Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2015: 1–15

DOI: 10.1177/2053951715595823
bds.sagepub.com
SAGE

Big Web data, small focus: An ethnosemiotic approach to culturally themed selective Web archiving

Saskia Huc-Hepher

Abstract

This paper proposes a multimodal ethnosemiotic conceptual framework for culturally themed selective Web archiving, taking as a practical example the curation of the London French Special Collection (LFSC) in the UK Web Archive. Its focus on a particular 'community' is presented as advantageous in overcoming the sheer scale of data available on the Web; yet, it is argued that these ethnographic boundaries may be flawed if they do not map onto the collective self-perception of the London French. The approach establishes several theoretical meeting points between Pierre Bourdieu's ethnography and Gunther Kress's multimodal social semiotics, notably, the foregrounding of practice and the meaning-making potentialities of the everyday; the implications of language and categorisation; the interplay between (curating/researcher) subject and (curated/research) object; evolving notions of agency, authorship and audience; together with social engagement, and the archive as dynamic process and product. The curation rationale proposed stems from Bourdieu's three-stage field analysis model, which places a strong emphasis on habitus, considered to be most accurately (re)presented through blogs, yet necessitates its contextualisation within the broader (diasporic) field(s), through institutional websites, for example, whilst advocating a reflexive awareness of the researcher/curator's (subjective) role. This, alongside the Kressian acknowledgement of the inherent multimodality of on-line resources, lends itself convincingly to selection and valuation strategies, whilst the discussion of language, genre, authorship and audience is relevant to the potential cataloguing of Web objects. By conceptualising the culturally themed selective Web-archiving process within the ethnosemiotic framework constructed, concrete recommendations emerge regarding curation, classification and crowd-sourcing.

Keywords

Selective Web archiving, Bourdieu, Kress, multimodality, ethnosemiotics, curation

Introduction: The practice and the theory

Through a combination of Bourdieusian ethnographic and Kressian semiotic principles, this article proposes a conceptual framework for the construction of a *small* corpus of thematically linked Internet objects within a *big* Web archive. The fundamental purpose of a Web archive is to retain a version of the fragile (Strodl et al., 2011: 8; Taylor, 2012: 2) and ephemeral (Day, 2006: 178; Gomes and Costa, 2014: 107; Masanès, 2006: 6) digital material found on the Internet for posterity, thereby providing a lasting record of Web objects deemed to be of intellectual and cultural value to current and future generations (Digital Preservation Coalition, n. d.; Kitchin, 2014: 30; Pennock, 2007: 1). As distinct from a digital archive per se, which preserves digitised copies of physical collections or born-digital documents never available in 'hard' form, a Web archive collects only 'material' found on the

Department of Modern Languages and Cultures, Faculty of Social Sciences and Humanities, University of Westminster, UK

Corresponding author:
Saskia Huc-Hepher, Department of Modern Languages and Cultures, Faculty of Social Sciences and Humanities, University of Westminster, 309 Regent Street, London W1B 2UW, UK.
Email: S.V.Huc-Hepher@westminster.ac.uk

'immaterial' Internet, regularly safeguarding it from future obsolescence as the on-line landscape evolves. In this sense, a Web archive, or collection therein, is not so much a record of born-digital data, and by no means an 'identical copy' (Brügger, 2014: 20) of the Internet, rather it is a reproduction, a created entity composed of digital material reborn and brought together in a technically and ontologically more restricted environment than in its original dynamic network.

Cognisant of the inherent limitations of Web archives in relation to the live Web (Pennock, 2013: 5; Spaniol et al., 2009) and with concerns over their long-term usefulness, or at least usability, as vast repositories of unwieldy Big Data,[1] this article ascribes several 'ethnosemiotic' principles to the practice of curating a smaller, thematically selected Web collection, which may arguably be a more manageable set of materials for present and future end-users (Brown, 2006: 32), as well as drawing attention to some of the problematics concerned, all the while from an 'ethnosemiotic' perspective. The collection under scrutiny is effectively an archive within an archive: for a Web archive refers to a vast agglomeration of resources harvested automatically from the entire World Wide Web (as with the US Internet Archive) or an entire national domain (as with the UK Web Archive or the Danish Net Archive; Jacobsen, 2008), whereas the 'micro' archive (Brügger, 2005: 10) under discussion is a targeted corpus of websites selected for their thematic coherence, presenting users with a clear pathway through the mass of 'messy' (Kitchin, 2014: 160; Mayer-Schönberger and Cukier, 2013: 12) data contained in the colossal, and ever-expanding, national UK Web Archive. Further, just as the collection of Web objects discussed here offers a defined, and necessarily small, route into the big data that arguably constitute Web archives, so the ethnosemiotic approach posited, which draws on the points of convergence between Bourdieusian and Kressian conceptualisations of *ethno*graphy and *semiotics* respectively, aims to offer a fine-grained theoretical route into the curation exercise.

The example taken is the London French Special Collection (LFSC),[2] housed within the UK Web Archive, which has been harvesting websites from the UK domain since 2004 and is itself hosted by the British Library. The Collection responds directly to the UK Web Archive's key mission to 'reflect the diversity of lives, interests and activities throughout the UK' (Pennock, 2013: 26) in its (re)presentation of one of London's most significant, yet comparatively invisible, minority communities: the French. In combining the theories of Bourdieu and Kress, and relating them to the LFSC, as curation process: selecting Web material that demonstrates the everyday existence of the London French in the spatial and temporal context of the here and now; as archival product: ensuring that the archived collection serves the social purpose of (re)presenting and preserving the multifaceted aspects of this community, from the institutional to the individual, through a variety of genres, discourses and modes, on a platform which is socially committed through its open accessibility (Kitchin, 2014: 55); and as analytical object: drawing on notions of field theory, reflexivity, objectivation and multimodality, the ethnosemiotic approach, integrated within a broader ethnography of the French community in London, finds its wider justification and use as a case-study here.

Constructing the Collection is one facet of an overarching doctoral project that seeks to reveal the everyday experiences and attitudes of a demographically diverse sample of London's French migrants, as recounted first-hand through a series of semi-structured interviews, focus groups[3] and a paper survey, as well as through other traditional ethnographic methods, such as participant observation and note-taking within London-French circles, and less traditional ones, including the social-semiotic analysis, or 'deconstruction', to adopt Kitchin's terminology (2014: 190), of community Web objects. To that end, in 2011, work began to appraise and collect Web material, or that which could be broken down into 'Web elements', 'Web pages' and 'Web sites' from the London-French 'Web sphere' (according to the five-tiered conceptualisation of the Web developed by Brügger, 2014: 5) to build a corpus of resources for the LFSC. Each Web entity was selected from the live World Wide Web, irrespective of domain (despite the standard UK TLD – Top Level Domain – scope of the UK Web Archive, since excluding <.com> and <.fr> domains, for instance, would have precluded a significant number of thematically relevant sites and pages) and was captured with the Web Curator Tool, which, like the majority of other tools, uses the Heritrix Web crawler, developed by the Internet Archive. In an effort to achieve consistency with the theoretical framework of the overarching project and to reflect the community as fully as possible – in keeping with the BL remit – the curation, construction and analysis stages were approached from a multimodal ethnosemiotic angle.

Although rarely united in a single investigative or analytical undertaking, with some notable exceptions (Bezemer et al., 2013; Dicks et al., 2006; Vannini, 2007), ethnographic and social-semiotic schools of thought share much common ground, such as agency and interest; habitus, practice and the insights of the imperceptible; the tyranny of language; dynamics and meaning-making; holism; reflexivity and social engagement. It is this hitherto unexplored common conceptual

ground that is seen as relevant to the practice of thematic, selective Web archiving and analysis. The branch of semiotics to which Kress subscribes, and by extension that adopted in the curation and examination of the LFSC, is the British school of *social* semiotics, in particular, multimodal social semiotics. Multimodality, in this context, refers to the multiple channels through which meaning is expressed in on-line environments, extending from the ostensible 'major' modes of written text, audio text or moving image – all of which can be embedded in the medium of a single Web 'page' – down to the finer-grained modes of gaze, layout or colour found within them. Each mode is capable of imparting meaning – however implicitly – and each acts intermodally (Jewitt, 2011: 11). Likewise, each mode is necessarily contingent on the socio-cultural context of its utterance (Kress, 2010: 8). All cultures or communities, in this case, the French community in London, as Lotman postulates, exist in their own 'semiosphere' (1990: 124–125), that is, the entire semiotic space of the culture in question, and it is the semiosphere of the London French on-line – itself a manifestation of the physical semiosphere they inhabit on-land – that informs the curatorial approach posited here and the prism to be relied upon when the corpus is transformed from a collection of Web objects to an object of analysis in its own right.

