

Towards a publication and research infrastructure for the Humanities

“Scholarly Editions on the Web” Hyper Training Workshop

Munich 22-25 June 2006

Paolo D’Iorio

1. Allow me to try to explain why we have gathered here today. We are here because we share a dream: to work together on a global scale for the preservation, analysis, and augmentation of our common intellectual and cultural heritage. We seek new and better ways to practice the time-honored art of the humanities scholar. We aim to study the great texts of literature and philosophy, and also the myriad other documents, images, and objects that constitute our heritage, in a manner that is more cooperative and open, but also more rigorous and comprehensive. This is the dream of Hyper. The word is an acronym for Hypermedia Platform for Electronic Research and it signals our shared goal of building an infrastructure for Humanities research and for the publication of critical literature and scholarly editions on the Web. Now, we all know what publication and research are, and we know what the Humanities are, or at least we assume we know, so let us consider what an infrastructure is, a web infrastructure.

2. So what is an infrastructure? It is a well-coordinated system of buildings, equipment, objects, services, procedures, administrative structures, conceptual models, etc. It is something that underlies (*infra*) and facilitates a certain activity and at the same time establishes a network of connections. As the Latin prefix suggests, an infrastructure is an underlying support, but it is also that which lies between different elements and links them together. In fact, in Italian the prefix *infra* expresses this duality: both the Latin sense of *infra*, *sotto*, which means "under", and the modern sense of *infra/intra/inter*, where *tra* indicates "between."

To build a new research infrastructure, one must first analyze the pre-existing structure. And a research infrastructure for the humanities does in fact already exist, one that has been developing over the course of 2000 years. It is a complex whole that can be separated into three parts.

3. First, there is the *Physical Structure*, composed of **objects of study** and **buildings** in which those objects are stored and which serve as work stations for the **people** who conduct research and preserve the objects. There are, in addition, **distribution systems** for the dissemination of knowledge and last but not least, **financial support**, which makes the entire system possible.

Second, there is an *Organizational Structure*. From a research perspective, the infrastructure has heretofore been organized into **disciplines**, has used **peer review** as a filtering mechanism, and has relied on the **impact factor** to evaluate researchers. From the legal-economic perspective, **copyright** has traditionally served to settle matters of intellectual property. **Market conditions** have also played a role in the

traditional model. Finally, from a political standpoint, various institutions have been responsible for the organization and financing of research (universities, the DFG, the European Commission, etc).

The third aspect of the traditional model we may call the *Logical Structure*, which I would in turn divide into the triad: *materials*, *scholars*, and *contributions*. *Materials* are the primary sources used by *scholars* in their research; *contributions* are the publishable results of scholarly research. This simple tripartite division is useful for indicating the various functions that a single object may serve. For example, a text written by Heidegger about Nietzsche would be a *contribution* to Nietzsche research, but it would be considered as *material* for Heidegger research.

I would also like to call attention to the **citation**, which may be viewed as the navigation system that links together the various elements of this logical structure.

The traditional infrastructure is a well-coordinated system that does attain its goal: to engender and disseminate knowledge. It is, however, not perfect. Let me enumerate three problems in particular.

4. Problems with the traditional infrastructure

A. This system is slow and expensive in regard both to access to primary sources and to the distribution of scholarly publications. If I want to study Nietzsche's manuscripts, I have to travel to Weimar, and if I want to disseminate the results of my research, I have to have it printed and the publication must then be transported hundreds of miles.

B. As the infrastructure grows ever more expansive, the research becomes increasingly less collaborative, cumulative, and cost-effective. If 100 scholars are working on Kant, the traditional infrastructure can establish contact between them fairly effectively. But when one is talking about 1000 scholars, traditional methods cannot easily ensure that, for example, a scholar in Germany will know that a colleague in Brazil or Japan has published an important paper on a particular passage or key concept in the *Critique of Pure Reason*. Admittedly, the apparatus does exist to communicate this information -- there are bibliographies, indexes, digests, etc. But as the field of research grows more complex, this apparatus becomes less and less efficient. As a result, different scholars may undertake the same research, or one scholar may remain ignorant of relevant and valuable investigations conducted by other researchers.

