Amélie Zöllner-Weber; Daniel Apollon

The challenge of modelling information and data in the humanities

2008

https://doi.org/10.25969/mediarep/19791

Veröffentlichungsversion / published version
Sammelbandbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Zöllner-Weber, Amélie; Apollon, Daniel: The challenge of modelling information and data in the humanities. In: Theo Hug (Hg.): *Media, Knowledge & Education. Exploring new Spaces, Relations and Dynamics in Digital Media Ecologies.* Innsbruck: Innsbruck University Press 2008, S. 118–137. DOI: https://doi.org/10.25969/mediarep/19791.

The challenge of modelling information and data in the humanities

Amélie Zöllner-Weber & Daniel Apollon

Abstract

Notions of information before and after "information society" imply important conceptual shifts in the humanities, which are reflected in a variety of approaches to texts and documents. Focusing on the evolving relationship between three related but distinct concepts – information, data and communication, – this article examines the diversity of knowledge management projects in humanistic disciplines.

Various humanistic approaches base their approach, implicitly or explicitly, upon a specific comprehension of the notions "information" and "data", and offer a rather opaque treatment of the notion of "communication". More than assuming a loose association between, on the one hand, information and data, and, on the other hand, knowledge goals in the humanities, the authors substantiate the claim that there exists a close functional relationship between these.

Introduction

In contemporary society, information and resulting knowledge seem to be as valuable as material goods. Information and especially its interpretation can influence us in a strong way. Daily life involves processing a mass of information. For many of our contemporaries the "stream of explicit information" starts with reading newspaper or watching morning news during breakfast. We gather rumours from people we meet, by reading or watching commercials for food in super markets, or surfing the Internet. Not infrequently, our days end in front of TV or reading a book, both media conveying information that we read, watch or hear and all of these activities involve absorbing and communicating some kind of information. It does not matter whether "information" is a philosophical theory, a narrative, an algorithm, or just loose news from our neighbour. Whether information can be used or not, depends on the given situation and on our decision. Not every kind of information appears to be useable though. But usually, we are able to rank the importance of information without noticeable efforts.

One can assume that human beings are able to extract and use information solely because they mobilize consciously the processing power of intellect. But also animals or small children are able to use information as they live. It does make sense therefore to anchor the principle of information processing in life itself. Endorsing the view that human information processing is in se not primarily an intellectual process, implies that information processing ability is essential for all life and, that, human production, processing and usage of informa-

tion may be treated as a function of life. Following freely Wittgenstein's terminalogy, human information may be seen as the expressing one of many possible "forms of life" linked with the concept of "language-game" which "is meant to bring into prominence the fact that the speaking of language is part of an activity, or of a form of life" (Wittgenstein, P.I., aphorism 23).

Meanwhile, for researchers, institutions, and businesses interested in information management and science, it remains a challenge to prepare, store, and share properly information in digital format. Part of the problem is associated with the imprecision and instability of what is meant to be personal and collective information. Therefore, do we really know what information is?

While one does not need to ask such questions in real life context, e.g. while driving a car, the issue of the nature of information arises when dealing with explicit and systematic digital information. If one does not know what "information" might be in a particular context, one is not able to handle or model it. Most obviously, gathering and processing of information is linked with the various ways to be shaped and shared. Therefore, human information (a form of life) and digital information (data and metadata) may share a common context.

The notion of "explicit information", as opposed to "informal information", has gained a tremendous momentum in industry, administration, culture, social life and science, thanks to information and communication technologies (ICT). Scientific activity in natural and life science involves gathering, digitalising, storing and processing such information. By processing and providing information, computing has become a vital part of technoscience (term coined by Hottois 1984). While technoscience has made its success dependent on epistemologically exploiting the potential of information processing, this has not been the case in the humanities where systematic use of information processing architectures remains restricted to a few fields of research.

One needs therefore to assess the role of digitalised information in the humanities, while resisting a growing tendency in the humanities to treat information structures as separated from communication processes. Dealing with the notion of information in the humanities poses a serious challenge: assuming a state of increasing diversification and even fragmentation of humanities, one may, postulate that, within this diversity, conflicting comprehensions of information may occur. Before outlining these different comprehensions of information in the humanities, an outline of information in contemporary research will be given below.

Creative, but problematic diversification of humanities

Scientific disciplines may be grossly classified in three categories:

The first category comprehends disciplines like chemistry, biology, and physics. In these disciplines, objects under investigation are essentially approached as manifestations of "nature". Driven by human curiosity, researchers identify, describe, measure, and compare the characteristics of these objects, ambitioning to exploit the results of their investigations to

understand the whole from its parts. Thereby, they develop and refine methods, which depend increasingly on measurement, mathematics and computation. The power of natural science is the ability of adding to data-gathering efficient processing mechanisms. The joint exploitation of "abstract" mathematical models and computer-driven large-scale simulations and predictions has undeniably opened new territories for science and technology. Increasingly, the usage of "abstract machines" and computational models is no more restricted to objects or domains that can be modelled in a nearly realistic way, but is applied to speculative simulations of increasingly exotic meta-real themes and pure structures. Hence, in contemporary science, information is not only something that is thought solely as a crude, but efficient reflection of observations, which can be verified "in the field". It has turned into, on one hand, a revolutionary ambitious theory used to represent observable as well as non-observable reality (e.g., in pure computational mathematics and metamathematics), and, on the other hand, as a powerful generalized production principle in the material and living world.

The spread of the information paradigm to new scientific domains may be illustrated by recent efforts in biology to exploit John von Neumann's theory of self-reproducing automata in order to create artificial cells in biophysics. In computational biophysics, information structures may not any longer be used solely as practical and inert representations of the intracellular world. The picture is reversed: intracellular processes are analysed and theorized as forms of computation. One recent example of such "transgressive" approaches is the biophysicist Albert Libchaber's effort to study and simulate protein self-assembly as computation and molecular evolution as "transcription machinery" (Noireaux 2005).

Natural science adopts increasingly various kinds of productive informational paradigms. This leads to the increasing blurring of the boundary consisting of three kinds of activities: firstly, pure registration and description of observations, secondly, pure structural simulation or fabulation and thirdly, engineering and production of artificial materiality, e.g., synthetic cells or bacteria.