The major traits of Bourdieusian ethnography bear a striking resemblance to the socio-semiotic aspirations of Kress, both of which helped to define the curatorial strategy adopted. Bourdieu insists that the logic of a theory of practice lies precisely, and exclusively, in its juxtaposition with, application to, and reflection on, the broader field and social space (Bourdieu, 1972[2000]: 263), in the same way that Kress believes that all modal communication and representation is a product of the prior social and cultural shaping of individuals and communities (Kress, 2010: 19), and should be seen in the (con)textual frames of 'discourse' and 'genre', as well as in the 'field of meaning as a whole (...) [and] across the range of modes in different societies' (2010: 11). Similarly, just as Bourdieu's theory of practice, notably his theory of habitus, seeks to find meaning in the ordinary habits and habitats of individuals and communities, in their embodied, habituated practices and tacit knowledge, so Kress emphasises the significance of the quotidian in revealing broader (socio-cultural) meanings (2010: 69). In other words, by shining a beam onto the minutiae of pre-reflexive, taken-for-granted, daily practices and activities of a specific population – that is, the ontological denotation of habitus – Bourdieu makes visible previously invisible, or at least undetected, social and cultural dispositions that he then attempts to translate into broader truths free from the 'objectivist' structuralism of Marx and Levi-Strauss (1972[2000]: 256). Thus, Bourdieu recommends a shift from the *opus operatum* to the *modus operandi* (1972[2000]) in order to unearth hidden realities, just as Kress believes it

> is the unnoticed, near invisible social and ideological effects of the signs of the everyday, the signs of ordinary life, of the unremarkable and banal, in which *discourse* and *genre* and with them *ideology* are potently at work – nearly invisibly – as or more effective than in heightened, clearly visible and therefore resistible instances. (2010: 69; original italics)

It is by applying these interrelated theories of Bourdieu and Kress to the LFSC that the construction of an entirely novel, mutually enhanced, conceptual, methodological and analytical paradigm has been achieved, of practical use to future curators and researchers alike.

Between curation and creation: Constructing a community

Empirical evidence gathered in the wider London-French study revealed that a resounding majority of this population recognises the existence of a French community in London; yet, as individuals, they do not conceive of themselves as belonging to it (Huc-Hepher and Drake, 2013: 402). For them, the French community in London is based in and around South Kensington and refers to a socio-economic elite with which they cannot identify (Favell, 2008: 125, 175; Huc-Hepher and Drake, 2013; Block (2006: 133) refers to them as 'free agents', whilst the very absence of a notion of French-community ties in Ryan et al.'s (2014) study of London-French social networks is telling). If this sentiment is considered to be applicable – hypothetically – to the London-French 'community' as a whole, it subsequently poses the question of the very validity of constructing a 'community' Web archive. For how can a community archive be created if the community does not exist in the eyes of its very 'members' and indeed has little visibility (Kelly, 2013: 436) in the eyes of the local population? Indeed, what are the effects of *objectifying* Web material which does not consider itself an object, less still a 'monument' (Brügger, 2005: 17)? Selecting and archiving a Web object which has hitherto functioned principally as a means of communication or display in the dynamic environment of the World Wide Web (although this functional notion is in itself complex, as the distinction between communication and representation is at best hazy in many on-line contexts; Kress, 2010: 191; Pennock, 2013: 10) systematically raises it to the status of aesthetic, historical or scholarly artefact through its very inclusion in a British Library archive. Surely, this transforms the

task of curation to one of creation: through the process of selection of on-line manifestations of the French community, the curator is in effect constructing both a culturally themed collection of Web resources reincarnated as rarefied objects of contemplation to be scrutinised by 'secondary' end-users, and a collective identity, or sense of community, of which the individuals themselves are devoid on-land, despite the unperceived commonalities of their shared cultural semiosphere. This could be deemed fitting in an Internet context, where the notion of 'community' is applied more frequently (Berthomière, 2012: 8; Bray and Donahue, 2010: 1; Casilli, 2010: 58; Miller and Wood, 2010: 1) than in physical settings, the term 'on-line community' referring to any group of individuals connecting to the same Web resource and often connected purely through this digital, physically disconnected, means (Rowley et al., 2010: 1), bearing direct witness to such a phenomenon. It can therefore be argued that the assemblage of culturally linked Web objects into a single 'community' collection has creative implications ontologically, imposing a collective identity on potentially disparately conceived websites and their creators, and epistemologically, since a parallel can be drawn here between the functional transformation which the final corpus has undergone, effectively taken from its born-digital dynamic, 'live' state and reborn as a static, thematically coherent, yet temporally and at times technically incoherent (Brügger, 2005: 23; Lepore, 2015: 18; Pennock, 2013: 12; Spaniol et al., 2009: 1), archived body.

Having acknowledged these caveats, Bourdieu's three-stage field analysis paradigm (Bourdieu and Wacquant, 1992) was borne in mind for the Web selection process. Strict adherence to Bourdieu's model involves: (1) positioning the field of study (in this case, the French community) in the overarching field of power (in this case, the French – and London – governing bodies); (2) identifying the objective structural relationships between competing individual and collective agents within the field(s) (for example, the relationship between French Londoners with official community groups or local schools); and (3) examining habitus and the effect thereof in the field(s) (in other words, the dispositions and practices of the London French) (Bourdieu and Wacquant, 1992: 80). Websites lending themselves to each of these analytical tiers therefore informed the LFSC selection methodology, thus allowing for a diverse (re)presentation of the London-French diaspora, rather than a monochromatic portrait which would crystallise the established (South Kensington) 'community' myth. Bourdieusian 'field' can be conceptualised as simultaneously comprising three denotations: field as (professional) domain, field as (power) game and field as (researcher) terrain, all of which are present in his 'field analysis' model (Bourdieu and Wacquant, 1992: 80; Grenfell, 2012: 222; Jenkins, 1992: 86). Consequently, the Franco-British Council, the French Institute, the French Lycée and the French Embassy websites, for example, were chosen to represent the field of administrative power, whereas sites such as Notre Dame de France (Roman Catholic Church), *Ici Londres* magazine or the Parti Socialiste were included to throw into relief the dominant field of power, as their respective religious, media and political influences could prove to counter that of the establishment, thereby potentially revealing field as game. Subsequently, these Web resources serve as empirical evidence at the level of field as terrain, in that they will become research objects at the final analytical stage of the undertaking. Web objects representing field as domain, such as Jean Michel Brun Ltd. (interior design), Les Editions de Londres (on-line publishing) or Echange Theatre Company (amateur dramatics) sites, were also collected, as they provided another perspective on the microcosmic social workings of the community within the macrocosmic social field of the 'host' culture. These Web objects, when selected in conjunction with other on-line material demonstrating the quotidian practices of the French on-land, and as such shedding light on migrant habitus, for instance, the 'Teatime in Wonderland' and 'Britishette' blogs or the 'Bastille Day Ball' Web page, help the researcher and/or end-user to understand the three-dimensionality of the migrant experience within the field (as domain and game). Furthermore, by embedding the LFSC at the centre of the broader London-French ethnography, itself an embodiment of the diversified data-gathering approach recommended in the Bourdieusian investigative paradigm, not only is the research triangulated, it is given greater (socio-political) meaning and validity (Kitchin, 2014: 147, 191).

The application of Bourdieu's field theory resulted in a rich dataset, not only regarding provenance, ranging from the official records of the established community to the informal displays of the unestablished 'non-community' (cf. the French diaspora's 'non-histoire', Berthomière, 2012: 1), but in the heterogeneous modes of expression presented, from the written and spoken word to the drawn, photographed and moving image. This selection method aimed to (re)present a cross-section of genres and discourses, allowing for the appreciation of field as terrain in the wider framework of field as domain(s) and game, abiding therefore by the objectivation strategies presented in the Bourdieusian model.

Whilst theoretically secure as a selection strategy, and successful in its manifestation of the London-French social field, the resultant corpus occasionally

falls short in its multimodal affordances due to the 'coherence defect' (Spaniol et al., 2009: 1) between the live Web and the 'surrogates' (Day, 2006: 178) archived in the collection, which at times – but inconsistently – lack the images, audio, video, layout and (hyper)links of the original Web pages. Despite the intrinsic technical shortcomings involved in the reproduction of the material at the present time, applying a relational, field-theory methodology not only enhances the comprehensiveness of the culturally themed corpus but also facilitates the task of selecting 'relevant' Web objects from the 'big data deluge' (Kitchin, 2014: 130; Mayer-Schönberger and Cukier, 2013: 70, both citing Anderson, 2008) that the Internet constitutes, which brings us to the question of 'value' and how to define it.