C. A third problem is the misuse, by publishers, of the otherwise legitimate mechanisms of peer review and copyright for the purposes of financial gain. The fact is that in the hard sciences there is a monopoly market that makes access to scientific information very expensive: a yearly subscription to "Brain Research" costs around 20.000 dollars. Presently, the best-funded libraries have to use 80 to 90 per cent of their budgets for the purchase of scientific journals and nevertheless will be able to afford only a small part of this literature.

The reasons for this phenomenon were explained, in an illuminating article by Jean-Claude Guédon, as a perverse consequence of a mechanism for the quantitative evaluation of academic work invented in the 1950s by Eugene Garfield: the Science Citation Index. The SCI comprised the definition of a collection of “core journals” which influenced the acquisition practices of libraries, thus creating the conditions for an inelastic market. At the beginning of the 1970s certain editors became aware of this process and tried to accelerate it, and at the end of the 1980s a new monopolistic market had come into being.

For scholarly publications in the Humanities there is not a monopoly market – there is no market at all. If libraries spend 90% of their budgets on journals in the hard sciences, not much is left for humanities journals and monographs. The alarm was sounded in 1999 in an article by historian Robert Darnton in the *New York Review of Books*: scholarly monographs are an endangered species, he argued. The only ones that will survive will be those written by authors who can themselves finance the publication, or those who work on themes that have broad public interest.

Fortunately, we now have the internet, which can rescue us from the effects both of the monopoly market and the non-existent market. Can the internet really solve the problems of the traditional infrastructure? Let us see what happens when computers and the internet are integrated into this system.

5. At first, computers were used in the humanities for lexical analysis and the creation of concordances, thesauri, etc, and so a new discipline came into being: computational linguistics.

Next, the PC (personal computer) and word processing programs came into use and the use of digital texts became common. But digital texts, which can be published directly on the internet, remain on the margins of the traditional system. Currently, scholars must choose between two alternatives. Either they submit their texts to a traditional publisher for peer review and thereby surrender their copyright and the option to publish on the internet, or they do publish their texts on the internet, but thereby do not receive the validation of peer review and as a result forfeit the benefits to their academic careers. This situation results in a charming paradox, one which is symptomatic of a dysfunctional system. Scholars produce texts in digital format that could, in principle, be published immediately on the Internet, but instead are published on paper, often in very small numbers. These pages are then re-digitized by projects like Google Books or, in Europe, by programs of "retrospective digitizing." Thus it happens that public funding pays once to support the original research (which is good), a second time to publish the results in book form (which is less sensible), and a third time to digitize the books (which, frankly speaking, is excessive.)

6. In sum, we see that the computer and the Internet do not solve the problems of the traditional infrastructure. What we have instead is the gradual creation of a global, googelized parallel system composed of a mass of digitized texts of all different kinds and with varying degrees of scholarly reliability and, to navigate all this, a simple search engine.

It is as if we had taken the complex traditional system (one which did in fact produce knowledge), placed it in a digital blender, and reduced knowledge to mere information. In this model, all notions of structuring information are lost: one searches for words and receives a list of occurrences, and that's it.

7. Beside this unstructured model, which has been very influential and has been used in many digitization projects and in humanities computing more generally, there does exist other initiatives that are considerably more structured. One model for the global structuring of information that I would like to call attention to is the Wikipedia, which, as the name indicates, is structured like an encyclopedia.

While this initiative is reasonably successful, it seems to me to be ill-suited for scholarly research. I cite three reasons:

- a. Lack of peer review
- b. Instability of the text, which can be modified at any time by anyone
- c. Instability of authorship: the articles in Wikipedia are the fruit of common labor in which it is difficult or impossible to distinguish individual contributions.

From the lack of peer review and the instability of the text, we can conclude that Wikipedia, while it is a formidable instrument for animating a community, is nevertheless fundamentally inappropriate as an instrument for administering information generated by a scholarly community. Or, alternatively, one must imagine that scholarship will become something very different from what it has been in previous centuries. The transformation would be dramatic, encompassing how knowledge is validated, how research is organized, how academic careers progress, etc.

In regard to the instability of authorship in the Wikipedia model, it appears to me that recent trends in the publication of material on the web signal not the death but rather the rebirth of the author. Think, for example, of the New York Times, whose subscription service (TimesSelect) is based on a close rapport between subscribers and the organization's Op-Ed columnists. And then of course there is the blog phenomenon, in which the author plays a fundamental role.

We may perhaps smile condescendingly at these models for the dissemination of knowledge; in fact, all good humanities scholars will probably look down on them. Such systems, we tend to think, do not provide the sources for scholarly research nor are they the proper venues for publishing our findings. But let us ask ourselves, we scholars, we repositories of certified, peer-reviewed, well-structured, and long-enduring knowledge: what is our model for the dissemination of knowledge? Is it a book, published two years after the research has been completed, distributed in 300 copies, after we've paid 4000 euro to supplement the publication? The problem is certainly not the book – everyone loves books and we all agree that it is still the best medium for reading. The problem is the 2 years, 300 copies, and 4000 euro. Faced with the incredibly efficient dissemination of knowledge made possible by Wikipedia, blogs, and an endless variety of web communities, what do we have to propose? Do we have

a model that preserves the complexity and structure of scholarly knowledge; a model that, while global and open, still supports meaningful peer review and preserves the concept of authorship; one that guarantees the stability of the text and provides a navigation system more sophisticated than lists of occurrences or encyclopedia articles? We have already noted that humanities computing, instead of constructing an integrated system capable of employing the new electronic medium for all forms of research, instead of creating a new infrastructure, merely produced a new niche discipline. Is this a sign that the humanities are destined to be forever resistant to large-scale coordination and condemned to constantly create new disciplines and sub-disciplines and ever narrower specializations?

8. With Hyper, we propose a model for a research and publishing infrastructure for the humanities. What will it look like?

In the first place, there is the physical structure: the foundation is no longer objects but electronic files, no longer buildings but computers; the dissemination is not accomplished by publishers, distributors, and bookstores, but rather, via the Internet.

Second, the organizational structure. Here it is necessary to invent organizational models for the future, and we must do this by taking into consideration the effective and time-honored examples of the past. We are calling for the founding of Scholarly Communities, to be modeled on the Science Academies that accompanied the founding of modern science (l'Accademia del Cimento, L'Accadémie des Sciences, the Royal Society). They are free associations of specialists who work on a specific author or area of research and who organize an internal peer review process. They themselves (and not the publisher) guarantee scholarly standards. They also preserve the stability of authorship and of the texts that belong to the HyperPlatform they administer.

I would like now to discuss the notion of open source in the humanities. As we define it, open source consists of two elements:

Public Archives: this means, the capacity to guarantee free access, via the Internet, to the primary sources for research in the Humanities (texts, images, sound, video, artworks, artifacts, etc.) in cases where such sources are held by public libraries. The legal scholars working on the Hyper project have drafted a model agreement between the scholarly communities and the archives to permit the free publication on the web of relevant primary sources for purely scholarly purposes. This model has been used for the publication of facsimiles of Nietzsche's manuscripts on HyperNietzsche.

Open Publishing, i.e., the capacity and in fact the duty of researchers to make freely available on the Internet the results of their research (in cases where it was funded with public money). This would make available on the Internet an enormous mass of scholarly contributions, without the need to pay 200 million euro to digitize them, because they are already in digital format.

Following the work of Richard Stallman, but before the advent of Creative Commons, we understood that in the case of scholarly work, it is necessary to distinguish the

aspect of copyright that matters to authors (respect for authorship, protection against plagiarism), from the aspect that matters to publishers (remuneration, restrictions on copying). For this reason, we use Copyleft instead of Copyright and we have drawn up a set of licenses called OpenKnowledge for the purpose of publishing scholarly contributions.

In Europe, most archives, libraries and research institutions are public and are financed with public money. It would greatly facilitate the dissemination of knowledge if European law also specified that everything supported by public funds must be made freely available to the public.

Finally, the logical structure remains unchanged, but it is now understood as what computer scientists call an "upper ontology." This upper ontology is compatible with the Dublin Core Framework but goes into more detail, especially with respect to the description of primary sources.

One might also say that these three parts form the ontology, the philosophy, and the technology of Hyper.

I would like to draw your attention to the **navigation system** of this hyper-infrastructure, which we call dynamic contextualization. Why has the traditional infrastructure been so useful? Because with a simple bibliographical reference at the bottom of the page, the author can refer to another document in a very precise manner: to a specific page in an article or a book. It is necessary to develop such a system for the Internet, but it must be one that employs all the powers of the Internet. To clarify this mode of scholarly navigation, I need only ask myself how a scholar in pre-digital times navigated the library. He did not follow a list of "hits" of the kind produced by Google. Scholarly knowledge is not structured like a list or a tree, but rather like a mathematical graph. Understanding this also helps to dispel a common misunderstanding, according to which the difference between printed books and hypertext is that a book ensures a sequential reading whereas hypertext introduces non-sequential reading. Nothing could be more false in the realm of scholarly research, because a key characteristic of scholarly reading is precisely that it is non-sequential. A classicist at work in the library is likely to have a dozen or more books open on the table and to jump from one to the other: he verifies, he looks for connections, he follows links made explicit through the venerable tradition of scholarly citation.

9. We must transpose this time-honored system of scholarly citation into an electronic environment so that when an user selects a critical essay, he will be presented automatically with a list of all primary sources cited in the essay, a list of all the articles cited by the selected essay, and, more importantly, *a list of all the essays in which other authors cite the essay currently being viewed*.

10. Likewise, if the user selects a manuscript page, the system should immediately make accessible, without the need for additional complicated searches, all the transcriptions and translations available for the page, as well as all the relevant text-genetic paths and critical essays that refer to the page.

Dynamic contextualization structures information in an intuitive way, much closer to the way in which knowledge is organized in actual scholarly practice. Indeed, scholarship is the capacity to analyze the same object with different criteria, and different objects with the same criteria. Not only that: from a cognitive point of view, **the oriented multigraph data structure** produced by dynamic contextualization is much closer to the operation of the human mind than the typical data tree structures in the vast majority of information processing systems.

11-12. It is easy for users to implement this system: they need only use one tag from the Hyper Markup Language, which suffices to create bi-directional links that the system updates automatically. (This distinguishes Hyper from Wikipedia and the rest of the web, where one is able to create and follow only unidirectional links.)

For our programmers, of course, this system is considerably more difficult to implement, because they are charged with building a scalable system that would be able to manage potentially millions of links. Michele will explain how this is possible.

13. Ultimately, the system will be a semantically structured peer-to-peer HyperNetwork of different communities of specialists using a unique and stable addressing system that supports citations and dynamic contextualization.

14. But it is now time to have a look at HyperNietzsche, the pilot project for the future federation of Hyperplatforms. Its motto is: : « Dahin wirken, dass alles Gute Gemeingut werde und den Freien Alles frei stehe » « To work to make all good things part of the common good and all things free to those who are free ».

You can use HyperNietzsche to navigate through the resources available on the site or to search for specific information.

In the navigation mode, as you page through Nietzsche's manuscripts, published texts, and other material, the contextualization bar to the left shows the scholarly contributions (editions, commentaries, articles) that are relevant to the primary source you are currently viewing.

Alternatively, you can search the HyperNietzsche database for specific information, for example, the occurrence of a word or phrase or the scholarly contributions of a particular author. You then obtain a list that corresponds to the criteria of your search.