DHQ: Digital Humanities Quarterly
2015
Volume 9 Number 4

The Data Sprint Approach: Exploring the field of Digital Humanities through Amazon's Application Programming Interface

David M. Berry <d_dot_m_dot_berry_at_sussex_dot_ac_dot_uk>, University of Sussex
Erik Borra <erik_at_erikborra_dot_net>, University of Amsterdam
Anne Helmond <a_dot_helmond_at_uva_dot_nl>, University of Amsterdam
Jean-Christophe Plantin <j_dot_plantin1_at_lse_dot_ac_dot_uk>, London School of Economics and Political Science
Jill Walker Rettberg <Jill_dot_Walker_dot_Rettberg_at_lle_dot_uib_dot_no>, University of Bergen

Abstract

This paper documents the results of an intensive "data sprint" method for undertaking data and algorithmic work using application programming interfaces (APIs), which took place during the Digital Method Initiative 2013 Winter School at the University of Amsterdam. During this data sprint, we developed a method to map the fields of Digital Humanities and Electronic Literature based on title recommendations from the largest online bookseller, Amazon, by retrieving similar purchased items from the Amazon API. A first step shows the overall Amazon recommendation network for Digital Humanities and allows us to detect clusters, aligned fields and bridging books. In a second step we looked into four country-specific Amazon stores (Amazon.com, Amazon.co.uk, Amazon.fr and Amazon.de) to investigate the specificities of the Digital Humanities in these four countries. The third step is a network of all books suggested for the Electronic Literature field in the four Amazon stores we searched, which offers a comparison to the field of Digital Humanities.

INTRODUCTION

In this article we seek to tentatively explore the field of digital humanities (DH) through the production of particular outputs of knowledge rather than the tools that are used. In this we are, perhaps, acting counter to the often remarked processual aspect of DH, that is, that digital humanities focuses not just on the outputs but also on the processes involved in producing those outputs, by, for example, creating data sets, digital tools, archives, etc. [Berry 2012]. However, as part of a rapid "Data Sprint", research outputs, in this case books and monographs, offer a means of producing quick datasets amenable to analysis under the tight restrictions on time that the Data Sprint form imposed upon us. We undertook this analysis to examine the epistemic connections represented through a network analysis of key texts in the field and we also aimed to explore the extent to which the digital humanities differ in comparable countries. By including an analysis of the field of electronic literature we sought to compare DH writ large to a possibly more specific discourse field. Electronic literature includes genres such as kinetic poetry, hypertext fiction, social media fiction and generative narratives, where the "the computer (or the network context) is in some way *essential* to the performance or carrying out of the literary activity in question" [S. Rettberg 2014], and is a strong subfield or related field to the digital humanities, depending on one's perspective. 1

This article brings two contributions to the field of DH. Firstly, we show the relevance of the "data sprint" method for DH inquiry, during which data are collected and analyzed over a short period of time, offering a mezzo-level of analysis between small and large datasets (or "big data"). Secondly, we show how web-based digital platforms can be repurposed to offer insightful data sources to investigate the field of digital humanities itself, therefore contributing to reflexivity in the community. Concerning the first point: the data sprint was part of a Winter School held at the University of Amsterdam in 2013[1] and was designed to facilitate short-form data analysis work in a rapid prototyping format, rather similar to hacklabs and hackathons. The term data sprint itself is drawn from the notion of a Book Sprint which is a "genre of the 'flash' book, written under a short timeframe, to emerge as a contributor to debates, ideas and practices in contemporary culture... interventions that go well beyond a well-written blog-post or tweet, and give some substantive weight to a discussion or issue...within a range of 20-40,000 words" [Berry 2012]. This rapid and collaborative means of writing is very creative and intensified, and tends towards the creation of texts that are appropriately geared to a specific subject or topic [Hyde 2013]. Indeed, participants in a Book Sprint, who are not always necessarily professional writers, are usually experts in a field but from a range of professions, backgrounds and interests. Book Sprints themselves are usually formed from 4-8 people actively involved in the writing process, and are facilitated by another non-writing member. However in the data sprint, which 2

created this paper, the team was made up of five members with varying degrees of competence in data analysis and programming. Nonetheless, the basic tenets of the Book Sprint were observed, in as much as the data sprint was carried out over a period of three days aiming towards final presentation of the results at a workshop organised at the University of Amsterdam. This compression of time naturally results in rapid prototyping and scoping aggressively in order to produce the work within the timeframe available, nonetheless it creates an extremely creative space for the use and reuse of data and algorithms in digital humanities-type work.

3 Due to the "data sprint" format of this project, several members were involved in multiple projects at the same time, or were not working co-locatedly. As a consequence, we heavily relied on online collaborative applications in order to work remotely together. A specific Skype channel was used for multiple purposes: after the tasks were divided between the various members, it served as a means to let the others know which step was done. It was also used to ask questions on a specific task. Furthermore, it served to quickly transfer lists of ASIN numbers from the seed books, or to transfer .zip files containing Gephi files resulting from requests to the Amazon API. Collaborative spreadsheets on Google Drive were a means to collaboratively write descriptive tables of seed books, but also to share first results from the graph (e.g. in a spreadsheets showing Eigenvector Centrality, indegree or outdegree - see below). Finally, a Dropbox was used to share .gexf (Gephi) files resulting from the crawling, and to share the graphs after working on the visualisation. Data sprints are based on reproducibility: the work done needs to be documented and shared online in order to foster similar work and further developments. In parallel to the collaborative online applications used to work together, we set up a website[2] from the beginning that served to document the methods, process, and results of our work. Similarly, crawl results and graphs were instantly uploaded on a specific Github page[3], as was the code accompanying the project[4].

4 Secondly, this article demonstrates how digital methods [Rogers 2013] can be useful to reach a reflexive account of the field of DH. By doing so, we rely on both traditional social network analysis and digital methods. One way of mapping a field, and the approach we take here, is by locating the books that are read by its scholars. Traditionally this has been done by citation analysis [De Solla Price 1965] for example by using Web of Science or, more recently, Google Scholar. There are also examples of network analysis of citations, for instance as in Dan Wang's analysis of key texts in sociology where two texts being taught in the same week of a syllabus is interpreted as a link between the texts [Wang 2012]. In this paper, we instead use a similar approach to Krebs, who used Amazon recommendations to uncover "emergent communities of interest on the WWW by examining purchasing patterns" [Krebs 1999]. We used data from Amazon's online bookstores in the US, UK, France and Germany to analyse buying patterns for the digital humanities and electronic literature, revealing a network of books connected by frequently being purchased together. This research aims to apply a form of algorithmic *distant reading* [Moretti 2005] of the books, in as much as we rely on the readings and purchasing patterns of others to inform the selections that are returned by the API. We have undertaken this extraction of data by retrieving it from the Amazon Product Advertising API.

5 Kaplan's typology of data sources in digital humanities scholarship [Kaplan 2015] is useful to describe the data used in our work: it falls in the category of "digital culture", as opposed to works based on "big cultural datasets" (such as the Google Ngram project mentioned above) or offering "digital experiences" (such as 3D virtual worlds). To this extend, we follow Borgman [Borgman 2015] who recently highlighted that more data is not necessarily more insightful than smaller data, as well as Schöch [Schöch 2013] who calls for a preference towards smaller yet more structured "smart data" instead of large and messy "big data." In order to fetch data, we use here a digital methods approach by repurposing an online device, the Amazon recommendation system, to see how we can make use of web-native objects such as recommendations for social and cultural research [Rogers 2013][Marres and Weltevrede 2013]. In other words, we seek "to deploy the logic of recommendation cultures" [Gerlitz and Helmond 2013] as put forward by Amazon's recommendations based on similar purchased items as an alternative consumer-based approach based on buying patterns to map a field. By moving beyond the traditional "editorial logic" which "depends on the subjective choices of experts" [Gillespie 2012] we explore the possibilities and boundaries presented by the rise of the "algorithmic logic" (idem) to retrieve, organize and present relevant information. Our research explores the "algorithmic structure of today's informatic culture" [Galloway 2006, 17] through the Amazon recommendation algorithm by combining both the editorial logic and algorithmic logic to provide an alternative way to map a field.

6 In this paper we call the Amazon API for different countries to show the relationships between different titles using the SimilarityLookup feature[5], an API operation that "returns up to ten products per page that are similar to one or more items specified in the request" to map the fields digital humanities and electronic literature. By focusing on country-specific versions of Amazon we can visualise national networks of book purchases and analyse differences and similarities for the fields per country or linguistic area. The introduction of geo-location technology on the web [Goldsmith and Wu 2006] has brought into being the notion of "national webs" which are demarcated by devices such as search engines that "go local" [Rogers 2013], as for example Google's country-specific websites like Google.ca, Google.za, Google.jp or Google.com.mx.

Amazon's recommendation engine also comes with ten country-specific websites: United States (amazon.com), United Kingdom (amazon.co.uk), Canada (amazon.ca), Austria (amazon.at),[6] Germany (amazon.de), Spain (amazon.es), France (amazon.fr), Italy (.it), Japan (.jp) and China (amazon.cn)[7] that we can use to compare the fields per country. This analysis across national webs allows us to see whether recommendations for the field of digital humanities would be different in different countries.