Future memory: Valuing habitus in the hinterland between the now and the not yet

Pennock (2007: 1) describes digital curation as 'maintaining, and adding value to, a trusted body of digital information for current and future use: in other words, it is the active management and appraisal of digital information over its entire life-cycle'. Yet, this definition fails to address the underlying complexity of both 'value' and 'appraisal', and the temporal implications of the 'current and future', as Dallas (2007: 3) astutely points out, inherent in the curation exercise. For, as with a physical archive, determining the value of a Web resource is not straightforward: According to which *criteria* can 'value' be defined and assessed? By what means can the longevity of 'value' be anticipated, when information deemed of value today risks not being held in equivalent esteem in future? The prospective assessment of value poses a major challenge to Web (and conventional) curators, all of whom are inextricably bound to their judgemental points of reference at the time at which they are making such assessments (Pennock, 2013: 10). Moreover, given the vastness of the data available on the Internet and, equally importantly, the lack of a long-standing Web-archival precedent, the difficulty of the task is multiplied for the curator of on-line material. Peters poses similar questions as those raised above (2011: 4), exacerbating the dilemma further by injecting the notion of *community* value and its appraisal, together with that of constructing a collective memory. He acknowledges that 'a collect-all approach (…) needs to be filtered and measured against criteria of demand: community memories that reflect communities' interests', but provides no solutions as to a reliable method of creating 'collective memories' or assessing 'valuable content' (2011: 4). He is not alone; the absence of a universal theory of digital curation (Flouris and Meghini, 2007; Hockx-Yu and Knight, 2008; Moore, 2008) and, by extension, an agreed theory of selective Web archiving, remains a challenge. With the exception of some persuasive, if technically focused, strategies put forward by Brown (2006), Brügger (2005) and Masanès (2006), theorising the practice of Web curation has been largely ignored. Flouris and Meghini (2007) have developed an objective, mathematically inspired theory of digital preservation for digital libraries, but this does not extend to the process of digital, or more specifically Web, curation. Furthermore, the curatorial applicability of this type of formulaic theoretical system to *selective* archiving is arguable in its very negation of the reflective, sensitive, informed and necessarily subjective curator from the curation process, boiling 'value' down to a set of lifeless equations and removing the 'aura' (Taylor, 2012: 8) and the 'subject-matter experts' influence' (Mayer-Schönberger and Cukier, 2013: 141) from the selection process.

If the traditional archivist's criteria for assessing value, as set out by the British Library prior to selections being made, are to be relied upon, those Web objects offering the most scholarly and verifiable information on the London-French community ought to have been favoured in this particular collection. Indeed, it was the specific remit for the curation of the LFSC that it should contain a substantial quantity of such material: 'Nominations and collections of archived websites that support scholarly research are therefore of particular interest' (Pennock, 2011: 1), which stands to reason given the UKWA's status as a 'trusted digital repository (TDR)' (Kitchin, 2014: 33). At this point, however, it would appear that the interests of the Web researcher-curator and those of the conventional or digital librarian-archivist may diverge. The 2013 UK non-print legal-deposit regulations constitute another point of departure: the British Library/UKWA and the traditional archivist seem to welcome the right to regularly crawl the UK domain and indiscriminately harvest big Web data, bypassing the need for temporally and financially onerous permissions (Jacobson, 2014: 2; Pennock, 2013: 9, 13), whereas the researcher-curator of the LFSC perceives the legislation in a less favourable light,[4] since any Web object selected and harvested for the collection under the licence-free framework would be housed in an ostensibly 'separate' collection, causing it to become 'stranded data' (Kitchin, 2014: 156, quoting Singh, 2012) accessible only on-site in one of the UK's six legal deposit libraries, thereby reducing the potential audience of the collection as a whole and jeopardising its socially committed founding principles (and therewith realising the interoperability and open-accessibility concerns voiced by Kahle (in Jacobsen,

2008: 4; Kitchin, 2014: 38, 55; Lepore, 2015: 7; Mayer-Schönberger and Cukier, 2013: 116)). In other words, with regard to this particular collection, which began its life pre-legal-deposit legislation and will continue to grow indefinitely, institutional 'Power and politics [may] continue to underwrite access' (Taylor, 2012: 8–9), just as they have in physical archives.

Power, politics and legislation aside, when adopting an ethnosemiotic theoretical model for 'valuation' and appraisal in culturally thematised Web curation, it is, arguably above all else, the habitus element of Bourdieu's three-stage field model which should take precedence; that is, the resources displaying the quotidian, taken-for-granted practices and spaces of the community under scrutiny. Kress states that 'communication is embedded in social environments, arrangements and practices' (2010: 35); similarly, Bourdieu gives prominence to a theory of practice (1965, 1972[2000], 1980, 1994), articulated through his concept of habitus. While Bourdieu's notion of field lends itself convincingly to the *selection* process, it is data embodying the habitus of the London French that is predicted to be of most *value* to future historians. Research conducted by the IRN on behalf of the UKWA supports this theory, as all scholarly 'users expressed the requirement for including more images and rich media, as well as more blogs' (Hockx-Yu, 2012: 1). The voice of the lone blogger is hence deemed of equal value to, if not greater value than, that of the political party; likewise, the objects and spaces, habits and practices, opinions and viewpoints of the blogger's on-line habitus are tantamount – in terms of their present analytical and prospective community/historical worth – to the official manifestations of London Frenchness, by virtue of the insights they provide into the cultural reality of the here and now. The survey alluded to in Ball's paper (2010: 24) confirms the perceived long-term value of blogs, with 71% of the 223 respondents believing their own blog should be preserved. Hank's empirical study also demonstrates that the majority of scholars who blog 'viewed their blogs as part of their scholarly record' and 'had an interest in preserving' them (Hank, 2013: 6). Given that blogs 'have the characteristics of personal journals' (Yoon, 2013: 175), Yoon too believes them to be of marked cultural and historical value for future scholars. The fact that they offer a privileged 'window into the past' (Yoon, 2013: 175, quoting O'Sullivan, 2005), providing future onlookers with rich evidence of the socio-cultural make-up of their time, since 'individual memory can only be recalled in the social framework within which it is constructed' (Yoon, 2013: 175, citing Halbwachs, 1992), confirms both their preservation worth and their status as convincing (re)presentations of the internal–external dialectics of Bourdieusian habitus. Thus, if the blogosphere is the closest the on-line environment comes to a window onto the habitus of London's contemporary French population, it can be argued that autobiographical Web data such as blogs should take precedence in the assessment of future value.[5]

The subjective self: Notions of authority, authorship, agency and audience

Although Bourdieusian habitus, as set within the structuring field, is helpful in constructing a 'valuation' framework for culturally themed collections, it remains difficult to avoid the 'selector bias' (Pennock, 2013: 10) inherent in selective 'micro archiving' (Brügger, 2005: 10) and, by extension, the curated product. Some have argued that it is this very subjectivity that distinguishes – positively – a curated collection from other so-called on-line 'archives', such as YouTube or Flickr, which are little more than 'vast reservoirs of materials' (Dawson, 2010: 12; Taylor, 2012: 2), 'data stores or back-up systems' (Kitchin, 2014: 30) because they are not subject to 'expert' appraisal or selection. However, the extent to which the 21st-century digital curator is an expert (Dicker, 2010: 3) in the field of Web archiving is questionable in view of the very 'openness' and 'democracy' (Casilli, 2010: 45; Taylor, 2012: 5) which has enabled access to the role in the first place. Many digital and Web curators receive little or no training, despite efforts to reverse this (Bromage, 2010: 1), and many on-line collections welcome user-generated content (Dicker, 2010: 1), user nominations of Web material (Gomes and Costa, 2014: 115; Lepore, 2015: 11; Masanès, 2006: 5) and user cataloguing information (Jacobsen, 2008: 3). Whilst this is in keeping with the open-access, collective ethos of the Internet and of institutional digital preservation initiatives (for instance, Bromage, 2010: 5; Dawson, 2010: 3), it is simultaneously somewhat paradoxical in its subversion of the 'valued' authority formerly invested in and associated with recognised archiving bodies, such as the British Library. As Dawson indicates, memory institutions should 'be conscious of the value that they bring (…) with respect to curation and quality of knowledge' (2010: 5); yet by outsourcing Web curation projects to benevolent 'interested-amateurs', they risk not only compromising the quality of their collections but also jeopardising their reputations. Despite these valid 'concerns about the quality and consistency of content and metadata created across diversely skilled/motivated individuals' (Kitchin, 2014: 155), among the advantages of loosening the hold over knowledge and information, is the economic gain of tapping into the services of willing researchers and other non-specialist parties interested in preserving cultural heritage (Masanès,

2006: 5), together with the opportunity it presents to begin to manage a minuscule proportion of the mass of data contained in the archives of the World Wide Web. The nascent age of big data promises multiple research opportunities, but its sheer volume could render it 'too big to handle', ultimately resulting in the UKWA becoming an underexploited 'dusty archive' (Meyer, 2011) or 'data mortuary' (Beagrie, 2006: 5, quoted in Dallas, 2007: 53), hence the necessity for targeted, thematic or otherwise, management of big Internet data in the form of smaller, selective collections curated by subjective subject-experts.

Whereas the curator of a themed Web collection is not necessarily a specialist in archival cataloguing or museum curation, it is likely that (Pennock, 2013: 10), or at least beneficial if (Gomes and Costa, 2014: 110), s/he has deep insider knowledge of the 'field' for which the collection has been created, which reintroduces the subjectivity-objectivity question from another angle. In keeping with Bourdieusian three-stage field analysis, 'insider' research, that is, an investigation which places the researcher at the boundary between external observer and internal participant, is ethically sound and scientifically valid, provided the researcher engages in the process *reflexively*, and is not, as Pennock fears, creating a collection that is 'unintentionally biased' (2013: 10). Likewise, it could be argued that provided the partial Web curator undertakes the process of appraisal and selection with an active awareness of this subjective position, s/he is equally justified in casting judgement over the potential value of a Web object, as opposed to making its 'research value constrained' (Pennock, 2013: 10). It is subject knowledge, or in this instance, the researcher's subjective knowledge of the research object, namely the French community in London, which validates the curator's agentive role and, in turn, endows him or her with due authority (Dicker, 2010: 9–10; Gomes and Costa, 2014: 110).

However, if the authority of the curator of a collection is subsequently dependent on (a) the institution's quality assurance and permissions systems and (b) permission being granted by the website holder for inclusion within the collection, the question of where the ultimate authority and agency dwell resurfaces. The Web researcher-curator is empowered to select and appraise data but is denied the authority to seek permissions actively and independently; similarly, the 'memory institution' (Dawson, 2010: 5) is authorised to accept or reject selections, but – until the 2013 amendment to non-print legal-deposit regulations – was refused the right to collect Web information without creator consent, thereby leaving the definitive authority with the producer of the content. Thus, the digital age brings with it a blurring of the lines of hitherto clear-cut distinctions between the established authority of the institution and the subordinate visitor (Dallas, 2007: 62), between the authority of the qualified curator and the lay selector, 'utilising the knowledge, expertise and interest of the community' (Holley, 2010: 2), and between the authority of the traditional author and the self-generated authorship of the on-line creator. As Kress underlines, 'formerly settled – quasi-moral, legal and semiotic – notions about authorship, text and property are now no longer treated as relevant; or are, more often than not, no longer recognised by those who engage in text-making' (2010: 21). Consequently, the authority of the untrained Web curator is jeopardised no sooner does the collection 'go live' and become accessible to any member of the on-line public, at which point any Internet user consulting it can nominate potential Web material, in the spirit of the crowd-sourcing, 'citizen science' (Kitchin, 2014: 97) era. It is precisely these redistributions of authorised and authorial power that Dallas addresses in his agency-oriented approach to digital curation theory and practice (2007), and which resonate with the technologically fuelled revolution in epistemological dynamics to which Kress refers (2010: 21, 134). Despite the doubt and uncertainty that such an overturning brings (Taylor, 2012: 2), akin to the 'dark side of big data' to which Mayer-Schönberger and Cukier ominously allude (2013: 170), it also offers new opportunities for the transmission and acquisition of knowledge, enabling users to become authors and giving curatorial agency to formerly passive visitors (Allen-Greil and MacArthur, 2010: 3; Kitchin, 2014: 188), and thereby serves the social function prophesied by Bourdieu and Kress.

Irrespective of the arguable socio-politically democratising role presented by new technologies, new authors necessarily imply new audiences. If it is conceded that Web collections blur former boundaries of authority and authorship, the resources they contain are also likely to be accessed by new users. Audience is a notion addressed both by Kress and Bourdieu, and is one that can be usefully applied to collections of Internet objects, in that the curator needs always to be mindful of the 'reborn' audience(s) the collection addresses, over and above the audiences of the born-digital objects, in the same way that big-data analysts should look beyond the primary function of the data and anticipate 'the value of information [. . .] in secondary uses' (Mayer-Schönberger and Cukier, 2013: 153) which may again affect the selection strategy adopted. Kress argues that all communicational and representational acts are interest-led (2010: 67), and whether the interest lies with the sign-maker or sign-recipient is largely dependent on the semiotic function of the resource as a whole. For example, Domingo et al. demonstrate that image is increasingly 'taking the place of

writing at the centre of the communicational stage' (forthcoming: 2), particularly in instructional food blogs, and that by designing Web pages in particular ways, making use of colour, spatial composition and (moving) images, for example, authors-cum-designers are conveying specific – though tacit – meanings to their 'readership'. The notion of audience is therefore intrinsically linked to that of design and authorship, and an understanding of multimodal socio-semiotic analytical principles equips the Web curator with the necessary skills to assess these initially imperceptible messages and agendas. Yoon (2013) and Technorati's (2010) empirical findings also confirm the centrality of audience and design in bloggers' motivations, as all Yoon's respondents declared having an intended audience (Yoon, 2013: 181), and it was cited as a major motivational influence for the bloggers of the quantitative Technorati study (Technorati, 2010). Audience, therefore, unlike the 20th-century *personal* journal (often fitted with a key to denote its privacy), shapes the content and provides the impetus for the 21st-century blogger.

Bourdieu's field theory is again pertinent in relation to audience: if all communication and action takes place within the broader framework of field (as game), questions over the respective agendas of key players in the special collection/Web archive/institution and their targeted users/audience come to the fore, and of the multiple audiences subsumed within the archive itself. That is, websites containing the official discourse of London-French 'authorities' will be designed to reach one audience, while blogs produced by French Londoners target quite another. Indeed, the multiple audiences envisaged by Yoon's respondents, decreasingly composed of friends, family, the general public, other bloggers, colleagues, professional networks and their selves (2013: 181), not only confirm the bloggers' target audience, as distinct from that of official sites, but also demonstrate the inadequacy of a singular notion of 'audience' when curating a stand-alone Web collection. Although the original, live-Web blog audience is intended to include all the above, it is possible that there is also an unintended, 'covert' (Murthy, 2008: 846) researcher audience in the born-digital environment, with yet another layer of present and future audience(s), coming at the material from very different perspectives, joining the strata when the new version of the Web object is reborn in the archive. The implicit heterogeneity of audience in born- and reborn-digital settings compounds the validity of the ethnosemiotic appraisal and selection process outlined above, since the methodology transcends the notion of 'audience' as a unified, homogenous whole, instead acknowledging and predicting the multiplicity of audiences implicated when on-line data is reborn in surrogate surroundings, in this case the LFSC/UKWA, and recognising the intrinsic infiniteness of meaning(s) through its dependency on audience interpretation (Kress, 2010: 37).

The implication of language: Naming and framing

Just as the notion of 'audience' is deceptively simple in the context of on-line curation, so language is superficially straightforward. When collecting Web objects for inclusion in a themed collection, the curator is required to engage in a process of naming and framing to give a sense of 'order' to the collection and increase its usability/accessibility. However, given the plethora of librarian standards for generic positioning and the allocation of metadata (Gill et al., 2005), as well as the discrepancies between archival, as opposed to Web-based, norms (relative to both structure and content), the activity of naming, defining, categorising and framing material is not straightforward; indeed, the recent admission in the Web community that reaching an absolute standard is unattainable means that Lyman's urgent call (in 2002) for a 'standard way of recording the metadata (...) to record the historical and technical context' (2002: 4) of Web objects harvested has yet to be achieved 13 years on. In addition to the pragmatic complexity of ordering and labelling originally *net*worked, uncategorised Web material – which is 'not discrete' in its born-digital form (Lyman, 2002: 2) – in a thematically rationalised, bounded *frame*work, are the deeper, ideological implications of the process. Both Bourdieu and Kress emphasise that language is not innocent; 'words *do* have power' (Jenkins, 1992: 155) and the researcher-cum-curator needs to be wary of their superficial 'naturalness', which is also a fundamental point made by Bourdieu, who urges the ethnographer to be suspicious of the implicit and symbolic power of language (1972[2000]: 227, 1982). He is emphatic on the repercussions of language in school and higher education fields, deeming insufficient linguistic capital, due to lack of exposure to socially valued language and rhetoric in the habitus of origin, to be the root of much academic underachievement and exclusion (Bourdieu and Passeron, 1964: 25).

The language employed by Web curators is no less innocent. Gomes and Costa highlight the positive role external researchers can make by 'generating additional meta-data' (2014: 110); yet, Dalton articulates concerns over the potentially conflicting interests of user-generated tags and the metadata of specialists, namely curators, who act as mouthpieces for the 'institutional voice' (2010: 5), while Hockx-Yu underlines the need for 'a hybrid of curatorial and technical skills' in order to address the challenges of naming and framing Web

data (quoted in Volk, 2012: 1). Ultimately, irrespective of who assigns the metadata to a website and frames it categorically, doing so is an implicated act: partial curators are implicated through their subjective perspective alone, and the language chosen for description has implications. It could be argued that this has always been the case when cataloguing physical collections, but the difference here is that a Web object is an innately *linked* entity, which in its born-digital state cannot be divorced from the network of which it is a co-dependent part, unlike a physical book which has a discrete physical existence in the world (irrespective of its potential abstract inter-textuality). Web objects are also intrinsically and fundamentally multimodal entities, or 'compounds of design elements' (Lyman, 2002: 4), again, unlike a book restricted by the physical limitations of its form, which complicates the naming and framing process further in the field of Web archiving. This casts doubt over the very applicability of 'cataloguing' archived Web material and may explain why metadata 'are often a neglected element of data curation' (Kitchin, 2014: 9), since 'precise systems that try to impose a false sterility upon the hurly-burly of reality, pretending that everything under the sun fits into neat rows and columns' (Mayer-Schönberger and Cukier, 2013: 43), are inexorably ill-suited to the inherently 'messy' data of the Internet. However, owing to the ever-increasing quantity of data contained in Web collections, selective ones included, and to the paradoxical fact that 'the excess of information can be transformed into a huge data paucity, over time' (Gomes and Costa, 2014: 120), as the 'huge volumes of data [. . .] make it difficult to interact and take advantage of them' (Gomes and Costa, 2014: 116), such designations are deemed in the interest of end-users (Dallas, 2007: 57), providing them with descriptive and contextual insights (as understood by the informed curator) which will assist their navigation through the big data of the Web archive, and thereby improve its research value and the credibility of the archival institution.

Nevertheless, compartmentalising material according to patent content characteristics (Abbot and Kim, 2008) implies classification, which in turn implies 'class'. Kress refers to classification as 'a social and semiotic process carried out by semiotic means', the result of which 'is to stabilize the social world in particular ways' (2010: 122) (which favours the usability argument). However, its 'seemingly innocuous character helps to make its political effects more effective' (2010: 122–123) (which supports the implicated argument). When choosing the terms to describe a selection in the LFSC (e.g. 'Website for guided London walks') or its generic classification (e.g. 'Arts & Humanities' from the seven umbrella subject categories provided, within which combinations of 18 sub-categories can be made, e.g. 'Languages'), the digital curator is performing an implicated semiotic act, at once restricting the meaning potential of the 'raw' material (Dallas, 2007: 58) by introducing an intermediary layer between the Web resource and the user, and allowing for 'unintended' meanings to be drawn from the associations between the resource and its framing genre or the websites and pages alongside it. In the case of the LFSC, these meanings could involve the fabrication of a sense of community (as discussed above) through the collective framing of thematically – but not necessarily socially, ontologically or hypertextually – linked Web objects.

Kress sees framing as a way of punctuating semiosis by fixing meaning in a specific spatio-temporal context and, more importantly, in a given mode, genre and discursive form (Kress, 2010: 122). In an effort to begin to construct a useful theory of culturally themed Web archiving, it is necessary to dwell briefly on Kress's conceptualisation of modal, generic and discursive framing of information. He posits that in any rhetorical process, 'meaning is *fixed* three times over – *materially* and *ontologically/semiotically* as *mode*; *institutionally* and *epistemologically* as *discourse*; and *socially* in terms of *apt* social relations, as *genre*' (2010: 121; original italics). Although the reliance on italics is somewhat obtrusive, it helps to clarify – multimodally – Kress's understanding of mode, genre and discourse. For Kress, therefore, mode corresponds to the channels through which meaning is conveyed, which traditional cataloguers might associate with the notion of medium (although modality functions on a considerably more granular level). Genre relates above all to commonalities between the texts/multimodal ensembles of a specific community or culture; conforming to the sociocultural norms of the genre gives a 'text' its identity and serves to position it within the said genre. In Bourdieusian terms, genre could be seen as the 'textual habitus' of a Web resource, emanating from social practices and interactions. Discourse, however, acting at a broader, external level of institutions and governing bodies (2010: 110), shapes and imparts knowledge. Bourdieu might have referred to discourse as 'textual field' therefore. Both can be considered to operate at the level of extra- and inter-textual coherence, rather than intra-textually, as is the case for modes, and both the (con)textual generic and discursive characteristics of a harvested Web object warrant consideration when fixing it an archive.

However, cataloguing Web material according to its perceived generic properties, as defined above, is a challenging and implicated task, requiring fine-grained multimodal analysis of the 'text' itself, coupled with knowledge of the cultural and structural framework of which it forms part. A sense of this complexity is

alluded to on the Digital Curation Centre (DCC) website, where genre classification is described as 'shrouded in ambiguity' (Abbot and Kim, 2008: 1) and a shift from a topical categorisation system to a text-typological one is advised. Indeed, the definition of genre provided by Abbot and Kim echoes Kress's words: 'Document genre, as with music, pertains to style and/or form. The style and form of a document is constructed to meet the functional requirements within the target community in realising predefined objectives of document creation' (Abbot and Kim, 2008: 1). Thus, a multimodal socio-semiotic approach to genre classification, which prioritises the implicit meaning-potential of mode (or text type, to employ DCC terminology) over thematic content (or topic for the DCC), is compliant with the expectations of the digital curation authority, namely the DCC, as well as being dependent on the expectations of its audience, predominantly, in this case, London Francophones or Francophiles. Furthermore, with on-line 'texts' deconstructing formerly fixed understandings of genre, through their simultaneous inclusion of a variety of modes, media, styles, forms and ultimately genres, a multimodal theory of classification, or information framing, is a convincing strategy. In practical terms, and if the recommendations of the DCC were applied, this might mean categorising content in the Web collection/archive according to its *on-line generic* typology, such as blog, website or pdf, as opposed to, or in addition to, its thematic specificity. Using *discourse* as a marker, for example, grouping together Web objects as a result of their pre-classified administrative (<.gov.uk>), institutional (<.ac.uk>), commercial (<.com>, <.co.uk>) or philanthropic (<.org.uk.) domain-name commonality, may present advantages over genre in a themed collection both in terms of its scope for automated classification (Warwick Workshop Report, 2005: 16), which in turn would reduce the subjectivity (and cost) of the classifying process, and its resolution of the ambiguity problematics posed by text typology. That is, in an Internet age where generic text-typological frames are increasingly porous, with a single London-French blog potentially corresponding to an on-land recipe book, diary, article, travel guide, photograph album and more, not to mention its modal variants, categorisation by genre or text type alone becomes a near impossible task, hence the need for discursive differentiation.

In on-line multimodal environments, generic naming and framing, as demonstrated above, is far from straightforward; just as modal boundaries merge in such settings, so fixing Internet texts in the wider socio-cultural and institutional frameworks of genre and/or discourse is challenging, particularly given the propensity of on-line media to encourage new genres to develop out of the medium itself (Domingo et al., forthcoming: 12, 21), and for them to be generically pluralist. This gives rise to the awkwardness and potential arbitrariness of assigning discrete genres to Web content solely for the purpose of facilitated cataloguing and searching, and to Mayer-Schönberger and Cukier's endorsement of organic, ad hoc tagging 'as the de facto standard for content classification on the Internet' (2013: 43). Nevertheless, a culturally themed collection, however small at its inception, is, like the host Web archive and the live Web, an 'infinitely' growing corpus, with additions to its original form being made with every scheduled capture of the on-line resources included (in 2012, the Internet Archive had collected 10 petabytes of data; two years on, the number had doubled, equating to over 'four hundred and thirty billion Web pages' (Lepore, 2015: 12)). In this way, although the cultural theme offers the end-user a coherent and more manageable set of materials within the big data of the Web archive as a whole, the ever-multiplying nature of the collection means that, for the sake of navigability and usability, further classification will doubtless be required in future. To this end, the discursive approach to the sub-categorisation of the on-line data, already framed generically within the cultural context of the London-French ethnological theme, is considered fittest for purpose, not least because it is facilitated by the inherent identity of the Web object's born-digital domain name.

As has been seen, the application of language to archived Web resources is open to misrepresentation and lends itself to oversimplification and/or partiality on the part of the curator (Gomes and Costa, 2014: 107; Pennock, 2013: 10–11), particularly if an inadequate text-typological rationale is adopted. Moreover, the very act of framing a 'set' of otherwise disparate Web objects in a Special Collection is in itself meaningful (Kress, 2010: 119). Through the housing of diverse London-French resources under a thematically homogeneous umbrella, those consulting the collection, today and in future, are likely to create conceptual links between sites and information that may have been unintended and, perhaps more importantly, that would not necessarily be created outside the collection in the born-digital environment; such associations allow for new meanings to be made, infinitely (in accordance with Peircean and Kressian semiotic theory), and for an imposed (by the curator) sense of coherence and inter-textual semiosis to be effected. 'Semiosis, the making of meaning', as Kress explains, 'is ongoing, ceaseless' (Kress, 2010: 93), and it is contingent on the dynamics of its materialisation through a given mode and its realization in the mind of the recipient (Kress, 2010: 93). Thus, regardless of the efforts of the Web curator – who is inescapably fixed in the time,

space and frame of mind at which the cataloguing process is undertaken – to ascribe defined nominal interpretations and generic/discursive classifications to the Web objects framed in the LFSC, their meaning potentials are limitless, and as dependent on the temporal and spatial framing of the end-user as on their integral positioning within the archive. In short, all meaning-making is dynamic and boundless, and as such, the curator's reliance on language to direct and contain it is innately problematic, warranting careful consideration. This leads to the final Bourdieusian and Kressian concept relevant to this discussion: dynamics.

Web-archival dynamics: On-land–on-line symbiosis

According to Taylor, 'the embodied, the archival and the digital overlap and mutually construct each other' (2012: 3), and Toyoda and Kitsuregawa refer to 'the Web as a projection of the real world' (2012: 1442). Thus, there is no distinction to be made between the London French on-land and on-line, the latter is the reflection of the former, and each affects the other symbiotically. The embodied presence of the French in London is displayed in digital form on-line which in turn feeds into the Web archive/collection. In a process of mutual construction, the collection will preserve and renew the physical and digital representations of the London French, the dynamics of which will intensify once French Londoners engage with the user-nomination functionality. The overt manifestation of this symbiosis will be the potential modification of the collection over its life-cycle, according to the nominations made by future LFSC users, many of whom are likely to be members of the London-French community itself, for, as Holley describes, the most successful crowd-sourcing initiatives have been those with which the public feel a direct connection, such as, 'history [. . .], personal lives [. . . or] genealogy', their contribution giving them a 'sense of public ownership and responsibility towards [their] cultural heritage collections' (2010: 2). Less obvious manifestations might take the form of modifications to the style and content of London-French blogs subsequent to the realisation that their once audience-specific material (Yoon, 2013: 181) is to be henceforth displayed and preserved in the official collections of the British Library, and as such transformed from personal log into the stuff of cultural heritage, deemed of lasting historical value to the nation. This inevitable elevation in status will undoubtedly have ontological and epistemological ramifications, measurable empirically only after the collection has been in the public domain for a sufficient period of time.

If members of the French community in London choose to nominate websites and if those already with a presence in the collection adjust, wittingly or otherwise, their behaviour on-land and on-line as a result thereof, they will be explicitly contributing to the dynamic process and product (Taylor, 2012: 4) characteristic of an ethnographic archive. Just as Bourdieu wrote, 'le réel est relatif' (reality is relative, 1994: 17), this dynamic Web archive, unlike traditional archival forms, invites physical, live beings to participate in the on-line curation exercise, thereby enlisting 'visitors as active subjects of knowledge construction' (Dallas, 2007: 59), indefinitely, blurring former divisions between the corporeal and the virtual, the lived and the represented, the present and the future.

Fittingly, dynamics are key to Bourdieusian and Kressian theories, the fact of which substantiates further the legitimacy of a multimodal ethnosemiotic conceptual framework for culturally themed Web curation. The fundamental overlap between the dynamic approaches adopted by Bourdieu and Kress lies in their shared belief that it is only by examining cultural practice through a relativist lens that true, and often hidden, meanings will be revealed. In this way, Kress extols methodological and analytical frameworks which compare modes in order to elicit semiotic substance: 'Depending on the *mode* and its *affordances*, relations and connections may have any number of forms (. . .) as a means of making meaning' (Kress, 2010: 156; original italics), and the LFSC provides an ideal, pre-selected and intrinsically multimodal set of data for comparing such modes. In a similar vein, Bourdieu's methodological and analytical recommendation, expressed through his three-stage field analysis model, that it is through the comparison of a variety of habitus practices (such as speech, posture, drinks, sports, food and so on) (1994: 21), and their positioning within broader field structures, that ethnographers gain an in-depth understanding of the social realities of ordinary people's lives (1972[2000]: 263), echoes the relativism advised by Kress. It is also an approach that has been recently advocated in *Google and the Culture of Search* as 'a helpful way to theorize the human dynamics at play in and across many [on-line/on-land] fields' (Hillis et al., 2013: 30). Grenfell recalls that 'much of Bourdieu's work demonstrates the way in which we should see *habitus* and *field* as mutually constitutive' (2012: 5; original italics), and it is only by scrutinising one that hidden truths of the other will materialise and *vice versa*. Likewise, the dynamic constitution of the LFSC, in its combination of field and habitus material, and in its dynamic spatial and temporal dimensions, is fundamental to its ethnosemiotic identity and validity.

This underscoring of the intermodal and inter-relational is akin to the concept of multimodality itself, in that different modes cannot be separated in their experiential effect, in spite of possible attempts to do

so for the purpose of analysis, for it is their very coalescence which completes the meaning-making. Consequently, all modes combine and mutually interact, in the same manner that all individuals, on-land as on-line, live multimodally, with layout (Kress, 2010: 88) or micro-gestures (Bezemer et al., 2013: 16) constituting equally telling modes as writing or speech, and their 'life lived offline [being] directly connected to online life' (Adami and Kress, 2010: 189). The Internet allows for modal and physical-digital interplay more than the printed text or the material archive has ever before permitted, hence the relevance of a multimodal ethnosemiotic approach to the construction of a Web collection.

Conclusion: Finding big meanings through a small approach to big data

This article has focused on the points of convergence between Bourdieusian and Kressian concepts pertinent to the field of Internet archiving, in an attempt to develop the basis of a theory of culturally themed Web curation. By examining the practice of constructing the ethnographically themed LFSC, as part of the big data that is the UK Web Archive, considerations and suggestions for selecting resources, assessing value, anticipating audiences, cataloguing and crowd-sourcing have been made through the prism of field, habitus, reflexivity, language and dynamics. The significance of the small-scale, micro-Web-archiving approach foregrounded lies in its deployment as a strategy for overcoming the 'data deluge' inevitably triggered by non-selective, catch-all repositories, such as JISC's UK Web Domain Dataset (1996–2010). National archives of the sort have therefore proven to be of limited use to researchers in the arts and humanities today, who are often unsuccessful in accessing the specific datum they seek within the big data archive consulted[6] or, in the words of Kitchin, 'extracting a meaningful signal from the noise' of big data (2014: 151). Until the development of more efficient search tools (Mayer-Schönberger and Cukier, 2013: 41), which nevertheless allow the researcher to have full access to unadulterated material (a delicate balance to achieve), selective archiving remains the most viable, user-friendly option.

The lasting output of *this* selective archiving experiment is the LFSC itself, which constitutes a unique and multifaceted representation of a particular migrant community, offering an exclusive window onto a largely invisible component of Britain's socio-cultural make-up at the dawn of the 21st century. The impact of the ethnosemiotically constructed collection extends from the present day to future users of the archive and covers a broad spectrum of interest and knowledge, from the inquisitive lay visitor, academic researcher or language teacher to the journalist, policy maker, historian or language learner. Furthermore, the approach posited here intends to make a valid contribution to the broader development of Web archiving, being potentially scalable from the community-themed level to larger on-line archives, whose themes may differ but whose selective principles concur.

Future research initiatives could be geared towards that very issue, namely, assessing the applicability of the small approach to bigger datasets and more diversified themes, yet remaining inside the boundaries of the selective approach. Within the scope of this Collection, however, forthcoming enquiry will endeavour to bridge the gap between big-data macro-archiving and small-data micro-archiving, by concentrating on the often-overlooked meso level. An analysis will thus be conducted of Web linkage between resources in the LFSC and <.fr> domain sites (or otherwise) to garner a more nuanced understanding of community interactions and cohesiveness, by supplementing the granular findings with the quantitative link data generated, and thereby 'tap the benefits of correlation' (Mayer-Schönberger and Cukier, 2013: 18).

Ultimately, in the context of big Internet data, the ethnosemiotic approach proposed here offers a qualitative alternative to that which Crawford terms 'data fundamentalism' (2013: 1). The ethnographic smallness and reflexivity – methodologically, archivally and analytically – allow the practices and narratives of individual migrant lives to give meaning to the vastness of the archived Web, which is 'why ethnographic work holds such enormous value in the era of Big Data' (Wang, 2013: 1).

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Notes

1. There is debate over whether Web archives constitute Big Data, but ventures such as the AHRC-funded, collaborative Big UK Domain Data for the Arts and Humanities project, and the 2014 Web Archives as Big Data international conference, together with the sheer volume of data (approximately 65 terabytes) held in the JISC UK Domain Dataset (1996–2010) or the 20+ petabytes in the Internet Archive (Lepore, 2015: 12)), support the definition.
2. Available at: http://www.webarchive.org.uk/ukwa/collection/63275098/page/1/source/collection
3. For further details of the participants, see the Appendix in Huc-Hepher and Drake (2013: 427–429).

4. See also Huc-Hepher's blog post at the British Library. Available at: http://britishlibrary.typepad.co.uk/webarchive/2014/07/researcher-in-focus-saskia-huc-hepher-french-in-london.html
5. Such value is examined in detail in Huc-Hepher's forthcoming articles titled 'The Material Dynamics of a London-French Blogger: A multimodal reading of migrant habitus as (re)presented on-line' and 'Searching for Home in the Historic Web: An Ethnosemiotic Study of London-French Habitus as Displayed in Blogs'.
6. See, for example, Peter Webster's blog post on the experience of researchers involved in the Big UK Domain Data in the Arts and Humanities pilot project, available at: http://buddah.projects.history.ac.uk/category/uncategorized/

References

Abbott D and Kim Y (2008) Genre classification. Digital curation centre briefing papers: Introduction to curation. Available at: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/genre-classification (accessed 18 November 2013).

Adami E and Kress G (2010) The social semiotics of convergent mobile devices: New forms of composition and the transformation of *habitus*. In: *Multimodality. A Social Semiotic Approach to Contemporary Communication*. London: Routledge, pp.184–197.

Allen-Greil D and MacArthur M (2010) *Small Towns and Big Cities: How Museums Foster Community On-line*. Toronto: Archives and Museum Informatics. Available at: http://www.archimuse.com/mw2010/papers/allen-greil/allen-greil.html (accessed 21 February 2012).

Ball A (2010) *Web Archiving (Version 1.1)*. Edinburgh, UK: Digital Curation Centre. Available at: https://www.era.lib.ed.ac.uk/bitstream/handle/1842/3327/Ball%20sarwa-v1.1.pdf;jsessionid=8D816FBF224D24F295D15525E80CB97-A?sequence=1 (accessed 26 February 2015).

Berthomière W (2012) *''A French What?'' A la Recherche d'une Diaspora Française: Premiers Eléments D'enquête au Sein de L'espace Internet*. Working Paper. Paris: Fondation Maison des Sciences de l'Homme. Available at: http://www.e-diasporas.fr/working-papers/Berthomiere-FrenchExpatriates-FR.pdf (accessed 24 March 2015).

Bezemer J, Cope A, Kress G, et al. (2013) Holding the scalpel: Achieving surgical care in a learning environment. *Journal of Contemporary Ethnography* 20(10): 1–26.

Block D (2006) *Multilingual Identities in a Global City: London Stories*. Basingstoke and New York: Palgrave Macmillan.

Bourdieu P (1965) *Un Art Moyen. Essai sur les Usages Sociaux de la Photographie*. Paris: Minuit.

Bourdieu P (1972[2000]) *Esquisse d'une Théorie de la Pratique*. Geneva: Librairie Droz; Paris: Le Seuil.

Bourdieu P (1980) *La Noblesse de l'Etat. Grandes Ecoles et Esprit de Corps*. Paris: Minuit.

Bourdieu P (1982) *Langage et Pouvoir Symbolique*. Paris: Fayard.

Bourdieu P (1994) *Raisons Pratiques – Sur la Théorie de L'action*. Paris: Seuil.

Bourdieu P and Passeron J-C (1964) *Les Héritiers. Les Etudiants et la Culture*. Paris: Minuit.

Bourdieu P and Wacquant L (1992) *Réponses: Pour une Anthropologie Reflexive*. Paris: Seuil.

Bray P and Donahue R (2010) *Common Ground: A Community-Curated Meetup Case Study*. Toronto: Archives and Museum Informatics. Available at: http://www.archimuse.com/mw2010/papers/bray/bray.html (accessed 21 February 2012).

Bromage S (2010) *Benedict Arnold Slept Here: New Life for Local History On-line and in the Community*. Toronto: Archives and Museum Informatics. Available at: http://www.archimuse.com/mw2010/papers/bromage/bromage.html (accessed 12 December 2013).

Brown A (2006) *Archiving Websites: A Practical Guide for Information Management Professionals*. London: Facet.

Brügger N (2005) *Archiving Websites. General Considerations and Strategies*. Aarhus: Centre for Internet Research. Available at: http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf (accessed 19 March 2015).

Brügger N (2014) The web of hyperlinks – Challenges for a historical approach. In: *Slides presented at BUDDAH research meeting*, IHR, London, 17 September 2014 (unpublished).

Casilli AA (2010) *Les Liaisons Numériques. Vers une Nouvelle Sociabilité?*. Paris: Seuil.

Crawford K (2013) The hidden biases in big data. *Harvard Business Review*. Available at: http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/ (accessed 9 March 2015).

Dallas C (2007) *An Agency-Oriented Approach to Digital Curation Theory and Practice*. Toronto: Archives and Museum Informatics, pp. 49–72. Available at: http://www.archimuse.com/ichim07/papers/dallas/dallas.html (accessed 21 February 2012).

Dalton JB (2010) *Can Structured Metadata Play Nice With Tagging Systems? Parsing New Meanings From Classification-Based Descriptions on Flickr Commons*. Toronto: Archives and Museum Informatics. Available at: http://www.archimuse.com/mw2010/papers/dalton/dalton.html (accessed 21 February 2012).

Dawson B (2010) *Think Globally, Digitize Locally: Charting an Institution's Course Toward the Digital Social Good*. Toronto: Archives and Museum Informatics. Available at: http://www.archimuse.com/mw2010/papers/dawson/dawson.html (accessed 21 February 2012).

Day M (2006) The long-term preservation of web content. In: Masanès J (ed.) *Web Archiving*. Berlin, Heidelberg: Springer-Verlag, pp. 177–199.

Dicker E (2010) *The Impact of Blogs and Other Social Media on the Life of the Curator*. Toronto: Archives and Museum Informatics, pp. 2–3. Available at: http://www.archimuse.com/mw2010/papers/dicker/dicker.html (accessed 21 February 2012).

Dicks B, Soyinka B and Coffey A (2006) Multimodal ethnography. *Qualitative Research* 6(1): 77–96.

Digital Preservation Coalition (n. d.) UK Web Archiving Consortium (UKWAC). Available at: http://www.dpconline.org/advice/web-archiving (accessed 6 March 2015).

Domingo M, Jewitt C and Kress G (forthcoming) Multimodal social semiotics: Writing in online contexts. In: Pahl K and Rowsell J (eds) *The Routledge Handbook of Contemporary Literary Studies*. London: Routledge.

Favell A (2008) *Eurostars and Eurocities: Free Movement and Mobility in an Integrating Europe*. Oxford: Blackwell.

Flouris G and Meghini C (2007) Some preliminary ideas towards a theory of digital preservation. In: *First international workshop on digital libraries foundations*, Vancouver, British Columbia, 23 June 2007. Available at: http://www.ics.forth.gr/isl/publications/paperlink/DLF107.pdf (accessed 23 March 2015).

Gill T, Gilliland AJ and Woodley MS (2005) Metadata standards crosswalks. In: Baca M (ed.) *Introduction to Metadata: Pathways to Digital Information*. Los Angeles, CA: The J. Paul Getty Trust. Available at: http://www.getty.edu/research/conducting_research/standards/intrometadata/crosswalks.html (accessed 4 March 2015).

Gomes D and Costa M (2014) The importance of web archiving for the humanities. *International Journal of Humanities and Arts Computing* 8.1(2014): 106–123.

Grenfell M (ed.) (2012) *Pierre Bourdieu – Key Concepts (Second Edition)*. Durham: Acumen.

Hank C (2013) Dispatches from Blog Purgatory. In: *Slides presented at CurateGear 2013: Enabling the curation of digital collections*, Chapel Hill, NC, 9 January 2013. Available at: http://ils.unc.edu/digccurr/curategear2013-talks/hank-curategear2013.pdf (accessed 28 March 2015).

Hillis K, Petit M and Jarrett K (2013) *Google and the Culture of Search*. London: Routledge.

Hockx-Yu H and Knight G (2008) What to preserve? Significant properties of digital objects. *The International Journal of Digital Curation* 3(1): 141–153. Available at: http://www.ijdc.net/index.php/ijdc/article/view/70 (accessed 23 March 2015).

Hockx-Yu H (2012) UK Web Archive in the eyes of scholars. In the British Library UK Web Archive Blog. Available at http://britishlibrary.typepad.co.uk/webarchive/2012/07/uk-web-archive-in-the-eyes-of-scholars.html (accessed 7 July 2015).

Holley R (2010) Crowdsourcing: How should librarians do it? *D-Lib Magazine*, March/April 2010, 16(3/4). Available at: http://www.dlib.org/dlib/march10/holley/03holley.html (accessed 18 October 2014).

Huc-Hepher S and Drake H (2013) From the 16ème to South Ken? A study of the contemporary French population in London. In: Kelly D and Cornick M (eds) *A History of the French in London: Liberty, Equality, Opportunity*. London: Institute of Historical Research, pp. 391–429.

Jacobsen G (2008) Web archiving: Issues and problems in collection building and access. *LIBER Quarterly (S.I.)* 18(3): 366–376. Available at: http://liber.library.uu.nl/index.php/lq/article/view/7936/82020 (accessed 6 March 2015).

Jacobson P (2014) Inside the struggle to preserve the world's data. *Newsweek*, 2 February 2014. Available at: http://www.newsweek.com/2014/07/11/inside-struggle-preserve-worlds-data-257020.html (accessed 19 March 2015).

Jenkins R (1992) *Pierre Bourdieu (Revised Edition)*. London: Routledge.

Jewitt C (ed.) (2011) *Routledge Handbook of Multimodal Analysis (Second Edition)*. London: Routledge.

Kelly D and Martyn C (eds) (2013) *A History of the French in London: Liberty, Equality, Opportunity*. London: Institute of Historical Research.

Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage.

Kress G (2010) *Multimodality. A Social Semiotic Approach to Contemporary Communication*. London: Routledge.

Lepore J (2015) The Cobweb: Can the Internet be archived? *The New Yorker*, 26 January 2015. Available at: http://www.newyorker.com/magazine/2015/01/26/cobweb (accessed 19 March 2015).

Lotman YM (1990) In: Shukman A (ed.), *Universe of the Mind: A Semiotic Theory of Culture*. London and New York: I. B. Tauris & Co Ltd.

Lyman P (2002) *Archiving the World Wide Web. Building a National Strategy for Preservation: Issues in Digital Media Report*. Washington, DC: Council on Library and Information Resources. Available at: http://www.clir.org/pubs/reports/pub106/web.html (accessed 23 March 2015).

Masanès J (ed.) (2006) *Web Archiving*. Berlin, Heidelberg: Springer-Verlag.

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Meyer E (2011) Web archiving: The state of the art and the future. In: Brack M (ed.) *The Future of the Past of the Web Report*. London: KCL. Available at: www.dpconline.org/component/docman/doc_download/662-oct2011fpwmeyer (accessed 12 December 2013).

Miller E and Wood D (2010) *Recollection: Building Communities for Distributed Curation and Data Sharing*. Toronto: Archives and Museum Informatics. Available at: http://www.archimuse.com/mw2010/papers/miller/miller.html (accessed 12 December 2013).

Moore R (2008) Towards a theory of digital preservation. *The International Journal of Digital Curation* 3(1): 63–75 (Available at: file:///C:/Documents%20and%20Settings/SAS/My%20Documents/Downloads/42-167-1-PB.pdf (accessed 2 April 2015).

Murthy D (2008) Digital ethnography: An examination of the use of new technologies for social research. *Sociology* 42: 837.

Pennock M (2007) Digital curation: A life-cycle approach to managing and preserving usable digital information. *Library and Archives* 1(1) Available at: http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf (accessed 19 March 2015).

Pennock M (2011) *The UK web archive and scholarly research*. Internal Report, British Library, London.

Pennock M (2013) *Web-Archiving – DPC Technology Watch Report 13-01 March 2013*. York, UK: Digital Preservation Coalition. Available at: file:///C:/Documents%20and%20Settings/SAS/My%20Documents/Downloads/dpctw13-01.pdf (accessed 26 February 2015).

Peters W (2011) The Arcomem project: Intelligent digital curation and preservation for community memories. In: Brack M (ed.) *The Future of the Past of the Web*

Report. London: KCL. Available at: http://www.ariadne.ac.uk/issue68/fpw11-rpt (accessed 23 March 2015).

Rowley S, Schaepe D, Sparrow L, et al. (2010) *Building an On-Line Research Community: The Reciprocal Research Network*. Toronto: Archives and Museums Informatics. Available at: http://www.archimuse.com/mw2010/papers/rowley/rowley.html (accessed 21 February 2012).

Ryan L, Mulholland J and Agoston A (2014) Talking ties: Reflecting on network visualisation and qualitative interviewing. *Sociological Research Online* 19(2) Available at: http://www.socresonline.org.uk/19/2/16.html (accessed 23 March 2015).

Spaniol M, Mazeika A, Denev D, et al. (2009) "Catch Me if You Can": Visual analysis of coherence defects in web archiving. In: *9th international web archiving workshop proceedings*, Corfu, Greece, 31 September 2009–1 October 2009. Available at: http://liwa-project.eu/images/publications/IWAW_09_Visual_Analysis.pdf (accessed 19 March 2015).

Strodl S, Petrov P and Rauber A (2011) *Research on Digital Preservation Within Projects Co-funded by the European Union in the ICT Programme*. Brussels: European Commission. Available at: http://www.ifs.tuwien.ac.at/~strodl/paper/Report%20-%20Research%20on%20Digital%20Preservation.pdf (accessed 23 March 2015).

Taylor D (2012) Save as (On the Subject of Archives). *E-misférica*, Summer, 9(1-2). Available at: http://hemisphericinstitute.org/hemi/en/e-misferica-91/taylor (accessed 6 March 2015).

Technorati (2010) State of the blogosphere. Available at: http://technorati.com/state-of-the-blogosphere-2010/ (accessed 26 February 2015).

Toyoda M and Kitsuregawa M (2012) The history of web archiving. *Proceedings of the IEEE* 100: 1141–1143. Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6182575 (accessed 19 March 2015).

Vannini P (2007) Social semiotics and fieldwork: Method and analytics. *Qualitative Inquiry* 13(1): 113–140.

Volk R (2012) Digital preservation: What I wish I knew before I started. Report for UCL, London, 2012. Available at: http://www.dpconline.org/advice/guides (accessed 18 November 2013).

Wang T (2013) Big data needs thick data. *Ethnography Matters*. Available at: http://ethnographymatters.net/2013/05/13/big-data-needs-thick-data/#more-4782 (accessed 23 March 2015).

Warwick Workshop Report (2005) Digital curation and preservation: Defining the research agenda for the next decade. Warwick Workshop Report, 7 and 8 November. Available at: http://www.dcc.ac.uk/sites/default/files/documents/Warwick_Workshop_report.pdf (accessed 7 July 2015).

Yoon A (2013) Defining what matters when preserving web-based personal digital collections: Listening to bloggers. *The International Journal of Digital Curation* 8(1): 173–192. Available at: http://www.ijdc.net/index.php/ijdc/article/view/8.1.173 (accessed 9 March 2015).