Source Criticism of Data Platform Logics on the Internet

Koch, Gertraud; Kinder-Kurlanda, Katharina

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Source Criticism of Data Platform Logics on the Internet

Gertraud Koch & Katharina Kinder-Kurlanda *

Abstract: *»Quellenkritik für Big Data Plattformen«*. Source criticism is an epistemological practice in social and cultural studies that is crucial for specifying the range and scope of the findings, or in other words their validity and reliability. In the context of big data, source criticism is not yet established in the fashion as it is known in other areas of social and cultural research. Currently emerging discussions in historical research emphasize the relevance of source criticism of digital objects or data. In the context of these discussions, this contribution suggests exploring the potentials of source criticism for platform logics. We focus on big data sourced from the internet. Nevertheless our results aim to be transferrable to other sources of big data. The inclusion of source criticism into big data analysis may in turn foster the integration of data-driven analyses into social and cultural studies research approaches. For an integration of source criticism, the paper proposes source critical analyses of information systems, in particular internet platforms, in big data analysis with regard to a) types of big data platforms, b) researchers as data makers, and c) mixed realities of platform usage practices. In analogy to source repertoires (*Quellentypen*) it suggests to classify internet platforms as providers of particular types of big data sources depending on their infrastructural materiality and ontologies for tracing the key issues of (external) source criticism: provenance, authenticity, and integrity.

Keywords: Epistemology, methodology, types of big data, data makers, critical data studies, source criticism, big data, internet platforms.

1. Introduction

Current discussions of big data highlight both potentials and pitfalls of analyses based in these data. The debate emphasizes a need for further specification of the concrete uses of big data analysis for specific areas of social research as well as for further epistemological reflection and elaboration of big data's

* Gertraud Koch, Institute of European Ethnology/Cultural Anthropology, Universität Hamburg, Edmund-Siemers-Allee 1 (West) 20146 Hamburg, Germany; gertraud.koch@uni-hamburg.de. Katharina Kinder-Kurlanda, GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany; katharina.kinder-kurlanda@gesis.org.

Historical Social Research 45 (2020) 3, 270-287 | published by GESIS
DOI: 10.12759/hsr.45.2020.3.270-287

potentials. They state a need for further contextualization of big data analyses through triangulation with other data sources, small data studies, historical records, or other methodological approaches for making sense of the patterns in big data sets. Beyond context relevance and meaning, integrity and messiness of data are topics discussed with respect to data quality (Boyd and Crawford 2012; Kitchin 2014; Mahrt 2015; Kinder-Kurlanda 2017). Critical voices speak about datafication as a trend within economy, politics, and the academic world. They claim as an initiating factor of the phenomenon of big data the data hunger of these stakeholders, who then study what they themselves co-create (Boellstorff 2013; van Dijck 2014). Still, it remains uncontested that big data comes along with computerization and demarcates a paradigmatic shift for social and cultural research, and thus needs further inquiry into how research epistemologies can adequately cope with the digital data turn by adapting methodological and theoretical approaches (e.g., Manovich 2011; Boellstorff 2013; Weller 2013; Venturini et al. 2018).

While our results are intended to be transferrable to other sources of big data, in this contribution we focus on big data sourced from the internet. Volume is not the only defining characteristic of such data, but rather the term "big data" refers to the possibility of searching, aggregating, and sorting through large data sets to assemble and place them in relation to one another (Boyd and Crawford 2012). Big data promises to allow timely analyses and prognoses as events are almost immediately reflected in social media platforms, and can thus be observed and analysed while they are happening, also allowing for the analysis of even global interconnections (Tinati et al. 2014). Researchers work with user-generated content – mainly social media data, e.g., from Twitter or Facebook –employing new, specifically tailored methods.

Science and Technology Studies (STS) perspectives on data, sociomateriality, and ethnography of infrastructure can help to disentangle the influences of various actors on platform data as a source. They contend that technologies (so, both platforms and data) are shaped by social processes (Bijker and Law 1992) and that there are networks of human and non-human actors that need to be considered in order to understand how discourses become inscribed in big data structures (Greenhalgh and Stones 2010). Such a perspective allows tracing individual properties of a dataset to specific interests in order to be able to assess their meaning as a research source.

Assessing sources critically is a basic epistemological practice, elaborated in historical research over centuries and applied broadly in social and cultural research today.[1] In historical research, source criticism is understood as a radi-

[1] There is a long history of source criticism and a large bandwidth of concepts as well as disciplinary uses of source criticism. A discussion of these various traditions goes beyond the scope of this contribution. In this paper we refer to source criticism as an element of the historical method (see Föhr 2018, 51ff).

cal change of scholarly practice, which emerged in the 19th century and was initiated by the German historian Leopold Ranke.

> Ranke encouraged a new generation of scholars to visit numerous archives, to scrutinize and compare documents, to trace back those who had created them, under which circumstances, at what moment, with what means, and for what purpose. This 'external' source criticism focuses on the creation, appearance and authenticity of a source. (Scagliola 2016)

It provides a specific focus for the next analytical step, namely the internal source criticism, which investigates the contribution and the statement of the source to the research topic (Shafer 1974).