The second category comprehends social disciplines such as sociology, psychology, ethnology, and political science. Their object of study (e.g., man, society, mind) is not any more "given by nature", but strongly manipulated by humans or an integral part of humanness. Therefore, such "social objects" may not be seen as completely natural. For example, many aspects of our modern society are based on the artificial monetary system. However, what is considered as observable is almost never context-free and involves complex interactions between sets of variables. To complicate matters, social science has bifurcated into two difficultly compatible types:

The first type emphasises philosophically inspired approaches, and ambitions to formulate theoretical models for human society and individuals. It proceeds towards its epistemic goal by exploiting observations of partial realities (behaviour, utterances, artefacts) in order to map whole systems. Pierre Bourdieu's sociology belongs here. Through systematic exploitation of metaphors inspired from classical mechanics (e.g., field, inertia, trajectory, position, multidimensional space), Bourdieu deployed a general sociocultural theory allowing understanding mechanisms of competition and domination. Bourdieu's attempt to let the whole (e.g., the notion of social field) and its parts correspond (e.g., particular aesthetic tastes) is not unlike

modern physics' efforts to extract latent structures and embedded order in physical phenomena.

The second type of social science does not nurture itself from the ambition to accomplish the kind of epistemological programme illustrated by Bourdieu, but rather seeks to exploit available methods in order to untangle the hidden informational structure, which systematically collected observations might yield. Social reality is thought as information mine, while diverse methods and techniques are thought and engineered as mining tools. Such mining research tends to be theoretically agnostic and utilitarian, ready to exploit, when deemed opportune, former "theories" as supporting discourse or "epistemic wrapping". Most applied social research belongs clearly to this second type: it partitions social reality into compact domains treated as if they possessed an existence as autonomous phenomenon. In Bourdieu's sociology, the overall theoretical system fills bare social facts with sense. In contrast, applied social research of the second type tends to look for interpretable structure embedded in the data within a particular domain.

Furthermore, within in a single discipline, say sociology, information and its treatment, as well as theoretical models tend to differ. In the first type of social science, information about an object might emerge from a theoretical study. The link between the elaborate discourse of the social scientist and empirical reality might be problematic. Research of the second type may face problems not unlike those outlined above in natural science, where, frequently the thickness and opacity of reality defies simple theorisation. While intensive data processing of collected observations may reveal significant causal phenomenal association, strong ties between variables, unstructured residual information, still, may challenge naive explanations. Unfortunately, not even demonstrating "significant" causal relations always appears to add sense to observational information.

The third category covers the fuzzily labelled "humanities", where the objects of study produced by human beings are completely artificial, but express various forms of life. E.g., a printed story may be conceived as a realistic object, containing information, which cannot appear without reading it. Use and perception, e.g., reading a book or watching a play in theatre, presupposes a usually complex context of past (e.g., genesis of the text by the author) and present human activities (e.g., expectations of readers or audience). Cultural objects may be studied as manifestations of intentional coding by humans, and not as immanent manifestation of physical nature or pure social reality. Works of art are, nowadays, communicated in multiple traditional and digital forms and formats. Cultural objects, when viewed as epiphenomena, express a rich and complex multilayer history of contexts, intermediaries, and actors. They offer intentional and interpretable representations of the world, while carrying along their "archaeological substrate". This substrate may accumulate the combined effect of various transactions on information, which took place before their users perceived these objects. Additionally, these cultural objects possess an "ecology" since they function within an intricate web of functional relationships within a larger contemporary production of meaning.

The study of cultural objects is bound to relate to this diachronic and synchronic complexity. As a consequence, cultural products cannot be approached solely as "naive" information transmitters: take, for example, the narrative strategy of an author of a criminal novel. A sub-

stantial part of the authoring process in that case, consists in covering on purpose information and deliberately misleading the reader towards erroneous conclusions, thus creating preconditions for a final revelation of the truth and dénouement. Also, philosophical texts cannot be tapped as pure "information mine" but as an open-ended discourse with various levels of interpretation. Historical sources also escape simplistic informational treatments. Cultural objects happen to be notoriously polysemic, seductive, opaque inviting their consumers to widely diverging conclusions because of multiplicity of possible contexts, e.g. the background of the observer or what the author may want to give the reader. They may be approached as a play between coding and decoding. Indeed, the humanistic character of the study of a cultural object may appear better in the analysis of and reflections on the variation of interpretations around such an object than in the objective and structural description of the object "itself". The dissimilarities between uses of cultural objects may, following Wittgenstein's concept of the language game, be more clarifying than mapping regularities or similarities. While educated discourse on such objects has dominated the traditional academic approach, recent efforts – to exploit information technologies to model such polysemic objects (objects that elude and delude) – remain a serious challenge. Early attempts to model the narrative, e.g., Greimas' structural semantics and actant model (Greimas 1996 & 1970; Nef 1976), focused solely on the assumption of a uniquely embedded latent differential dynamics (the semiological square) exploiting de Saussure's notion of "differential quality" fell into reductionism.

This paper operates therefore on the assumption that the humanities, more than social and natural science reflect a state of fragmentation, which affects the overall epistemic ambitions of these disciplines and, as a consequence, the use of various informational paradigms. The inherent multiplicity of interpretation possibilities of empirical material in the humanities makes it particularly difficult to envisage predictive generalisations. Heterogeneous material like books, films, or pictures resists the type of generalisations encountered in natural and social sciences. While methods and interpretations abound, many humanists, conscious of the vicious nature of the "hermeneutical circle" (e.g., the double lock of inferring the whole from its parts and the parts from the whole in Staiger 1955) and of the impossibility of escaping from the multiplicity of interpretation (Ingardten 1968; Ricœur 1969) abandon "natural" ambitions to infer the whole from its parts and lean towards one of three analytically approaches:

The first approach accepts fully the view that the study of cultural objects is synonymous with *ars interpretandi*. The analytical perspective is moved from the inherent structure of the object, say, a novel, a painting, a film, to the multiple interpretations generated in the environment of this object. The notion of "horizons of interpretation" and of "merging of horizons" (see Gadamer 1960) have, since, influenced considerably textual disciplines (Eco 1962 & 1979). The empirical object is treated as a platform allowing dealing with hermeneutical practices. Ultimately, hermeneutics transcends the notion of method as conceived modern (natural) science (Gadamer 1962; Betti 1955). This affects particularly the treatment of imprecision in literature and art: while imprecision in measurements is an obstacle to be overcome in science and engineering (or addressed statistically), imprecision, vagueness, lacunae,

descriptional inconsistencies, are, following Ingardten (1931), crucial parts of the meaning production of literary works.

