Defining Level and Scale as Socio-technical Operators for Mining Digital Traces

Quentin Lobbé, David Chavalarias, Alexandre Delanoë

Quentin Lobbé, David Chavalarias and Alexandre Delanoë

Defining Level and Scale as Socio-technical Operators for Mining Digital Traces

Abstract: In this article, we investigate the epistemological dimensions pertaining to the notion of *scale* in Digital Humanities (DH). We first echo the growing concerns of digital tools makers who call for separating the notion of *complexity* from the metaphor of *scale* frequently used in the DH literature. We then harvest a corpus of 825 DH papers related to the notion of *scale* and we build a semantic map on top of them to highlight the various ways DH scholars make use of *scale* in their contributions. By reviewing this map, we show that *scale* acts as a blurry concept in DH literature along with *level* used as a sibling. We then argue for distinguishing between *level* and *scale* to reconstruct and visualize the complexity of the social sphere. We redefine *level* and *scale* as operators of a *socio-technical algebra*. This algebra aims at increasing our collective capacity to mine digital traces. We then give practical examples of the joint use of *level* and *scale* with the *phylomemy reconstruction process*. We finally introduce *GarganText* a free text mining software heir of de Rosnay's *Macroscope*. We explain how *GarganText* embeds our definitions of *level* and *scale* considering mathematical foundations, programming technologies and design principles.

Keywords: level, scale, phylomemy, semantic map, macroscope

1 Introduction

The metaphor of *scale* has spread widely in the literature and daily vocabulary of Digital Humanities (DH). As scholars in Social Sciences (SS) are more and more interested in the scientific potential of digital resources, they look for new terminologies to revisit classical concepts and define innovative approaches. However, the epistemological definition of *scale* remains blurry and each sub-domain of DH has its own interpretation. In the best-case scenario, it can either refer to the promises of digital data or to some exploration mechanisms implemented within software; otherwise, *scale* is simply used as a stylistic formula. Yet, other research domains – especially the field of Complex Systems (CS) – have deeply investigated the notions of *scale* and have also introduced the sibling concept of *level*. This vocabulary issue is of great importance: by clearly shaping the notions of *level* and

https://doi.org/10.1515/9783111317779-015

scale we might reinforce the epistemological connections between Social and Computer Sciences. Indeed, we argue for the creation of a socio-technical algebra able to translate any research question of Social Sciences as a mathematical proposition interpretable by computer scientists and implementable in a software. Regarding the current confusion around the metaphor of *scale* in DH, we have chosen to focus on both *level* and *scale* to turn them into the first two operators of our future socio-technical algebra. As part of the *Zoomland* effort, our paper will thus argue for re-defining *level* and *scale* (section 4) by using arguments from digital tool-makers (section 2) and by conducting a wide review of the DH literature (section 3). We will then give practical examples of how *level* and *scale* can be used to reconstruct socio-historical processes from digital resources (section 5) and introduce *GarganText* a free text mining software that implements these two notions (section 6), which makes it possible to combine these operators for investigating a research question.

The evolution of social sciences

Social Sciences have been deeply impacted by the 2000's revolution of the Information and Communications Technologies (Borgman 2003). Over the past 20 years, the unprecedented flow of digital data produced by communication devices and electronic networks has induced an epistemological shift among SS: as new research questions arose, fieldwork and experimental practices evolved to study the digital world. SS had to investigate the inner nature of digital data, explore the scientific capacity of such new materials and create dedicated tools (Edelmann 2020). Scholars then organized themselves and gave birth to the wide domain of *Digital Humanities* (Mounier 2012) and sibling fields: *Digital Studies* (Stiegler 2016), *Digital Sociology* (Boullier 2015), *Digital History* (Kemman 2021), *Digital methods* (Rogers 2013), *Cultural Analytics* (Manovitch 2016), etc. Within the scope of DH, researchers now consider the digital world as a new reflexive way to study our societies and collective memories.

The promises of digital resources

In the same time, archiving and knowledge institutions (libraries, museums, etc.) invested in state-of-the-art infrastructures to store, curate and browse large data-

bases of digital resources (born-digital data or late-digitized data[1]). By doing so, they allowed for diving into the richness of these catalogs and mine digital traces. Indeed, digital resources are *charged* with an *evidential power* (Ginzburg 2013) that bear witness to some socio-historical processes and can thus be considered as *traces* of the social sphere. But investigating such data is not a trivial task and DH scholars have to tackle various socio-mathematical issues when they face digital traces: volume, heterogeneity, discontinuity, incompleteness, etc. Yet, these efforts are worthwhile as the power of connectivity of digital resources (hypertext connections, dynamical properties, multi-dimensional relationships, etc.) enables more complex and valuable studies (Boullier 2015).

The point of view of digital tool-makers

The mutation of part of Social Sciences into Digital Humanities is intrinsically connected to scholars' comprehension of the nature and potential of digital resources. Consequently, digital tool-makers[2] have become central within the scientific ecosystem of DH and their voices should be taken into account and integrated into our own discussions. So far, the notion of *scale* in DH has mostly been empirically investigated and materialized through the uses of software designed by digital tool-makers: *scale* is an *in situ* feeling whose potential is still largely unexplored. Recently, tool-makers have been interested in the mathematical, technical and visual *explainability* of their own tools (Jacomy 2021). They aim at preventing *complexity questions* from being hidden behind the metaphor of *scale* and the illusion of continuity induced by software between original data sets and experimental outcomes. This idea of *not hiding the complexity* will now guide our reasoning and help us to re-define *scale* and *level*.

2 A false feeling of continuity

In what follows, we will focus on data exploration tools used by scholars in Digital Humanities as they all share a common *undertone*: they are powered by *scale* mechanisms. Tool-makers usually denominate exploration tools as *datascapes*; that is, ad hoc exploratory environments used by scientists to study digital mate-

1 With some exceptions, we will mainly use *digital resources* to refer to both *born-digital* and *late-digitized* data.