In his investigation of historical source criticism in the digital age, Pascal Föhr (2018) understands source criticism within the historical method as the most crucial element of historical research. At the heart of source criticism are the provenance,[2] the authenticity, and the integrity of the source. Föhr comes to the result that, in general, the processes of the historical method are still relevant in the digital realm but need to be complemented by computational approaches. Moreover he refers to digital objects' embeddedness in particular information environments as a crucial factor for source criticism and the need to emphasize this dimension.[3] These various information systems provide different contexts for the creation of big data, so that source criticism requires different considerations across platforms. The understanding of the conditions of data creation and the pitfalls for data quality due to platform logics are not only a necessary complement for the epistemologies of big data analyses but a question of quality of the outcome.

2. Definitions and Criteria for Source Criticism on Internet Platforms

Big data has become a buzzword and lacks a clear definition. In this contribution we follow the understanding that, in addition to volume, big data is further characterized by velocity (as it is created in or near real time), variety (various formats, both structured and unstructured), and its capacity for linking to other data sets thus promising overall, timely analyses of global connections (e.g., Boyd and Crawford 2012; Kitchin 2014; Mayer-Schönberger and Cukier 2014; Tinati et al. 2014).

[2] While the terms "provenience" and "provenance" are often used synonymously, "provenience" usually stresses an artefact's place of origin (similar to a place of birth) and "provenance" the journey of an artefact since its origin (similar to a curriculum vitae; Price and Burton 2011).

[3] Big data analyses are characterized as future work fields of historians by Föhr but are not discussed explicitly in respect to source criticism.

An internet platform as provider of big data is a term that also may address a large variety of different web applications with varying functions and infrastructures in the back end. Most commonly, social media platforms are addressed in social research with Twitter being the most favored one due to its accessibility of big data. Still a number of other big data sources exist, which provide unique options for social research, such as self-tracking or internet of things (IoT) platforms. Although they are not yet so much in the focus as a source because of research practicalities (access, retrieval, etc.), these sources need to be considered as well in research. Given that each of these internet based platforms is particular in respect to data creation, a consideration of this variety allows learning about their differences and commonalities by contrasting them with each other. We thus consider a broader spectrum of internet-based platforms as relevant for source criticism: beyond social media platforms we include IoT, self-tracking, citizen science, social media metrics, data of public administrations, and science repositories as big data providers for social research.

The question now is what provenance, authenticity, and completeness mean for big data on internet platforms and how this can be assessed.

Provenance refers to the origin of a source and thus the question of who has created it with what intention, in which institutional and socio-cultural context. This information about a source is relevant for the meaning given to the source within particular institutional or everyday life contexts and thus is needed for the interpretation and understanding of the content of the source in research. From an archival perspective important principles are to arrange and describe the sources materials to their original purpose and function, the evaluation of the creator's social role and power, and what is known about ownership and uses of the source (Society of American Archivists 2019).

In her reflection about the value of archival perspectives in the digital environment, Anne J. Gilliland-Swetlands (2000) lists a number of issues for the transfer of archival principles to the digital environment. Even though these issues refer to archival tasks they also reflect issues of source criticism in digital environments for addressing the digital nature of materials. For a source critical reading of big data from the internet, we gain a list of items relevant for the further specification of what digital source criticism should comprise in respect to provenance:

- "life cycle control of high-volume, dynamic multimedia collections of born-digital and digitized materials, from creation through final disposition";
- "identification and preservation of the evidential value of digital materials through design, description, preservation, and evaluation of information systems";
- "exploitation of context and hierarchy in the design and use of digital materials";

- "elucidation of the nature, genesis, and use of digital materials by their creators" (Gilliland-Swetlands 2000, vi).

Another crucial challenge is already hinted at by Gilliland-Swetland when she mentions that collections are not only high-volume but also *dynamic*. When applying source criticism to digital objects this challenge arises out of the objects' volatility and concerns the question of *authenticity*. Copy-paste allows for endlessly cloning digital objects which also challenges the transmission of the idea of authenticity of sources to these objects. Furthermore for most big data providing internet platforms the manipulation of digital objects and content are not evident. Mostly changes of digital objects are not tracked and documented or – if so – are not accessible for researchers from outside. Only in particular cases we have version controls, e.g., Wikipedia, as an integral element of the platform infrastructure. In some fields, which are highly affine to authenticity of sources, new approaches for creating mechanisms to prove authenticity of digital objects are increasingly installed, e.g., digital object identifier (DOI), version management, and researcher's identification systems as elements of research repositories or platforms. Furthermore, the question of authenticity touches upon the interlinkage of online and offline sources, thus the "identification and exploitation of the interdependencies among digital materials, related nondigital materials, and their metadata" (Gilliland-Swetlands 2000, vi), in particular when digital copies are provided as another format of representation of non-digital objects and sources.

