Andrea Scotti (Firenze)

Pinakes: Structuring and Destructuring Documentation in the Humanities. A Project for Modelling Data in History Research[1]

1. Introduction

1.1 A short history of Pinakes

In 1996 *Pinakes* came into existence as an application to manage and publish on the web projects within the field of *Documentary History.* What *Documentary History* is and in which way *Pinakes* contributes to the development of such a field of research will be discussed later on in this article. It might help briefly to reflect on how *Pinakes* was initially planned and how it developed through the years. At first *Pinakes* was designed to solve the problem of cataloguing in a relatively short time (two years) a large number of scientific manuscripts located at the National Central Library of Florence. This project[2] was born digital but the researchers and the institutions with whom it was carried out could not at that time offer the digital tools which could ease the work, nor had they suggestions in this respect. Therefore there was the need to create, both in the input and in the output phase, an application which satisfied the classical codicological methods as well as the research needs of the historian of science. The only application available in Italy at that time was *Manus* a DOS application unable to run on a network and notably lacking an interface for publishing the resulting data on the web. This tool required an enormous effort both on the part of the input user and – in view of the cost – on the part of the tax payer. The result was that the data population on it was (and still is) even poorer than the limited one available for many years on paper.

The first problem that arose in building *Pinakes* concerned the definition of the logical schema. The question was how to avoid simply copying the methodologies used to classify documents (in its broad sense) on paper. The object was to draw a generalized logical model that could help solve the primary and interdisciplinary classification problems of the humanities. Bearing this in mind the target *Pinakes v. 1* developed a sequence of indices whose aim was to cross-index and unify all the common attributes of all physical and semantic objects stored. It was on this basis that the first *Pinakes* application for the scientific manuscripts catalogue was designed and built.[3] Still this was an experimental model and offered a base for further developments.

Meanwhile while working at the Institute and Museum for the History of Science in Florence and on other leading research projects around Europe,[4] together with Mar-

[1] I would like to thank here: Prof. Michael Stolz (Universität Bern, Institut für Germanistik), Prof. Thomas B. Settle (Polytechnic University, New York & IMSS) and Vera Hupfauer (University of Munich) for their attempt to give to my written English a form that could be considered readable.

[2] Andrea Scotti: *Scientific manuscripts catalogue. General Catalogue of the scientific manuscripts at the National Central Library in Florence (Italy)*. Supported by the Italian Ministry for Cultural Heritage, the National Central Library, Florence, hosted and managed at the Institute & Museum for History of Science, Florence. 1996–1998.

[3] See: <http://www.pinakes.org> on the chapter »Current Projects« first project from the page bottom.

[4] See the list on http://www.pinakes.org.

co Beretta and Daniele Nuzzo, the *Pinakes Group* was founded. On the basis of previous experience and with the new possibilities offered by the web and the growth of technology, the aim of this group was to create a free application for use within projects of *Documentary History.* The new modelling of this second generation *Pinakes* started in 1998/9 with two research projects.[5] Thousands of objects with different physical and semantic morphologies offered the possibility of re-analyzing and re-thinking the model used in *Pinakes v. 1*. The details concerning such changes will be discussed below. Nowadays, after a substantial cataloguing campaign and having seen a very large number of different objects – which may occur in any research project – *Pinakes* is entering a third generation phase. The next chapters refer to this version of *Pinakes*, unless stated differently. The main differences between *Pinakes v. 2* and *v. 3 beta* are: the former was conceptually flexible but explicitly defined in its semantic structure; the latter[6] explicitly defines a logical structure and semantically requires only the minimum attributes to describe objects, leaving the user the freedom of customizing his own families of new attributes. This solution offers the possibility of building a common taxonomy concerning the different families of objects thereby ensuring, at the same time the chance of displaying and exchanging of information following international standards and the defining of data granularity following the needs of each project. The formal/logical structure has been thought of as a work in progress. That means that it is possible to define in some standard classes a customized subclass of attributes for any given object. This is independent of the fact that an object can be a family of concepts, a group of texts or again a bunch of catalogue records. The hope that such an application will work depends mainly on a transformation of the reasoning methods that should take place within the research community of the humanist and implicitly from the fact that a very large community will use it. An attempt to contribute to that reasoning transformation should be given here.

1.2 Computing and the Documentary History

A discipline called *Documentary History*[7] has never been registered in the curricula of the humanities. Something similar could be one called *documentarist* which is a subset discipline of library and archive studies. This subset revolves around both the different methods of manuscript classification, book classification, print classification etc., and the exchange of records within libraries. Mainly the *documentarist* is concerned with the problem of the development of national models of classification developed to preserve paper material. Here, *documentary,* means something else. In fact the *documents* which *Documentary History* is referring to are all objects resulting from mankind's activity. In this sense *Documentary History* could be described as a history that carries out research

[5] See <http://www.pinakes.org> on the chapter »Current Projects«: *Parnasuss Scientiarum* by Andrea Scotti and Marco Beretta, *Panopticon Lavoisier* by Marco Beretta, data manager Andrea Scotti.

[6] From now on *Pinakes v. 3.*

[7] If at an institutional level there is no recognition of such a discipline, a search on Goggle produces the following matches: 148.000 for *Documentary History*. That means that *Documentary History* is accepted as a disciplinary practice of the humanities but, as well as *Computing and Humanities*, has no institutional body able to sustain its particular disciplinary status. This is even more so when documentary history is exchanged with library and archive documentation.

at the primary record level: selecting, ordering and therefore recognizing artefacts. These objects can be books or manuscripts but also instruments or simply dry plant collections, and even facts, events, concepts, theories etc. Due to the fact that all the necessary knowledge for carrying out such a work can not be handled by one person alone, the *documentarian* can only be an amateur in each field. By creating a community application, each specialist could have access to the results in other fields and use them as references, avoiding the risk of redundancy and recompiling information. If no generalization model of describing objects is implemented, the information resulting from this kind of miscellaneous cataloguing and selecting is not going to be uniform. Moreover, an other and more likely result of such a practice, would be that each given class of objects would become a specialized *Documentary History* case. This is currently the state of the art. In fact, each historical essay has as background a specialized documentation for its own area of reference. The only way in which it is possible to intersect different areas of specialization, is the interpretation written by the historian himself. There is no existing documentation project or catalogue with printed results, in which the relation between objects of different scholarship isn't either set or grouped or ordered in an index. This is due to the fact that in print the same notion of *Documentary History* is barely affordable. A catalogue of this kind would need to have a large number of pages devoted exclusively to the cross-referencing of indices. In such kind of work the field of primary source cataloguing would have to be pretty small and well defined. This implies that the largest proportion of this research would have been that of creating the indices rather then that of finding, selecting and recognizing on a large scale the objects as such. In this way, due to the small amount of catalogued objects, the effectiveness of classifying at a primary record level would be lost. This implies that the only reasonable way of carrying out *Documentary History* research consists in conceiving it in a computational form. A considerable proportion of the digital projects in *Documentary History* that have been published on the web in the last decade have been keen to present information within static navigation trees. On the one hand, this way of publishing offered the possibility of re-establishing in the computational world the *discursive method* already existing on paper. On the other hand this choice did not solve the indexing problem and did not give an obvious advantage in creating relations among different items of one or more indices, nor did it offer the possibility of reading the information in a new way (by overcoming the sequential reading of a book). The relations among objects were not used to group items and their properties. In such projects the notion of *link* was understood simply as a reification of the humanistic reference *footnote*. These projects have introduced a notion of scholarly tag (more on this below) which has some advantages *per se* thus undergoing the risk of creating specialized tag families that cannot easily be cross-indexed.

In contrast to the use of static hypertext, the implementation of relational data bases offered a more economic way of managing documentary data. This would be true if instead of building dedicated databases, scholars had thought of generic models offering the logical elements required to describe particular semantic families based on common disciplinary taxonomies. If the definition of *Documentary History* as collecting objects of

different kind and creating relations among them is accepted,[8] then the key points, in order to exploit such a discipline, are both the index coherency and the flexibility in representing the attributes/properties/relations of different kind of objects *per se*. This means that what is needed is an application that offers tools enabling the researcher to describe his own and/or other classes of objects within a common structuring logic. In this sense *documentary history* is the background against which the basic research problems concerning *Humanities & Computing* have been discussed. At this point these problems have not been fully solved, but great progress has been made by focusing on how the different disciplines of the humanities could share documentary results for the benefit of research in each different field.

2. Sources and Background problems of Humanities & Computing[9]

2.1 Defining the Humanities, its curricula and theoretical foundations

Hereafter are quoted among others some sources concerning the Italian debate[10] on *Humanities & Computing* as a discipline. It is not that other countries have been able to define the academic curricula of such discipline in a better way, but Italy is certainly the country where the lack of acknowledgment of its contributions to humanities research is more pronounced than anywhere else. For this reason many well-known personalities in the field of *Humanities & Computing* have come together, at first to define and then to defend, this area of research. Father Busa – one of the early promoters of this discipline in Italy – has written a short note to define the problem:

> 1. by the ›Humanities‹ we don't just mean traditional literary disciplines concerned with elegant expression (however important such like still are), but rather that they are to be understood, fundamentally and crucially, as the statistical and scientific approach to human written communication that only the recourse to the computer has made possible, today; and that only such an approach can deal with the challenge of electronic globalisation, which is turning out to be a new form of infrastructure for management, commercial and technological communication.
>
> 2. that there are three kinds of textual computing, namely:
>
> 2.1. documentary data banks accessible remotely,
>
> 2.2. publishing distribution of new kinds of book, cd, dvd, multimedia approaches, and their ongoing development,

---

[8] This does not mean that within documentary research is not possible or that it is pointless carrying out monographic or local projects. On the contrary, such researches are a lively part of historical research. The question here is how the results of such projects can be re-used by others or even how they can be referred to larger projects if their data modelling is local and not generic.