2 Engineers, computer scientists or social scientists acculturated to digital technologies.

rials (Girard 2017). These tools are always built in relation to a specific data set, an issue or a research topic– in the same way as R. Rogers issue-driven methodology for building digital tools (Rogers 2013). Datascapes can thus be seen as a *monadic* way to investigate digital traces.[3] (Tarde 2011) There, what matters most is the upstream harvesting of an input digital material and the initial hypothesis formulated by the researchers that will later drive the making of the datascape. Sometime datascapes become standalone software such as the graph analysis software *Gephi* (Bastian 2009): a datascape that emancipated itself from ad hoc constraints and became generic. Overall, every datascape conveys a shared feeling, the impression that one can zoom within the data, navigate the digital resources, change the *"scale of analysis"*.

Scaling mechanisms are inherent to datascapes. For instance, some researchers use *zoom* features to explore a citation network with Gephi, other scientists apply a change of scale by annotating individual documents to reveal – from a bird's eye view – interactions that structure a body of texts, some scholars take advantage of geographical data to analyze the dynamics of an historical process in various places, etc. In the context of digital humanities, tool-makers rely on the nature of digital resources to unblock the navigation from individual elements to collective structures. *Zooming in* or *out* thus appears to be a natural metaphor to explain how the majority of exploration tools work (Boullier 2016). However, *change of scales*, as a DH notion, has never been investigated. This is a major issue because *zoom* mechanisms induce a fake feeling of *continuity* between the original digital resources and their future explorations. Tool-makers hide a lot of un-natural tasks behind *multi-scale* features (Boullier 2016): filtering, aggregations, re-processing, etc. The intrinsic complexity contained within the original digital resources is thus reduced before being explored through any interface. We think that this is the source of the paradox described by M. Jacomy (tool-maker and co-inventor of Gephi) in his thesis:

> A Gephi user once told me: "*Gephi understands the network, but I do not understand Gephi.*" I understand this statement as an acknowledgement that the visualization is correct despite being incomprehensible. (Jacomy 2021: 190)

By hiding the complexity of the digital resources behind a false feeling of continuity, implicit *zoom* mechanisms induce *visual explainability* issues. The notion of *scale* needs to be defined and separated from the notion of *complexity* to improve

3 In Metaphysics monad means unity, in Mathematics monads are used to define sets of rules between categories sharing a common space (ie, adjunctions), in functional programming monads are used to abstract control flows and side-effects.

exploration and analysis processes in DH. Our goal in this paper is to follow a *complex systems* approach to define the concept of *scale* along with the sibling concept of *level*. We think that this clarification can contribute to the conceptual corpus of Digital Humanities and guide the making of innovative tools as suggested by M. Jacomy to study social and historical phenomenon:

> We cannot see into the complex as if it was simple. We must switch metaphors and build our scientific apparatus from a different perspective. We must build something else, for instance, *complexoscapes* — composite visualization systems where inevitable reductions are counterbalanced by the possibility of navigating between complementary views and visualizations. (Jacomy 2021: 190)

3 Scale and level as undertones in Digital Humanities

To complete the point of view of digital tool-makers (see section 2) and support our thought, we will now conduct a wide review of a corpus of Digital Humanities papers that somehow use the words *scale* and *level* in their arguments. By analyzing the many ways *scale* and sibling expressions are used in the DH literature, we will understand how the DH community organizes itself (or not) around this notion. We will use bird's-eye visualizations of the whole corpus and review individual contributions to improve our analysis.

We first harvest a corpus of 825 scientific papers' metadata[4] (titles and abstracts) extracted from both the *Web of Science* and *Scopus* and matching the query *("digital history" or "digital humanit*") and (scale or level or multi-level* or multi-scale* or macroscope or "scalable reading" or "deep mapping")*. With the free text mining software *GarganText*[5] we then extract a list of terms and expressions used within the papers by the researchers. GarganText next reconstructs connections between these terms by computing the conditional probability of having one term written in a paper jointly with another from the list. The resulting map (Figure 1) shows the semantic landscape of our corpus. There, terms are dots and semantic relationships are edges. Colors highlight communities of terms more frequently used together, these groups represent the main subjects of research and communities of interest hidden within our 825 papers. We count 5 distinct communities: *digi-*

4 The corpus, the list of terms and the resulting map can be downloaded at https://doi.org/10.7910/DVN/8C1HKQ. Accessed July 10, 2023.

5 See https://cnrs.gargantext.org/. Accessed July 10, 2023.

tal history and the detection of patterns in literature (purple), *the issue of digitization of cultural heritage* (pink), *modeling as goal* (brown, Figure 2), *digital library and the quality of metadata* (orange) and *the issue of visualization* (blue, Figure 3)