The history of the provenance of a source – the usual way of coping with questions of authenticity in source criticism – thus cannot be applied to digital objects directly. The check of authenticity of a digital object is a field of experimentation; it needs new approaches and enhancement through computational methods (Föhr 2018, 72ff, 186ff). An approach, suggested by Föhr (2018), is "authenticity approximation," which is a subjective evaluation of the single parameters of source criticism by the researcher combined with a weighting factor for the relevance of a parameter for this source. This method is promising because it considers the particularity of digital media (Föhr 2018, 250ff). Parameters for assessing the authenticity of a source are integrity (bitcode, IT-related information), persistence (where published, long-term accessibility, etc.), dating (date of publishing, changes, etc.), authorship (identifiability, intention), addressee (recognition, relation to author), content (indicators for authenticity), and relations (references, indications, hyperlinks). However, addressing all these parameters demands a source critical reading of each digital object on a platform by a researcher and thus is not feasible for automated big data analyses.

A crucial dimension of source criticism is the authenticity of data, which is maybe best translated into information science by the concept of data integrity, i.e., the maintenance of, the assurance of the accuracy, and the consistency of data over its life-cycle. In the discussion about data integrity of big data, from

the perspective of a library professional Carl Lagoze brings up a discussion on information stewardship and the fracturing of the control zone of information quality.

> The notions of selection, intermediation, bibliographic description, and fixity that are core principles of the library meme stand at odds to the web information meme. These contradictions become sharper as the web has moved over the past decade into the web 2.0 era and beyond. (Lagoze 2014, 6-7)

The requirements of open knowledge and collaborative knowledge production challenge traditional ideas of information and data quality. Epistemologies and methodologies of data production are blurred through the participation of lays, amateurs and users in data creation and thus contests how to still do credible science (Lagoze 2014). The high affinity to copy-paste and mash-up culture of digital media (Bleicher 2017; Schönholz 2017) raises the question of what this means for empirical social and cultural analyses and how this particular quality of digital objects can be considered in source criticism. So far, we do not know about reflections on this aspect for big data analyses even though it can be an important dimension for example as re-tweeted and/or modified citations on Twitter or other social media platforms. Moreover, for digital objects this question is also a technical one. In computer science the concept of integrity refers to data and the accuracy of their recording and retrieval, i.e., the prevention of unintended changes when data are stored or accessed. Differently from the interpretation of analogue sources, big data analyses demand the preparation of data for automated analyses. The source has thus undergone a modification by the researcher before it will be analysed, which is necessary because of the operation mode of tools for structural analyses. The integrity of a source thus has a different meaning in the digital realm and in automated big data analyses. Understanding in which ways data preparation creates changes, how this affects the content and what this means for the interpretation of the source thus are further relevant dimensions of source criticism in big data analyses.

In the following we investigate how data are made and thus become a matter of source criticism by different types of platforms, researchers who employ such data as research data and platform users who both generate the data and about whom the data is supposed to reveal research insights. We do this by sorting out how platform providers, infrastructures and interfaces, researchers through the use of tools and their lens as researchers, and users as communities of practice all act as data makers, with different capacities for data-source creation.

3. Source Criticism I: Platforms as Big Data Makers

Creating data about social and cultural phenomena is always a reduction of social reality, methodologically guided by the intention of condensing reality.

This is also the case for the modelling of social reality on internet platforms who have found different ways to facilitate and represent social interaction, resulting in very specific datasets being made available for researchers. Elements of social practice are reduced by being modelled and re-configured in programmed online systems. The ways in which this works are often opaque as "We know nearly nothing about how people approach and build *ad hoc* information systems to understand their own audiences" (Baym 2013, emphasis in original text). While the scientific reduction of data is affine to epistemological critique this engineered reduction of social reality in computer systems and on internet platforms is mostly not reflected in this respect but assessed with regard to its functionality for particular tasks. Scholarly discussions in critical data studies highlight the need for criticism of these algorithmically created realities as a source for research data. Topics of critique are 1) the role of platforms as data "owners" in making data, 2) their often highly structured and structuring models for social interaction (such as, e.g., hashtags or likes), and 3) the interfaces and tools for data sourcing that are provided to researchers, such as programming interfaces (APIs) for sourcing data.

3.1 The Role of Platforms as Data "Owners" in Making Data

The political economy of platforms has already been analysed in detail (e.g., van Dijck 2014) and while we acknowledge that powerful actors such as large internet firms aim to manipulate and monetize user behaviour on a large scale (e.g., Willson and Leaver 2015), we will not go into detail here on the complex ecosystems of exploitation and appropriation that are created.[4] Rather, we are interested in the effects of this setting on internet data as a research source.

There are various ways in which researchers may gain access to internet data, involving various types of collaboration (e.g., public private partnerships, as an "embedded researcher" or as a regular user bound to the platforms' terms of service) with platform providers (see Breuer et al., forthcoming). As data is "owned" by internet platforms these platforms decide how much of it has to be given to researchers. All of the possible collaborative relationships that researchers may enter into in order to gain access to big data from the internet give the internet platforms a critical role in shaping the data, e.g., by selecting subsets or by only providing samples to the researchers. This way of determining the data's completeness is distinct from the way in which platform interfaces and their structures (covered below) shape what internet data looks like. In addition to the data being the result of a specific platform as a context the (problematic, restricted, often costly, unequal) way in which it is provided fundamentally determines a dataset's integrity. However, in the following we are looking at the technical frames and capacities of a programmed interface

[4] See also Diaz-Bone and Horvath 2020, in this HSR Forum.

which only allows for certain interactions or affordances (Bucher and Helmond 2017).

