SAME DIFFERENCE: IDENTITY AND DIVERSITY IN LINKED OPEN CULTURAL DATA

SUSAN BROWN

Abstract *Linked Open Data (LOD) was designed to respect heterogeneity in source datasets. However, the fundamental mechanisms of interlinking require sameness without nuance, so Linked Data is at risk of the problems associated with lack of diversity in big data generally. This article investigates the tension between difference and sameness specifically as it relates to asserting the identity of entities. It links ambiguity in natural language and cultural expression to Derrida's notion of différance and the foundation of Linked Data structure in Peircian semiotics. Representing entities so as to foreground rather than suppress subtle differences or ambiguity is a challenge given the lack of anything in between owl:sameAs and owl:differentFrom that can be supported by formal logic. The reuse of Uniform Resource Identifiers (URIs) recommended as a best practice for data interlinking is troubled by confusion over the relationship between URIs and representation, whereas owl:sameAs suffers from a range of forms of misuse. Despite these challenges to representing nuance and ambiguity, however, there are several ways in which humanities researchers and cultural institutions can pursue better means of representing diversity and difference using LOD, particularly through interdisciplinary and multisectoral collaboration.*

International Journal of Humanities and Arts Computing 16.1 (2022): 1–16
DOI: 10.3366/ijhac.2022.0273

www.euppublishing.com/ijhac

Keywords: Linked Open Data, Semantic Web, knowledge representation, description logic, cultural heritage, semiotics, deconstruction, ambiguity, difference, diversity, *sameAs* problem

Similarity is an institution.
Mary Douglas, *How Institutions Think*[1]

Linked Open Data (LOD) is about surmounting differences. It is about overcoming silos – the result of different publishing locations, differences of data formats, and the myriad data renderings and interfaces that block effective access and reuse – to create an interoperable Web. The Web reflects a diverse cultural record, past and present, both in creators' backgrounds and among cultural artefacts themselves, but the mechanisms of LOD are inimical to representing the differences and variety of which diversity consists,[2] bridging differences by asserting sameness through ontologies and vocabularies. The Web Ontology Language *owl:sameAs* relationship is a paradigm of LOD in invoking the primary meaning of sameness: 'identity'.[3]

Difference matters when it comes to digital artefacts: cultural differences among creators, distinctions of media, form or content among artefacts, particular contexts and conceptual frameworks, and local or specific forms of knowledge are significant components of many scholarly approaches. How might a humanities approach to Linked Data highlight and investigate cultural difference and diversity, given that the Linked Data promise for opening up the wealth of Web content rests on sameness? I argue that foregrounding rather than suppressing difference matters profoundly when it comes to Linked Open Cultural Data, outline how Linked Data's underlying mechanics impede the representation of diversity, and suggest that interdisciplinary and multisectoral initiatives are needed to address this major challenge.

I. BIG DATA AND DIVERSITY

LOD, as a form of Big Data, furthers an epistemic shift which we are only beginning to comprehend, note danah boyd and Kate Crawford:

> Just as Ford changed the way we made cars – and then transformed work itself – Big Data has emerged a system of knowledge that is already changing the objects of knowledge, while also having the power to inform how we understand human networks and community. 'Change the instruments, and you will change the entire social theory that goes with them,' Latour reminds us.[4]

Big Data typically normalizes or erases diversity, or, worse still, exacerbates social inequality.[5] As Big Data is harnessed by commercial and anti-democratic interests, it lends itself to attacks on civil society and human dignity.[6] One review of the legal and ethical implications of Big Data concludes that 'additional work is needed to support diversity in a data-responsible society'.[7]

LOD extends the early idealistic vision of the Web by Tim Berners-Lee and colleagues in the World Wide Web Consortium, offering a tool for tackling thorny political challenges such as the fight against climate change.[8] The open data movement is growing worldwide, but efforts often founder due to high overhead. Rob Kitchin insists we need to study further the broader impacts of open data projects as 'complex sociotechnical systems with diverse stakeholders and agendas' as well as 'the messy, contingent and relational ways in which they unfold'.[9]

The Linked Data cloud keeps growing, but much Linked Data is closed and not much open data is cultural.[10] Forays into Linked Data for both humanities research and cultural heritage projects are often constrained by infrastructural and resource limitations.[11] Few established Linked Data infrastructures operate across large numbers of distributed cultural datasets,[12] and a big challenge remains of how to make LOD simply usable, let alone diverse, although the two arguably go hand in hand.[13] LOD remains the best means to support collaborative knowledge creation, but to enable inquiry into cultural factors such as social identities or historical processes while interrelating large bodies of data poses significant challenges.