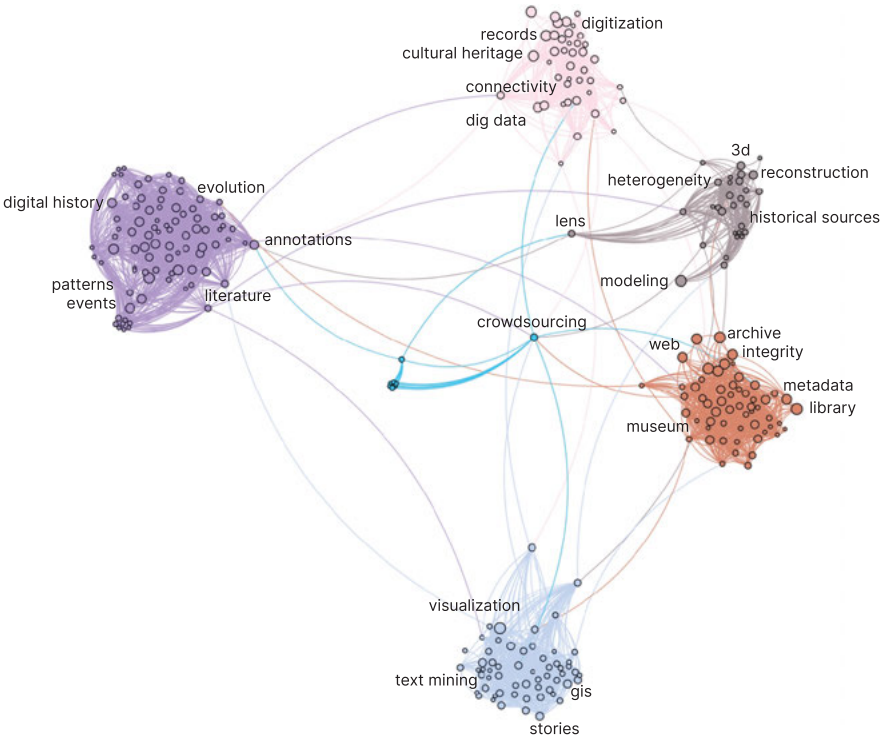

Figure 1: Semantic map of 825 scientific papers extracted from both the Web of Science and Scopus. Built with Gargantext and spatialized with Gephi.

The semantic map (Figure 1) shows that *scale* is not a unified notion in the DH literature as our corpus of papers organized itself in very distinct and distant communities (a unified literature would have produced a single, large and central community). Yet, *scale* lives through the entire DH literature, not as a well-defined concept, but more as blurry undertone. There, *scale* conveys various meanings among different communities:

– promises for future works (“how to scale up the solutions based on collaborative research efforts” (Tolonen 2019))
– range of actions over a digital data set (“micro-scale uploads” (Mcintyre 2016), “fine-grained annotation” (Wang 2021))

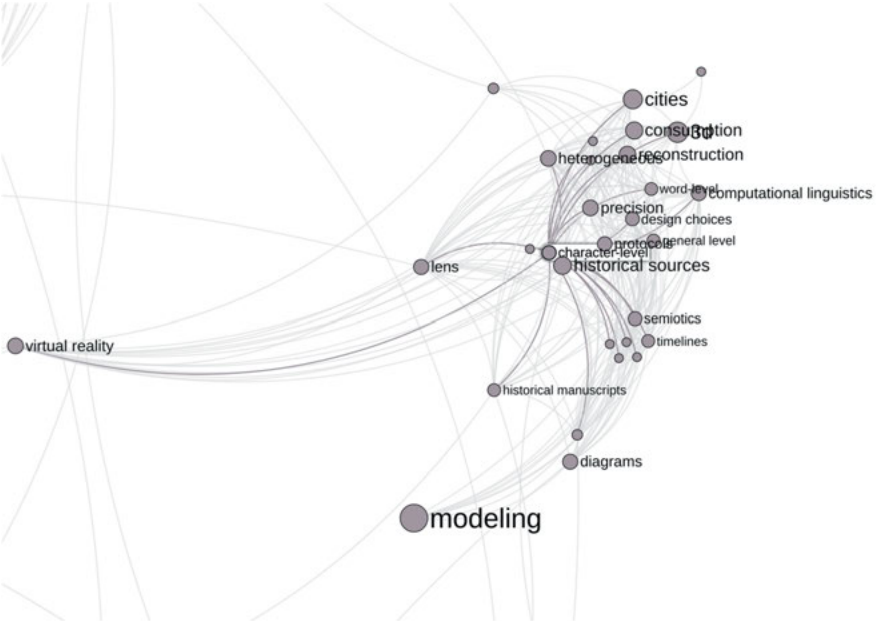

Figure 2: Detail of the community modeling, reconstruction and virtual reality from the map of Figure 1.

- impact, quality or context of an historical event / object ("the rise of revolutionary movements made manifest through large-scale street actions" (Sakr 2013), "a center of high level of artistic production" (Boudon 2016), "the reconstruction of macro- and micro-contexts" (Nevalainen 2015))
- quantity of data analyzed in a paper ("large scale analysis" (Risi 2022), "visualizing information on performance opens new horizons of significance for theatre research at scales" (Bollen 2016))
- scalability issues for data engineering ("regarding the overall archiving storage capacity and scalability" (Subotic 2013))
- scope of the observed patterns or temporal motifs ("macro-level patterns of text and discourse organization" (Joulain 2016))
- choice of an analytical layer in the case of multi-dimensional or cross-domain analysis ("every object that is catalogued is assigned entry-level data, along with further data layers" (Edmond 2017))
- exploration and visualization tasks ("interfaces are a valuable starting point to large-scale explorations" (Hinrichs 2015), "screen-based visualization has made significant progress, allowing for complex textual situations to be captured at the micro-and the macro-level" (Janicke 2017)) see Figure 3;

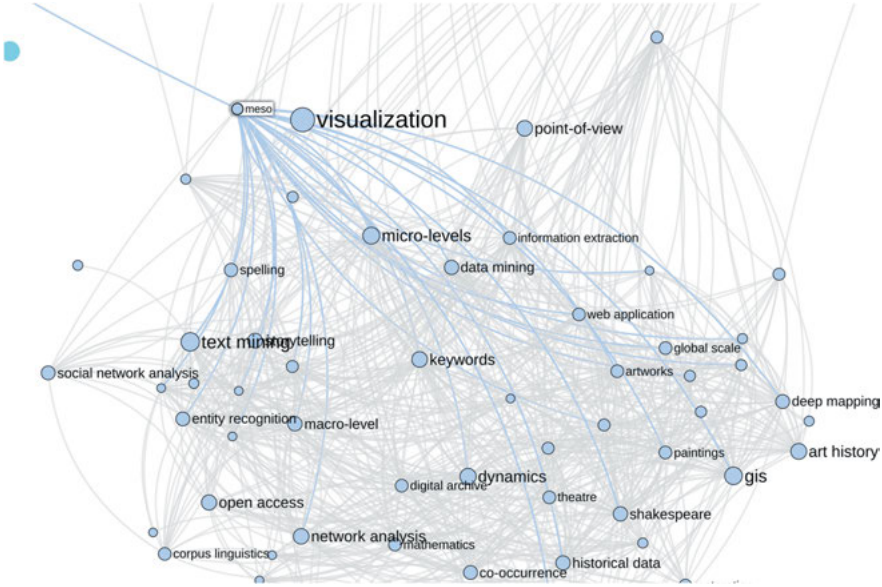

Figure 3: Detail of the community data visualization from the map of Figure 1.