3.2 Platforms as Models for Interaction

Big data diverge in a much wider range than from structured, and semi-structured to un-structured types. Data are made within complex contexts of knowledge production. The contexts of data making shape the nature of data, thus not all big data are alike but vary in respect to the processes in which they are made (Boellstorff 2013; Gitelman 2013; Ribes and Jackson 2013). This insight of critical data studies is in line with the epistemological questions of source criticism of creation, appearance, and authenticity as a most crucial starting point for further analyses. Although source criticism usually studies single objects, big data analyses refer to a multitude of such single pieces, for which as a totality a source critical reading hardly can be done. The reflections of Föhr suggest that it is still possible and fruitful to apply source criticism as an epistemological practice to information systems as a whole. He distinguishes the information technological *source* from the single object as a *re-source* for analytics in humanities, using the example of communication with computer scientists (Föhr 2018, 74). This rhetoric figure can be made useful for source criticism in big data analyses because it refers to the circumstance that big data analysis depends on the provision of data on particular digital information systems (i.e., usually internet platforms) and the relevance of these grounds for the single data objects themselves. The particularity of each information system sets conditions for the creation, appearance, and authenticity of data[5], which also has consequences for the integrity of data on these platforms.

Depending on the information system, the modes of how data are made, manipulated, and shared – or, to put it differently, from a computer science perspective: how data are stored, processed, and retrieved – varies widely. It is obvious that these modes of data production matter for the epistemic quality of data on these platforms. The particular quality and integrity of big data is beyond a series of other factors also relational to the platform and differs from information system to information system.

All platforms have in common, however, that change is inherent (Karpf 2012). Changes to a platform can be implemented that may have substantial effects on data at any point in time and disrupt a research setup. The researcher has no influence over whether or how her research setup changes. One could

[5] Even though authenticity is a critical concept, even more in the context of digital information systems with their high affinity to cloning objects by copy-paste actions, the notion of authenticity can be taken up here in the sense of the originality of an object in opposite to objects faked by bots, trolls, and other modes of manipulating communication on internet platforms / information systems.

say that she is using a research setup or "infrastructure" created by the platform with completely different and opaque interests over which she has no control.

A source critique of information systems in a STS fashion of how data are made will contribute to a better understanding of how social reality is drafted and which modes of knowing social reality are fostered on this specific information system. Moving on from there the understanding of how big data on particular information systems are created may be complemented by further methodological and theoretical sources.[6]

Still, the epistemic quality of an information system and the data provided here cannot be defined independently from the research question and thus needs an assessment for big data analyses, so beyond generic features of source criticism of information systems (including the change of their feature over time) more specific analyses are required considering the specific research context.

3.3 APIs and Other Interfaces for Data Sourcing

As mentioned above, there are various ways in which researchers may gain access to internet data. Data may have been obtained from an API provided specifically for researchers by the platform, it may have been bought as a "complete" dataset or it may have been "scraped" from the public website. Both visible data (such as followers or likes) and – at least at the surface – non-visible data created from digital traces is accessible for scraping by accessing the backend of websites. Due to the limitations of the programming interfaces (APIs) offered to researchers for gathering data, it is often unclear how complete a given data set is, or which sampling methods the platform operator has already applied (Driscoll and Walker 2014; Felt 2016).

4. Source Criticism II: Big Data Researchers as Data Makers

Guided by specific research interests, researchers are seeking to reduce social reality into analysable datasets. For social media data this means that they make decisions about when (which timeframe?) and where (which platform[s]?) to collect what data (which metadata, e.g., of a text posted online?), thus taking a large role in "making" data. The analysis of collected social media data is also always influenced by philosophical or epistemological assumptions as "[...] a researcher's subjective judgments can become deeply infused

[6] Moreover, from such a source critical analysis of information systems it is likely to collect some knowledge about sources of messiness of data specific for this platform.

into a data set through sampling, data cleaning, and creative manipulations such as data masking" (Ekbia 2015, 15). Arguing against a "data gold rush," Mylynn Felt (2016) suggests the combination of critical social-media analytics with traditional, qualitative methods for creating thick descriptions through multi-methods approaches (Felt 2016). For various reasons, researchers are, however, often unable to work based on the epistemological foundations demanded also, e.g., by Critical Data Studies, namely situational, reflective, interdisciplinary, theory-driven, explorative, and utilizing computer-intensive methods (e.g., Ruppert et al. 2013; Kitchin 2014; Schroeder 2014). Finding methods that are adapted to the changing field is difficult at the level of big data analysis, as it requires particularity and attention to detail rather than comprehensive, overarching approaches. Most frequently, however, various problems that arose from characteristics of the data other than their size were addressed in a study of social media researchers' everyday practices (Kinder-Kurlanda, forthcoming): the often difficult access to the data, the ethical problems involved in their use, and the sophisticated technical knowledge required. Beyond these practical dimensions it turned out that challenges of interdisciplinary work, current publication formats and practices, and the problems of academic career planning in a new field of research had a particular influence on approaches.