2. LINKED DATA ENTITIES AND IDENTITY

Five-star LOD, the ubiquitous benchmark articulated by Tim Berners-Lee, requires adoption of a uniform data model, the Resource Description Framework (RDF), alongside other web standards.[14] The data model presumes heterogeneous sources and differences among datasets: 'RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time' unlike relational databases.[15] To achieve the fifth and ultimate star of five-star Linked Data, these datasets must link to other data, a primary mechanism for which is asserting identity or sameness between entities, the things that are the subjects or objects of Linked Data statements. This sameness involves either 1) reusing external Universal Resource Identifiers (URIs) for subjects or objects of Linked Data statements, indicating that that the data refers to the same entity as other datasets that use that URI, or 2) using *owl:sameAs* relationships to link internal and external URIs for things (e.g. persons, places, cultural objects) or concepts (artistic movements, techniques or colours): 'an owl:sameAs statement indicates that two URI references actually refer to the same thing: the individuals have the same "identity"'.[16] Linking is also achieved by reusing ontologies and aligning properties from different ontologies, but that is not the focus here.[17] Asserting the sameness of entities across disparate datasets is a fundamental underpinning of Linked Data, and often functions unproblematically to enable interconnection and contextualization. URIs for entities are frequently adopted for reuse, and identifiers from different knowledge systems often refer unproblematically to the same thing, like names within natural language.

However, Linked Data has limited capacity for representing difference. The Web Ontology Language (OWL) supports complete differentiation of entities with *owl:differentFrom* and the separation of classes of entities with *owl:disjointWith*, but nothing in between.[18] This is because the querying, reasoning and inferencing associated with OWL ontologies rely on the description logics of knowledge representation. Description logics are hostile to ambiguity:

> DLs are logics (in fact, most DLs are decidable fragments of first-order logic), and as such they're equipped with *formal semantics*: a precise specification of the meaning of DL ontologies. This formal semantics lets humans and computer systems exchange DL ontologies without ambiguity as to their meaning and also make it possible to use logical deduction to *infer* additional information from the facts stated explicitly in an ontology . . .[19]

Not for nothing, then, is data disambiguation a huge component of producing LOD.

3. AMBIGUITY

Ambiguity abounds in and, for many, enhances natural language and thus expressions of human qualities and activities.[20] Philosophers from Heidegger to Kierkegaard, Sartre, Beauvoir and Lacan consider ambiguity inherent to human language and existence. It is perhaps the most unsettling form of difference.[21] The *Oxford English dictionary* defines ambiguity as: 'the fact or quality of having different possible meanings; capacity for being interpreted in more than one way'; 'A nuance which allows for an alternative reading of a piece of language'; 'The fact or quality of being difficult to categorize or identify'.[22] Ambiguity in all of these senses undermines the identity assertions required for machines to process relationships using an ontology based on description logics.[23]

Ambiguity results from the complex significatory ability of natural language to convey difference within a single statement. The problem is not outright difference, which is handled well by Linked Data, but the ambivalence and multivalence of meaning generated by Derridean *différance*, the perpetual play of language. While the structures that permit play previously boasted a metaphysical centre, 'a reassuring certitude, which itself is beyond the reach of play', that metaphysical or transcendent presence that grounded meaning was thrown into doubt in the twentieth century, causing the 'rupture' that produced poststructuralism.[24] Derrida marks ethnology as a privileged discourse in the turn towards decentred language due to its relationship with cultural difference: 'ethnology could have been born as a science only at the moment when a decentering had come about: at the moment when European culture – and, in consequence, the history of metaphysics and of its concepts – had been *dislocated*, driven from its locus, and forced to stop considering itself as the culture of reference.'[25] Understanding and representing cultural difference in Derrida's analysis is inextricably entangled in the *différance* of signification that challenges the premise of LOD that an entity's identity is certain or exact.

Perhaps the most sophisticated attempt to unpack signification, one studied by Derrida, is the semiotic theory of Charles Sanders Peirce which underlies John Sowa's influential work on knowledge representation, conceptual graphs and ontologies for the Semantic Web. Unpacking Peirce's theory of signs, Sowa notes the diversity of representation in language: 'we can talk about the same phenomena at different levels of detail from different perspectives with different choices of types and different numbers of existential quantifiers. There is no limit to the variety of perspectives, purposes, questions, answers, decisions, actions, social interactions, and metaphysical explanations.'[26] Peirce and Derrida differ over metaphysics. Peirce dismisses metaphysics as 'a puny, rickety, and scrofulous science',[27] developing a pragmatist theory of signification as a phenomenological process grounded in the logic of pure mathematics. Thus, Sowa stresses that 'Peirce's categories [of signs] are determined by phenomena that are observed or inferred from observations, not on debatable essences, substances, or natures'.[28] Derrida considers metaphysics inescapable and pronounces empiricism 'non-philosophy', despite conceding Peirce's progress towards 'the de-construction of the transcendental signified' and celebrating his refusal 'to bind linguistics to semantics'.[29] Yet their understandings of language converge in many respects,[30] and Peirce's approach to analysing, classifying and interpreting signs from a quite different framework nevertheless leads one semiotician to conclude that 'indeterminacy … is no less than the fulcrum point of the life of signs and hence of their meaning'.[31] That this sense of language, through Sowa, informs the construction of computational or applied ontologies bodes well for the representation of ambiguity computationally.