The ultimate meaning does not reside in a general theory with optimal explanatory power. It would be, however, unfair to see hermeneutical approaches in contemporary textual discipline as pure relativism. Doubrovsky (1966) fought against total deconstruction, arguing that any work, beyond a seemingly anarchistic polysemy, points in a direction, an "intimate unity" and a "ultimate meaning".

The second approach, while not necessary rejecting the validity of hermeneutical relativism, favours an ecological perspective on cultural objects and analyzes them as contextual manifestations with various degrees of stability and plasticity (Cerquiglini 1989). While Stanley Fish's reader-response model (Fish, 1976) and notion of "interpretative communities" (Fish 1980) reflects a clear ecological perspective, Frederic Jameson's Marxist approach (Jameson 1971 & 1991) points towards more explicative models.

The third approach abandons ambitions to address multiple interpretations, and most of the contextual issues. It seeks refuge in isolating and simplifying the cultural objects as a reality in se amenable to "objective" description and modelling as if these objects could be "naturalised", but without the kind of theoretical ambition of, e.g., modern physics. While modelling, encoding and "informationalising" cultural objects, analysts of the third type do not come close to Chaitin's minimalist informational definition of "comprehension as compression": "The simpler the theory, the better you understand something" (Chaitin, 2006). "Hard structuralism" (see Brémond 1966; Greimas 1966; 1970 & 1973) depending heavily on the Saussurian linguistic paradigm, operates the reduction, but fails to deliver the comprehension.

One should not be surprised then, why subjects, methods, epistemic programmes and scientific raison d'être in the humanities appear to be seriously fragmented.

The introduction of machines and external conceptualisations in the humanities

The multiple disciplinary perspectives outlined in the previous section involve conflicting understanding of what the adjective "humanistic" is thought to cover. More specifically, it highlights conflicting understanding of which methods (in the strict sense) or approaches (in a loose sense) are constitutive of the intellectual efforts deployed within the humanities. Within this context, information and communication technologies, and correlated concepts of information, seem to serve many interacting purposes ranging from purely "engineering" motives to more explorative, productive, creative, and philosophical horizons.

Within the vast and potentially unbounded intellectual domain of humanities, many research activities have focussed on products of human activity and thought, e.g., speech recordings, pictures, documents and many more artefacts. Among all possible artefacts, documents and texts, however, have remained privileged objects of study and are still the object of predilection of method-centred approaches. Procedure-centred approaches using various kinds of

formalisms on texts and records, it should be remembered, have been established in academia long before the introduction and generalisation of ICT (e.g., genealogical reconstruction of ancient text families, early syntactic analysis, textual semiotics, archaeological excavations).

The spread of computer tools, networking and meta-descriptive standards (e.g., mark-up languages like XML) has added a new level of expertise on top of earlier research practices. Observations gathered from physical artefacts (the manuscript, the lithography, the handwriting, etc.) are now currently computerised and organised as data sets. Such computerisation of observational data serves both practical and theoreticcal purposes: to identify, retrieve, represent embedded information structure contained in the artefact. This recent interest to extract, structure, and, represent embedded information using meta-descriptions (to be distinguished from layout descriptions) as metadata, has spawned a new interest for modelling in the humanities. The most obvious issue is the inherent bias of representing artefacts together with digitalised meta-structures in a form that is guided by the choice hic et nunc of the researcher who encoded it. While adopting a single, isolating, rigorous perspective on the meta-description of an artefact may sound "scientific", it may well hide efficiently other potentially valid perspectives. Explorative access to artefacts may thus be obstructed, while creators of these artefacts might have wanted people to explore their products more unrestrained.

The authors will not discuss in-depth in this paper possible reasons behind the historical preference given to documentary artefacts in contrast with other seemingly less "cultural" products, nor launch a full-blown discussion on the theme of the dematerialisation of artefacts. We derive, though, two particular angles of view from this sketchy appraisal of view the landscape of humanities. Firstly, we assume that the textual preference builds up a preconception among humanist scholars that "literacy" and "literariness" somehow reflect more the quintessence of humanity, than, say, transient short lived rough by-products of everyday activities, which may be left to ethnologists, or to archaeologists if these products are old enough. Secondly, we argue that there exists a functional relationship between, on the one hand, the reverential attitude of some scholars towards highly-valued documents and, on the other hand, a penchant towards substantialism, as witnessed by an exclusive concentration on intratextual features. The restricted world approach of metadescription of embedded information structures offers a powerful platform for cultivation of what may be tagged as "reverential substantialism". Again, we argue that ICT has not eradicated substantialism from humanities, but, rather, has recontextualized this epistemological position as "information substantialism".

The rise of the new paradigm of information substantialism in the humanities, may not have been possible, without two important transformations taking place during the last half of the 20th century:

Firstly, the number and types of artificial products have expanded constantly unbounded by the former norms that defined the canonical status quo of academically acceptable collections of objects and cultural expressions deemed worthy of interest. Such a proliferation of new instances of cultural objects is historically linked with the rise of the Internet. Within a short time span, new forms such as weblogs, multimedia combining text, video, and audio, and socialware have emerged.

Secondly, an extension of the notion of empirical reality from pure physical materiality of cultural objects to integrate their digital information structure (e.g., files in computer systems) has taken place. The epistemic thinking, which made this extension credible and possible, appeared long before the spread of various computing practices in humanistic disciplines and served later as a fertile ground for a more radical digital realism.