Automated processes and algorithms are "imbued with particular values and contextualized within a particular scientific approach" (Kitchin 2014, 5). Big data prepared by collecting and storing must then first be cleaned, usually again with automated tools and scripts, in order to make analyses possible at all. Decisions are made as to which parts and which properties of the data are considered. These processes of data generation are little documented and often remain opaque (Helles and Jensen 2013), resulting in an imperious opacity of data-driven approaches to science (Ekbia 2015). The tools for gathering and handling big data and the specific characteristics of the data itself advantage certain ways of analysing such data over others, independently of which methodology would be required to answer a specific research question. For example, Tufekci (2014) has criticized the widespread practice of sampling via dependent variables, such as compiling a data set on the basis of Twitter hashtags. Similarly, Busch (2014) points to the problems arising from convenience sampling in big data research and from the simplification of large data sets necessary to allow analysing the big datasets, which favours some aspects to the detriment of others. As we have pointed out above, every data analysis is a reduction. A source critical approach requires documenting these decisions, which here means revealing the seemingly "natural" flows of data through available tools that allow processing them and suggesting specific ways of making certain characteristics visible and encourage specific types of results over others.

5. Source Criticism III: Platform Users as Data Makers

Platforms are made up of different layers, some of them visible to every user, some to those able to access and make use of the html code of a website. However, some layers, such as "the databases and algorithms within each platform that process and translate the activities of users into data structures and interfaces we see when accessing the platform" (Walker 2017, 16), are hidden from view although they constrain what information researchers can access. For example, most platforms quantify user actions and offer the results of analyses as part of the user interface (Grosser 2014). These underlying features and structures of the platforms both enable and constrain communication and expression (Plantin et al. 2016). The platforms thus offer affordances that allow both users and researchers to generate and interact with data held in the (programmable) data structures and algorithms (Bucher and Helmond 2017; Walker 2017).

The ephemerality of platforms and the volatility of platform data shows different aspects for different data from different platforms. For example, for self-tracking data, individual user interactions with the collected, individual data play a much larger role in data's volatility than the interactions of other users with the data. The individual interface and its functionality thus must be taken especially into account when assessing provenance and authenticity of the data; a comprehensive record of interactions would be an essential element for completeness. Therefore, the relation of platform and source critical properties of big data can only be sketched here roughly and should be seen as a short justification for the types of platforms we want to introduce here as crucial for big data platforms. With reference to provenance, authenticity, and integrity as well as further variables such as the obscurity of data creation, data stewardship, or the expertise for data creation, we suggest seven different big data source types: 1. Social media platform data; 2. IoT data; 3. Self-tracking data; 4. Citizen science data; 5. Social media metrics data; 6. Scientific data; and 7. Data provided by public institutions. These big data source types of internet platforms are a first approach and need systematical inquiry in future research and possibly revisions and differentiation based on empirical studies of how society and social media platforms relate.[7]

Social Media Platform Data

Data shared, communicated about, and socialized within the technical frames provided by social media platforms are highly problematic for a source critical

[7] A great example of such research is the book "Twitter and Society," edited by a team of researchers, which gives an impression of what information is needed for building a typology of internet platforms as big data sources (Mahrt et al. 2013).

reading and thus also in respect to high-quality social research. Since a lot of research focusses on these platforms, relatively more information for source criticism is available here. This gives an indication of the relevant objectives and questions to be asked of internet platforms to come to a source critical reading of big data, thus introduced more in depth here than for the other types of big data platforms.

Sherry Turkle (1997) already pointed out the complex role played by performance on the internet: Users and avatars may differ as various types of performance come into play. The data thus does not necessarily reveal the "fantasies, intentions, motives, opinions and thoughts of people" (Manovich 2011). Rather, social media data are interfaces between people and the world that show only some aspects of their real lives and fantasies and also contain data that is supposed to produce a certain image (ibid.). Big data can thus only show to a limited extent that human systems are complex and paradoxical and that people behave in unforeseen ways (Kitchin 2014). By analysing social media big data, only the online behaviour of specific individuals using a particular platform is observable and also only becomes visible to a certain extent due to the various problems of access and data quality mentioned above. The data is comparable, for example, with the content of written letters plus some additional data on their senders and recipients (Schröder 2014), but not with a survey of users about their motivations and intentions. Only the linking of data about online behaviour with offline demographics and the like enables a more complete picture of the user (ibid.).

The example of social media data shows that the authenticity and completeness requirements for these data pertain to content, surrounding metadata, and linked context and are challenged by the ephemerality of platforms and the volatility of platform data. For social media data this means that a source critical reading may hinge on an exact recording of not only the posted text itself but also of information about user interactions at a specific point in time or about surrounding metadata (such as a user profile) and linked data (such as external websites linked to in a post). However, it is currently neither established how such a recording is to be performed and what information exactly should be recorded. More crucially, it is not clear how to deal with the volatility of content per se. Depending on the research question it may not suffice to look at a digital object only at one point in time.