– choice of a fragmented / elementary reconstruction approach ("our method relies on character-level statistical machine translation" (Scherrer 2015), "a relatively novel technology which allows to effectively represent the intrinsic word-level uncertainty" (Prieto 2021))
– an ad hoc question of complexity ("levels of granularity" (Allen 2011), "levels of complexity" (Rose 2016)) see Figure 2.

Thanks to the Figure 1 and details, we reveal that the epistemological frame of *scale* in DH literature is not yet stabilized. For our part, we think that *scale* should be separated from all notions of *complexity* to clear its definition and help tool-makers to create the next generation of datascapes (see section 2). After having reviewed the corpus used for building the map, we are able to clear two distinct proto-definitions of *scale* and *level* as notions:

1. connected to *visualization* and exploration tasks, *scale* can be seen as a zoom mechanism. Here, *scale* becomes a design principle, a feature of interactivity. By allowing the user to navigate through *multiple* scales, the ergonomic choices behind the datascapes become an *empirical and analytical algebra* for social scientists;

2. the word *level* is used as a sibling for *scale* but not as synonym. In fact, *level* is mostly associated to the notion of *complexity* of the research subject; that is, a *reconstruction* choice made upstream from the visualization task.

In what follows, we will use these two proto-definitions enriched with Complex Systems concepts to introduce our own definitions of what we call: level of observation and scale of description.

4 Level of observation and scale of description

In order to distinguish between the notions of *level* and *scale*, we now use elements of knowledge from the Complex Systems' literature (CS). As a research domain, CS naturally addresses the questions of *scale*, *level* and *complexity* as it aims at studying how collective structures and global dynamics emerge from individual interactions and how natural or artificial phenomena evolve, connect or enrich themselves throughout time. Furthermore, the two proto-definitions outlined in section 3 echo the most recent outcomes of CS' literature.

4.1 Level

CS scholars have recently made a clear distinction between *level* and *scale* (Chavalarias 2021). *Level* is generally defined as a domain higher than *scale* and *scale* refers to the structural organization within a given *level*. Choosing a level of observation means making a choice of complexity. In Biology, for example, the choice of a given *level* determines what the main entities under study (organs, cells, genes, etc.) are. Applied to Digital Humanities, choosing a *level* means choosing the intrinsic complexity of the processes we want to analyze. In the case of quantitative Epistemology for instance, choosing a micro *level of observation* means choosing to reconstruct the evolution of a given scientific domain (from a corpus of scientific publications) by looking at the way this domain has resulted from the temporal and internal combinations of many sub-research fields. There, these small fields of research can be considered as elementary entities of analysis. But choosing a macro *level of observation* instead means choosing to consider larger research domains (for instance, Sociology or Biology) as elementary entities of analysis. By doing so, we will focus on the evolution of the inter-disciplinary interactions between wide research domains.

4.2 Scale

For its part, the choice of a *scale* determines the resolution adopted to describe a phenomenon at a given level. The *scale* can be seen as an exploration principle used for zooming through a given visualization. By re-scaling this visualization, we explore different scales of description and we navigate among layers from individual interactions to macro structures (Lobbe 2021). The choice of a given *level* occurs once and for all during what we call the *reconstruction* step upstream from the visualization task. The goal of the reconstruction task is to model a phenomenon from a collection of harvested, curated and annotated digital traces. We then use these traces to create a mathematical approximation for the phenomenon's structure and behavior (a network for instance). This reconstructed object is then projected in a visualization space where it can be explored and described throughout different *scales*.

4.3 Socio-technical algebra

Levels and scales can now be combined as distinct operators of a socio-technical algebra. Such an algebra aims at translating any social science research questions into mathematical formulas comprehensible by computer scientists. These formulas will eventually be implemented within software or graphically translated through interactive user-interfaces. So, let's call φ a generic data analysis process (represented by a triangle in Figure 4). φ consists of three standard steps: the data collection (*digital resources* in Figure 4), the reconstruction (i.e., the computer-based modelling of the targeted socio-historical phenomenon) and the visualization (i.e., the exploration of the reconstructed phenomenon through an interface by a researcher). The choice of a level occurs during the reconstruction step. The choice of a scale occurs during the visualization step. According to our socio-technical algebra, φ can be seen as a function of both levels x and scales y such as $\varphi = f(x, y)$ (see Figure 4). Thus, DH scientists first start by harvesting digital traces related to their research question. Then, they choose a level of observation that will determine the complexity of the modelling of the phenomenon under study. Finally, they visualize the complex object (a map for instance) reconstructed by the computer through an interface. By interacting with this interface, they will move from one scale of description to another and base their upcoming analysis upon this choice of scale. In the future, this socio-technical algebra will be enriched with additional operators. For instance, in Section 6 we will introduce the *order 1* and *order 2* metrics: two extra operators that can be combined with level and scale.

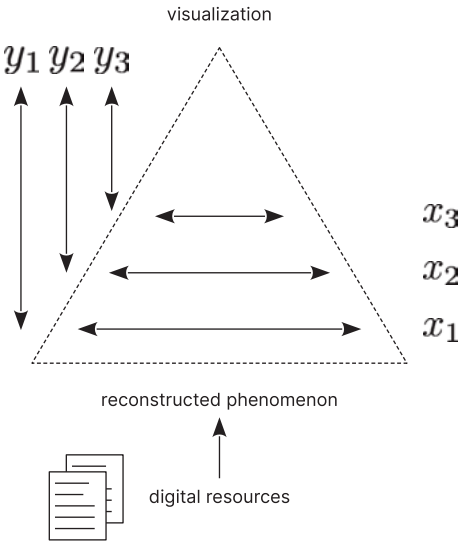

Figure 4: Diagram of the generic digital data analysis process φ seen as a function of levels x and scales y.

5 Using level and scale with phylomemy reconstruction

In this section, we will give a concrete example of how *level* and *scale* can be used by DH scholars to reconstruct and explore socio-historical processes. We address the following epistemological research question: how can we reconstruct the evolution of the scientific landscape of a given research domain throughout time? This question assumes that the evolution of science can be reconstructed from the main traces left by researchers and scholars; that is, publications and scientific papers. As a constraint, we won't use any pre-existing structured resources (citations graphs, ontologies, etc.) apart from an initial corpus of scientific publications: the dynamic structure will emerge from the co-occurrence and co-use of scientific terms and expressions in time. In fact, this research question can be extended to all type of timestamped corpus of textual documents: how can we reconstruct the evolution of a corpus of letters? How can we reconstruct the evolution of a literary genre? How can we reconstruct the evolution of online debates? How can we visualize the coverage of a targeted public event by newspaper through time? etc.

To answer this question, we will make use of a new scientific object: the phylomemy (Chavalarias 2021; Lobbe 2021). Phylomemies can be reconstructed on top of any timestamped corpora of text data. The *phylomemy reconstruction process* is part of the larger family of co-word analysis approaches: a type of text mining

techniques used to analyze the co-occurrences of words, terms or expressions within texts (Callon 1986). More precisely, Phylomemies are designed to reconstruct the dynamics of *terms-to-terms relationships* through time and visualize the evolution of *semantic landscapes*. Phylomemies are thus inheritance networks of textual elements of knowledge. Applied to the analysis of the evolution of science for instance, phylomemies can be seen as genealogical trees of scientific fields that structure themselves in evolving branches of knowledge; that is subdomains of research contained within a given discipline.

5.1 Methodology

Phylomemies have first to be *reconstructed* from a collected set of documents before being *visualized* and explored through dedicated software (see Section 6). The reconstruction process can be divided into four subsequent steps:[6] 1) *Indexation*, we first frame a corpus of texts and then extract its core vocabulary (terms or expressions). We next choose a temporal resolution (e.g. 3 years) that chunks this corpus among ordered sets of equal periods. Within each period, we compute the terms' co-occurrence their co-presence in the original documents. 2) *Similarity measures*, within each period we use the co-occurrence of terms to compute a similarity measure. It results in graphs of similarities potentially containing meaningful groups of terms frequently used together. We call these groups *fields*. 3) *Field clustering*, a clustering algorithm is then used to detect coherent fields of terms within each period. 4) *Inter-temporal matching*, an inter-temporal matching mechanism reconstructs the kinship relations between fields from one period to another. It assigns each group of terms a set of parents and children by using a semantic similarity measure. By doing so, we highlight elements of semantic continuity over time called *branches*.

5.2 Level and scale with phylomemies

The phylomemy reconstruction process already considers the notion of *level* and *scale* as defined in section 4: the level can be set up during the reconstruction step; the scale can be set up during the visualization step. There, the *level of observation* has been modelled as a continuous variable $\lambda \in [0,1]$ and a quality function F_λ has

6 See Chavalarias 2021 for details concerning the algorithms and text mining techniques used in the phylomemy reconstruction process.

been designed to control the intrinsic complexity of the phylomemy (Chavalarias 2021). F_λ aims at answering the following question: "What should be the global shape of the phylomemy, so that for any term x I could be able to find an *informative* branch of knowledge dealing with x?". By choosing a level λ between $[0, 1]$, we influence the informativeness of the branches of the resulting phylomemy: they thus might vary from very precise branches to very generic branches. For a low *level* ($\lambda \to 0$) all fields and terms will be connected within few but large branches; For a high *level* ($\lambda \to 1$) the phylomemy will look like an archipelago of specific and accurate branches. Then, once the phylomemy has been reconstructed, it is projected in a visualization place – a dedicated datascape – where researchers can explore the evolving structure from term relationships to branch similarities and choose the good scale of description and resolution to analyze the whole object (Lobbe 2021).

As we have already harvested a corpus of scientific publications in section 3, we will now reuse it (along with its list of terms) and reconstruct its semantic evolution. This will give us clues of how the domain of Digital Humanities has positioned itself regarding the notion of *scale* in the last 15 years. This will enrich our analysis of Figure 1 with a temporal perspective. We first reconstruct the phylomemy for a *level* $\lambda = 0$.

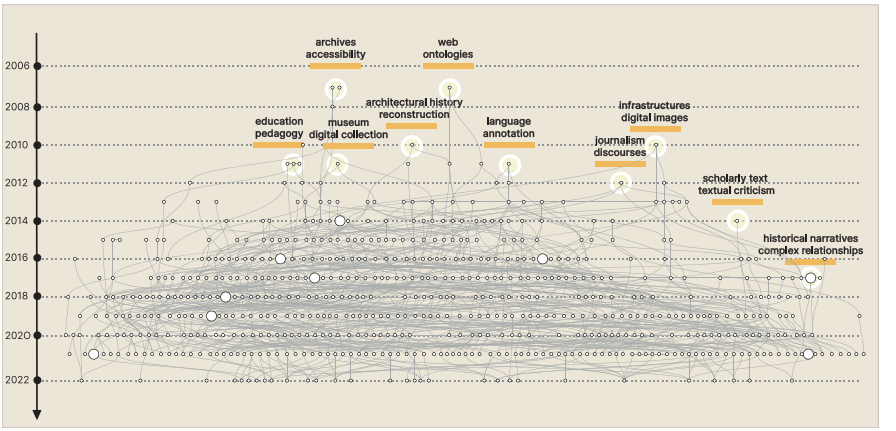

Figure 5: A phylomemy of the scientific literature of digital humanities reconstructed for $\lambda = 0$.

In the resulting Figure 5, the phylomemy must first be read from top to down. Circles represent groups of terms jointly used together in a set of papers at the same time. The bigger the circle, the larger the number of matching documents. For readability motive, we here choose not to display the textual content of the

circles. We highlight in yellow the origins of significant sub-research domains of DH motivated by classical pre-digital research subjects (*archiving*, *education*, *textual criticism*, etc.) or new type of digital resources (*web data*, *digital images*, *digitized scholarly texts*, etc.).

But this phylomemy (Figure 5) cannot be considered as informative: its complexity is too high, we need to simplify it (ie, remove weak kinship links) to reveal more structured shapes. To that end, we now reconstruct the same phylomemy for higher *levels* $\lambda > 0$.

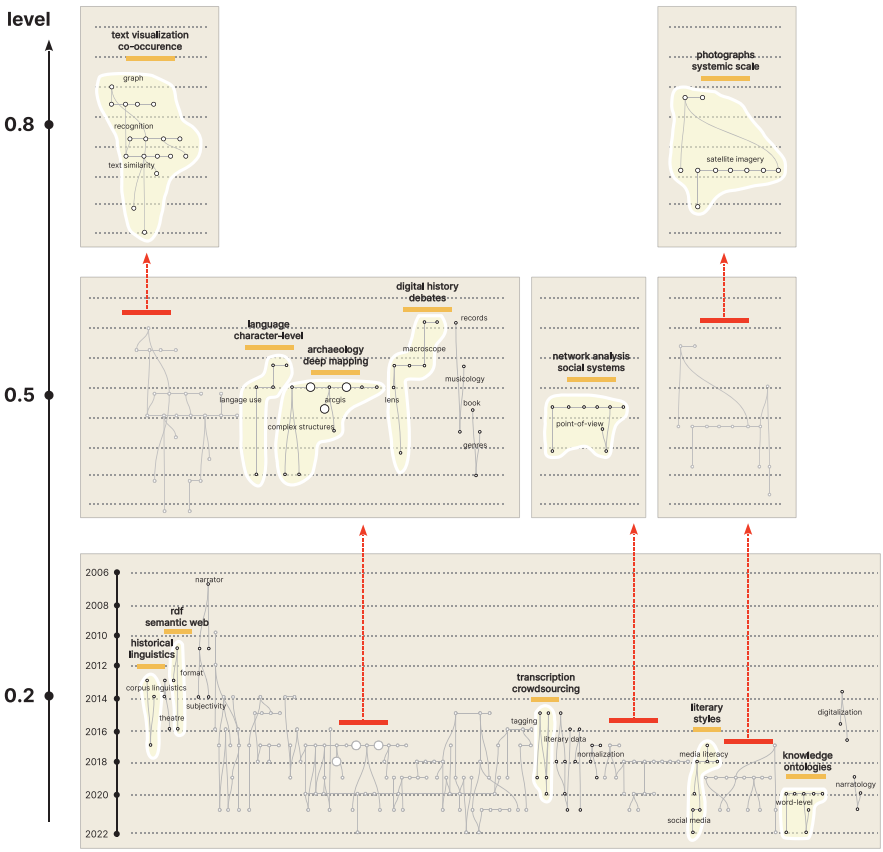

Figure 6: The progressive specialization of phylomemies of the scientific literature of digital humanities reconstructed for $\lambda = 0.2$, 0.5 and 0.8.

The resulting Figure 6 shows how the phylomemy specialized itself for subsequent *levels* $\lambda > 0$. To build this visualization, we have first tested many values of λ and we have then taken on the values for which the complexity of the phylomemy significantly changed that is: $\lambda = 0.2$, 0.5 and 0.8. Thus, the scientific landscape of DH – regarding our original corpus of papers – starts to structure itself at $\lambda = 0.2$: the single and large branch (Figure 5) breaks into smaller branches that represent satellite sub-research domains of DH like *historical linguistic*, *literary style analysis*, *crowdsourced transcription and annotation*, etc. We then need to reach $\lambda = 0.5$ to see the emergence of the main Digital Humanities' research branches: *digital history*, *archeology*, *network analysis*. Finally, very specific branches like *text visualization* appear and stabilize themselves at $\lambda = 0.8$.

By choosing the example of phylomemies, we see how our definition of *level* can be used to analyze the same process – the temporal evolution of DH literature – at different degrees of complexity and how we influence the intrinsic structure of the resulting visualization. We finally invite the readers to explore the Figure 6 in our online datascape[7] to experiment some *multi-scale* navigation mechanisms. Indeed, phylomemies embed an endogenous *scaling* mechanism that builds on the kinship links of each branches. The weights of these links (ie, the semantic similarity measure) are sorted and distributed among increasing ranges that result in a finite number of *scales* per branch. By moving from one *scale* to another within our datascape, scholars can choose a suitable resolution for each branch and aggregated groups of terms whose link weight is inferior to the selected *scale*. In the Figure 7, we show how this mechanism can be used on the unnamed grey branch of the phylomemy $\lambda = 0.2$ of Figure 6. This branch embeds about twenty different *scales* of description.

6 Beyond level and scale, introducing GarganText

Recent technical reviews (Chavalarias 2021; Lobbe 2021) have shown that no software or datascapes are today able to implement the notions of *level* and *scale* as defined in Section 4. That's the reason why we have decided to create our own *complexoscope* (see Section 2). We here want to introduce GarganText[8] (Delanoe 2023) a free text mining software, heir of De Rosnay's *macroscopes* (Derosnay 2014); that is,

7 The phylomemy reconstructed for level 0.2 can be explored at http://maps.gargantext.org/phylo/zoomland/. Accessed July 10, 2023.

8 GarganText has been invented and is developed by A. Delanoë at the ISCPIF CNRS, see https://gargantext.org/. Accessed July 10, 2023.

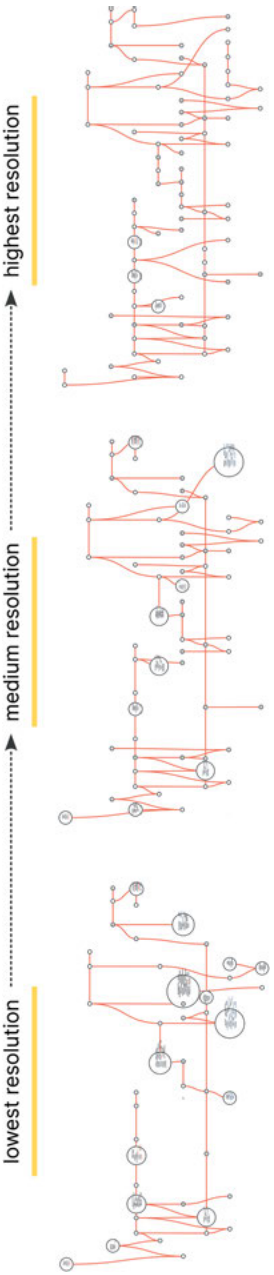

Figure 7: Multi-scale mechanism for choosing a suitable resolution for a given branch of knowledge.

a software designed to reveal the global structure and dynamics of corpora of textual documents by computing individual interactions (co-occurrence,[9] conditional probability,[10] temporal similarity,[11] etc.) between terms and expressions. Contrary to classical datascapes, GarganText is *data agnostic*, it takes any textual elements as input data (short texts, novels, corpus of thousands of papers, etc.) and produces three types of visualizations: basic charts, semantic maps and phylomemies. Between the initial corpus and the resulting visualizations, GarganText uses the functional capacity and mathematical stability of the programming language *Haskell*[12] to materialize an agile data analysis process. By using a functional programming language, GarganText allows scholars to control and customize their analysis at will by choosing between different strategies (implemented as standalone functions in the source code) and even by going backward to previous steps. The researchers can thus choose a first *level* of complexity, jump to a phylomemy, explore its *scales* and then go back to another *level*, and so on. By doing so, GarganText creates a true continuum of data exploration, from backend functionalities to design principles. Finally, GarganText enables collaborative and decentralized analysis.

Discussion. Can we go beyond *level* and *scale*? Can we imagine additional variables to our data analysis process $f(\varphi)$? We think that the answer is yes and GarganText already implements a *linguistic order* when one wants to reconstruct a semantic map. Using the graph functionality of GarganText after having selected together the right terms to throw light on, we can analyze two types of graphs, the one of order 1 and the one of order 2. Each type of graph has its own interpretation regarding our research purpose.

The **order 1** graph is used to approximate the global quality of the corpus. Analyzing its clusters gives a simple idea of the main picture of the corpus by detecting eventual noise in it. The order 1 graph results in semantic clusters from association of terms in conjunction (i.e., terms A and terms B are in the same textual context). The central clusters show the main topics of the corpus and its peripheral clusters describe the secondary themes.

9 For instance, in the corpus of, the terms *micro-digitisation* and *sustainability* are jointly used three times by various authors so their co-occurrence count is 3.

10 For instance in the Figure 1, the weight of the link (ie, the conditional probability) between *knowledge graph* and *rdf* is 0.8. It means that we read *knowledge graph* in a paper, it will be very likely associated to *rdf*.

11 For instance in the Figure 5, the temporal similarity between the group of terms *language, debates, corpus linguistics, historical linguistics & discourse structure* (in 2014) and *term annotations, language, software, corpus linguistics & historical linguistics* (in 2017) is 0.65 as they share 3 out of 5 terms, knowing that these terms are weighted regarding their specificity in the corpus.

12 See https://www.haskell.org/. Accessed July 10, 2023.

The **order 2** graph shows the clusters built from the graph of association of terms in disjunction (i.e., terms A and terms B that can be interchangeably used in same textual context). As a consequence, the clusters throw light on the main concepts of the current corpus. For instance, the Figure 1 is an order 2 graph.

Hence the order 1 and order 2 present two different types of interpretation. First order graph shows the subjects to improve both the quality of the set of documents and the selection of the terms under study. Second order graph shows the main concepts highlighted in the corpus for the research goal of the team working together in the GarganText collaborative working space.

To conclude, we want to call to mind that *level* and *scale* – as defined in Section 4– are two orthogonal notions of a much more complex socio-mathematical algebra meant to mine digital traces. In the future, we invite DH scholars to enrich this algebra by the defining new notions or objects (such as *order 1* and *order 2* graphs) along with interoperability operators: for instance, *the level is a higher order entity than the scale*. GarganText already lays the groundwork for more innovative analytical variables that will take advantage of the power of reflectivity and connectivity of digital resources.

References

Allen, Robert. “Visualization, Causation, and History.” Association for Computing Machinery (2011): 538–545.

Bastian, M., Heymann, S. and Jacomy, M. “Gephi: an open source software for exploring and manipulating networks.” In *Proceedings of the international AAAI conference on web and social media* 3.1 (2009): 361–362.

Bollen, J. “Data Models for Theatre Research: People, Places, and Performance.” *Theatre Journal* 68.4 (2016): 615–632.

Borgman, CL. *From Gutenberg to the global information infrastructure: access to information in the networked world*. Cambridge, Massachusetts: MIT Press, 2003.

Boudon-Machuel, Marion and Charron, Pascale. “Figurative Art of the Loire Valley through the Prism of the Digital Humanities: ARVIVA and Sculpture 3D Projects.” *Actual Problems of Theory and History of Art* 6 (2016): 425–432.

Boullier, D. « Les sciences sociales face aux traces du big data. » *Revue française de science politique* 65.5 (2015): 805–828.

Boullier, D., Crépel, M. and Jacomy, M. « Zoomer n’est pas explorer. » *Réseaux* 1 (2016): 131–161.

Callon, M., Rip, A. and Law, J. Eds. *Mapping the dynamics of science and technology: Sociology of science in the real world*. London: Palgrave Macmillan, 1986.

Chavalarias, D., Lobbe, Q. and Delanoë, A. “Draw me Science-multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies.” *Scientometrics* 127.1 (2022): 545–575.

Delanöe A., Chavalarias D. *Mining the digital society – Gargantext, a macroscope for collaborative analysis and exploration of textual corpora*. forthcoming.

De Rosnay, J. *Le macroscope. Vers une vision globale*. Média Diffusion, 2014.
De Tarde, G. *Monadology and sociology*. Melbourne: re. press, 2011.
Edelmann, Achim, &Wolff, Tom, Montagne, Danielle and Bail, Christopher."Computational Social Science and Sociology." *Annual Review of Sociology* 46.1 (2020): 61–81.
Edmond, Jennifer & Folan, Georgina. (2017). "Data, Metadata, Narrative. Barriers to the Reuse of Cultural Sources." In Communications in Computer and Information Science. 2017. 253–260.
Ginzburg, C. *Clues, myths, and the historical method*. Baltimore, Maryland: JHU Press, 2013.
Girard, P. "Studying and exploring digital traces through datascapes: the interdisciplinary experience at Sciences Po médialab." *Datavisualisation in Sociology and Social Sciences* (2017). Accessed July 10, 2023. https://sciencespo.hal.science/hal-03567429.
Hinrichs, Uta, Alex, Beatrice, Clifford, Jim, Watson, Andrew, Quigley, Aaron,Klein, Ewan and Coates, Colin. "Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration." *Digital Scholarship in the Humanities* 30.1 (2015): 50–75.
Jacomy, M. (2021). *Situating Visual Network Analysis* (Doctoral Thesis, Aalborg University, 2021).
Jänicke, Stefan & Wrisley, David. "Visualizing Mouvance: Toward a visual analysis of variant medieval text traditions." *Digital Scholarship in the Humanities* 32 (2017):106–123.
Joulain-Jay, Amelia. "Developments in English: Expanding electronic evidence." *ICAME Journal* 40 (2016): 173–178.
Kemman, M. *Trading Zones of Digital History*. De Gruyter Oldenbourg, 2021.
Lobbé, Q., Delanoë, A., & Chavalarias, D. "Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge." *Information Visualization* 21.1 (2022): 17–37.
Manovich, L. "The science of culture? Social computing, digital humanities and cultural analytics." *Journal of Cultural Analytics* 1.1 (2016): 1–14.
Mcintyre, Julie. "Blank pages, brief notes and ethical double-binds: micro digitisation and the 'infinite archive.'" *Archives and Manuscripts* 44 (2016): 2–13.
Mounier, P. Ed. *Read/Write Book 2: Une introduction aux humanités numériques*. Marseille: OpenEdition Press, 2012.
Nevalainen, Terttu. "What are historical sociolinguistics?" *Journal of Historical Sociolinguistics* 1.2 (2015): 243–269.
Prieto, Jose, Bosch, Vicente, Vidal, Enrique, Alonso Villalobos, Carlos, Orcero, M. and Marquez, Lourdes. Textual-Content-Based Classification of Bundles of Untranscribed Manuscript Images. *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, 3162–3169.
Risi, Stephan, Nielsen, Mathias, Kerr, Emma, Brady, Emer, Kim, Lanu, Mcfarland, Daniel, Jurafsky, Dan, Zou, James and Schiebinger, Londa. "Diversifying history: A large-scale analysis of changes in researcher demographics and scholarly agendas." *PLOS ONE* 17.1 (2022). https://doi.org/10.1371/journal.pone.0262027.
Rogers, R. *Digital methods*. Cambridge, Massachusetts: MIT press, 2013.
Rose-Steel, T. and Turnator, E. "Medieval Music in Linked Open Data: A Case Study on Linking Medieval Motets." *International Journal of Humanities and Arts Computing* 10 (2016): 36–50.
Sakr, Laila. "A Digital Humanities Approach: Text, the Internet, and the Egyptian Uprising." *Middle East Critique* 22 (2013): 247–263.
Scherrer, Yves and Erjavec, Tomaž. "Modernising historical Slovene words." *Natural Language Engineering* 22.6 (2015): 881–905.
Stiegler, B. *Digital Studies Organologie des savoirs et technologies de la connaissance-Bernard Stiegler*. FYP éditions, 2016. Accessed July 10, 2023. https://www.fypeditions.com/digital-studies-bernard-stiegler-al/.

Subotic, Ivan. “A Distributed Archival Network for Process-Oriented Autonomic Long-Term Digital Preservation.” *Association for Computing Machinery* (2013): 29–38.

Tolonen, M., Roivainen, H., Marjanen, J. and Lahti, L. (2019). “Scaling up bibliographic data science.” In C. Navarretta, M. Agirrezabal, & B. Maegaard, Eds., Digital Humanities in the Nordic Countries: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, 450–456.

Wang, X., Song, N., Liu, X. and Xu, L. “Data modeling and evaluation of deep semantic annotation for cultural heritage images.” *Journal of Documentation* 77.4 (2021): 906–925.