[9] Quotations in this chapter will be separated from the text and with smaller characters to underline the fact that these are the key arguments in the field. To understand how broad is the range of problems involved into the definition of this discipline see the review/discussion group on: <http://www.princeton.edu/~mccarty/humanist> and <http://www.computerphilologie.de/>. To the problem of the discipline definition see in the latter: <http://computerphilologie.uni-muenchen.de/jg02/ orlandi.html#fn1>. In particular on the curricula problem see: <http://www.uni-duisburg.de/FB3/ CLiP2001/abstracts/Thaller_en.htm>.

[10] See <http://193.205.145.117/lingue/docenti/informatica/appello/index_e.htm>. On this matter you can read the full text of all quotation used here.

> 2.3. critical. This is ›my‹ area: comprehensive statistical and classificatory examination of written linguistic material, using large-scale corpora of natural texts, which are up-to-date and representative; their mark-up; their collation into lexicological systems etc.
>
> 2.1 and 2.2 constitute merely the reproduction and distribution, in a searchable form, of ›digital images‹ of words and punctuation and images and sounds. 2.3, on the other hand, adds to each word many internal hypertexts, which then provide a semantic qualification context by context.[11]

In the same document we can read the definition given to *Humanities & Computing* by Willard McCarty.[12]

> In essence the goal of humanities computing is to refurbish scholarship for the electronic age in order to strengthen and extend, rather than dilute or pervert, its traditional aims and methods. Humanities computing seeks in this way to help the scholar and to preserve the scholarly way of life. It also seeks to make scholarship more accessible, its products more readily available to students without compromising the integrity of these products. It builds bridges outward from the core of scholarship to applications well beyond it, for example in the language industries. […] Humanities computing teaches a critical attitude, focused as much on what the computer cannot do as what it can do for us. In this and many other ways, the field is thus of as well as in the humanities, in spirit and basic aims as traditional as any discipline taught in […] most distinguished universities.[13]

These ideas presented schematically here have been augmented in detail elsewhere.[14] The main problem in the *Humanities & Computing* is not only the definition at an institutional level of its status, but particularly the understanding of what is called its new hermeneutics. The discussion of this theme has been extremely challenging in the last years. The quotation of some sources listed and summarized in an article by Tito Orlandi[15] will give an overview of the state of the art. The following is, among others, the opinion in the field of Willard McCarty.

> *Just a tool*: otherwise intelligent colleagues refer to the computer as *›just a tool‹* or *›simply a bunch of techniques‹*, as if ways of knowing did not have much to do with what is known. Because the computer is a meta-instrument – a mean of constructing virtual instruments or models of knowing – we need to understand the effects of modelling on the work we do as humanists. *Creative expression and mechanical analysis*: What is the relationship between creative expression and mechanical analysis? What scholarly role can the algorithmic machine play in the life of the mind as practising scholars live it, and how might this role best be carried out? The effects of computing may easily be overemphasized, and often they are, but we have good reason to suspect that fundamental changes are afoot. *Mediation of thought by the machine*: From the beginning it has been quite clear that humanities computing is centred on the *mediation of thought by the machine* and the implications and consequences of this mediation for scholarship. We are reminded by the cultural sea-change of which the computer is a most prominent manifestation, that our older scholarly technologies, such as alphabetic writing, the codex, and printing, are technologies, and that they also

---

[11] See footnote n. 10.

[12] On Willard McCarty's research see: <http://www.kcl.ac.uk/humanities/cch/wlm/index.html>.

[13] See footnote n. 10.

[14] Willard McCarty: *We would know how we know what we know: Responding to the computational transformation of the humanities*. 1999 Available: <http://www.cch.kcl.ac.uk/legacy/staff/wlm/essays/ know/>. See also: Koenraad de Smedt e.a. (Eds.): *Computing in Humanities Education. A European Perspective*. Bergen: University of Bergen 1999 (Socrates/Erasmus Thematic Network Project on Advanced Computing in the Humanities). Available: <http://gandalf.aksis.uib.no/AcoHum/>.

[15] <http://computerphilologie.uni-muenchen.de/jg02/orlandi.html>.

> shape our thinking. *Methodologies*: What jumps immediately into focus is the importance of methodologies. When you teach humanities computing, what immediately becomes obvious is that the only subject you have to talk about is the methodology. *Computing and the humanities not separated*: That computing and the humanities are fundamentally separate is an illusion caused by a lack of historical perspective and perpetuated in the discipline-based structure of our institutions. *Philosophical training*: In the broad sense, philosophical questions naturally arise out of a machine that mediates knowledge and whose modelling of cognition reflects back on the question of how we know what we know. Philosophical training would seem a *sine qua non* because of its disciplined and systematic focus on logic and critical thinking skills, as well as a concern with how to interpret diverse representations of knowledge, including what philosophers and literary critics jointly refer to as hermeneutics. *Computing not purely utilitarian*: The assumption that computing mimics what we already do, that it is purely utilitarian would mean that projects were thoughtlessly undertaken, software then written and put out into the field, but it seems that we can save much grief by prior thought about the questions we would want to ask. *The labour-saving myth*: We know this myth to be silly; we know that only the dull, unimaginative scholar would not be inclined to do a better job with the time liberated from mechanical. We also know that the computer does not so much save labour as change the nature as well as scope of what we labour at. *Research methods*: We must objectify our research methods before we can compute the artefacts we study, and in so doing we bring out into the open what has formerly been hidden from view. Part of the problem has been the attitude in the humanities by which the physical bits and craftsmanship of research, its technology, are relegated to a lesser status.[16]

Furthermore another attempt to define the impact of the new media on the reasoning within the humanities research is the one expressed by Roly Sussex.

> [W]hat is interesting about computational methods is that these methods are providing us with both a new methodology and a new epistemology. The notion of ›data‹ is undergoing a reworking. Humanists are learning to interpret statistical reports on what our software says the text is doing. This whole process is tending to bring some areas of the Humanities closer to questions of methodology in other disciplines, and indeed to make the Humanities more scientific.[17]

Finally the point of view of Manfred Thaller is included in Tito Orlandi's paper.

> We are dealing with methods, that is, the canon (or set of tools) needed to increase the knowledge agreed to be proper to a particular academic field. Computer science is a very wide ranging field. At one extreme, it is almost indistinguishable from mathematics and logic; at another, it is virtually the same as electrical engineering. This, of course, is a consequence of the genealogy of the field. Having widely different ancestors in itself, computer science in turn became parent to a very mixed crowd of offspring. The existence of this wide variety of disciplines, related to or spun off from computer science in general, implies two things. First, there must be a core of computer science methods, which can be applied to a variety of subjects. Second, for the application of this methodological core, a thorough understanding of the knowledge domain to which it is applied is necessary. The variety of area specific computer sciences is understandable from the need for specialized expertise in the knowledge domain of each application. The core of all applied computer sciences is more than the sum of its intellectual ancestors, which may themselves be inextricably associated with particular knowledge domains. If we accept the assumption that the successful application of computational methods strongly depends on the domain of knowledge to

[16] See footnote n. 10.
[17] See footnote n. 10.

> which it is applied, then we also have to accept that applying computational methods without an understanding of that domain will be disastrous.[18]

This first and very short summary, which has not included many other points of view for obvious reasons, can be concluded with the words of Orlandi:

> We conclude that it is pointless to teach computer science to humanities scholars or students unless it is not directly related to their domain of expertise. We conclude that humanities computing courses are likely to remain a transient phenomenon, unless they include an understanding of what computer science is all about.[19]

Another key element in the building of *Humanities & Computing* is that concerning the modelling and formalizing of problems in this discipline.

2.2 Modelling, sharing and re-using information

In the humanities research modelling is not a common practice. With the exception of logic and linguistics, where modelling has been a required method for a long time, in other fields such as history, modelling is resumed simply as the sum of the known methodologies. The difference between modelling and the latter consists in achieving or not a certain degree of formalism in representing knowledge. Why is it important to achieve a clear and easily understandable formalism? The answer to part of this question has been given by the above quoted sources. Formalism should be intended as a shareable modelling method that enables humanist scholars to pursue the building of a representation of knowledge in a machine readable form. This means that independently of whatever the field of research or study is, it will be possible to represent the semantic structure of objects and/or concepts in a shareable way. In order to achieve that, likewise the *Text Encoding Initiative*, the debate on how the modelling should be carried out and what kind of targets should be reached at first should focus on the generalization methodology. What should be shareable at first is this formalism. There are at least two ways to understand this formalism: the first, the mark-up structuring, that is a grammar that represents both the *nodes* of any text and the rules of how you could use that grammar; the second, the logical abstract data structuring, that is the meta-language environment in which information in natural languages is its semantic population.

Willard McCarty gives a significant definition of the mark-up: »So far I have been using the term ›mark-up‹ broadly to denote the act and product of recording a textual entity in computationally tractable form. In the more usual, specific sense, what mark-up languages primarily have to offer to my research is proximity to the source text«.[20] Language analysis and morphological tools are nowadays available for linguists, semiologists and philologists. This is because their primary sources are mainly texts. No equivalent effort has been done for *Documentary History* because the abstract modelling has suffered from its ancillary position within historical methods and studies. To build a

---

[18] See footnote n. 10.

[19] See footnote n. 10.

[20] Willard McCarty: *Depth, Mark-up and Modelling*. In: *Computing in the Humanities Working Papers* 25 (2003). [Jointly published with *TEXT Technology*, 12 (1, 2003)], p. 3. Available: <http://www.chass. utoronto.ca/epc/chwp/CHC2003/McCarty_b2.htm>.

modelling tradition in the Humanities, a more abstract and generic language is needed. To quote John Lavagnino: »In […] humanities-computing projects, the theory that matters most often turns out to be new theory developed to support the work of modelling: the existing system of definitions in many fields proves to be inappropriate for the kind of work that computers can do well, in which things need to be definite and discrete to an extent that isn't usually necessary for human readers. It is not so much that the theory behind the activity of modelling takes precedence, though we do find that we need to take that into account; it is that we need to rework our fundamental approach to the field in order to create data that computers can work with«.[21] Modelling in the humanities should be understood as a generalization method. This means that whatever object or concept and whatever relation is to be stated between those objects or between those concepts or moreover between objects and concepts should be representable as a logic flow in a model. If such a definition is accepted, then the method of generalization should focus on how it is possible to gather from the different disciplines non ambiguous taxonomies that define semantically all properties, attributes and relations of objects/concepts. Taxonomies express the way in which different disciplines organize/represent their knowledge and are specific to each study/research field. They are a treasure for those who want to build computational environment models for the humanities. At the same time for each field of research these taxonomies represent the uniqueness of each area of study, in other words, they represent the personality of each discipline. Therefore there is a difficulty to overcome, when attempting to build upon these taxonomies a generalist formal model. Some hypothesis how to overcome these difficulties will be discussed in the next chapter. The main problem is apparently the self image that each discipline in the humanities has developed and the prudent conservatism that characterize the way these disciplines explore the computational world. Manfred Thaller has described such an attitude as a *timidity* of the humanities: »After all, the Humanities have a very long tradition in the usage of complex, fuzzy and vague information, which is extremely relevant in overcoming the information glut much complained about – much more so, than the elegance of purposefully produced information as processed by our colleagues in the hard sciences. That the Humanities in general, are much too timid at the moment to claim their proper relevance for the solution of the problems of an information society is something the confessing Humanities' computer scientist can only diagnose; he can not be required to share that timidity«.[22]

3. Understanding the approach: examples from the History of Sciences

3.1 Why History of Science?

History of Science is a relatively young discipline and includes a great number of specialized fields of research such as for example: mathematics, astronomy, physics but also art history and sociology.[23] In fact if a certain project has as its target to reconstruct the life

[21] John Lavagnino: *Forms of theory: some models for the role of theory in humanities-computing scholarship.* Available: <http://www.uni-duisburg.de/FB3/CLiP2001/abstracts/Lavagnino_en.htm>.
[22] See: <http://www.uni-duisburg.de/FB3/CLiP2001/abstracts/Thaller_en.htm>.
[23] A more detailed list of disciplines that underpins the History of Science can be found in the *Cumulative Isis Indices.*

and works of a given scientist it is likely that the required primary sources for such research will be of many different kinds. Some examples from the projects *Panopticon Lavoisier*, quoted at the beginning of this article, could help to explain the problem. Here are analysed objects that stretch from mineralogy to the history of art, chemistry, printed and manuscripts texts, scientific instruments and life records etc. The variety of disciplines, that pertain to the study of all these kind of objects, have their own harvesting, cataloguing and study methodology. How is it possible to set a common index among these objects? How is it possible to let specialists in all these different fields work with a formal input model and at the same time enhance the semantic diversity of each object typology? The answer to these questions can be found by rendering an abstract generalist model of the environment where humanistic data are produced, edited and displayed. Currently data coming from some disciplines, such as bibliography, are *well formed* only when there formal data definition and storage satisfies the given national record model required for those objects. For example: the art historian, at the level of the prime record cataloguing, uses a typification of objects recognized by national/institutional taxonomies for art objects; an historian of mineralogy does the same kind of work when classifying his items and so on. These different controlled vocabularies, known as disciplinary taxonomies, explain the required descriptive attributes for any kind of object that should be classified.

The following examples[24] will show how different the methods of cataloguing different objects are. *Pinakes* has been applied mainly in projects concerning science history. Therefore in order to understand the methodology applied and the theoretical background of this application the samples that will be used here come from its implementation environment.

3.2 What does a multi-morphology documentation *imply*

The seven different objects listed hereunder are chosen to show what kind of research problems would arise if such documentation had not been indexed homogenously. Moreover, the following examples have been chosen to explain what methodological impact the generalization of multi-morphological objects description has in the structuring activity of documentary data. The objects in question are:

[24] These examples can be found on the web at the address: <http://www.pinakes.org>, under the chapter »Hosted projects« in the Panoprticon Lavoisier Project. Go into the project and start from the menu »Iconography«. Choose David's portrait and follow the relation navigation system.

Portrait by David of Lavoisier and his wife – a painting portraying Lavoisier with some of his scientific instruments and his wife – detail of an aerometer.

The aerometer/collection of aerometer – An instrument for measuring the specific gravity of fluids; a form of hydrometer.

The manuscript with the first drawing of the aerometer and explanation of its functions.

The first drawings for the publication of the articles on the aerometer in the *Memories* (part of a manuscript).

The transcription: the digital text.

The printed article in the *Memories*: the printed text as image and its transcription.

Bibliographical reference/s: Marco Beretta, Imaging a Career in Science. The Iconography of Antoine Laurent Lavoisier, Canton (Mass.), 2001, xvii + 126 pp. Ill.

The chronology event recording the invention and use of this instrument (1768): Lavoisier guides the construction of several aerometers; Paris.

Pyrite arsenicale
Normalized text: Echantillon minéral
Comment: Le numéro 9 bis correspond au catalogue manuscrit Lavoisier (inventaire 10, 2ème genre: »Pyrites arsenicales marcassites«). Cet échantillon fait partie de l'échantillon n°9: Quartz gris pénétré de pyrite arsenicale en poussière, veiné de spaths calcaires blancs écailleux.

The fact that these objects require different cataloguing methods is not only due to the their semantic and morphological difference but depends also on the history that has codified these descriptions into different cataloguing traditions. Especially at a national level, but in some cases at an international level, compulsory descriptive patterns have been accepted as common practice. The best way to miss the target of reaching an overview of all the cultural heritage holdings of a country is to assign this task to a ministry. Take for example the case of the first object on the list: a painting and its details. To describe such an object the Italian state requires the completion of a form containing 79 fields for each element catalogued. Seeing the density of the of Italian cultural heritage, no worse choice could have been made.[25] The specialization of the description should be a subset of its generalization as a logical entity. All objects here above listed have an individual history and a combined one. How is it possible to index them in such a way that by finding one of them it is also possible to reach in a »narrative« way the others as well? In other words it is necessary to identify which attributes of these objects are shared among each other and which are not. Thereupon we need to analyze the existing taxonomy of each discipline that has been studied and use that reference terminology in order to benefit the existing researches on these objects. If no taxonomy is given, we need to implement one based on the self evident expressivity of the natural language. Subsequently we need to decide in which way relations should be declared taking into account the rules of the relational theory. The latter defines relations as predicates that can be *active* or *passive* depending upon the position in which the members of the relationship are. This preliminary analysis, carried out to build *Pinakes v. 2,* brought to the definition of all the minimal attributes that physical/logical and semantic objects always have. Some of them are compulsory others not. In this way, at a primary record level, the problem of a multi-morphology was bypassed but not solved. In *Pinakes v. 2* the logical presupposition was that all information could refer to two main areas: *Object* and the *Component*. *Object* can be a physical one such as a book, an instrument, a building etc. but also abstract objects such as the chronology of a scientist's life. In this way, in the case of a physical object, its material and location information could have been sto-

[25] See the Italian standard at: <http://www.iccd.beniculturali.it/download/OAC.pdf>.

red and for each object there would have been a *Component* i.e. a description or a title etc.; in the case of a logical *Object* – such as the container of the life line of a scientist – the only required information would be its type (life-line) and the *Components* would be all the events on that life line.