This means that beyond the fact that datasets are the result of a changeable platform and infrastructure (changes which neither users nor researchers can control), data itself is subject to continuous change, or to put it differently, inextricably characterised by its volatility. For the example of social media data the information scientist Shawn Walker (2017) has shown that latency plays an important role. Posts, their accompanying metadata, and other content linked to in such posts (such as videos, images, and web pages) are all linked to a specific point in time: "Viewing these items disconnected from that bound moment in

time may result in viewing a different post and content than the user intended" (Walker 2017, 48). Social media posts, surrounding metadata, and links embedded into them change over time. For example, a user looking at a tweet once and then again a year later may find that the profile picture of the account user has changed, that there are more or less "shares," "likes," and replies, that embedded links no longer point to the same external website, or even that the tweet itself has been deleted. Platform changes, changes in platforms that the post connects to, or user interactions by the original tweeter or by others may all have changed. Authenticity of an internet platform dataset thus requires defining a specific point in time at which a dataset is captured and to define which characteristics (such as metadata, linked data, surrounding context of a user profile, etc.) need to be captured. The criteria for this decision depend on what qualifies as "complete" to satisfy the requirements of a specific research question.

Internet of Things Data

In the internet of things (IoT) data emerge from machine-to-machine communication, for example in settings of ubiquitous computing or smart city approaches. While the provenance of data can be clearly connected to a particular machine or object, authenticity and completeness need to be specified with regard to, first, technical errors which may occur in machine-to-machine communication and algorithms for processing these data and, second, by specifying at which point to draw a line around a specific network of machines communicating with each other.

Self-Tracking Data

Self-tracking data may be collected automatically on internet platforms without much activity of the user, for example when data are sent and collected automatically from a smart phone app or measuring point while people are jogging. In other approaches data collection depends more on user activity, e.g., the manual input of personal data into apps. Furthermore both modes can be combined. The degree of automatization thus may vary and with it also the options of how data collection can be manipulated, accordingly provenance, authenticity, and completeness of data will be different and need to be considered in source critical reading.

Citizen Science Data

Citizen science data usually combines individual interactions with social interactions, sometimes introducing scoring or gaming elements to motivate via competition, with user reflection of collected data (similar to self-tracking but looking at collectively generated data). Again provenance, authenticity, and

completeness depend on the modes of how these data are gathered, which vary widely in a range from IoT approaches on the one side (e.g., citizen fine dust measuring networks initiated by the Open Knowledge Foundation) and individually collected, qualitative data in smart phone apps on the other side (e.g., on mental health issues).

Social Media Metrics/ Audience Data

In a critical approach, internet researcher Nancy Baym (2013) addresses audience data providers, such as Tweetdeck, Facebook demographics, Google Analytics, and Top Spin. Metadata on audience behaviour are data which are not accessible to everybody on the internet, but only to platform providers from their backend. Baym assesses the quality of social media metrics using the example of music media metrics and highlights the process of making audiences by definition and how visible media metrics are applied to as measures with particular fallibility and ambiguity. She thus highlights the value systems which shape how data are collected, stored, and analysed, and how they are interpreted from an economic point of view. Beyond the making of big providers all internet platform metrics are biased through the lens of the observer. Moreover, fallibility of social media metrics needs to be mentioned and is also listed by Baym: a) skewed by algorithms that manipulate rankings and other representations; b) non-representativeness of populations in social media; c) deception of bots and through purchased actions (bought pay views); and d) ambiguous meaning (Baym 2013).

Scientific Data

Repositories provide access to scientific data and also relevant metadata for source criticism. These data platforms are made for the requirements of source criticism according to scholarly quality standards by offering diverse processes for the management of data, such as digital object identifiers (DOI), identifiers for researcher, versioning of data sets, etc.

Data Provided by Public Institutions

Many institutions, such as administrations, public bodies, and governmental organizations also provide big data in a growing variety, such as the data and metadata of cultural objects provided by memory institutions and aggregated by Europeana. Due to the official background of these institutions and the big data stored here, source criticism may follow established paths as known from the non-digital world. Even though not entirely unproblematic – for example, data are often distributed, they may be offered on different levels of aggregation or in not-compatible formats – it should be possible for researchers to qualify the fallibility of these data sets in a reliable way.

6. Conclusion: Platform Epistemologies

Analogous to diverse types of historical sources such as court documents with particular properties, we can distinguish types of big data, which are emerging from the diverse modes of how data are made on internet platforms by the providers, data researchers applying tools (and the tools themselves), and users of the platforms. In the triangle of these "data makers" each platform provides a specific mode of data creation that affects provenance, authenticity, and integrity of the data collected for social research. So far big data have not been considered in this perspective, and research literature does not distinguish big data types or reflect properties of big data systematically in the sense of source criticism; such considerations are more or less occasionally integrated into methodology chapters.