This paper is presented as a partial and tentative means of mapping a field, but also as a moment in the developing field of digital humanities [Jameson 2006]. As such we are fully appreciative of its limitations as a study: indeed this is a crucial part of the approach which stresses the "hermeneutics of screwing around," as Ramsay [Ramsay 2010] called it, as our research is the result of a "data sprint". Thus, we present this work as suggestive of the formation of a discursive network crystallising around the notion of "digital humanities" and the possibilities of further research in this vein. To this extent, we follow the critical tradition of DH towards "tools, data, and metadata" [Liu 2013] used and analyzed in this study. To put it in other words, our work "extend[s] reflection on core instrumental technologies in cultural and historical directions" [Liu 2012, 501], by mixing new media studies resources with DH. 7

METHOD

In this paper we use a form of social-network analysis that visualises the relationships between the different entities, in this case books, in our networks. As Alan Liu explains, 8

> the premise of social-network analysis is that it is not individuals or groups but the pattern of relations between individuals or groups that is socially significant. Such an approach commonly produces analyses in the form of social-network graphs composed of nodes and connecting edges (also called ties) accompanied by metrics of degree, distance, density, betweenness, centrality, clustering, and so on. The goal is to describe a topology of social relations that allows researchers to understand, for instance, which nodes are pivotal to connections within communities [Liu 2013].

One of the interesting outcomes of the practices involved in undertaking this form of digital method is the "hermeneutics of screwing around" [Ramsay 2010]. Interpreting the data of a network visualisation is an iterative process of adapting the viewing perspective, changing the data filters, editing colours, layout, depth, degree, and relative "importance" of particular nodes. In other words, adapting the way the data is presented to view it from different perspectives is a method that lets us surface patterns and interesting features of the graph. 9

In our case study we used ten seed books[8] for each of the fields of digital humanities and electronic literature (see Table 1 and 2) as a starting point to request up to ten similar books per book using the SimilarityLookup operation in the Amazon.co.uk Product Advertising API. According to the API documentation, "Similarity is a measurement of similar items purchased, that is, customers who bought X also bought Y and Z. It is not a measure, for example, of items viewed, that is, customers who viewed X also viewed Y and Z.[9]" Initially we repeated the request to reach a depth of three, which includes the results, the subsequent results and subsequent result to create a broad overview over the Digital Humanities (see Figure 1). However, for subsequent analysis we decided to limit our request to the results and the subsequent results (a depth of two) in order to limit the scope of our dataset and complexity of analysis. 10

In other words, we fetched a maximum of ten recommendations per book to a depth of two degrees. This generated a maximum of one hundred book titles for each of the subject areas of digital humanities and electronic literature[10]. This data set then allowed us to do a first analysis with the Gephi software [Bastian et al. 2009], which allows sophisticated data analysis and visualisation, and based upon this we constructed the next data phase which we could scope in relation to the first data set. 11

The second data phase was organised around a comparative approach in relation to data requests to the Amazon API for recommendations for the digital humanities books in each of the following countries: "ca", "cn", "de", "es", "fr", "it", "jp", "co.uk", "com". We requested our data for all ten countries via the Amazon API, but only retained the four (US, UK, FR, DE) that returned results for our requests. 12

Seeds

These are the subject domain expert "seeds" that were used to generate the initial data output. The word "seed" here refers to a computational notion of a piece of data that is used to generate other data, such as in the snowball technique described by Issuecrawler.net [Rogers 2010]. A seed is an expert-determined value[11], in this case the Amazon Standard Identification Number (ASINs) for a specific book on digital humanities or electronic literature, which is fed into the Amazon 13

Related Product Graph, created by Erik Borra, calling the Amazon API[12]. We used ten seeds per subject area that returned up to ten "similar" books per item, and then for each of these results the process was repeated. Using Gephi, the ten data sets were then "appended" into one master data set that was used for the generation of the subject area visualisations.

The selection of seeds was somewhat heuristic, and based on discussion between subject domain experts on the most important books in each field. For the field of electronic literature, the selection took into account frequently referenced books as documented in the "ELMCIP Knowledge Base of Electronic Literature". 14

Below are the initial seeds for the subject areas. 15

ASIN	Title
262018470	Digital_Humanities
230292658	Understanding Digital Humanities
1405168064	A Companion to Digital Humanities
816677956	Debates in the Digital Humanities
1856047660	Digital Humanities in Practice
472051989	Hacking the Academy: New Approaches to Scholarship and Teaching
226321428	How We Think: Digital Media and Contemporary Technogenesis
252078209	Reading Machines: Toward an Algorithmic Criticism
26251740[13]	Mechanisms: New Media and the Forensic Imagination
26212176[14]	The Digital Word: Text-based Computing in the Humanities
1409410684	Collaborative Research in the Digital Humanities

Table 1. Digital Humanities Seeds

ASIN	Title
801842816	Hypertext
801855853	Hypertext 2.0
801882575	Hypertext 3.0
801855799	Cybertext: Perspectives on Ergodic Literature
816667381	Digital Art and Meaning
268030855	Electronic Literature: New Horizons for the Literary
262517531	Expressive Processing: Digital Fictions, Computer Games, & Software Studies
262631873	Hamlet on the Holodeck: The Future of Narrative in Cyberspace
262633183	Twisty Little Passages: An Approach to Interactive Fiction
1441115919	New Directions in Digital Poetry
1441107452	Cybertext Poetics: The Critical Landscape of New Media Literary Theory

Table 2. Electronic literature Seeds

Gephi Procedure

The following section describes the steps taken in creating network visualisations using Gephi from the GEXF result files as output of requesting the recommendations for the seed books with a depth of two. These steps describe how a "master set" was created from ten individual files (corresponding with the seeds) for each of the queried local Amazons. Described here is the process for creating the Digital Humanities .co.uk "master set". Any deviations from these settings for creating the other master sets are noted below. 16

1. For each seed book we have 10 .gexf files, each containing the recommendations for one ASIN with a depth of two.
2. Combine those ten .gexf files to create one graph by appending them in Gephi
    1. Open the first file as a New Graph. Graph Type: Directed
    2. Open the next file and choose "Append Graph" to add it to the first.

3. Repeat until done
4. We now have one graph with 155 nodes and 257 edges.

3. Run spatialization algorithm Force Atlas 2 [Jacomy et al. 2011].[15]
    1. Settings: Scaling 100, Gravity 5
       These settings are not fixed settings, rather, they are adjusted per graph to increase readability and are adjusted during the process in an iterative process to produce the best readability. Scaling represents repulsion and increasing gravity prevents islands from drifting away.
4. Rank nodes according to InDegree to relatively scale node size according to the number of recommendations it receives from other books.
    1. Settings linear scaling: Min size: 4, Max size: 40.
5. Color code clusters to algorithmically detect communities in the graph using modularity statistics[16]:
    1. Statistics: Run Modularity with the following settings: randomize 1, use weights 1, resolution 1.0
    2. Partition nodes with Modularity Class partition parameter[17].
6. In the Data Laboratory copy data from "title" to Label to only show the book title as the node's label.
7. Show labels (in our case book titles) of the nodes scaled to node sizes in overview and run the "Label Adjust" layout algorithm to prevent overlap for readability
8. Preview: Default Straight presets
    1. in order to improve readability, shorten labels if they are longer than thirty-four characters
    2. custom label color (#333333)
    3. custom edge color (#BDBDBD) to further increase readability of the graph.
9. Save graph as PDF
10. Calculate network statistics for tables 1 and 2. In the "statistics" panel
    1. Click "Average Degree" and note the resulting number
    2. Click "Graph Density", check "Directed", and note the resulting number
    3. Click "Network Diameter", check "Directed", leave other options empty, and note the diameter, average path length, and number of shortest paths (rounded to three decimals)
11. Design graph further in Adobe Illustrator:
    1. Adjusted community cluster colors using the tool "I want Hue" [18] a tool for data scientists which generates palettes of optimally distinct colours. These are the colours we generated and used in our graphs:
        - #7CD5B0
        - #CA52CC
        - #C44639
        - #CAC943
        - #503A60
        - #50733C
        - #CB99A2
        - #C0477D
        - #70D151
        - #8675CB
        - #C1823A
        - #533B2E
        - #7BA7BD
        - #C4C488
12. Highlight seed URLs
13. Put graph in pre-cooked country template

Digital humanities Amazon.de same but Force Atlas 2 Settings: Scaling 300, Gravity 1 17

Digital humanities Amazon.fr same but Force Atlas 2 Settings: Scaling 300, Gravity 5 18

Electronic literature Amazon combined same but Force Atlas 2 Settings: Scaling 300, Gravity 5 19

Electronic literature Amazon.com same but Force Atlas 2 Settings: Scaling 300, Gravity 5 20

VISUALIZATIONS OF DIGITAL HUMANITIES

This first graph was created by combining the ten master data sets from the ten digital humanities seeds (see Table 1) for Amazon.com with a depth of three and shows the overall Amazon recommendation network for Digital Humanities. 21