Anticipating, the discussion, in the next sections of this paper, of the relationship between the concepts of "information" and "data", we may already note that the above-mentioned extension may not have taken place without an increasing tendency among method-focused researchers to equate "information structure" with "data structure". Equating "information" with "data", for all practical purposes, favours approaches, which computationally offer verifiable and retrievable coding of such "information structure". After it is established, such information structure tends to be self-imposed as if it constituted a scientific neutral ground enabling an objective description of empirical reality. Additionally, thinking of and believing in information structure as a neutral ground which allows a multiplicity of subjective, alternative, less consensual interpretations that may be applied to the underlying material. As a consequence, techniques, methods and ultimately frameworks of analysis and knowledge tend to be treated as interchangeable, pluggable components. But as stated before, the question still arises whether information modelling of an artefact, e.g. encoding literary texts, deserves a special treatment as "objective".

Most computational applications and methods were not developed from the perspective of the humanities, but were imported from neighbouring domains and applied with varying success to these disciplines. Increasingly, techniques, methods and models of reality and knowledge are imported from computer science. Various techniques, loose ideas and paradigms from e.g. the extremely heterogeneous domain of artificial intelligence (A.I.) penetrate into traditional humanistic disciplines, e.g., philosophy, literature and media studies, and function as powerful bridging frameworks between these domains. The term informatisation may be contrasted with the more usual term digitalisation: while "digitalisation" usually refers more to the concrete transformation of analogue contents into digital files by means of some encoding, "informatisation" (Hardt & Negri 2000) designates more broadly the application of some information paradigm to production systems: "We might call the passage from […] the domination of industry to that of services and information, a process of economic postmodernization, or better, informatization."(p. 280). A powerful side-effect of the "informatisation" of central activities in the humanities is the appearance of "human engineering" within the humanities, e.g., speech recognition systems for telecommunication, universal access solutions, dynamic publishing, sellable learning objects, and most recently "real-world" ontologies.

The rise of informationalism in the humanities

Humanities tend to bifurcate in two directions:

The first direction may be characterised by an increasing level of generalisation of some humanistic scholar's horizon, quite often, with the help of increasingly exotic mixes of sources of inspiration and experiments to exploit language creatively (as witnessed, e.g., by the work Jacques Lacan, Paul Virilio, Gilles Deleuze, Julia Kristeva, Roland Barthes, Marshall McLuhan, or Donna Harroway to mention a few salient examples). These thinkers, presenting often-incompatible approaches, deliberately blur the borders between art, philosophy, and traditional scholarly work.

The opposite direction is characterised by an intentional narrowing of the field of study of humanities. One opts for programmatic specialisation and restriction of the scope of research options of methods and techniques, ambitioning to deal with describable objects of study. One evicts both the elaboration of ambitious theories, and ultimately the idea that thinking is, inescapably, *ars interpretandi*. Such approach tends therefore to focus on practical projects, which may involved highly specialised engineering, e.g. establishment of digital corpora and critical editions (the Canterbury Tales Project see Robinson 2003) large-scale semantic networks (e.g., WordNet, see Fellbaum 1998), or more ambitious, speculative enterprises (e.g., Cyc, see Witbrock et al. 2005). Such projects may be characterised as attempts to produce homomorphic mappings of specific domains and impose, intentionally or not, a de facto hermeneutic closure. One should note, however, that the notion of precision and accuracy involved in such enterprises might differ from natural science. Commonly, the accumulation of accurate information (e.g., on variant readings in manuscripts) does not imply necessarily the slightest notion of measurement, but rather, reflects a disciplined registration or mark-up of "data", as it is the case in contemporary text encoding practice and establishment of digital corpora. Occasionally, the accuracy of registration is thought to lead to some form of measurement and validation to be carried out on the mapped representation, as it is the case in the exploitation of linguistic corpora to simulate, test and possibly falsify former linguistic models. Such enterprises build upon the beliefs that the phenomenological dimension may be accessed through such structural mapping. The problem, though, may well reside in this leap of faith.

The dichotomy proposed here might be somewhat sketchy, but has the advantage to position roughly the diversity encountered in the humanities relatively to representations of "precise science". Our claim is that the strong influence from technoscience, and a general perception of the humanities as being imprecise, vague, "more art than science", and sometimes an imposture, has exerted a significant pressure on these disciplines to go for more precision and adopt engineered procedures in order to regain credibility in the academic. A possible scenario is the growing success of designing informational models of cultural objects, making it academically, politically, and economically credible to treat these new "models" or "mappings" (this term is used here as 'mapping complex structure) as crucial sources of objectivity and reality. From the perspective of research strategies, these new informational cultural objects (e.g., "terms" or "syntactic structures" in diverse digital corpora) may be offer the

same kind of research value as, e.g. "genes" do in bioinformatics, where the real selling value resides in the genomic mapping. The trustfulness of the activities applied to these objects profits increasingly on the precision, verifiability and technical sophistication of the methods applied to carry out the mapping. Ultimately, the object of study, methods, technology, and the prospect of real life application form a bundled package, which may be competitive in research. As we will discuss below, information technology and new developments in human society have opened totally concrete opportunities to carry the idea of "humanities as precise science". As we will argue later, this idea or vision, while opening new territories for humanities, may also involve problematic trade-offs and loss of horizon for the activities deployed.

Of course the two directions outlined in the first paragraphs of this section are not exclusive and, obviously intermediary positions are possible. There are still cases where the humanistic scholar sees himself or herself as author of a "work" which calls for a necessary, but demanding coherent monumentality. The double imperative, the first being to carry out empirical analysis which may be recognized by critical readers asking for accurate facts or at least consistent descriptions, and the second being to unfold a critical, creative perspective and possibly challenge earlier systems of thoughts, still remains at the heart of some humanistic projects.

The discussion above offered as a sketchy, but, in the view of the authors, necessary overview, which may highlight the impact of diverse informational paradigms amplified by ICT on the humanities. Indeed, the use of computers in the humanities through more than four decades seems to cover a varied research landscape, spanning from, census collection by historians, to simulations of human speech or analysis of syntax in computational linguistics, to collection of large corpora of heterogeneous cultural artefacts, to text encoding or, even, to simulations of symbolic processes. Only the imagination seems to limit the number of possible combinations of applications of information technologies with epistemological perspectives on human and cultural content. However, not always, the apparent sophistication of the computer-based techniques spreading into diverse humanistic domains seems to be matched by a corresponding epistemological in-depth reflection on the context and prospects of ongoing practices. In our view, such reflection may not be carried out, in a contemporary setting, without addressing in some details the embedded notions of "information", "data", and "communication" which may be at work in the humanities, with or without the concrete usage of ICT.