It then becomes necessary to find new criteria, concepts, and tools to assess provenance, authenticity, and completeness of big data. Most prominently, a reconfiguration of these concepts is required that allows the changeability of big data and its platforms over time. For example, tools are needed that facilitate a documentation of functional changes on platforms. In addition, issues of boundedness need to be addressed. Platforms, data, and algorithms are more and more connected, requiring a) a source criticism that goes across multiple platforms and b) to create criteria for making decisions about completeness, which will usually require to "cut off" data collection at a specific point as there is seemingly endless interconnection between platform systems. Finally, digital communication requires consideration with its specificities of crossing online and offline, of allowing for mashup culture, and for making copy-paste the norm, thus challenging notions of authenticity that are rooted in uncomplicated ideas of origin. The required, fundamental rethinking of criteria, concepts, and tools for provenance, authenticity, and completeness of big data may eventually allow reassessing the ones used for assessing these qualities for more traditional data sources, making visible hidden assumptions about the nature of text and language.

References

Baym, Nancy K. 2013. Data not seen. The uses and shortcomings of social media metrics. In *First Monday* 18 (10). Chicago: University of Illinois at Chicago Press.

Bijker, Wiebe E. and John Law. 1992. Shaping Technology/Building Society. In *Studies in Sociotechnical Change*. Cambridge, Massachusetts: The MIT Press.

Bleicher, Joan Kristin. 2017. The manifestation of mash-up categories. In *Digitisation. Theories and Concepts for Empirical Cultural Research*, ed. Gertraud Koch, 132-158. New York/London: Routledge.

Boellstorff, Tom. 2013. Making big data, in theory. In *First Monday* 18 (10). Chicago: University of Illinois at Chiago Press.

Boyd, Danah, and Kate Crawford. 2012. Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon. In *Information, Communication & Society* 15 (5): 662–679. London: Taylor & Francis.

Breuer, Johannes., Kinder-Kurlanda, Katharina E., and E. Bishop. Forthcoming. The practical and ethical challenges in acquiring and sharing digital trace data: negotiating public-private partnerships. In *New Media and Society*. Chicago: University of Illinois at Chicago Press.

Bucher, Taina, and Anne Helmond. 2017. The Affordances of Social Media Platforms. In *The SAGE Handbook of Social Media*, ed. Jean Burgess, Thomas Poell, and Alice Marwick, 233-253. Thousand Oaks: SAGE Publications.

Busch, Lawrence. 2014. A Dozen Ways to Get Lost in Translation: Inherent Challenges in Large-Scale Data Sets. In *International Journal of Communication* 8: 1727–1744. Los Angeles: University of Southern California Press.

Diaz-Bone, Rainer, Kenneth Horvath, and Valeska Cappel. 2020. Social Research in Times of Big Data. The Challenges of New Data Worlds and the Need for a Sociology of Social Research. *Historical Social Research* 45 (3): 314-341. doi: 10.12759/hsr.45.2020.3.314-341.

Driscoll, Kevin, and Shawn Walker. 2014. Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. In *International Journal of Communication* 8: 1745–1764. Los Angeles: University of Southern California Press.

Ekbia, Hamid R., Ali Ghazi, Inna Kouper, Michael Mattioli et al. 2015. Big Data, Bigger Dilemmas: A Critical Review. In *Journal of the Association for Information Science and Technology* 66 (8): 1523-1545. Hoboken: John Wiley & Sons.

Felt, Mylynn. 2016. Social media and the social sciences. How researchers employ Big Data analytics. In *Big Data & Society* 3 (1). Thousand Oaks: SAGE Publications.

Föhr, Pascal. 2018. Historische Quellenkritik im Digitalen Zeitalter. Dissertation. Basel: Universität Basel. <https://edoc.unibas.ch/64111/1/F%C3%B6hr_Pascal-Historische_Quellenkritik_im_Digitalen_Zeitalter-2018.pdf> (Accessed May 19, 2020).

Gilliland-Swetland, Anne J. 2000. Enduring paradigm, new opportunities. The value of the archival perspective in the digital environment. Washington, D.C.: Council on Library and Information Resources.

Gitelman, Lisa. 2013. "Raw data" is an oxymoron. Cambridge, Massachusetts: The MIT Press. <http://ieeexplore.ieee.org/servlet/opac?bknumber=6451327> (Accessed May 19, 2020).

Greenhalgh, Trisha, and Rob Stones. 2010. Theorising big IT programmes in healthcare: Strong structuration theory meets actor-network theory. In *Social Science & Medicine* 70 (9): 1285-1294. Amsterdam: Elsevier.

Grosser, Benjamin. 2014. What do metrics want? How quantification prescribes social interaction on Facebook. In *Computational Culture: a journal of software studies* 4: 1-19. London: Goldsmiths, University of London Press.

Helles, Rasmus, and Klaus Bruhn Jensen. 2015. Making data – Big data and beyond. Introduction to the special issue. In *First Monday* 18 (10). Chicago: University of Illinois at Chicago Press.

Karpf, David. 2012. Social science research methods in Internet time. In *Information, Communication & Society* 15 (5): 639-661. London: Taylor & Francis.

Kinder-Kurlanda, Katharina E. 2017. Big Data. In *Digitisation. Theories and Concepts for Empirical Cultural Research*, ed. Gertraud Koch, 158-176. New York/London: Routledge.