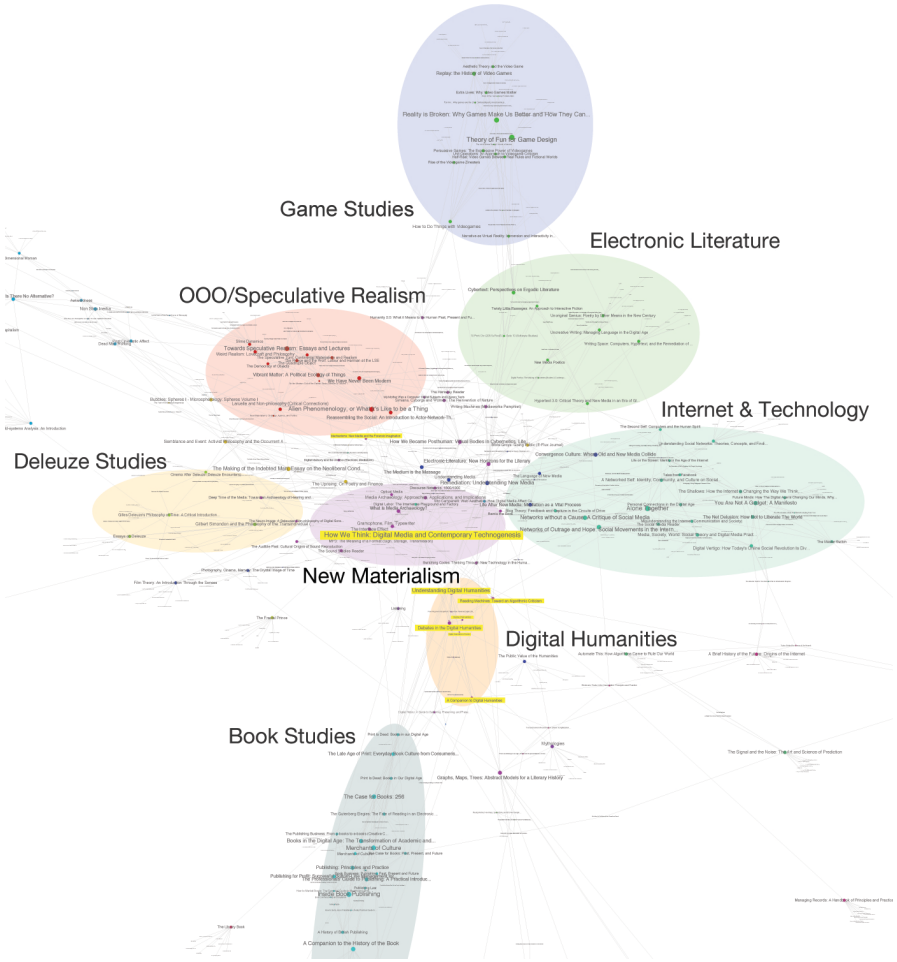

Figure 1. Part of the full Digital Humanities Amazon.com similar items purchased graph with 853 book titles with color-coded communities: Game Studies, Electronic Literature, Internet & Technology, New Materialism, Digital Humanities, Book Studies, Deleuze Studies and OOO/Speculative Realism. Node size scaled according to InDegree. Seeds indicated in yellow. Full image: https://wiki.digitalmethods.net/pub/Dmi/Winter13AmazonRecommendations/DH_Full_ClusterColorCoded_A4.pdf

Within the graph visualisation (Figure 1) each node represents a book. It is a directed graph which means that the edges between the source (book A) and the target (book B) are directed (A points to B). In our case the edges represent recommendations so the source (book A) points to the target (recommended book B, C, D etc.). We retrieved recommendations to a depth of three, meaning that our data set includes the seed books (i.e. depth 0), books recommended in relation to the seeds (i.e. depth 1), books recommended in relation to the books in depth 1 (i.e. depth 2), and books recommended in relation to the books in depth 2 (i.e. depth 3). 22

What we see in Figure 1 is the clustering of particular books into more or less distinct genres or disciplinary groups. For example, we see that there are many connections between books on internet and technology, electronic literature and game studies, but fewer between game studies and book studies, or between book studies and speculative realism. A closer reading of this visualization by a topic expert on the Digital Humanities, David M. Berry, revealed that the books cluster around particular fields within the Humanities such as Digital Humanities, Media Studies, Literary Studies and related areas. We identified the following clusters from the visualisations produced: Game Studies, Electronic Literature, Internet and Technology, New Materialism, Digital Humanities, Book Studies, Deleuze Studies and OOO/Speculative Realism. Some books appeared to form bridges between the fairly distinct clusters, such as Hayles' *How We Think*. 23

We also see here how, while some books are densely interconnected leading to the clustering in the graph, others simply lead away from the initial seeds. An example is Bate's *The Public Value of the Humanities* (2011), which is bought by many Amazon customers who bought Gold's *Debates in the Digital Humanities* (2012). However, as you see from Figure 2, people who bought Bates' book don't tend to buy many other digital humanities books. Instead they are more likely to buy other books about universities and the current crisis in academia. This suggests that digitial humanists are interested in the 24

crisis in academia, but those interested primarily in the crisis in academia are not, on average, particularly interested in the digital humanities.

We also see many connections between the seed books, which suggests that our initial choice of seeds was reasonably representative of the field. 25

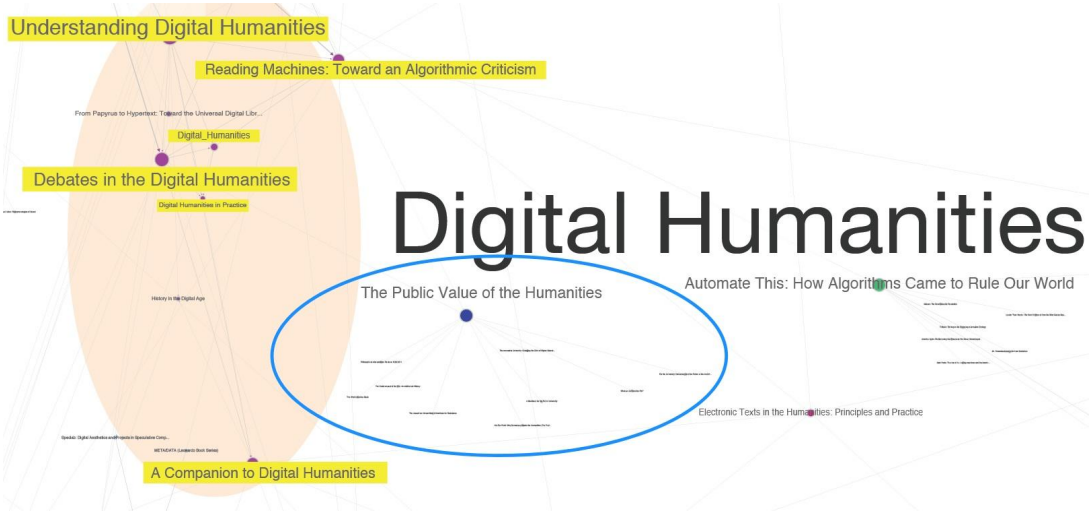

Figure 2. Zoom into the full Digital Humanities similar items purchased graph with 853 book titles with one visible color-coded community: Digital Humanities and a related cluster around “The Humanities/The University in crisis” on the right (circled in blue).

Figure 1 and 2 show the network as generated from Amazon.com with a depth of three. If we look at the networks generated from individual country Amazon stores, we find important differences between the countries. In order to be able to compare the US, UK, French, and German Amazon we limited the crawl depth for recommendations to two. This means that we only looked at similar items suggested for our seed books as well what was recommended for those similar items. 26

Feeding the seed books for the digital humanities into the US Amazon (.com) generated a densely interlinked graph showing ninety-five individual books (Figure 3) with all seed books clustered in the middle. We also see that all the books bought by people who bought the seed books are connected to more than one seed book. 27

Figure 3. The Digital Humanities Amazon.com similar items purchased graph with 95 book titles. Node size scaled according to InDegree. Clusters are color-coded using modularity. Seeds indicated in yellow. Full image: https://wiki.digitalmethods.net/pub/Dmi/Winter13AmazonRecommendations/DH_COM_FA2_ScaledIndegree_ModularityColorCoding_ColoredSeeds.pdf

Here a group of central digital humanities books have clustered in the center of the graph. Eight of the seed books are visible here[19], but we also prominently see four new books not on the list: Bartscherer and Coover's anthology *Switching Codes*, Fitzpatrick's *Planned Obsolescence*, Moretti's *Graphs, Maps, Trees* and McGann's *Radiant Textuality*. Their size, scaled according to InDegree, indicates that they are often bought together with other books that are recommended with the seeds. These additions can be found around the digital humanities cluster, suggesting more affinity (or more precisely, more links or edges) between these books than some of the other digital humanities books. We see adjacent fields beyond this central cluster. 28

The first related field is electronic literature on the right of the graph in Figure 3. McGann's *Radiant Textuality* connects to Hayles' *Electronic Literature*, which connects on to other books on electronic literature. In the bottom of the graph, circled in blue and green in Figure 4, we see a cluster on materiality, software and code studies, alongside books on speculative realism/object-oriented philosophy, which again lead to posthuman studies. 29

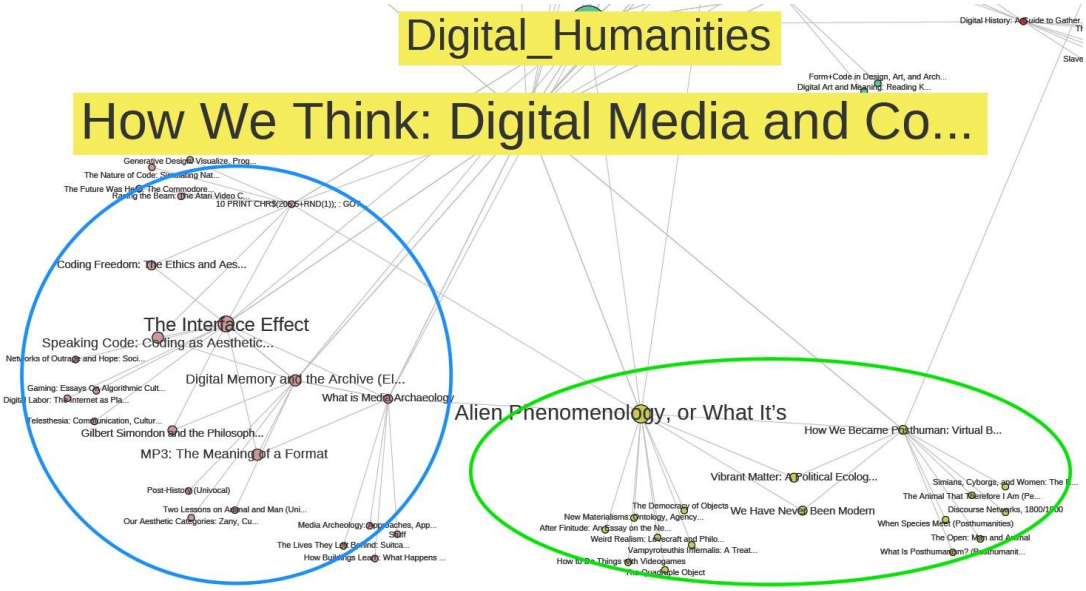

Figure 4. Detail of the Digital Humanities Amazon.com similar items purchased graph.

Ian Bogost's *Alien Phenomenology* is notable for being a bridge to speculative realism, circled in green in Figure 4, and you can see books linked through Hayles' *How We Became Posthuman* are also part of this cluster, which on the Hayles' side includes cybertheory and posthumanist theory. At the lower left of the graph, in blue, we see Alex Galloway's *The Interface Effect* is an important hub, with a fairly disparate group of books surrounding it. 30

Hayles' *How We Think* has the highest InDegree (that is, it is referenced by the most other nodes: it is recommended as a book frequently bought by purchasers of the highest number of other books in our sample) but is positioned on the edge of the main digital humanities cluster and not centrally within this digital humanities cluster. Instead, this book appears to function as a bridge or broker between different fields such as speculative realism/object oriented ontology and new materialism. Very interestingly, it has the same function in the electronic literature networks, as we will discuss later in this paper. 31

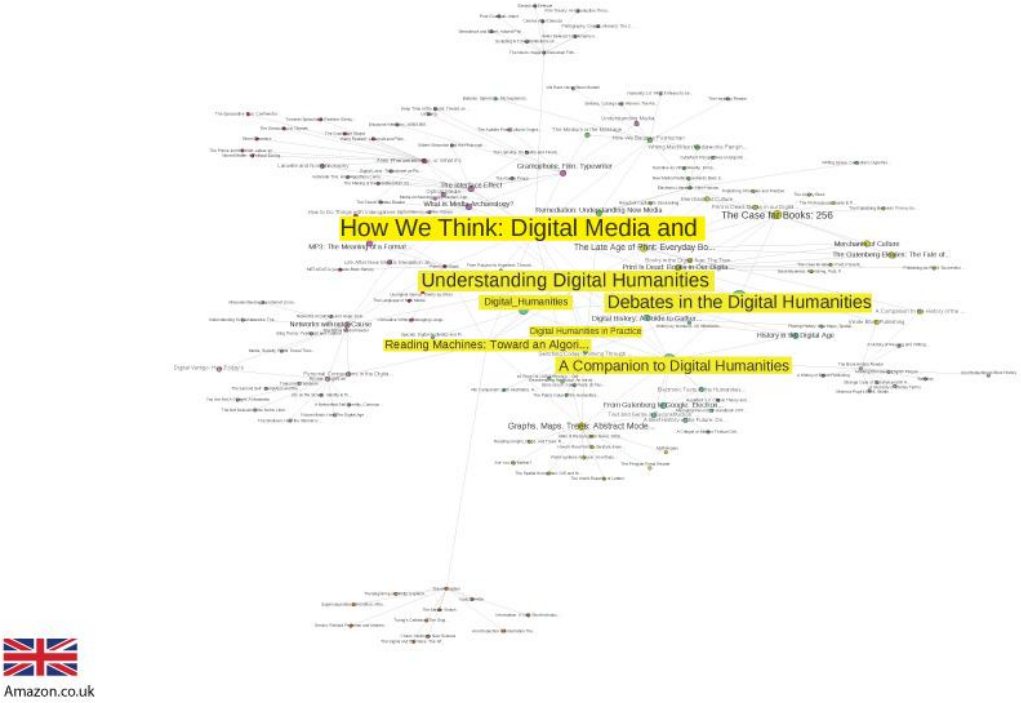

Figure 5. The Digital Humanities Amazon.co.uk similar items purchased graph with 155 book titles. Node size scaled according to InDegree. Clusters are color-coded using modularity. Seeds indicated in yellow. Full image: https://wiki.digitalmethods.net/pub/Dmi/Winter13AmazonRecommendations/DH_UK_FA2_ScaledIndegree_ModularityColorCoding_ColoredSeeds.pdf

The UK Amazon graph shows 155 book titles with seven seeds clustered in the middle[20]. Similarly to the US Amazon, Hayles' *How We Think* functions as a bridge to the fields of object oriented ontology and new materialism and Bogost's *Alien Phenomenology* functions as a bridge to speculative realism on the top. However, there are three new clusters visible that distinguish it from the US Amazon graph. First, we see a cluster of (popular) new media theory books on the left with Geert Lovink's *Networks without a Cause*, Nancy Baym's *Personal Connections in the Digital Age* and Sherry Turkle's *Alone Together* forming bridges. Second, a cluster on digital history in the bottom middle and third, a cluster on book history and book publishing on the bottom right. 32

Figures 6 and 7 show the networks generated by feeding our English-language seed books into the French and German Amazon stores. In both cases, the seed books almost all disappear. 33

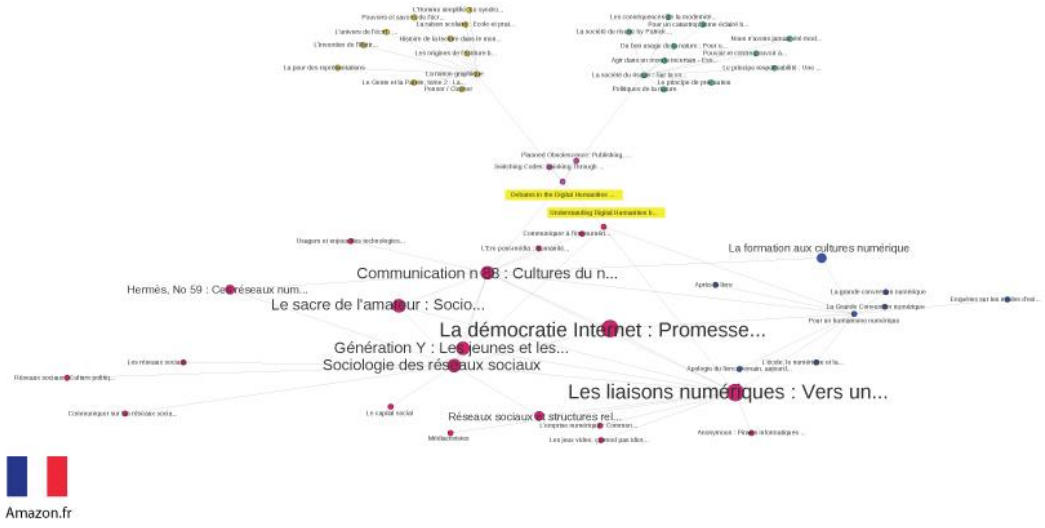

Figure 6. The Digital Humanities Amazon.fr similar items purchased graph with 62 book titles. Node size scaled according to InDegree. Clusters are color-coded using modularity. Seeds indicated in yellow. Full image: https://wiki.digitalmethods.net/pub/Dmi/Winter13AmazonRecommendations/DH_FR_FA2_ScaledIndegree_ModularityColorCoding_ColoredSeeds.pdf

On the graph for Amazon.fr, only two books from the seeds remain: *Debates in the Digital Humanities* and *Understanding Digital Humanities*. It immediately becomes apparent that besides these two seeds there are only two other English language books in the graph: *Planned Obsolescence* and *Switching Codes*. The seed *Debates in the Digital Humanities* acts as a bridge to four clusters, which, unlike the other graphs, solely consist of French-language books. These books are both original French books and translations. A possible explanation for this lack of English books in the French-speaking DH network is the current trend in translating DH into French terms, either "*Humanités numériques*"[21] or "*Humanités digitales*"[22]. 34

The first cluster with the largest number of nodes (in pink/purple), is constituted by French sociology and media studies books, mostly written by academics but published by mainstream publishing houses which aims to foster general public debates about technology in society. On the right side, the cluster with blue nodes revolves around two books written by Milad Doueihi, a French-speaking classical historian: *Pour un humanisme numérique* and *La grande conversion numérique*. These books on "numerical humanism" deal with the "numerical turn" and the future of books and education after digitalization, which echoes the general purpose of the previous cluster. Books about philosophy of technique constitute the cluster on the top left side and all these books are connected to a French edition (*La raison graphique*) of Jack Goody's *The Domestication of the Savage Mind,* which acts as a bridge to the digital humanities. Books about sociology, philosophy of risk society, and citizen science constitute the last cluster, with green nodes on the top right side, and they are all connected by the French edition of Ulrich Beck's *Risk Society: Towards a New Modernity*. 35

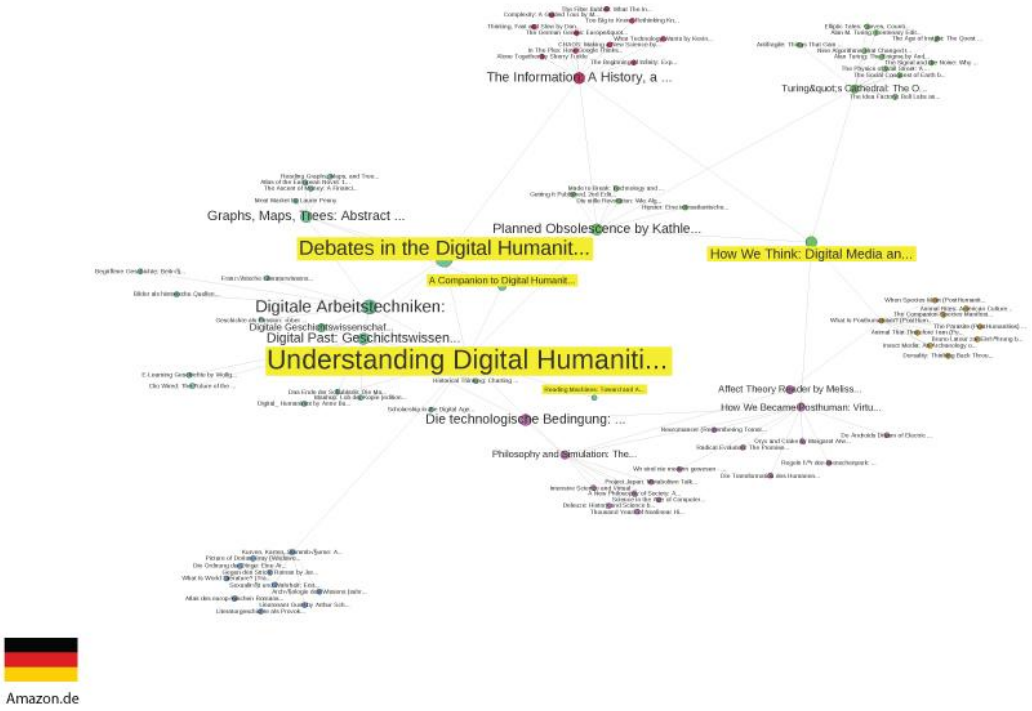

Figure 7. The Digital Humanities Amazon.de similar items purchased graph with 92 book titles. Node size scaled according to InDegree. Clusters are color-coded using modularity. Seeds indicated in yellow. Full image: https://wiki.digitalmethods.net/pub/Dmi/Winter13AmazonRecommendations/DH_DE_FA2_ScaledIndegree_ModularityColorCoding.pdf

On the graph for Amazon.de, five books from the seeds remain. The various clusters are less thematized than the other countries, and topical drifting appears. The cluster in the center, in green, contains digital humanities literature, and media studies books. The cluster in purple, in the bottom center, goes from media studies to philosophy and French theory, which echoes the close cluster on the right (in brown nodes) on posthuman theories. The two clusters on the top are related to information science (red nodes on the top left side) and to computer science books, mainly about Turing (green nodes, top right). The last cluster, at the bottom left with blue nodes, contains literary theory and novels. 36

Unlike the French graph, there are many English language books in the German graph, but interestingly enough the graph introduces German books on digital humanities such as *Digitale Arbeitstechniken: für die Geistes- und Kulturwissenschaften*, *Digitale Geschichtswissenschaft*, and *Digital Past: Geschichtswissenschaft im digitalen Zeitalter* on the topics of digital history and digital tools for the humanities and cultural studies. It also includes *Die technologische Bedingung: Beiträge zur Beschreibung der technischen Welt*, a compilation of translated essays including Galloway, Hansen, and Hayles translated into German. As an introduction into the topic of "the technological condition" in German it 37

acts as a bridge to related books in English.

Further comparing the various local Amazon domains under study we can see that .com has the most densely interlinked recommendations (see density and average degree in Table 3) while .fr is the sparsest. This of course also reflects how quickly other books are recommended (average path length). 38

data set	nodes	edges	density	avg degree	diameter	avg path length	number of shortest paths
.co.uk	155	257	0.011	1.658	5	3.139	3808
.fr	62	70	0.019	1.129	3	1.884	207
.de	92	118	0.014	1.283	5	2.910	1052
.com	95	206	0.023	2.168	5	2.716	1870

Table 3. Network statistics for the networks of each of the local Amazon domains under study.

VISUALIZATIONS OF ELECTRONIC LITERATURE

Our analysis suggests that electronic literature is a far less cohesive field than digital humanities is in the USA, at least in so far as printed books can be said to represent the field. Figure 8 shows the network of all books returned for our electronic literature seed books in the four Amazon stores we searched, and you can see that a large number of books were returned for the seed books. The seed nodes, which are marked in yellow, are scattered around the graph rather than clustered in the center as in the US digital humanities graph in Figure 3, and they are not heavily recommended, as is indicated by the small size of most of the seed nodes. The two seed books within electronic literature that Amazon.com notes readers of other books buy are Hayles' *Electronic Literature* and Aarseth's *Cybertext*, but as Figure 8 shows, the Force Atlas 2 algorithm in Gephi does not pull them close together, although they are directly linked to each other. This is because people who buy these two books also buy books in distinctly different fields: *Electronic Literature* connects to DH, book culture, and disucssions of materiality, and *Cybertext* to game studies. There are very few links directly between game studies and DH or game studies and book culture. Between *Electronic Literature* and *Cybertext* we see a range of key texts in new media studies that are also important in electronic literature, such as Manovich's *The Language of New Media*, Bolter and Grusin's *Remediation*, and Jenkins' *Convergence Culture* as well as core media studies texts such as McLuhan. 39

Further comparing the DH and electronic literature graphs from a more numerical point of view (see Table 4), we can see that although the electronic literature graph has fewer nodes, it is actually more densely connected than the DH one (see density and average degree). Looking at the average path length, one can see that it is easier to reach other books given a book related to EL according to Amazon. 40

data set	nodes	edges	density	avg degree	diameter	avg path length	number of shortest paths
DH full	853	1402	0.002	1.644	13	5.807	58965
EL full	464	977	0.005	2.106	9	4.111	33944

Table 4. Network statistics for the full Digital Humanities (DH full) and Electronic Literature (EL full) networks.

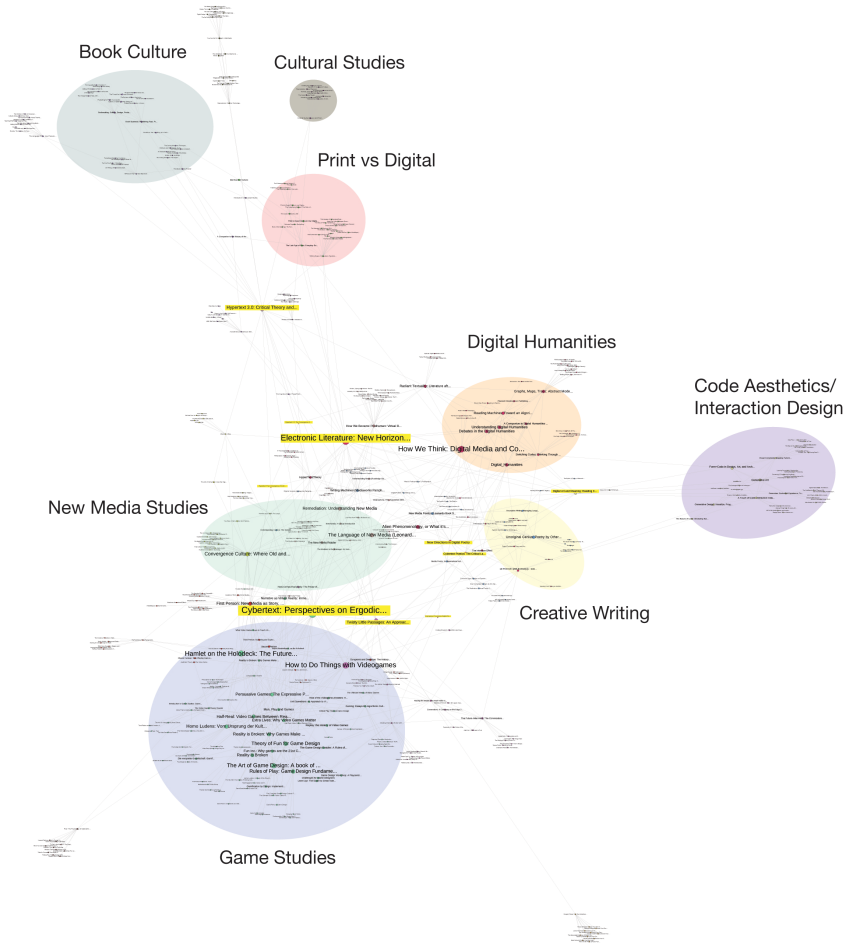

Figure 8. The Electronic Literature Amazon (.com, .fr, .de, .co.uk combined) similar items purchased graph with 464 book titles. Node size scaled according to InDegree. Clusters are color-coded using modularity. Seeds indicated in yellow. Full image: https://wiki.digitalmethods.net/pub/Dmi/Winter13AmazonRecommendations/EL_FULL_FA2_ScaledIndegree_ModularityColorCoding.pdf

The games studies cluster below *Cybertext* is far more cohesive and interlinked than the rest of the network, suggesting that this is a more clearly defined field than electronic literature, at least in terms of having an established set of printed books that are frequently bought together.

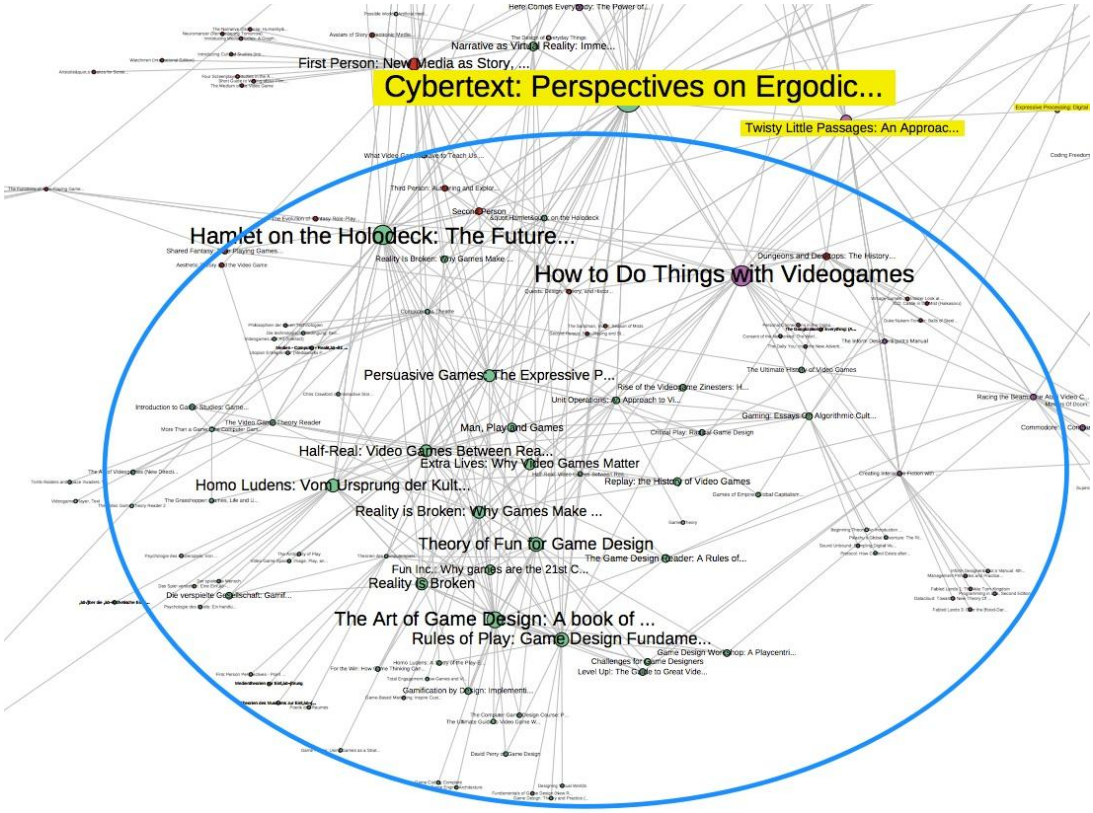

Figure 9. Detail of the electronic literature graph in Figure 8, showing the game studies cluster.

At the top of Figure 8, we see links leading away from electronic literature into discussions of the role of the book in a networked society and further away into book history and general discussions of bookmaking and the book business. This shows how one field drifts towards another, but we also see that this section of the graph is not interlinked and that there are few or no connections back into the more centrally placed books that are more closely related to electronic literature. 42

The upper right of Figure 8 shows a digital humanities cluster very similar to that generated by our digital humanities seeds, while the lower right hand side shows an interesting loosely connected cluster of works on conceptual poetry and writing and on digital poetics. We see Perloff's *Unoriginal Genius: Poetry by Others* is a hub here, and while Perloff's book is not specifically about electronic literature, its focus on remixing and cut-and-paste as literary techniques is clearly relevant to electronic literature. Another hub, seen towards the bottom right of Figure 10, is Montfort et.al.'s *10 Print Chr (205.5+Rnd(1))*, a book about a one line BASIC program that generates a simple graphic maze consisting of two randomly repeated characters in the Commodore 64 computer's character set. This is again not exactly electronic literature, but it is closely related to electronic literature, and in the book the short program is analysed as closely as literature ever was by several important critics in the field of electronic literature. Perhaps we must conclude that some of the most important books for the study of electronic literature are not about electronic literature per se? 43

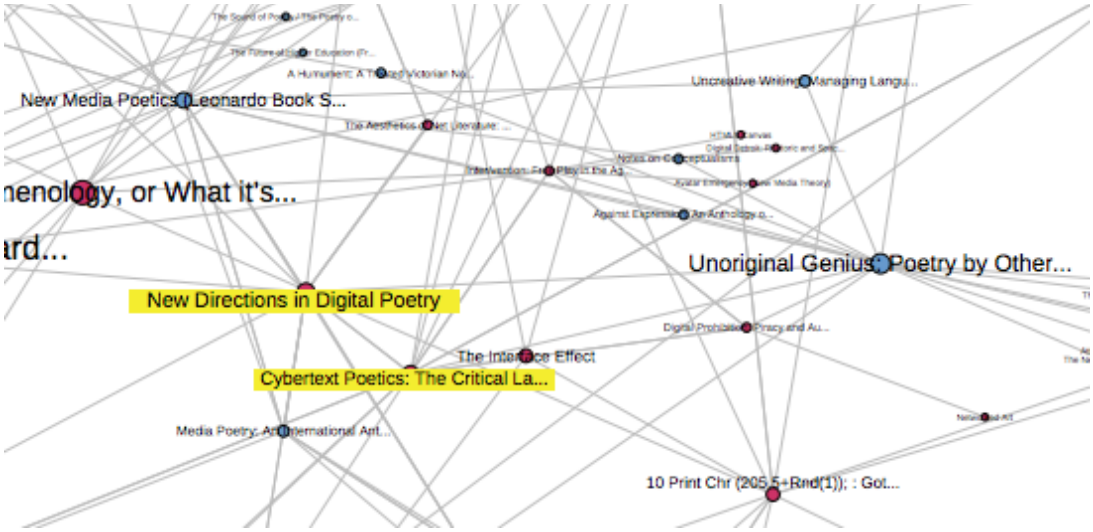

Figure 10. Detail of the electronic literature graph in Figure 8 showing a cluster of books on conceptual writing and digital poetics.

Far out to the right in Figure 8, we see a cluster of books about generative art and code art (Figure 11). 44

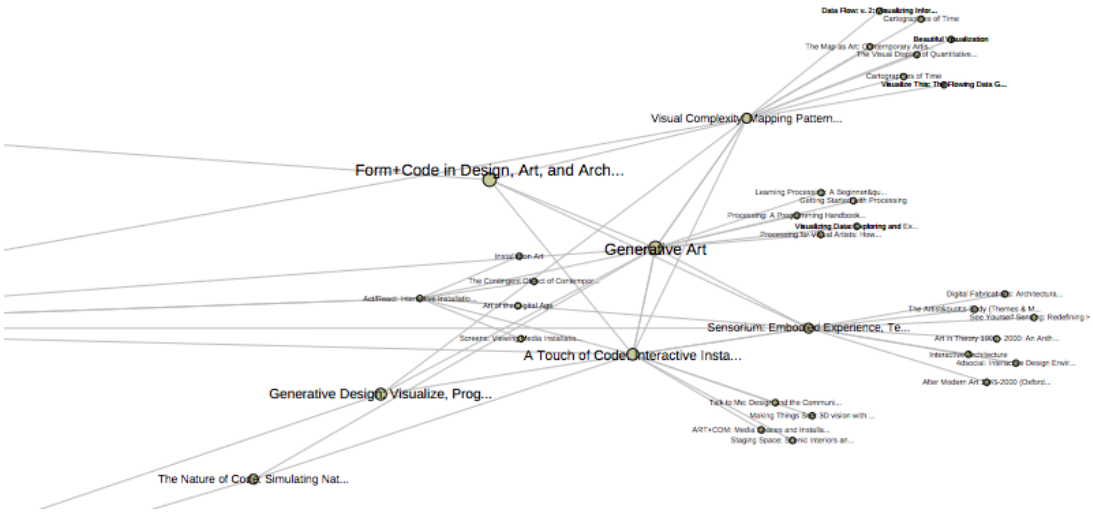

Figure 11. Detail of the electronic literature graph in Figure 8 showing a cluster of books on code and art.

As you can see in Figure 8, this cluster is quite distant from the rest of the network, and is connected to it by a few clear brokers: Bartscherer and Coover's recent anthology *Switching Codes*, Simanowski's *Digital Art and Meaning: Reading Kinetic Poetry, Text Machines, Mapping Art, and Interactive Installations*, and Montfort et.al.'s *10 Print*. *Switching Codes* also shows up as a very central book in the US digital humanities network as shown in Figure 3, and appears to function as a bridge between different communities related to digital humanities, electronic literature and digital art. 45

Despite the fact that electronic literature research is scholarship about creative works of electronic literature, only three works of electronic literature show up in the graph, and these are works published on CD-ROM by Eastgate systems in the early 1990s. Most electronic literature is published online and is not part of Amazon's database. We see in Figure 12 that Joyce's seminal hypertext fiction *afternoon, a story*, shows up and is connected to the seed books Hayles' *Electronic Literature* and Landow's *Hypertext 3.0* as well as Murray's *Hamlet on the Holodeck*, and also to a relatively unrelated book, Shirky's *Here Comes Everybody*. 46

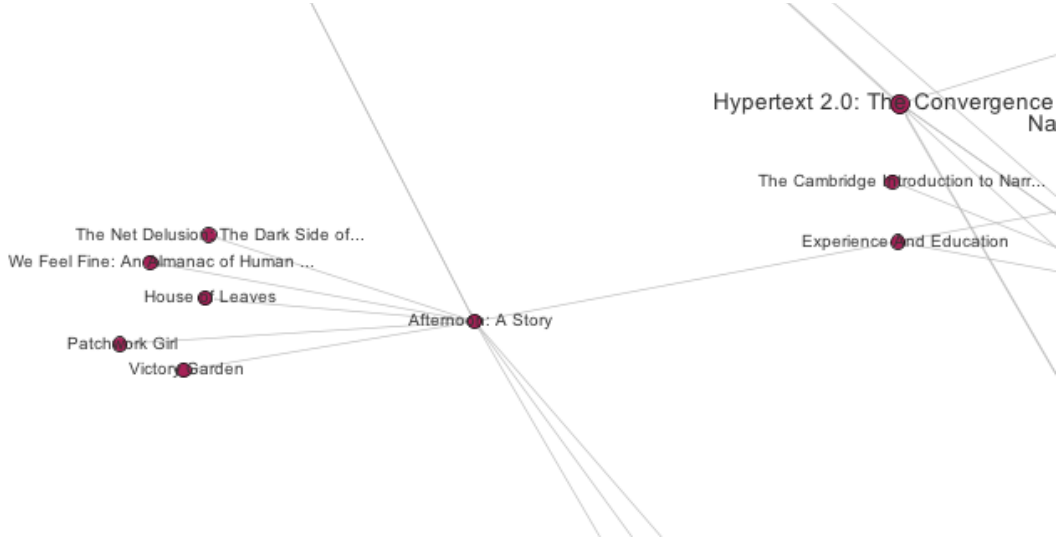

Figure 12. Detail of the electronic literature graph in Figure 8 showing hypertext fictions *afternoon*, *Patchwork Girl* and *Victory Garden*.

We see that people who bought *afternoon* also bought two other frequently cited hypertext fictions from the early 1990s: Jackson's *Patchwork Girl* and Moulthrop's *Victory Garden*. Research in the field of electronic literature tends to reference a broad range of creative works in the field [J. W. Rettberg 2014] and while using our methodology on a traditional field of literary studies might capture literary works as well as scholarly books, literary works of electronic literature are not sold on Amazon. This may be one reason why the field of electronic literature appears less cohesive than that of the digital humanities. 47

LIMITATIONS

Studying a field only from the point of view of Amazon book recommendations is clearly not going to tell the whole story about digital humanities or electronic literature. This type of study of the field excludes other book sellers and traditional scholarly resources such as journals and journal articles and also excludes the outputs of digital humanities projects such as archives, TEI projects, websites, tools and code archives, and also works of art and literature which are central to the field of electronic literature. Also, it bears the temporal limitations imposed by the time constraint of the three-day data sprint format. Finally, there are also technical limitations using the Amazon API in our survey of the fields of electronic literature and digital humanities, such as a limit of ten recommendations per book and a maximum of 3600 requests to the API per hour. 48

Nonetheless, it does provide a new way of looking at the field. The sheer size of the Amazon database allows us to see interesting connections between the digital humanities and adjacent fields. Also, even with these caveats it is notable that the results, broadly speaking, do reflect clusters of what we might think of as fields of study, and the connection between them. 49

It is also important to realise that while the data does, according to Amazon, give information about what other books are bought by customers who buy book X, book sales don't necessarily mean that the books are read, cited, used, or influential. In addition, customers who buy digital humanities books also buy other books on Amazon and if books are frequently bought together they will be marked as "similar" by Amazon as recommendation engine. This leads to the inclusion of *The Portrait of Dorian Gray* in the German digital humanities graph and, as a classic novel, it is the most sold item in the graph. 50

Our data from the French and German Amazon stores differed slightly from the English-language Amazon stores with fewer results. One reason is that we use the same, English-language books as seeds in all the national Amazon stores. While we could have chosen French and German language books, we thought the results yielded were still interesting, and many non-English titles rapidly appeared in these graphs. The seeds introduced similar language-specific items on the topic of digital humanities. This may point us to French and German-specific subfields of the digital humanities in their native language. 51

CONCLUSIONS AND FURTHER RESEARCH

Based on these network graphs of different disciplines, we would conclude that they appear to have different styles of communication, at least in terms of the importance of printed books in the field. The field of digital humanities in the USA as 52

viewed through Amazon's SimilarityLookup is cohesive, with a relatively small number of books that are bought together. In France and Germany, on the other hand, we see that the field is far less well defined, and based on the books that are bought together, it is hard to say clearly what the digital humanities are or are not in these countries. Britain offers an intermediate position, where the digital humanities are understood in a less precise or perhaps broader manner than in the US, and where we see relationships to many more fields.

Electronic literature, as viewed through print books and Amazon's SimilarityLookup, is a field that is far less cohesive than the digital humanities in the USA, and we see instead that books on electronic literature intermingle with books in related disciplines: new media studies, game studies and the digital humanities chief among them. Game studies in fact comes to the fore in the graph drawn from our electronic literature seeds, and appears to be a field almost as cohesive as the digital humanities in the US. 53

We mentioned earlier that there is some irony to analysing the digital humanities by looking at the books published in the field rather than at the digital projects and tools developed. But at least in the US, it appears that the field of digital humanities is very clearly defined by its books. Our diagrams would be an excellent starting point for a reading list for a newcomer to the digital humanities. Electronic literature, on the other hand, is not as easily described by this method. Perhaps more of the publications on electronic literature are entirely digital, whether as creative works or as shorter articles in online journals and other online publications, and thus they are not visible to Amazon. Or perhaps electronic literature is more interdisciplinary by nature, and thus people who read books about electronic literature read more broadly rather than focusing on that topic alone. 54

This research raises further questions that we think could be explored in relation to the research questions. 55

How can a more formal and defensible seeding strategy be developed? In the project, in common with other digital methods projects, a domain subject expert is used to generate initial data sources, seeds, and links. It would be interesting and useful to reduce or eliminate the seeding process such that once the general thematic area is identified, a standard seed generation methodology can be followed. This may still use knowledge elicitation from the subject domain expert, but would be formalised. 56

How can validation of the API results be implemented so that the API does not always return identical data for search queries? This is a result of how Amazon handles data fields that contain a super-set in relation to the capacity of the API return values. One method might be multiple requests and a smoothing algorithm to average the results from the API. 57

The ability to create the graphs from the API is extremely powerful and although we undertook some limited secondary data generation, such as querying the Kindle Highlights database[23], it would be useful to formalise this method and generate sufficient data which when linked to the primary graph data enables interactive exploration of the graph output. 58

As an example of further research we ran some preliminary data requests to the Kindle Highlights database, however, the number of highlights in our texts was very low. Most digital humanists either do not read the digital (or at least, the Kindle) versions of the texts, or they do not highlight their digital versions. Nonetheless, with the growth in e-readers, iPads, tablets, and the like, we can expect this database to be of increasing interest to researchers undertaking similar projects to this in the future. For example, these are the most popular highlights in Kirschenbaum's *Mechanisms: New Media and the Forensic Imagination,* which is important to both the digital humanist and electronic literature subject areas. 59

> forensic materiality rests upon the principle of individualization (basic to modern forensic science and criminalistics), the idea that no two things in the physical world are ever exactly alike.
>
> The point is to address the fundamentally social, rather than the solely technical mechanisms of electronic textual transmission, and the role of social networks and network culture as active agents of preservation.
>
> a digital environment is an abstract projection supported and sustained by its capacity to propagate the illusion (or call it a working model) of immaterial behavior:

Each of these sections has limited information associated with it, although it is noticeable that no page number is given – Kindles do not have page numbers as part of the product. 60

As an exploratory approach to mapping a field or disciplinary area of research, this approach has much to recommend it. It provides a useful entry point for drawing up an initial map of the field and for developing understanding of the way in which books provide a structure for a field's development. Whilst we wish to reiterate the limitations of this approach, particularly in view of the digital nature of the two fields we chose for comparison, digital humanities and electronic literature, and the resultant absences in the data and visualisations that are created, we nonetheless think that used appropriately it is a 61

method that is very amenable to an exploratory method of field-mapping.