The schema – here simplified – is:

a) the Object attributes are:

a. it is of certain type: type *Required* (list of types from multidisciplinary taxonomies) (**index**);

b. has a natural name or has a shelf mark/inventory number which is of an institution (public, private etc.) in a given collection (**indices**);

c. is made of certain material, and parts can be of different ones (**index**);

d. relates to other objects in a certain way (ex.: is part of, contains etc.) (**index**);

e. has physical dimensions (if physical);

f. contains a number of parts (pieces, folios, pages etc.)

g. has at least one component (without a semantic definition no object can be defined) *Required;*

b) the Components attributes are:

a. it is of certain type: type *Required* (list of types from multidisciplinary taxonomies) (**index**);

b. has a text that can be a natural text, a description, belongs to a corpus and therefore has a normalized text (text taxonomies); one of all these is *Required;* without text no *Component* can be defined (**index**);

c. can have one or more Persons related with different responsibilities (index);

d. can have one or more city (**index**);

e. can have (within the object) one or more positions (such as folios, pages, locations different from urban area);

f. can have one or more languages (**index**);

g. can have one or more digital resources (**index**);

h. can have one or more subjects (discipline, topic, argument from subject indices of multidisciplinary taxonomies) (**indices**)

i. has to have a time definition independently of the known granularity of this information (from the range c. XII- 1st half to dd/mm/yy) and the time has to have an attribute defining to what that time is referred within the *Component* (print, discovery, delivery what ever – these values are taken from the known taxonomies of attributes definition of time within history) (**index**) *Required*;

k. can relate to one or more components in a certain way;

l. can belong to one or more macro family (meaning a defined group such as »the bibliography on Lavoisier«, »the Iconography« etc.) (**index**).

m. can have one or more text transcriptions with attributes are: person (responsibility) name, city, subject, digital resource (**index**).

Pinakes v.2 is able to tell the following: that to understand one object it is necessary to look up all the others. In other words there is a semantic nucleon from one object to

another that will narratively explain where the relations are driving the reader. This navigation sets all objects in a row and they are visible two by two so that what is told by means of the relationship declaration explains at the same time the unity of the group and the logic of the navigation. An illustration is the following: There is a painting that is a portrait of the Lavosier couple in which some instruments have been represented in a given order. These instruments are not there to fill – as has been said – the figurative space of the portrait. They represent, on the contrary, the career of a scientist who has discovered and built a certain number of tools to find out the basic laws of modern chemistry. But how do we know that? Because there is an instrument located at the CNAM in Paris that has been identified as Lavosier's and there is one of his manuscript in the library *Academie de Sciences* where the instrument is sketched. Moreover, there are drawings that have been made for the printed edition of some of his *Memories* and these can be found in the first national French edition of Lavosier's works. The text, describing the functions of that instrument and the digital copy of the works, together with the studies done on it – quoted as bibliographic records – explain the unity of all these objects that once were located in the same place and therefore belong to one event: the chronological event describing the creation and thinking of the instrument. Other reconstructions are possible. *Pinakes v. 2* is able to represent an unlimited, but semantically controlled, number of relationships between all these objects and gives even the possibility of telling contradictory or conflicting histories in the same environment. Such contradictions are no any longer a matter of the computational model: they are the result set of the semantic population using its logical features.

Nevertheless, there are some limits to such features that should be explained. In *Pinakes v.2* the objects of the real world are still represented in tables. This means that, given the analysis required to build this application, a high formalization of the dynamic between attributes and objects has been partially reached. In fact the generalization used by *Pinakes v.2* limits the possibility of distinguishing the description of both the physical/logical objects and of the components. Still the possibility of setting, under index constraints, different objects by means of common attributes has set the initial logical background for a further development of this application. This means that, despite the lack of specialization of both the physical/logical object and the semantic one, different objects can be grouped by the given set of attributes. This, together with the explicit declaration of relations gives the possibility of building a very large subset of »narrative« interpretations of concepts and objects of the documentary history. The use of a declarative method for the relation definition creates a »library« of predicates that can be reused in order to navigate also by means of how objects/concepts involve each other. Such structure offers the possibility of recreating entire collections whose objects are today scattered around the world.

3.3 Object displacement and reconstructing collections on the web

In the presentation of Panopticon Lavoisier Marco Beretta briefly sketched the aim of the project: »Panopticon Lavoisier aims at creating a virtual museum of the collections of the French chemist Antoine Laurent Lavoisier (1743–1794) scattered throughout the world. A detailed chronology of Lavoisier's life and works, the catalogue of Lavoisier's manuscripts (ca. 6000 items), laboratory apparatus (ca. 500 items), library (ca. 3000 i-

tems) and minerals (ca. 4000 items), the digital edition of Lavoisier's collected works, the bibliography on and of the French chemist (ca. 2000 bibliographic records) as well as his complete iconography are integrated in one relational database, *Pinakes*, and made available to remote users.«[26] Each collection can be investigated alone but the relations established among the members of each collection offer the possibility of reaching an overview of the whole. This method has been applied also to other scientists and can be expanded to any typology of objects. The focus is not on tailoring digital reifications of the real world objects but on offering the possibility of reconstructing their cultural and ›sense‹ contexts by means of an architecture of semantic defined relations. It has been above mentioned that *Pinakes v.2* has a fixed number of possible attributes for all kind of objects. This on the one hand, from the formal point of view, is a limit; on the other hand such strong definition results in it being very efficient if a large number of projects are using *Pinakes v.2.* By setting index constraints, all *Pinakes* projects can be indexed crosswise. This implies that relationships can be created between large subsets of data from different projects offering a navigation not only through the morphological diversity of one project but also through all the different conceptual architectures of all available ones. In this way the reconstruction of a collection can turn into the reconstruction of broad historical contexts whose study involves many scholars. The agreement on the rules imposed by the use of *Pinakes* offers everyone the same methodology and shared values to define: objects attributes, persons, dates, cities etc. so that all indexes are formally one. All relationship type definitions are stored in one place and can be re-used in different projects.

3.4 Sharing disciplinary taxonomies and international structuring/encoding standards

The definition of taxonomy in the Oxford Dictionary is expressed as follows: at first »the study of the general principles of scientific classification« and then »orderly classification of plants and animals according to their presumed natural relationships.« Any attempt to search on the web for expressions such as »humanistic taxonomy«, »taxonomies in the human sciences« or »taxonomy modelling in history« give no answer. Whereas if the search uses expression such as »shared taxonomies« the answer will return a list of many projects that support the idea of building their terminology heritage to achieve common descriptions of experiments, objects and phenomena. It is not true that human sciences, because founded on the natural languages, cannot effectively assemble common taxonomies. As in the natural sciences, specialized taxonomies which share a dictionary of terms and type definitions ensure the possibility of cross-indexing many different fields of research and open up a comparison strategy among scientific results;[27] the human sciences should focus on this problem.

To start building all of this is a very considerable effort and so much so because cooperative research in the humanities – with the exception of sociology, anthropology, linguistics and some others – is a very recent phenomena. Secondly, within the discipline of text studies and history there is very little space for sharing due to the very nature of the enterprise which sets authorship in the centre of research activity. Sharing

[26] See: <http://moro.imss.fi.it/lavoisier/entrance/projbox.html>.
[27] See: <http://www.dsi.dk/projects/cpp/project.htm>.

implies compromising with the notion of authorship. In the existing humanities such a compromise could be fatal. Some attempts to create a common terminology through disciplines has been made. The *RAK* and *RAK-NBM*[28] experience in Germany has been a forerunner in the field for years. Nevertheless, much effort is required to diffuse the idea that the disciplinary lexica are the key knowledge tools for the communication and representation of scientific results. This is even more true for history disciplines because a great deal of ambiguity is still present in the attempt to give a common definition of the basic attributes of objects. The main target should be to achieve a generic model of taxonomy metadata producing the possibility of building common repositories such as that of the Getty-TGN[29] in order to develop, for example, an ongoing and shared list of historical people or scientific instruments name list etc.

Moreover, in this way it would be possible to achieve in the humanities a common methodology of expressing time ranges, sharable within a chronological concept definition of non-specific-historic disciplines. These achievements can be reached if the urgency of standardizing *descriptive data* is understood. Unfortunately nowadays such urgency is not always – and not only for financial reasons – recognized. This problem is closely connected to the international standardization of digital born data. In recent years many projects, like the *Text Encoding Initiative* (TEI)[30] or the *Open Archive Initiative* (OAI)[31] together with the DCMI (*Dublin Core Metadata Initiative*)[32] commission with its successful attempts to standardize the minimal requirements for interoperable online metadata standards, have been the avant-garde in the field.

The DCMI's activities that have particularly supported consensus-driven working groups, global conferences and workshops, standards liaison, and educational efforts to promote widespread acceptance of metadata and practices have made, together with the WC3 commission, the largest contribution to the basic assessment of the standardization problems. These initiatives have introduced the problem of the construction of *standard descriptive attributes* which leads to a second level of the taxonomy problem. On the one hand there is the need to reduce the ambiguity of semantics. On the other hand, to succeed in such an enterprise, it is necessary to have logical descriptors that generalize that ambiguity into a super-semantic formal group of tags. This implies that the descriptor is a normative type, like that expressed in a natural language, becoming in this way a type descriptor *per se*. This re-establishes the problem of the semantic definition at the level of meta-data without suggesting a solution for the basic data ambiguity, that is, re-introducing a problem of taxonomy within the environment that should store the *taxa* i.e. the data as such.[33] In other words the entire logical architecture that

---

[28] *Regeln für die Alphabetische Katalogisierung* von Personenamen, etc. and *Regeln für die Alphabetische Katalogisierung von NichtBuchMaterialien*.

[29] See: <http://www.getty.edu/research/conducting_research/vocabularies/tgn/> that is a Getty Thesaurus of Geographic Names – Online. In this site there is a good exaple how to offer a shared txaonomi of geographical references. For example by searching the city name »Augusta« the answer will be »Augsburg«. This name will be geo-referenced by means of a hierachy able to display all historical names of that current one. For each one an historical profile is furnished so that a full identification of the name reuqeste can be retrived.

[30] See: <http://www.tei-c.org/P4X/>.

[31] See: <http://www.openarchives.org/>.

[32] See: <http://dublincore.org/>.

[33] See: <http://chicagoschoolmediatheory.net/projectstaxonomy.htm>.

should manage the raw data, its definition terminology, is submitted to a taxonomy.[34] This should develop a shared vocabulary of functions, methods etc. for all possible environments. The attempt in *Pinakes v.3* is to overcome this loop by establishing a logical model that is both independent of the terminology in which the model is expressed and applicable to any set or subset of raw data information coming from the *real world*. This choice is based on the presumption that at the moment the semantics of the natural languages – that is what historians deal with – is not computable but in essence can be logically represented by means of a generalization model.

There are two different approaches: data structuring with arbitrary names of logic units displaying the »natural« (existing) taxonomies and the encoding (tagging) data structuring that attempts to take control of both the semantic of the tag – its hierarchy – and that of the data. Still the tag »<author>« that requires a person's name and implies the responsibility of that person to a given object, does not necessarily imply that the used name is that of the author and not that of, for example, the editor which has an other type of responsibility. Implicative actions are computable only at the level of a logical flow not semantically. The procedure »if a person's name is set then a responsibility is required« can be automated; the procedure »if the person name set is ›Isaac Newton‹ then his responsibility as ›author‹« cannot be automated. In order to succeed in automating the latter would be needed to have already digitalized all information concerning the works where Newton is an author as well as the works where he has other responsibilities. If such information would be available there would be no need to automate information.

4. Pinakes *current model and its functionalities*

4.1 The first generalization: definition of the physical/logical and of the semantic objects

The generalization is the foundation methodology of building a relational database. It is within the generalization model that the relation between entities is built. The entities are logic units that in a relational database can be represented by a table or more tables. The relations are also logic units that in a database exists in a table and that can exit between tables. The generalization as method is expressed as follows:

- given the entity E, known as entity *father*, and one or more entities E_1, …, E_n called *daughter* entities, of which E is more general, meaning that includes E_1, …, E_n as specific case.
- in this case it is said that E is a generalization of E_1, …, E_n
- and that the entities E_1, …, E_n are specializations of the entity E.

In *Pinakes v 2* and *v 3* there is a common strategy of generalization. The difference between the two consists mainly in the formalization approach to data storage (more later on this). The idea is that – even abstracting from the real structure of a database – any experienceable/knowable physical/logical object, beyond its natural attributes, has to have a description (a simple or complex text) or a name. The generalization model used

[34] The use of UML (Unified Modelling Language) would help to avoid the need to introduce natural languages definitions which eventually could compel the growth of ambiguity rather then deploying a shared functional definition. See: <http://www-306.ibm.com/software/rational/uml>.

in *Pinakes v. 2* applied to the *Panopticon Lavoisier Project* could be explained as follows: David's work is a physical object *of type* painting. As physical object it has a number of attributes from *height* to *bright*, to *executing technology* etc. As semantic object (what was called previously component) is of *type portrait*, has or a *description* or a *title*. This semantic object has, as well as the physical one, a certain number of attributes such as – the person who has the executed the work, or who made the frame and or again who furnished the paint – and many others such as the historical period in which it was painted and who or what was depicted etc. So where is the point? By setting as the main entity the physical/logical object it is implied that all semantic objects thereon dependent are a specialization of that entity and that this is its generalization. If this is accepted then it should be accepted also that the definition of this relationship is itself a generalization of all possible representations of relation between objects of reality and their known significance. Why separate the two? There are cases where the known name of an object does not explicitly explain its semantic density. In other words, take the case of our portrait. It has a number of details which should be explained. These are not of physical nature but belong to the raffiguration as a whole. Each one does not have a title (meaning a natural self explicit name) and therefore requires a description. By creating two different models that can have their own specialization it has been made possible to reach – if needed – a very fine granularity of the information concerning both the physical/logical and the semantic objects. This granularity offers the possibility of setting into relation all types of objects belonging to the same semantic range. That means it is possible to set into relation a physical/logical object with another one, as well as being possible equally to do this operation between semantic objects. The main generalization model established a formal separation between groups of attributes/properties that define if an object can belong to the physical/logical or the semantic class. The generalization of the main class established that – given a minimal number of attributes/properties – each object can undergo an attributes/properties specialization which depends upon the information available. This specialization should follow the taxonomy of the field of origin. Moreover each relationship explains the logical position of the relation members by means of a transitive predicate. This means that in the given example the instrument represented in the painting *is a detail of* the portrait and at the same time *is a raffiguration* of a real instrument located at the Museè d'Arte et Metiers in Paris. This *was thought of and drawn* in one of Lavosiers's manuscripts, and so on. For the logic model all physical objects are at the same level as well as the semantic ones. The relations among objects (physical/ semantic) is established on the basis of the historical interpretation that the author of the project is able or wants to express. The index of both are independent from the typology of the relations. The relations are the narrative index of the raw phenomena there testified by the objects and their significance.

This procedure follows the given properties of all entities represented in a database:

- Every occurrence of a *daughter* entity *is* an occurrence of the entity *father*. For example: the occurrence of a text (name or description) in the semantic object is always at the same time an occurrence of the physical object. In other words no physical object is known without a name or a description.

- Every *property* of the *father* entity (attributes, identifier, relations etc.) is also *property* of the *daughter* entity. For example: the shelf mark (attribute) and owner institution (attribute) of the physical/logical object are also attributes of the semantic object/s. This property is called *inheritance* property.
- A generalization is considered *total* if every occurrence of the father entity is *at least* an occurrence of the daughter entity otherwise this generalization should be considered *partial*.
- A generalization is considered *exclusive* if every occurrence of the *father* entity is *almost* an occurrence of the *daughter* entity, otherwise this generalization should be considered *overlapping*.

Therefore in *Pinakes v. 2* and *v. 3* the generalization should be defined *total* because there is no experienceable/knowable object that eludes a semantic denotation. The generalization should moreover be considered *exclusive* because there is no semantic denotation which is not given a physical/logic object. The generalization method expressed here is that of the prepositions theory which consists in finding for each given preposition a second that contains the first as one of its particular cases.

4.2 The second generalization: definition of required minimal generic attributes for objects in Pinakes v. 3

The second generalization is the result of a transformation. The *overlapping* generalization can be easily turned into an *exclusive* one. New entities have been added representing the intersections of the overlapping ones. This analysis started to verify which entities could represent the mandatory attributes of the physical/logical objects and semantic ones. The information granularity that physical objects can have by means of its description is representable as follows. A physical object of a given type, height, brightness, depth, is located somewhere, belongs to someone, is made of a given material. These minimal requirements, given the example, can be found in all the objects quoted. Among these requirements for physical objects, only two are mandatory: *is of a given type* and *belongs to some one.* This is because a physical object, no matter if public or private, if existing or destroyed, has (or had) an identification and is (or was) placed somewhere. Independently of the object type these attributes are always given. Also the semantic object has some minimal attributes. Among all other attributes (such as person, place, position etc.) there is one that is mandatory. This is time. A special chapter should be devoted to such attribute. In fact, even if apparently obvious, time is the only attribute in historical studies that gives gnoselogical foundation to all the possible information in the field. Without time (a precise date or a span of time) information in history does not represent knowledge of a given phenomena or object. Therefore, for the semantic object – required of any physical/logical one – the only chance of being retrievable and acknowledgeable is to be chained to a time line. The ambiguities of timeline representation (calendars) do not make such a mandate easy. On the contrary they introduce a fundamental variable that can be cleared out only on the basis of a larger knowledge of time representations conflicts and time representation intersections.

A question could be raised against the assignment of the time attribute only to the semantic object: »Is the physical/logical object not time defined?« A possible answer could be that the physical/logical object does not have a ›native‹ time attribute without

a semantic description of its value/function/significance etc. Therefore at the level of the attributes, time definition does not relate to the object as such but its required declaration of significance i.e. the semantic object. This might be more obvious than apparent if the fact that history is always a semantic view is taken into account – in a manner of speaking – of the real objects acting in its narration. In fact history does not deal with the objects as such but with their representation communicated throughout the telling that is *per se* a semantic unit. It is such unit to which history refers and it is such unit that needs to be set within the time representation in order to gain sense. Yet at first its identity is time, not sense. In other words any semantic unit has sense only within a time frame.

There are some implications in all of this. At first, due to the fact that there are different physical/logical objects which eventually require different sets of descriptive attributes it is clear that the storage flow should be managed from the data modelling. It is also clear that an abstract model should not represent all possible reifications of known classes of physical/logic/-semantic objects but their logic flow. This implies that the method of specialization of these two super classes (physical/logical and semantic objects) should be carried out by the deployment of logic control sets into the models. This allows the management of the predictable needs of specialization of the model. There are two ways to design the architecture of such a dynamic. The first is to introduce a hierarchical system of logic nodes that specify the attributes of each object. The second is to manage the object model as made from type descriptors and value descriptors. The first introduces a sub model for each specialization, the second represents it by means of typification. Practically, the first method needs to introduce new tables to represent the specialization of each new attribute of the objects. The second needs to introduce new values in a given table describing the model of the object. The advantage of having a logical management instead of a direct representation in tables of all possible values is that: to any time and rate a new specialization of each attribute can be added without the need to introduce new tables in the database. The latter, in fact, implies the use of a forbidden practice called »alter db« which is the setting out of synchrony all interfaces working and representing data for that database. It is possible to introduce, at the interface level, a procedure that checks the synchrony and thereafter produces an input form and output method to co-ordinate the two by means of including new tables and relations. But the development of such interfaces is not predictable because in the database there is no logical description of their growth. For this reason, to allow such a practice would mean introducing an heuristic method of growth that on large data risks being unmanageable.

The definition of this dynamic and the creation of its logical representation is what has been called here the second generalization, in *Pinakes v. 3*. This generalization, independently of all attributes of all objects, states a logical relationship between two super classes representing the basic information required to define and to communicate all objects of the *real world*.

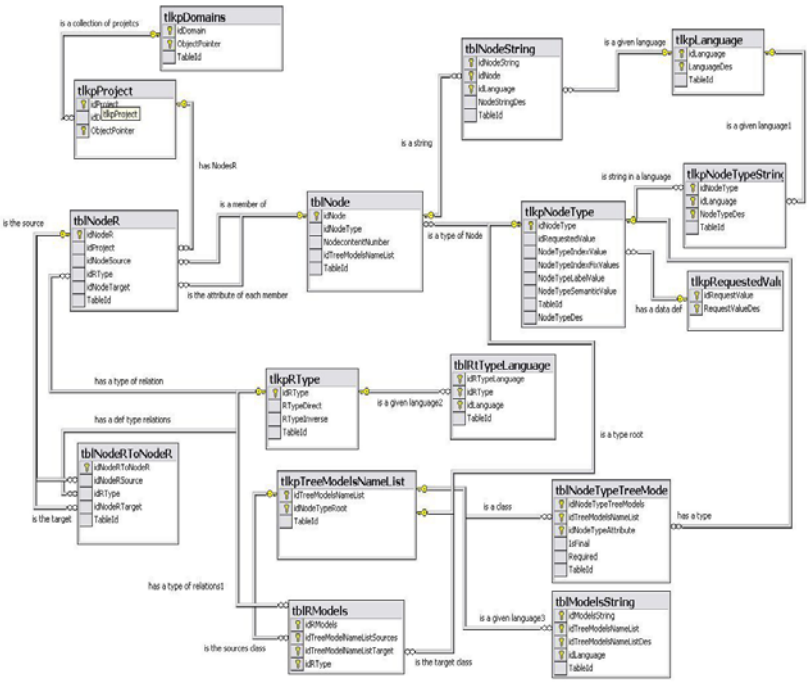
4.3 Defining the models and their method of specialization

The modelling itself is the attempt to formalize the ways in which data belong to each class. Each model has a name, a definition of a given set of fields with a required defini-

tion of value. Theoretically, it is possible to introduce for each kind of attribute a new model. The consequence of such practice has been explained above. *Pinakes v. 3* accepts only a specialization of value members, not of new representations (i.e. new tables). Therefore to characterise a model it is necessary to have the nomenclature for each new piece of information that needs to be added to an existing one. This means that into a model should be added the name of a field and the definition of its corresponding value. In order to have control of such procedure the values defining the field name and those defining the value type should be derived from existing taxonomies of the different research fields. Once the new attributes names and values have been defined, it is necessary to declare if a field is an index or not, and if a field is required for the entire model or only for a sub set of information of a given object type. The declarative method chosen here of building the models and their specializations implies a coercive connection with controlled vocabularies (disciplinary vocabularies) which are the only source for semantically defining the value and name of each class member.

The method of modelling specialization is theoretically a different formal set from that concerning the storage of data. Normally the logical abstraction and the data storing strategies belong to two different management phases. In *Pinakes v. 3* the logical abstraction manages the storage methodology and the model formalism at the same time. The method of specialization of the models is made in such a way that the existing taxonomies could represent and communicate their objects through the logic tools of this environment. This does not imply that such an environment would not be able to describe objects not already described by the disciplines that study them. This would be true if each discipline had accepted, or more accurately expressed, had built not a local but a shareable taxonomy.

If some disciplines have been successful in building such taxonomies then it should also be recognized that their impact within a great number of research fields in the humanities has been very poor. In history, writing is still a matter of taste and individual interpretation. The choice regarding the method of describing objects of the *real world* is also individual. Nevertheless the *topic* of each discipline is very well recognizable by means of the tradition that expresses a given method. The population of data that eventually would be managed by *Pinakes v. 3* would have the advantage of being driven by such traditions and not by arbitrary ways of naming and defining object descriptions. Supposing that such problems could be solved by methodological agreements within the humanities – which is hard to believe – then the general and peculiar taxonomies would at most be the key skeletons of arguments.



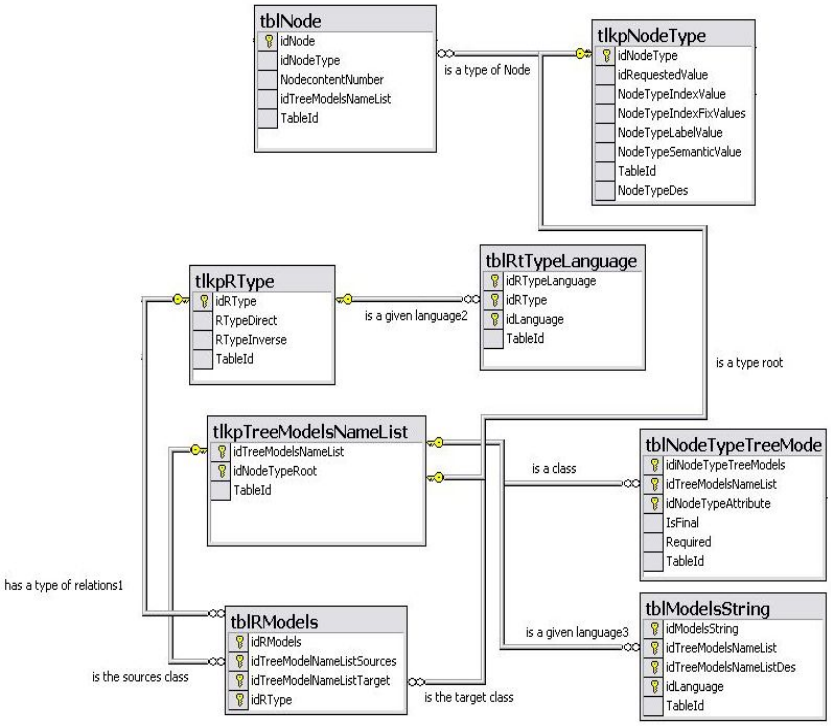
General logical schema

This would be a schema of concepts and thematic arguments which *per se* could help to standardize the resulting sets of historical research. At the same time such a skeleton would work as a primary semantic background which denotation would be of significant impact only if it supported the building of strong generalization models. A proposal for such a generalization can be given here. The table and relation schema adopted in *Pinakes v. 3* generalization methodology store both data concerning the model definition and the semantic information.

4.4 How and what data are stored in the schema: the data flow.

The above architecture represents the schema of the metadata and data management/storing. In order to understand the models and their method of specialization we need to analyze in detail the flow that creates the models. The following illustration shows its logical engine:[35]

[35] This series of numerals should be not understood as a »sequence«. They are meant exclusively to identify a table within the flow.



Schema 1: View on the model and its structure

Legenda of the schema 1:

(1) *tlkpTreeModelsNameList*: is a table with unique values where all models are defined by a name representing a given class of attributues of the physical/logical/semantic object;

(2) *tblNodeTypeTreeModel*: is a table where all attributes i.e. fields name are retrieved in order to determine which kind of fields are required to populate a given model;

(3) *tlkpNodeType*: is a table with unique values where all types of strings and numerals (integers) have to be defined by means of an international or local taxonomy;

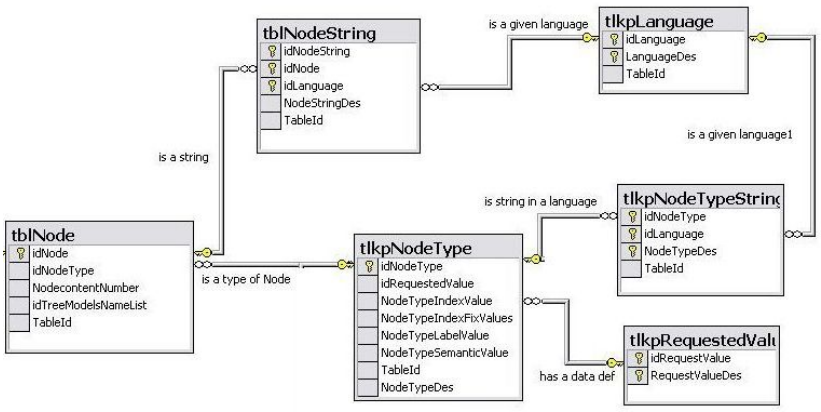
(4) *tblNode*: is a table where all strings and numerals (integers) concerning all information (raw data and meta-data) have to be stored.

(5) *tlkpRType*: is a table with unique values that stores all relation definition and/or declaration concerning both the meta-data and the raw data population;

(6) *tblRModels*: is table where all relations between given models are declared using a type of relation stored in (5).

These tables (1, 2) perform a descriptive and controlling role upon the raw data. This means that before storing data – semantic data into the database – it is necessary to declare what kind of object/concept is to be described there. This implies that there is a superimposition of the meta-data on the semantic ones. Such superimposition is by the way, common practice in the humanities. In fact, in humanities data are at first descri-

bed as reference and then argued, not the other way around, meaning that their formal status is the guarantee of their semantic validation in given contexts. Similarly, the database accepts new raw data (3, 4) if these are successfully validated by the logical engine i.e. the models. But how does raw data come through? In *Pinakes v. 3* we have decided to store all raw data as alphanumeric strings and/or numerals (integers). So that in order to have a string that is used as descriptor within (2) we needed to store that data in (4). But to store that data in (4) it is necessary to define its type in (3). Once these data have been stored they can be retrieved in (2) to the group of attributes of (1). In other words, to be able to define the attributes of a given model (here called Tree) it is required to have stored their descriptors in (4) which allows such storage only if these data have been defined by a type in (3). Each model has to have an explicit type declaration of its existing relationship, if any, with other models in (5). The relationship is always expressed with a transitive verb which has to enforce the action declared in the predicate (here declared as: »has a type of relation«) connecting two tables. This information is stored in a table with unique values and used in the table where two models are set into relation in (6). Given the definition of the name of the attributes and their values (for each model) and given the definition of the relation among the models it is possible at any time and any rate to add new attributes to them. The logical flexibility of this model is independent of the semantic density of the data population. In such way it is given the chance to determine both a self-defined metadata structure or to follow that offered by the international standards. Normally the data flow is never presented within the discussion of the logic model. Nevertheless it would be of some help to understand in which way the quoted flexibility has been achieved and in which way the control over the formal data and over the raw data structure has been built and carried out. The control over the raw data storage is the main task of the database. Its structure and flow definition is hereafter described by means of its architecture:



Schema 2: View of the »Node«

In *Pinakes v. 3* the »Node« is the logical location in which all possible data are stored. To understand the granularity of such data one needs to abstract from the logic of the

information as experienced from the common humanistic way of communicating a description of a given object of the *real world*. The kind of information which the humanist is used to is that which represents the entire mass of minimal knowledge required to identify an object of the *real world* by means of a logical sum[36] of different typified strings and/or numerals. An attempt to translate such concept into, so to speak, *natural language*, would be: to know something about any given object (a description existing in a catalogue, the object itself or a bibliographical source etc.) it is necessary to have some minimal information, each part of which is defined as a type of information for that kind of object (i.e. the title string is the title not the content). These could be in the case of a manuscript: the title of the work, its incipit, its explicit, the compiler, the author, the date of compilation of the work, the date of compilation of the manuscript, the shelf mark of the manuscript which implies the name of the owner (institution or private) that has it located somewhere in the world. How is all this information stored in *Pinakes v. 3*? At first a model should be chosen to describe the physical and semantic information of an object or logic unit. How to do this has been explained above. But it could happen that not all fields (meaning attributes of our model) that are required for the needed description are available. Some attributes (meaning field names) may not allow the input of data values that are needed to describe the object? All these problems that are common to any application used within the documentary digital activity of the humanities have been solved here by means of the possibility of a self-definition of the attributes i.e. a specialization of each input form. The self-definition of attributes can be carried out only after having filled the values concerning the required ones of the chosen model. Using the example quoted above, the storage would have the following flow:

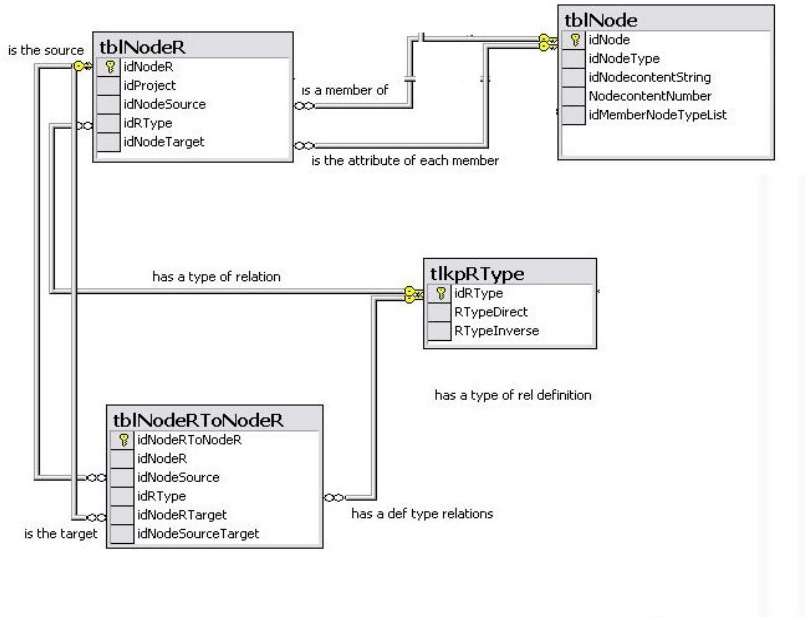
- choose the model concerning the *physical object*:
  a) define a type: manuscript (required – **index**);
  b) give a name of an institution and define the institution (required if there is a shelf mark – **index**);
  c) give the shelf mark concerning the object (required if there is an institution – **index**);
  d) give the dimension of the object and choose a unit of measure for it;
  e) give material of the object (the whole or/and different parts) (**index**)
  f) … etc.

- choose the model concerning the *semantic object*:
  a) same procedure and define the attributes needed …
  b) … etc.

The procedure for adding attributes is: give a name to the attribute (or better retrieve one from a disciplinary taxonomy), this implies compiling the table called *tlkpNodeTypeString*. Once the name of the attribute is defined it is required to declare, in *tlkpNodeType*, if this name is a semantic value, a label etc. When the name is of a field such string will be defined as *label*. Declare what kind of value should be accepted in that field (string or integer). Moreover if this attribute should be or not an index, if it is required or not and which model is using it. Having done that, the model to which this

[36] On this expression see: <http://www1.odn.ne.jp/~slc/algorithm_e.htm>. Normally the *logical sum* is a method used in executing condition IF in binary procedures. It is an addition of many signals by the *logical operator* OR opposed to the operator AND which result is a *logical product*. In the same way if we de-construct semantic information into atomic units in order to be able to re-represent it by means of indices the method can be called precisely *logical sum.*

new attribute is assigned will recognize it as a field of its input and output forms. The value of this field will be written (depending if string or integer) in the *tblNode* table (if number) and in the *tblNodeString* table (if string). This procedure is applied also to the *semantic object* model or to any other model existing. If the new attribute is a *structured information*,[37] such as that concerning the building of a subject index (Discipline, Topic, Argument), then there should be defined a new model instead of creating single attributes for a given one. The definition of a new model implies the creation of a set of attributes and a set of relations between them. That means to create a new sub form for a main one. Once each string and/or number (integer) is defined by a type means, this is the field name or the semantic value of a field (attribute) of a given form (model). In this way the most atomic semantic unit has been reached both at the metalanguage and natural language level.



Schema 3: View on the relation between nodes and the relations between node sets

By de-structuring all information into atomic parts the result is the creation of a multiple series of *n*-tuples.[38] Such series offers the possibility of navigating through different

[37] Dan R. Olsen Jr. defines structured information as follows: »Structured information consists of atomic items of information that stand on their own and compositions of information in various ways to form larger structures.« To understand what kind of logic object such definition is related – that involves great part of conceptual objects that humanities normally deals with – see: <http://icie.cs.byu.edu/ice/ StructuredInformation.html> and <http://icie.cs.byu.edu/ice/>. Here this expression simply defines any information that is atomic sum and that has one or more level of semantic implication needed to be an information. In other words, everything is structured if is to be defined by more then one single type of »Node«.

[38] *N*-TUPLE: This is a mathematical term for a finite sequence of n terms. For example, the set {1, 2, 3, 4} is a four-tuple. The set {Frank, Jane, Ed} is a three-tuple. Any time there is a list of n things in a certain order, you can think of it as an *n*-tuple. Detailed: The type $\alpha * \beta$, where α and β are of any type, is the type of orde-

cross-referencing indices and enables the recomposition of all information, thereby establishing a discourse throughout the relations given with other multiple series of *n*-tuples This procedure – called reification of *real world objects* – is controlled and stored within the schema here above presented. The Node (*tblNode+tblNodeString*) – single strings or numerals – are set into relation and written into the table *tblNodeR* where a definition of relationship[39] makes explicit, in a manner of speaking, who is who, in the representation of the information given. The table *tblNodeRtoNodeR* represents the relations among objects resulting as a construct that is the first semantic recomposition of their atomic units (*tblNodeR*). In this way the first recomposition (*tblNodeR*) identifies the object by means of single units and a predicate. The second (*tblNodeRtoNodeR*) identifies the given relations between objects by means of constructs and a predicate. The second level represents the narrative interpretation required for building any historical discourse and depends on an arbitrary set of predicates that can not be foreseen nor set into given taxonomies. The first level establishes the catalographic description of each object following the rules given by the discipline out of which the object comes and its predicates are the *taxa* of that discipline. The second establishes a net of semantic relations that cannot belong to the same set of information as the first nor can they undergo the same standardization. This second set is arbitrary and has an heuristic development. This means that the relations among objects, once given their semantic identity, can be managed only by means of logic-computational constrains but not by means of their significance.

5. Technological research and the Humanities – Conclusions

5.1 Comparing reasoning (1): technology as a library for humanities

The question now is: how is it possible that technological reasoning could come into a more than simple functional intersection with that of humanities? How does the one – the humanistic tradition – grow with the achievements of the technological reasoning, and how does the latter as well gain a broader sky-line by means of the problems set by the first? A full answer to these questions can not yet be given. Nevertheless an attempt to suggest a projection supports the possibility that within the near future the intersection of technological reasoning and that of the humanities could contribute to the creation of a so called »middleman« hermeneutics. Such expression attempts to define an area of knowledge that for the present is still fuzzy. The implication of such a definition apparently does not guarantee that the evolution of such knowledge will find a categorical foundation. But let's explain what is meant by »middleman«. Such a name comes

---

red pair whose first component is of type α and second component has type β. An ordered pair is written as (e_1, e_2) where e_1 and e_2 are expressions of any type. Similarly we can define n-tuple $(e_1, \ldots, e_n)$, where each expression e_i, $(1 \leq i \leq n)$ is of type α_i, $(1 \leq i \leq n)$ and each expression is separated by comma. The type of *n*-tuple is represented as $\alpha_1 * \ldots * \alpha_n$.

[39] In *Pinakes v. 3* the definition of a relation type implies the explicit use of the *direct* and the *inverse* form. So to speak, seeing that all relationships can be expressed only by transitive verbs, the direct relationship will be represented by the *active* form of the verb and the inverse by the *passive* form of that same verb. In this way it is made clear who is father and who is child in the relation representation. In other terms the declaration of the relation vectorial status clears the semantic of the relation but has no control on its sense.

from the world of security software and usually identifies an application whose function is to check what a member of a LAN is doing on the web. So much so, that this application is able to see if a given user by means of his/her navigation is ›sitting‹ in a wrong net place or not. If the behaviour of this software is a valid metaphor for describing the very nature of the technological misunderstanding of the humanities and the semantic misunderstanding of computing science is not clear. Nevertheless, the logical aim of such applications could be used, gnoseologically speaking, to represent the current state of the art: the one user who sits in the wrong place with wrong currency in front of the offering agent who is there with the wrong offer for what the user is asking.

The difficulty of defining Humanities & Computing has been described above. Apart from all the quoted problems, it remains to clarify whether there is a will to introduce technological thinking into the practice of the humanities studies and research, as well as whether technological studies, with the understanding of the humanities, could have an impact for the same development of that technology. These last notes do not aim to give answers but attempt to draw patterns derived from both fields respectively and thereupon paint a possible architecture of a given intersection.

When through the words of an historian a library of the 14th century is presented to a member of the IT community the latter has at first no perception of the fact that the texts there are in no way comparable to an article available on the web. The complexity of the information net that lays mute there remains not depicted. This twilight zone would have the same shape if the problem were observed from the other end. A humanist who finds himself reading what an IT specialist is expressing in the code written for an application is as lost as if the characters of that writing were in an unknown language. If the reciprocal exatraneity of these communities is so big then why attempt to set a »middleman«, a conceptual common *library* between the two? What the first has lost during his career of studies is the perception of the layers that time sets upon the reading and understanding of the facts and thoughts of the far and recent past. The work and the focus of an IT researcher is naturally projected in the present even when not in the near or distant future. On the contrary, the historian – no matter if he/she is a documentary historian or not – is focused on the reading, in its broad sense, of what eventually can be understood of something which happened years ago. So at first it looks as if the only difference in the reasoning could be the *target-set* of the different focus methods. Well to some extent this is the main difference. It is true, apparently, that both the historical studies and those of the technological curricula do not have, by the very nature of their methods, any object that could be shared. But both talk about objects, properties and functions. Both have a need to represent a reasonable formalism concerning events that happen and that have to be described within the language they are using. Meanwhile the *formalism* of the one does not deal with *real objects,* the *naturalism* of the other principally does not deal with the *formal coherence* of the description used to represent *real object*. Therefore there is between these two ways of approaching the problem a lack of intermediation due to the nature of each reasoning method. In current times the crossing of the two is no longer exceptional but is a common practice for any humanist and is at the same time the larger market for any IT enterprise. So time is ripe to find common ground to compare reasoning and *libraries* that eventually belong to each other. Despite the academic absence and disinterest in supporting such curricu-

la, the humanities and the IT community along with their transforming gnoseology should come to a common mind where the development of the one is the economic engine of the other and the economic growth of the latter, by means of transforming languages and applications, the knowledge growth of the first.

5.2 Comparing reasoning (2): the development of discursive reasoning.
The starting point of this so called *meeting* could be the notion of *knowledge hybridization*. We could understand, by such term, anything and its contrary. Thus we need to attempt to define gnoseologically the notion.

Humanistic reasoning is mainly based on the *discourse*. The logic and contradictions within this method of representation can be found in the development of its tradition and cannot be discussed here.

Here on the contrary we need to bring to the surface that humanistic *discourse* as a knowledge representing method is currently lacking strength because of the weakness of its own targets. One could say that if humanistic research would make a reasonable contribution to the advancement of learning, then this would have sense only if that contribution is not based exclusively on the *nuances* of good reasoning. More specifically such contribution should be based on the quality and quantity of information capable of being an *add-on* to current knowledge. Not that good reasoning should be forgotten along with the results that such practice can furnish today. No, that way of grasping ideas is still and precisely one of the main targets of the humanistic tradition. But its development should seriously consider a growth path in the light of the new circumstances that today's communication media require. This means that such reasoning should undergo a transforming method in order to succeed in developing proper formalism for its *descriptive data*. The important point is exactly this: *descriptive data*. This type of information is the only one which could undergo the automation procedures that eventually IT reasoning would be able to manage. Still the division within the *discursive method* between *descriptive* and *interpretative* data has not gained a conceptual citizenship. Such a failing is due to the fact that any *discourse* has to have a documentation to be set as the proof for what is stated there-so much so that the documentation is the *book in the book* rather then simply being a reference. The growth of such practice has made it impossible to state anything without having an entire library that eventually can prove what is stated there. Paradoxically the best way of bringing a statement into the open is to furnish all possible references to what has been said has already been argued, so that the reference becomes the *discourse* and the *discourse* simply a comment on the reference. The more this practice has taken place the less the very nature of a *semantic index* has had the chance of being exploited. The more structured references were required to state new ideas, the less the humanist was able to provide a logical formalism for that reference system. A deep analysis of such habits would require taking into account the medieval *auctoritas* rhetoric, for which we have no space available here. Furthermore, today within *discursive reasoning*, as a consequence of such growth, there is a need always to go back to what has been already said and comment upon that – over and over again. In this sense is meant the expression *book in the book*. In this sense the recursive modelling of the IT gnoseology – just to take one example – could be of some help to frame out a new way of making that literature alive instead of having it as the mute

proof of the *discourse* itself. Following the idea that the best map of the world would be one which, by means of its precision and details, could represent one to one the world as such, this kind of map would not help anybody out there but would be simply as foreign as the world. An attempt to generalize the *descriptive knowledge* in the humanities should be done by means of a representation synthesis whose aim should be to understand and communicate the new knowledge in order that the formalism required by the new media could carry that already known and produced. And this in such a way that the formalism required by the media could shape the conceptual nature of the semantics required for the *discourse* to argue the sense of the *real world objects* that remain, in any case, the matter of the humanistic disciplinary foundation.

Bibliography

Lavagnino, John: *Forms of theory: some models for the role of theory in humanities-computing scholarship.* Available: <http://www.uni-duisburg.de/FB3/CLiP2001/abstracts/Lavagnino_en.htm>.

McCarty, Willard: *Depth, Mark-up and Modelling*. In: *Computing in the Humanities Working Papers* 25 (2003). [Jointly published with *TEXT Technology*, 12 (1, 2003)]. Available: <http://www.chass.utoronto.ca/epc/chwp/CHC2003/McCarty _b2.htm>

McCarty, Willard: *We would know how we know what we know: Responding to the computational transformation of the humanities*. 1999 Available: <http://www.cch.kcl.ac.uk/legacy/staff/wlm/essays/know/>.

Scotti, Andrea: *Scientific manuscripts catalogue. General Catalogue of the scientific manuscripts at the National Central Library in Florence (Italy)*. Supported by the Italian Ministry for Cultural Heritage, the National Central Library, Florence, hosted and managed at the Institute & Museum for History of Science, Florence. 1996–1998.

Smedt, Koenraad de e.a. (Eds.): *Computing in Humanities Education. A European Perspective*. Bergen: University of Bergen 1999 (Socrates/Erasmus Thematic Network Project on Advanced Computing in the Humanities). Available: <http://gandalf.aksis.uib.no/AcoHum/>.

Empfohlene Zitierweise:

Andrea Scotti: *Pinakes*: Structuring and Destructuring Dokumentation in the Humanities. A Project for Modelling Data in History Research. <http://www.germanistik.ch/publikation.php?id=Pinakes_Structuring_and_Destructuring_Dokumentation>