Kinder-Kurlanda, Katharina E. Forthcoming. Big Social Media Data als epistemologische Herausforderung für die Soziologie. In *Soziale Welt. Sonderband „Digitale Soziologie - Soziologie des Digitalen?“*. Baden-Baden: Nomos.

Kitchin, Rob. 2014. Big Data, new epistemologies and paradigm shifts. In *Big Data & Society*: 1-12. Thousand Oaks: SAGE Publications.

Lagoze, Carl. 2014. Big Data, data integrity, and the fracturing of the control zone. In *Big Data & Society*: 1-11. Thousand Oaks: SAGE Publications.

Lagoze, Carl, Jean-Christophe Plantin, Paul N. Edwards, and Christian Sandiv. 2016. Infrastructure studies meet platform studies in the age of Google and Facebook. In *New Media & Society* 20 (1): 293-310. Chicago: University of Illinois at Chicago Press.

Mahrt, Merja. 2015. Mit Big Data gegen das "Ende der Theorie"? In *Digitale Methoden in der Kommunikationswissenschaft* (Band 2), ed. Axel Maireder, Julian Ausserhofer, Christina Schumann und Monika Taddicken, 23-37. Berlin: Freie Universität Berlin.

Mahrt, Merja, Cornelius Puschmann, Axel Bruns, Jean Burgess, und Katrin Weller, eds. 2013. *Twitter and Society*. New York: Peter Lang Inc.

Manovich, Lev. 2011. Trending: The Promises and the Challenges of Big Social Data. In *Debates in the Digital Humanities* 2, ed. Matthew Gold, 460-475. Oxford: Oxford University Press.

Mayer-Schönberger, Viktor and Kenneth Cukier. 2014. *Big Data. A revolution that will transform how we live, work and think*. London: John Murray.

Price, T.Douglas and Burton, James H. 2011. Provenience and Provenance. In: *An Introduction to Archaeological Chemistry*, ed. T. Douglas Price and James H Burton, 213-242. New York: Springer.

Ribes, David and Steven J. Jackson. 2013. Data bite man. In *"Raw data" is an oxymoron*, ed. Lisa Gitelman, 147-166. Cambridge, Massachusetts: The MIT Press.

Ruppert, Evelyn, John Law and Mike Savage. 2013. Reassembling social science methods: the challenge of digital devices. In *Theory, Culture & Society* 30 (4): 22-46. Thousand Oaks: SAGE Publications.

Scagliola, Stefania. 2017. *Digital Source Criticism in the 21st century: reconsidering Ranke's principles in the digital age*. Luxembourg Center for Contemporary and Digital History. Online: <https://www.c2dh.uni.lu/thinkering/digital-source-criticism-21st-century-reconsidering-rankes-principles-digital-age> (Accessed May 20, 2020).

Schönholz, Christian. 2017. ‘A brilliant copy every time!’: aspects of a cultural proportion. In *Digitisation. Theories and Concepts for Empirical Cultural Research*, ed. Gertraud Koch, 117-132. New York/London: Routledge.
Schröder, Ralph. 2014. Big Data and the brave new world of social media research. In *Big Data & Society*: 1-11. Thousand Oaks: SAGE Publications.
Shafer, Robert Jones. 1974. A guide to historical method. Homewood, Illinois: Dorsey Press.
Society of American Archivist 2019: <https://www2.archivists.org/glossary/terms/p/provenance> (Accessed May 2, 2019)
Tinati, Ramine, Olivier Phillipe, Catherine Pope, Leslie Carr, and Susan Halford. 2014. Challenging Social Media Analytics: Web Science Perspectives. *Proceedings of the 2014 ACM conference on Web science 23.06.2014*: 177-181. Southampton: Web Science Institute.
Tufekci, Zeynep. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *ICWSM: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media 02-04.06.2014*. Detroit.
Turkle, Sherry. 1994. Constructions and Reconstruction of Self in Virtual Reality: Playing in the MUDs. In: *Mind, Culture, and Activity* 1 (3): 158-167. London: Taylor & Francis.
van Dijck, José. 2014. Datafication, dataism and dataveillance. Big Data between scientific paradigm and ideology. In *Surveillance & Society* 12 (2): 197-208. Chapel Hill: University of North Carolina Press.
Venturini, Tommaso, Liliana Bounegru, Jonathan Gray and Richard Rogers. 2018. A reality check(list) for digital methods. In *New Media & Society* 20 (11): 4195-4217. Chicago: University of lllinois at Chicago Press.
Walker, Shawn. 2017. The Complexity of Collecting Digital and Social Media Data in Ephemeral Contexts. Dissertation. Washington: University of Washington. <https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/40612/Walker_washington_0250E_17763.pdf?sequence=1> (Accessed ?).
Weller, Toni, ed. 2013. History in the digital age. New York/London: Routledge.
Willson, Michelle and Tama Leaver. 2015. Zynga's FarmVille, Social Games, and the ethics of Big Data Mining. In *Communication Research and Practice* 1 (2): 147-158. London: Taylor & Francis.

Historical Social Research
Historische Sozialforschung

All articles published in this Forum: