of specific print objects' (27) [2]. Digitisation projects and the visual representation of primary sources became the prevailing interest in humanities computing.

## 4 The Form of Attention of Digital Humanities

The 'forms of attention' of digital humanities – see [7] – shifted then from processing to representation. And the new developments in technology encouraged that process. In computer science, besides the so-called data processing or database community a large and authoritative document community grew up and established itself. [8] Both groups suffered from the problem of having their data 'trapped' in proprietary systems. The dissatisfaction of the document community with its early systems led to 'generalize its markup' and to endorse the ISO SGML standard, a markup language that was accessible to the writer and allowed to encode not only the 'presentational aspects of documents,' but also 'more general properties of texts' (p. 26). Since 'for the document community, the factor of most permanence was the document,' that community 'chose to standardize the representation of data.' On the other hand, 'for the database community, the factor of most permanence was the semantics of applications,' and so that community 'chose to standardize the semantics of data.'

These different leanings proved decisive for the choices of the scholarly community. Three foremost humanities computing associations, the Association for Computational Linguistics (ACL), the Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and the Humanities (ACH) decided to promote the Text Encoding Initiative (TEI) and to adopt the ISO SGML standard for the encoding of texts. 'Data semantics was not irrelevant to the document community,' but the definition of semantics 'did seem to be a difficult problem' (p. 27). And also 'attempts to define semantics in the scholarly community, most notably the Text Encoding Initiative, similarly met with resistance.' Thus, 'the route proposed by SGML' seemed 'a reasonable one' and the scholarly community conformed to it:

> promote the notion of application and machine independence, and provide a base on which semantics could eventually be developed, but avoid actually specifying a semantics (p. 28).

As a consequence, the centre of attention moved over from processing information content to mere data representation.

## 5 The Web and Its Languages

The same repercussions can be noticed by observing the expansion of the Web. It is not by chance, that the languages employed in the construction of the Web, HTML and now ever increasingly XML, are basically data representation languages. They express the structure of the representation, not the structure of what is represented, unless the two structures match and can be put into a one-to-one correspondence. The processing of the documents accessible on the Web depends on the structure these languages assign to them, and thus on the constraints of a hierarchical tree structure. XSLT, the language introduced to process data in XML format, 'takes a tree structure as its input, and generates another tree structure as its output.' [9] It preserves the structure of the document and what it can process is not the structure of the information content it conveys.

Since it allows easy access and excellent visualisation, the Web has been confidently envisioned as a potential universal library. In this conviction, a number of large-scale digitisation projects were begun, such as the Million Book Project (also known as the Universal Library), led by Carnegie Mellon University and started in 2002, the Google Book Search Project started in 2004, and the Microsoft's MSN Book Search, announced in 2005 and subsequently discontinued. But, as it has been observed by Deegan and Sutherland [2], 'the paradigm for the universal library' they enforce 'is not a library at all, it is the Internet' (p. 151). And the Internet really is a different kind of information space from a library. In the Internet the 'professional organisational principles' that belong to the library science tradition 'do not appear to be carried over' (p. 150); in the information space created by the Internet, 'order' is virtually neglected and so 'one of the major benefits that libraries bring to the almost boundless intellectual space that is our literate culture is lost' (p. 149). All in all, as Deegan and Sutherland maintain, 'Google "Book" Search (note our inverted commas) is not providing electronic text, it is providing books' (p. 147). The emphasis is again on the document – the book – and not on its information content – the text. Mass digitisation projects show, once more, that in the Web 'the potential of the computer as visualisation tool' has probably overtaken its analytical and, for many humanists, more appropriate 'computational' uses (p. 75).

## 6 Major Technological Innovations

How, then, can we explain that major technological innovations such as the introduction of personal computers and the expansion of the Web produced almost paradoxical

effects on humanities computing? How could they hold back the development of its methods and research practices? We may assume an evolutionary point of view to look for a possible answer. What matters more for the evolution of biological organisms are not so much their external features, but rather their physiological capabilities and functions. In a similar way, if a digital object can now be visualised as the reproduction of a corresponding physical object, it can also be evaluated for its functionalities and the available facilities to process the information content it conveys. Functional as opposed to visual features are what really matters.

Now, on the one hand, 'what humanities computing has been doing, implicitly, for years' can in many ways be described as 'knowledge representation.' [10] But, on the other hand, if knowledge representation is legitimately seen as 'a medium for pragmatically efficient computation,' [11] representing information and processing information cannot be regarded as separate activities, each one opposed to the other. The form of a knowledge representation can actually be thought of as depending on the computational procedures aimed at processing its information content.

It is precisely for their concern over processing information content that humanities computing research practices are now aligning with those of relevant neighbouring fields. In the specific domain of knowledge organisation and subject indexing, Vanda Broughton, an expert in faceted classification systems and thesaurus construction, observes:

> Current co-operative work with scholars in the area of humanities computing suggests that, in combination, facet analytical and text encoding methods may offer a solution to improving the usability of metadata tools and providing more subtle and sophisticated means of subject representation (p. 193). [12]

Here knowledge organisation and humanities computing concur expressly on the analysis of information content, which is exactly what the new Semantic Web technologies are aiming for. Thus it is indeed the technological evolution of the Web what can help recover that partially neglected aspect of humanities computing which its nascent construction momentarily and almost paradoxically contributed to obscure.

## 7 The Semantic Web

With the help of these new technologies humanities computing can get back to its original inspiration: the 'attention' that in a successive phase of its development was mostly directed to the 'representation of primary source materials' goes back again to building 'tools' for processing their information content. [4] Humanities scholars too recognise 'the semantic web' as their 'future' and humanities computing is thus bound to produce 'formal representations of the human record' suitable for automatic processing. For, as John Unsworth again reminds us,

> those representations – ontologies, schemas, knowledge representations, call them what you will – should be produced by people trained in the humanities. Producing them is a discipline that requires training in the humanities, but also in elements of mathematics, logic, engineering, and computer science […]. There is a great deal of work for such people to do – not all of it technical, by any means. Much of this map-making will be social work, consensus-building, compromise. But even that will need to be done by people who know how consensus can be enabled and embodied in a computational medium. [13]

New developments induced by Semantic Web technologies can also be observed in the field of digital libraries. An interesting example is offered by the so-called semantic digital libraries, [14] whose declared purpose is to integrate Semantic Web and social networking technologies into a digital library management system. The basic assumption, here, is that 'semantic technologies can offer more efficient solutions for building robust, user-friendly ways of accessing content and metadata.' [15] Semantic technologies, it is averred, can supply 'efficient discovery techniques in the new, interconnected information space' of digital resources accessible on the Web. The use of ontologies produces new forms of information and knowledge organisation, that do not reduce themselves to a 'mere specification of metadata schemata' previously established, but allow 'metadata to become more open, unstructured, and what is most important, highly interlinked' (p. 78-79). [16] The use of lightweight tag ontologies 'provides the possibility for machine-processable representations that can be shared across social tagging systems.' [17] The practice of social tagging can then usefully help to integrate valuable sources of semantic annotations in a digital library platform that provides linked data services.

The application of semantic annotation technologies both in digital library systems and humanities computing applications clearly shows that in both fields a need for common functionalities is actively felt. The case could easily be generalised and may wishfully prompt a closer reflection on the prospects of a common research agenda for digital libraries and digital humanities.

## References

1. Arms, W.Y.: Key Concepts in the Architecture of the Digital Library. D-Lib Magazine 1(1) (1995), http://www.dlib.org/dlib/July95/07arms.html
2. Deegan, M., Sutherland, K.: Transferred Illusions: Digital Technology and the Forms of Print. Ashgate, Farnham (2009)
3. Thaller, M.: Defining humanities computing methodology. In: de Smedt, K., et al. (eds.) Computing in Humanities Education: A European Perspective, ch. 2.3, University of Bergen-HIT Centre (1999), http://www.hd.uib.no/AcoHum/book/fm-chapter-final.html
4. Unsworth, J.: Forms of Attention: Digital Humanities Beyond Representation. Paper Presented at the 3rd Conference of the Canadian Symposium on Text Analysis (CaSTA). The Face of Text: Computer-Assisted Text Analysis in the Humanities, McMaster University (2004), http://people.lis.illinois.edu/~unsworth/FOA/

5. Buzzetti, D.: Masters and Books in 14th-century Bologna: An edition as a database. In: Bocchi, F., Denley, P. (eds.) Storia & Multimedia, Proceedings of the 7th International Congress of the Association for History and Computing, August 29-September 2, pp. 642–646. Grafis Edizioni, Bologna (1994)
6. Buzzetti, D.: Image Processing and the Study of Manuscript Textual Traditions. Historical Methods 28(3), 145–154 (1995)
7. Kermode, F.: Forms of attention. The University of Chicago Press, Chicago (1985)
8. Raymond, D., Tampa, F., Wood, D.: From data representation to data model: Meta-semantic issues in the evolution of SGML. Computer Standards & Interfaces 18, 25–36 (1996)
9. Kay, M.: What Kind of Language is XSLT? An analysis and overview (2005), http://www.ibm.com/developerworks/library/x-xslt/
10. Unsworth, J.: Knowledge Representation in Humanities Computing. Paper Presented at eHumanities. NEH Lecture Series on Technology & the Humanities, Lecture I, Washington, DC, vol. 4 (2001), http://people.lis.illinois.edu/~unsworth/KR/
11. Davis, R., Shrobe, H., Szolovits, P.: What is a Knowledge Representation? AI Magazine 14(1), 17–33 (1993)
12. Broughton, V.: Finding Bliss on the Web: Some problems of representing faceted terminologies in digital environment. In: Gnoli, C., Mazzocchi, F. (eds.) Paradigms and Conceptual Systems in Knowledge Organization, pp. 188–194. Ergon-Verlag, Würzburg (2010)
13. Unsworth, J.: What is Humanities Computing and What is Not? Jahrbuch für Computerphilologie 4, 71–84 (2002), http://computerphilologie.tu-darmstadt.de/jg02/unsworth.html
14. Semantic Digital Libraries: Bringing Libraries to Web 3.0, http://semdl.info/
15. Kruk, S.R., Westerski, A., Kruk, E.: Architecture of Semantic Digital Libraries. Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Work Package, vol. 4, pp. 1–12 (2008)
16. Kruk, S.R., Westerski, A., Kruk, E.: Architecture of Semantic Digital Libraries. In: Kruk, S.R., McDaniel, B. (eds.) Semantic Digital Libraries, pp. 77–85. Springer, Berlin (2009)
17. Kim, H.L., et al.: The state of the art in tag ontologies: A semantic model for tagging and folksonomies. In: DCMI 2008: Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications, Dublin Core Metadata Initiative (2008), http://dc2008.de/wp-content/uploads/2008/09/kim-scerri-breslin-decker-kim.pdf