Notes

[1]This paper is the result of a data sprint at the Digital Methods Winter School, "Data Sprint: The New Logistics of Short-form Method", 22-24 January 2013, Amsterdam, the Netherlands. See https://wiki.digitalmethods.net/Dmi/WinterSchool2013 for more info. Thanks to Richard Rogers and other participants at the Winter School for helpful comments and suggestions.

[2] https://sites.google.com/site/whatisdigitalhumanities/

[3] https://github.com/dmberry/Digital-Humanities-and-Electronic-Literature

[4] https://github.com/digitalmethodsinitiative/arpg

[5] http://docs.aws.amazon.com/AWSECommerceService/latest/DG/SimilarityLookup.html

[6]Amazon Austria (.at) redirects to Amazon Germany (.de)

[7] http://www.amazon.com/gp/feature.html?ie=UTF8&docId=487250

[8]These ten seed books were selected by two topic experts: David M. Berry selected ten key books in the field of digital humanities and Jill Walker Rettberg selected ten key books in the field of electronic literature. The concept of seed used here is drawn from the computational use of pseudo-random numbers to begin a computational process, it is a term also used in "Minecraft" to indicate a starting number to generate a procedural world in the game. A seed in the context of this paper is a set of values, in this case book ASINs, which can be used to bootstrap the API call process. Using 10 seeds rather than just one increased the space of results which could be produced from the Amazon API and the rapidity of the data set collection, which, due to the compressed time available in the data sprint, was extremely useful.

[9] http://docs.aws.amazon.com/AWSECommerceService/latest/DG/SimilarityLookup.html

[10]As some books may have the same recommendations and some books have less than 10 recommendations, the actual number is much lower.

[11]In determining the seeds we relied on the editorial logic of subject experts David M. Berry for choosing the Digital Humanities seed set and of Jill Walker Rettberg for the Electronic literature seed set (see previous footnote).

[12]The source code for this tool is available at https://github.com/digitalmethodsinitiative/arpg

[13]The lookup of these books failed because the actual ASIN ends with an X (26251740X and 26212176X). This does not mean that the book is automatically excluded from the analysis because if it is recommended by one of the other seed books it will once again be included in the data set.

[14] Ibid.

[15]It is scaled for small to medium-size graphs, and is adapted to qualitative interpretation of graphs [Jacomy et al. 2011].

[16] see: Blondel et al..

[17]"This structure, often called a community structure, describes how the the network is compartmentalized into sub-networks. These sub-networks (or communities) have been shown to have significant real-world meaning." http://wiki.gephi.org/index.php/Modularity

[18]http://tools.medialab.sciences-po.fr/iwanthue/

[19]Two seeds are not visible on the map because they are not connected to the graph: *The Digital Word* and *Collaborative Research in the Digital Humanities*. There were no similar items suggested for these books.

[20]Three seeds are not visible on the map because they are not connected to the graph: *The Digital Word*, *Collaborative Research in the Digital Humanities*, and *Mechanisms: New Media and the Forensic Imagination*. For the first two books there were no similar items suggested. *Mechanism* is included in the graph because lookup of the incorrect ASIN failed. It did show up in the previous graph because it was a similar item to one of the other books.

[21] cf. example here: http://books.openedition.org/oep/238

[22] cf. example here: http://cdh.epfl.ch/digital

[23]https://kindle.amazon.com/search

Works Cited

Bastian et al. 2009 Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." *ICWSM*: 361-362.

Berry 2012 Berry, D. M. *Understanding Digital Humanities*, Basingstoke: Palgrave Macmillan, 2012.

Berry and Dieter 2012 Berry, D. M. and Dieter, M. "Book Sprinting", accessed 14/08/2013, http://www.booksprints.net/2012/09/everything-you-wanted-to-know/, 2012

Borgman 2015 Borgman, C. L. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press, 2015.

De Solla Price 1965 De Solla Price, Derek J. "Networks of Scientific Papers." *Science*149, no. 3683: 510–15, 1965.

Galloway 2006 Galloway, Alexander R. *Gaming: Essays on Algorithmic Culture*. Minneapolis, MN: University of Minnesota Press, 2006.

Gerlitz and Helmond 2013 Gerlitz, Carolin and Helmond, Anne. "The Like Economy: Social Buttons and the Data-intensive Web". *New Media & Society,* 2013

Gillespie 2012 Gillespie, T. "The Relevance of Algorithms." In *Media Technologies*, edited by T Gillespie, Pablo Boczkowski, and Kirsten A Foot, 2012.

Gold 2012 Gold, M. K. *Debates in the Digital Humanities*, University of Minnesota, 2012.

Goldsmith and Wu 2006 Goldsmith, J.L. & Wu, T. *Who controls the Internet?: illusions of a borderless world*, Oxford [u.a.: Oxford Univ. Press], 2006.

Hyde 2013 Hyde, A. "Book Sprints", accessed 14/08/2013, http://www.booksprints.net, 2013.

Issuecrawler n.d. Issuecrawler (n.d.) Issuecrawler.net, accessed 08/07/2013, http://www.govcom.org/Issuecrawler_instructions.htm

J. W. Rettberg 2014 Rettberg, J. W. "Visualising Networks of Electronic Literature: Dissertations and the Creative Works They Cite." *Electronic Book Review*. 2014-07-06. http://www.electronicbookreview.com/thread/electropoetics/analyzing

Jacomy et al. 2011 Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software." *PLoS ONE*, 2011, 9 (6): e98679. doi:10.1371/journal.pone.0098679.

Jameson 2006 Jameson, F. "Postmodernism or the Cultural Logic of Late Capitalism", in Kellner, D. Durham, M. G. (eds.) *Media and Cultural Studies Keyworks*, London: Blackwell, 2006.

Kaplan 2015 Kaplan, F. "A Map for Big Data Research in Digital Humanities." *Frontiers in Digital Humanities*, 2015, 1.

Kirschenbaum 2011 Kirschenbaum, M. "Digital Humanities As/Is a Tactical Term", in Gold, M. K. (ed.) *Debates in the Digital Humanities*, University of Minnesota, 2011.

Krebs 1999 Krebs, V. "The Social Life of Books. Visualizing Communities of Interest via Purchase Patterns on the WWW." Orgnet.com, 1999, accessed 14/08/2013, http://www.orgnet.com/booknet.html

Liu 2012 Liu, A. "Where Is Cultural Criticism in the Digital Humanities?" In M. K. Gold (ed.), *Debates in the Digital Humanities*. Minneapolis: U of Minnesota P, 2012, pp. 490-509.

Liu 2013 Liu, A. "From Reading to Social Computing," in Price, K. and Siemens, R. (eds.) *Literary Studies in the Digital Age: An Evolving Anthology*, 2013, New York: MLA P / MLA Commons, accessed 11/02/2013, http://DLSAnthology.commons.mla.org

Marres and Weltevrede 2013 Marres, Noortje and Weltevrede, Esther "Scraping the Social?" *Journal of Cultural Economy*, 2013, 6, no. 3: 313–335.

Moretti 2005 Moretti, Franco. *Graphs, Maps, Tree: Abstract Models for Literary History.* London: Verso, 2005.

Ramsay 2010 Ramsay, Stephen. "The Hermeneutics of Screwing Around; or What You Do with a Million Books." Presented at Brown University, April 17 2010. http://www.playingwithhistory.com/wp-content/uploads/2010/04/hermeneutics.pdf

Rogers 2010 Rogers, Richard. "Mapping public Web space with the Issuecrawler." *Digital cognitive technologies: Epistemology and the knowledge economy*: 89-99, 2010.

Rogers 2013 Rogers, Richard. *Digital Methods*. Cambridge, MA: MIT Press, 2013.

Rogers et al. 2013 Rogers, Richard, Esther Weltevrede, Erik Borra, and Sabine Niederer. "National Web Studies: The Case of Iran online." In *A Companian to New Media Dynamics*, edited by J. Hartley, J. Burgess, and A. Bruns, 2013.

S. Rettberg 2014 Rettberg, S. "Electronic Literature." In Ryan, Marie-Laure, Lori Emerson and Benjamin J. Robertson. *The Johns Hopkins Guide to Digital Media*. Baltimore: Johns Hopkins UP, p 169-172, 2014.

Schöch 2013 Schöch, C. "Big? Smart? Clean? Messy? Data in the Humanities." *Journal of Digital Humanities,* November 22,

2013. http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

Wang 2012 Wang, Dan. "Is there a Canon in Economic Sociology?" in *ASA Economic Sociology Newsletter* 11(2), May 2012.