A brief history of the notion of information and data

The notions of information and data have undergone a considerable evolution and diversification in the period after the Second World War. We assume that humanistic disciplines reflect in various aspects this historical development and choose to focus on three important levels of understanding the notion of information. The first level being the socio-cultural notion of information as "mediated rumour", the second level covering information theoreti-

cal approaches to digitalised information, and, finally, the third level involving the notion of "informationalism".

Information as mediated rumour

"[The Greeks] marched in order by companies to the assembly, and Ossa (Rumour) walked blazing among them, Zeus' messenger, to hasten them along." – Homer, Iliad 2.93

Greeks had even a dedicated goddess or daimon Ossa or Feme (Φημη) of rumour and gossip. The short quotation above serves to remind us that the first level of our construction refers to cultural uses of the word "information as credible rumour" with deep roots in preliterate oral cultures, that antedates by far information technology. A general definition of the cultural, mediated understanding of information is «mediated rumour with varying degrees of trustfulness». Thus the first level, stretching from preliterate orality (Ong 1982) to contemporary literate cultures involves diverse "technologies of the mind" (Goody 1968,1977,1986 & 1987), e.g., procedures, devices, and sign systems by means of which people can share some messages which are meaningful to them. One should note that this first level does not disappear, or leaves traces as some kind of archaeological ruin, but rather, finds its niche in the new enriched space offered by the Post-Gutenberg digital galaxy, which more than previously achieved in societies, allows massive transmission of and transactions on such "trustful rumours".

Well before the spread of computers and networks, industrial societies, have constantly undergone a slow evolution from the trading of "trustful rumours" from mouth to ear, to contemporary techniques of written and audiovisual media. The "technologies of the mind" and "literacies", following Goody, have exhibited a considerable historical and geographical diversity. A possible phase shift might have occurred in the passage from a pure social validation of news and rumours to a methodological and reproducible dimension, exploiting mechanical or digital automatisms. By methodological validation of information, we imply the relatively recent appearance and all kinds of information processing and social practices, which leads to an increasingly dissociation of the message from the messenger, the meaning from the coding, and the transmission from the transaction. Contemporary media, both digital and traditional, may reflect this dissociative approach to reality.

So, what does this historical survey explain? Firstly, it places the historically and socially multilayered information representation of trustfulness at the heart of today's widespread representation of 'information as data' and 'data as trustful facts' in information and communication technologies. The "trustful rumours" in Homer's Iliad or television news can be interpreted as more or less trustful to reality. Contemporary notions of "data" discussed in the next section, however, can be seen as a more radical departure from traditional trustfulness. The emerging data-grounded approach to reality and truth, this is our view, involves a double set of embedded knowledge constructs:

Firstly, "data", which may act as a realistic container and representation for different kinds of real-world or imagined objects, (e.g., data encoded in XML, an ontology in AI, a conceptual map);

Secondly, "data-processing", which may, exploiting the combined power of digital storage and algorithms, offer a method to describe, validate, and enhance the space of "complete" reality. With reference to Greek mythology, these informational ambitions may call upon beliefs in the eudaimonic power of data.

Shannon's information theory and its reinterpretation

In contemporary technoscience, the notion of information builds upon the notion of "quantity of" and privileges information transmission at the expense of message content (Durand-Richard 2004). Semantic content of messages is disregarded. Superficially, "mediated rumour" should be dissociated from the notion of "quantitative scientific" information. Such a view tends to break down, when a historical perspective is adopted, because it emphasizes the social, cultural, and personal nature of information, and points at the difficulty of treating information as dehumanised autonomous content out there.

"It is common to think of information as a much later arrival on the evolutionary scene, as something that depends on the interpretative efforts, and, hence, prior existence of intelligent life. According to this view, something only becomes information when it is assigned significance, interpreted by a sign, by some cognitive agent. Beauty is in the eye of the beholder, and information is in the head of the receiver. To speak of information as out there, independent of its actual or potential use by some interpreter, and antedating the historical appearance of all intelligent life, is bad metaphysics" (Dretske 1981, p. vii).

Paradoxically, the ideology of information as being something that is outside, criticised above by Dretske, seems to be inspired by early communication and information theory. But early information theory, as exposed by Claude Shannon, never intended to create an exterior, secluded and somehow autonomous space of information. In Shannon's theory, signals are only non-human entities, which tend to die slowly or quickly due to the entropic effect of noise. Indeed, in Shannon's early work, information and entropy were interchangeable terms. The mathematical description of ensuring and describing the survival of signals in a noisy environment is called information theory. Information remains exclusively a quantification and compaction, which can be measured. Shannon never implies or presupposes include that efficient or well-compacted information is meaningful. Shannon's theory is essentially dealing with guaranteeing and improving transmission of signals, without any ambition to map content. For Shannon "The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point." "Coding" in Shannon's terms is not synonymous with the term "encoding" as used in recent Semantic Web settings. More precisely: a signal is understood as coded information for the sole purpose of safe transmission. The coding has no ambition to represent a source combined with semantics, using a grammar, which describes its content so that humans or machines might understand it. Again, coding serves only the purpose of creating a situation allowing for the efficient transmission of any content. In Shannon's approach, communication amounts to the successful transmission of signals by means of coding through a channel. It is not necessarily a transaction on semantics. Human meaningfulness belongs indeed to another realm. Decod-

ing is the proof that transmission has taken place with measured efficiency. For Shannon, the enemy is noise, not non-sense.

We may note that information, either taken as diffusion of mediated rumours (e.g., in audio-visual media, and computers), or as, Shannon's theory of information, both deal with the transmission of some content. However, stricto sensu, the relation between the two notions of transmission and content cannot be equated.

It appears, however, that Shannon's ideas, have been feeding a much broader notion of socio-cultural communication, disregarding its strict theoretical self-imposed limitations. Shannon's theory has been exploited metaphorically in order to shape a new discourse on human (digital or social) communication.

Since, key concepts and academic fashions have, dispensed with the limitations of Shannon's model and slowly evolved toward a new hybrid information ideology merging aspects of the information theoretical discourse with older strand of mediated rumours. The result is, in the Internet age, not a new augmented media space offering "more of the same", but, in our view, a more ambitious world-view dominated by a belief in the production of trustful reality by data structures.

The argument of this merging outlined above, is of particular importance for encoding practices in the humanities and the emergence of the Semantic Web during the last two decades. We refer specifically to the emergence of the beliefs that encoding (to be distinguished from "coding", which "encoding" presupposes) of text or any other human artefact, based on a meta-lexicon and encoding grammar offers a verifiable and reusable source representation. While, text encoding, with all its technical sophistication and automation is historically related to information theory, and presupposes it as its technical substrate, its functional relationship to Shannon's restrictive model appears to be more of a metaphorical than formal nature. Indeed, text encoding does not offer to guarantee and improve physical text transmission (even if it may offer some side-effect in this direction), but, alternatively, offers a metarepresentation of a source material with various degrees of ambition to convey meaningfulness and trustfulness. This metarepresentation may be viewed as a hybrid activity mixing the first domain's mediated rumours model with a metaphorical use of the second domain information theory. The potential as well as the problems of current encoding practices reside in this mixture.

Extending "data" and the rise of Informationalism

A progressive shift from a strictly mathematical notion of quantitative information in information theory towards a broader, more ambitious, but increasingly less rigorous notion of "data" has taken place in the Internet age. This shift combines both hard aspects inherited from information and automata theory with soft symbolic aspects. Thus, the notion of "data" has extended its domain and connotes today much more than efficiently coded and compressed information. It seems to cover, at least, six interrelated, but not equivalent functions:

1. Data, as machine-readable data. Data, in its simplest function, provides an efficient, transmittable virtual container of any encodable source. Data, taken in that sense, is synonymous with "coded information".
2. Data as semantically enriched annotation. Recent advances within encoding and mark-up systems (e.g., XML) appear to complicate matters. "Data" is increasingly present both in the coding and in the channel: "data" codes not only the content but also, the modalities of the communication of the content. As a consequence, a rigid distinction between information as data vs. communication is no longer tenable. A new distinct function provides meta-representations of e.g. texts, digital video, sound, etc., which can be verified (e.g. XML files against their DTD). The source, together with its meta-representation, may form a new unit, as it is the case in detail encoding of critical text editions.
3. Data as mediating data. Data may add a communicational syntax to present, mediate, and transmit these meta-representations with their source. E.g., the application of thematic filtering mechanisms exploiting mark-up "mediates" meta-representations to recipients.
4. Data as coded human transaction. Data codes, embeds semantics, communicates, expresses and unfolds virtual, but nonetheless, real, individual, and social communication.
5. Data collections as immaterial commercial and cultural goods. This category applies to data, data flows and collections, which can be stored, shared, transmitted, and exchanged as autonomous objects of transaction.
6. Data as a posthumanist project. Data and information-based dynamics concretise a larger semi-automation vision ambitioning to extend humanity beyond the limits of normal biology and psychology.

The work of Castells is of particular relevance to illustrate the interlinking of the six data functions outlined above. Castells (2004) clams that informationalism constitutes a hegemonic paradigm pervading all culture and society through the "informatisation" of all techniques and through the increasing virtualisation of social organisation. Castells' vision supports the claim that, such a pervasive evolution takes place, leading from the rather esoteric world of early cybernetics to broad ideological perspectives.

From Castells' point of view, the levelling power of informationalism resides in its material and symbolic power to recapitulate, recontextualise, and, ultimately, reduce of all previous industrious, biological, and soon socio-cultural activities to instances of some automaton. The emergence of the paradigm of informationalism is, in his view, due to the capacity of ICT based technologies "to self-expand their processing power because of their recurrent, communicative ability" and to "recombine information on the basis of recurrent, interactive communication" (p.10).

From Logic to Artificial Intelligence to the Semantic Web

Information is thought to offer a huge potential for all kinds of human activities. One can even say that most of activities, nowadays, cannot be performed without processing information. Categorisations and descriptions of things and ideas in the world are transformed into information. Categorisation or description systems are, however, not a new development of the modern world. All kinds of intellectual technologies have been developed and refined in the past. These have been and are still essential to classify all kinds of information. In particular, semantic ambitions, or rather, utopia, are not new and may be reminiscent of earlier intellectual endeavours, e.g., Cabbalistic science, Novalis' Allgemeine Brouillon, Encyclopaedism, etc. But after the appearance of computer systems, humans are now able to transfer information to new media and to store it there in order to manipulate it. This creates new challenges: today, more and more data can be stored in computers. But, due to the huge growth of the available mass of information, it may be difficult to handle and retrieve stored information. Therefore, data and information have to be represented and modelled to exploit the capacity of computer systems.

Information or knowledge representation in AI has exerted a strong influence on other disciplines, and on the industrial and public sectors. The perceived value of this kind of modelling has been increased by the various visions of the Semantic Web. This idea is a driving force fuelling ambitions to transform large parts of the "unsorted" Internet into a Semantic Web so that advanced retrievals may be made possible. In a Semantic Web environment, terms are related to each other semantically so that a network can be generated: "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." (Berners-Lee et al., 2001). Due to the increasing popularisation of its formats and ideas, views on information, which first appeared in the philosophical logic, are re-introduced in an applicable and computational way in the humanities. By preparing and modelling cultural objects or artefacts (using for example RDF or OWL), the description of these objects takes place exclusively on a formal and structured level, while other types of discourse are excluded. Indeed, it is the descriptions, which are formalised, rather than the objects.

The extension of the domain of "data" fuels the visions of the Semantic Web. While inheriting older methods of representing knowledge, it introduces new ambitions to produce a semantically enriched structure of a given content. Obviously, the word "semantics" is being used with varying degrees of ambition, sometimes restrictively as "thematic representation" (hence offering a little more and a little better than traditional catalogue, tables of content and indexes) or, downright as, "knowledge representation" (hence offering, e.g., meaningful representation of a presidential election, of a given character in a play, of Plato's philosophy).

Frequently, semantic ambitions have been interpreted as "logical reasoning" and information processing as "logical manipulation of symbols" (Bechtel and Abrahamsen 2002, p.8f, with reference to Hobbes and Leibnitz). Eventually, the possibility of symbolic manipulations may justify hopes to carry out cognitive operations by means of algorithmic information processing, of propositions and rules. Formal logic systems can only operate on the so called "closed

world assumption" so that conclusions can only become true if a proposition is already true and if such a proposition is included in the system (see Charniak & McDermott 1985, p. 504). Extensions of truth logic to, e.g., modal logic (Blackburn & al. 2001), possible worlds semantics (Hintikka 1962; Herrick 1999; Dolezel 1976 & 1978; Divers 2002), Kripke's semantics (Kripke 1963), or fuzzy logic (Zadeh 1965; Halpen 2003) do not offer any radical alternative.

As computer science expanded during the last half of the 20th century, ideas and theories, which were already prefigured in philosophy and mystical speculation, were now treated as engineering problems in search of algorithmic solutions. Rapidly, the burgeoning discipline of artificial intelligence bifurcated in alternative directions. Experiencing soon the limitations and intractability of logical reasoning machines, new hopes were raised by heuristics: "[…] heuristics […] are often more useful than algorithms" (Bechtel & Abrahamsen 2002, p. 10). In addition, the research on cognition and neuroscience encouraged radical departures from the straightjacket of reasoning by formal logic. Exploiting fresh advances in neuroscience and overcoming the initial obstacles of neural network theory, a connectionist approach was developed encouraging the idea that "knowledge" in humans exploits self-organising. In connectionism, the "basic idea is that there is a network of elementary units or nodes, each of which has some degree of activation. These units are connected to each other so that active units excite or inhibit other units." (op. cit. p.2; see also Kohonen 2001).

AI nowadays aspires to provide solutions in a wide range of applications ranging from linguistics, and cognitive modelling to robotics, "intelligent" information processing, and representation. The common scientific and engineering programme of AI underlying all these activities, following Charniak and McDermott: "[…] the study of mental faculties through the use of computational models." (Charniak & McDermott 1985, p. 6). Adequate and efficient acquisition, representation, and processing of knowledge using algorithm treatment of information structures remain the goal of AI.

As emphasised by Bechtel and Abrahamsen, the different approaches to AI still compete against each other and justify radical doubts on the possibility of a general theory of computational cognition. Human information and included knowledge resist attempts to hard coding and simulation by computer systems.

Recurring attempts are made to engineer so-called knowledge representation, exploiting ideas originating in formal logic. The data structures produced are interpreted as models of the real world. The knowledge engineering process exploits natural language descriptions of the real world (e.g., artefacts, collections, decision making) with the intention to construct a formal description, which offers a trustful homomorphic mapping of the real world (Puppe et al. 2000, p. 600).

Once the description architecture is considered suitably implemented, the knowledge engineer can hope to realize the crucial ambition of the whole enterprise: to simulate human transaction on knowledge using computation. Obviously, why algorithms cannot replace human knowledge, computerised knowledge representation allowing user-driven or user-independent goal-oriented interaction is considered beneficial. For example, "information" (meaning here: reality-mapping using data structures and rules) is architectured and packaged as expert sys-

tems to provide diagnosis and problem solving strategies (Charniak & McDermott 1985, pp. 455–456). Furthermore, produced information can be integrated in Knowledge Maps so that human beings can navigate through large fields and retrieve results more easily.

Conclusions

In this paper, we have defended the view that information should be treated as a function of life. We have also emphasized that the comprehension and usage of information and data differs from discipline to discipline. After the introduction of computational methods in the humanities, different efforts to represent and model cultural artefacts and their information structure have been deployed.

These methods reflect standard approaches common in other disciplines, which concentrate on other subjects of research, e.g., genetic research. While the data and meta-data modelling techniques are adopted, neither the theoretical goals of experimental and observational science (most commonly variants of explanatory reduction), nor the hermeneutical project, are pursued in the wake of information modelling. The result may be some sophisticated craft applied to cultural artefacts, building upon undeclared assumptions that the "real nature" of artefacts is to be found in their digital (re-)presentations.

By using computational approaches in the humanities, researchers or other involved experts, e.g. librarians, should be aware of the often-loose notion of information and the changes in the history of this term. It might be useful to inspect carefully methods derived from other disciplines before introducing them in humanistic approaches. Furthermore, humanists should attempt to define concepts of information, which may fit within the theoretical horizons of the humanities. One should also ask critical questions about the ideological, sociological, and epistemological context of the demand for highly detailed, precise engineering when modelling cultural artefacts and information. For the humanities, informational high-precision modelling is not an absolute prerequisite in many analytical situations. Furthermore, one can oppose to such quasi-exhaustive norms of precision the alternative vision of exploiting various computational approaches to produce rich, and occasionally precise descriptions of artefacts exposing various, possibly incompatible perspectives. Multi-perspective descriptions may rescue the research material from the grip of the single expert's dominance and involve a wider community of experts and users. It may offer new experimental opportunities and offer a wider research space to expose, map, and express differences and disagreements, which, after all, fuel the core of humanistic research.

References

Bechtel, W. & Abrahamsen, A. (2002) *Connectionism and the Mind*. Oxford: Basil Blackwell Publishers.

Berners-Lee, T. et al. (2001) The Semantic Web – A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities**.** *Scientific American*, May 2001.

Betti, E. (1955) *Teoria generale della interpretazione.* 2 vols., Milano: Dott. A Giuffré.

Blackburn, P., de Rijke, M. & Venema,Y (2001) *Modal Logic.* Cambridge: Cambridge University Press.

Brémond, C. (1966) *La logique des possibles narratifs.* Communications 8, L'analyse structurale du récit. Communications, 8, Paris.

Castells, M. (2004) Informationalism, Networks, and the Network Society: A Theoretical Blueprint. In: Castells, M. ed. *The Network Society, A Cross-cultural perspective.* Northhampton, MA: Edward Elgar, pp. 3–48.

Cerquiglini, B. (1989) *L'éloge de la Variante, Histoire critique de la philologie*. Paris: Seuil.

Chaitin, G. (2006) The Limits of Reason. *Scientific American*, 294, No. 3 (March 2006), pp. 74–81.

Charniak, E. & McDermott, D. (1985) *Introduction to Artificial Intelligence.* Reading (Mass.): Addison-Wesley.

Dretske, F. (1981**)** *Knowledge and the Flow of Information.* Oxford: Basil Blackwell.

Divers, J. (2002) *Possible Worlds***.** London: Routledge.

Dolezel, L. (1976) Narrative Modalities. *Journal of Literary Semantics*, No. 5, pp. 5–14.

Dolezel, L. (1976) Narrative Semantics. *PTL*, No. 1, pp. 29–51.

Doubrovsky, S. (1966) Pourquoi la nouvelle critique: Critique et objectivité. Paris: Mercure de France.

Durand-Richard, M.-J. (2004) Des lois de la pensée aux constructivisme. *Intellectica*, n° 39, Paris, 2004/2.

Eco, U. (1962**)** *Opera Aperta.* Milano: Bompiani.

Eco, U. (1979) *Lector in Fabula.* Milano: Bompiani.

Fellbaum, C. ed. (1998) *WordNet, An Electronic Lexical Database.* Cambridge (Ma.): The MIT Press.

Fish, S. (1976) Interpreting the Variorum. *Critical Enquiry*, No. 2 (Spring 1976), pp. 465–485.

Fish, S. (1980) Is There a Text in This Class? The Authority of Interpretive Communities. Cambridge (Ma.): Harvard University Press.

Gadamer, H.-G. (1960) *Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik.* 5. Aufl. 1986, Tübingen : J.C.B. Mohr (Paul Siebeck).

Goody, J. ed. (1968) *Literacy in Traditional Societies.* Cambridge: University Printing House.

Goody, J. (1977) *The Domestication of the Savage Mind.* Themes in the Social Sciences. Cambridge: Cambridge University Press.

Goody, J. (1986) *The Logic of Writing and the Organization of Society.* Studies in Literacy, the Family, Culture and the State. Cambridge: Cambridge University Press.

Goody, J. (1987) The Interface between the Written and the Oral. (Studies in Literacy, the Family, Culture and the State). Cambridge: Cambridge University Press.

Greimas, A.J. (1966) *Sémantique Structurale.* Paris: Larousse.

Greimas, A.J. (1970) *Du sens.* Paris: Larousse

Greimas, A.J. (1973) Les actants, les acteurs et les figures. Paris: Larousse.

Halpen, J.Y. (2003) *Reasoning About Uncertainty.* Cambridge (Ma.): MIT Press

Hardt, M. & Negri, A. (2000) *Empire.* Cambridge (Ma.): Harvard University Press.

Herrick, P. (1999) *The Many Worlds of Logic.* Oxford: Oxford University Press.

Hintikka, J. (1962) *Knowledge and Belief.* Ithaca, NY: Cornell University Press.

Hottois, G. (1984) Le signe et la technique. La philosophie à l'épreuve de la technique. *Collection «Res – L'invention philosophique»***,** p. 59–60. Paris: Aubier Montaigne.

Ingardten, R. (1931) *Das literarische Kunstwerk.* Halle: Niemeyer.

Ingardten, R. (1968) *Vom Erkennen des literarischen Kunstwerkes*. Tübingen: Niemeyer.

Jameson, F. (1961) *Postmodernism, or, The Cultural Logic of Late Capitalism*. Durham: Buke University Press.

Jameson, F. (1971) Marxism *and Form: Twentieth Century Dialectical Theories of Literature.* Princeton: Princeton University Press.

Kohonen, T. (2001) *Self-Organizing Maps.* Springer Series in Information Sciences, Vol. 30, 3rd extended edition.

Kripke, S. (1963) Semantical analysis of modal logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, No. 9, pp. 67–96.

Leibniz, G.W. (2001) *Theodicy*. Original ed. 1701. London: Wipf & Stock Publishers.

Nef, F. ed. (1976) *Structures élémentaires de la signification.* Paris: Éditions Complexe.

Noireaux V. et al (2005) Toward an artificial cell based on gene expression in vesicles. *Phys. Biol*., No. 2, 3, September 2005, pp. 1–8.

Neumann, J. Von (1966) *Theory of Self-Reproducing Automata.* Edited and completed by A. W. Burks. Urbana: University of Illinois Press.

Ong, W.J. (1982) *Orality and Literacy: The Technologizing of the Word*. London and New York: Routledge.

Puppe, F., Stoyan, H. & Studer, R. (2000) *Knowledge Engineering***.** In: Görz, G. ed., Handbuch der Künstlichen Intelligenz. München: Oldenburg, pp. 559–641.

Ricœur, P. (1969) *Le conflit des interprétations.* Paris: Seuil.

Robinson, P. The History, Discoveries and Aims of the Canterbury Tales Project. *The Chaucer Review,* No. 38, 2, 2003, pp. 126–139

Staiger, E. (1955) *Die Kunst der Interpretation.: Studien zur deutschen Literaturgeschichte*. Zürich: Atlantis.

Witbrock, M. et al. (2005) Knowledge Begets Knowledge: Steps towards Assisted Knowledge Acquisition in Cyc. In: *Papers from the 2005 AAAI Spring Symposium on Knowledge***,** Stanford, California, March 2005, pp. 99–105

Wittgenstein, L. (1965) *Philosophical Investigations.* New York: The Macmillan Company.

Zadeh, L.A. (1965) Fuzzy Sets. *Information and Control*, No. 8, 1965, pp. 338–353.

Other electronic sources [accessed 1 September 2007]:

The Canterbury Tales Project. Available from: <http://www.canterburytalesproject.org>

The Cyc Knowledge Base TM. Available from: <http://www.cyc.com/cyc/technology/-whatiscyc_dir/whatsincyc>

Libchaber, A.J., Home page. Available from: <http://www.rockefeller.edu/research/abstract.php?id=93>