Yet Sowa's invocation of Peirce in outlining three types of computational ontologies 'for any particular domain' sidesteps the complexities of signification. The ontology most apt for Cultural Data, the 'descriptive' scientific ontology, is to be 'judged by the same criteria as any theory of science: it must make testable predictions about the domain'; the example is Newtonian mechanics.[32] Perspectives and interpretative processes are absent from a conceptualization of a predictive ontology that seems ill-suited to multifaceted questions about the past, such as whether Newton, Leibniz or the Kerala school invented calculus.[33] Sowa's reductiveness here simply reflects what current ontological structures support: 'Low-level, task-oriented modules have been the most successful in science, engineering, business, and everyday life.'[34] He knows that 'ambiguity is inevitable'[35] and that to approach the flexibility of language an ontology would require a sophisticated, modular, and dynamic approach.

Theories of signification, then, including those underlying computational ontologies, stress ambiguity and *différance* in ways that highlight unproblematized assumptions of the achievability of disambiguation within Linked Data, and in practice this pertains perhaps particularly to non-positivist domains. Cultural Data demands more than low-level tasks, but to understand why ambiguity represents such a challenge requires considering some basic mechanisms of Linked Data. I move now to an exploration of how these challenges manifest in two pivotal forms of sameness related to entities: the reuse of URIs and the use of *owl:sameAs*.

4. REPRESENTING ENTITIES WITH URIS

Some things seem more amenable to disambiguation than others. Named entities in particular seem like they should be straightforwardly either the same or different. However, this assumption quickly becomes problematic, for reasons including the confusion of identifiers and representations. In contrast to the Internet of Things, in which an identifier denotes a specific appliance, for instance, Linked Data for cultural heritage typically uses URIs to represent concepts of things – of people, paintings or abstract ideas – opening the door to ambiguity in the reuse of URIs.

Computational ontologies as knowledge representations are avowedly representational, modelling the world for specific purposes. But that representationality itself causes ambiguity. The self-evident difference between an entity and a representation thereof is regularly blurred in the use of URLs for reference entries *about* a person as URIs to *denote* the person. This is evident, for instance, in the *Wikidata* data type *external-id* (English label 'External identifier'), for 'strings that represent identifiers used in external systems (databases, authority control files, online encyclopedias, etc.)'.[36] This data type is used for *P6745* ('Orlando author ID'), a property that is an instance of Q55650689 'Wikidata property for authority control for writers, described as a 'Wikidata property for authority control for authors | Wikidata property to identify writers | Wikidata property to identify authors'.[37] Orlando author IDs are six-letter strings taken from the paths of URLs for author profiles within the textbase *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*; *Wikidata* treats these as IDs for the authors themselves. Thus *Wikidata* uses 'hallra' as an external-id for the author Radclyffe Hall based on the URL of a *representation* of Hall,[38] and likewise the '*Oxford Dictionary of National Biography* [ODNB] *ID*' and the '*Encyclopædia Britannica Online ID*' also derived from URL paths for entries about Hall. By contrast, Hall has only a single P1343 'described by source' property for a 'work where this item is described'.[39] Meanwhile, the external-ids (55 in July 2021) conflate such works with actual person identifiers from authorities such as *Europeana*.

The slippage is understandable. These URL components are a convenient means for *Wikidata* to provide handy cross-links to information in common sources, which benefits *Wikidata* users and contributors alike. Moreover, entity IDs or URIs for persons can be hard to distinguish from URLs for texts about a person. Online reference entries often resemble the human-readable HTML pages for entities that should be provided along with RDF for machines, according to best practices for dereferenceable URIs. The *Orlando* entry for Hall is narrative, but it begins with a data-like list of Hall's various names. Conversely, the National Portrait Gallery page from which *Wikidata*'s external-id is derived is apparently a record for a person whose ID is 'mp01984'; its summary of Hall's life, importance, available images and related entities is not unlike some online reference works. This page also resembles Hall's OCLC Worldcat Identities page, which harvests copious information (links, book covers, associated subjects).

The generic signals given by various web resources, as shown in Hall's case, make it difficult to distinguish pages *about* entities from Linked Data entity pages that contextualize them. The spectrum signified by the five stars of LOD is also relevant here. The *Orlando* textbase, because paywalled, earns no stars.[40] Like most memory institutions with legacy metadata that can stretch back more than a century, the National Portrait Gallery uses databases to manage collections information, while adopting Linked Data standards such as Resource Description and Access (RDA);[41] much of that metadata is open on the Web, but not accessible as Linked Data.[42] OCLC publishes Linked Data but is still working to stabilize the related infrastructure for its members,[43] and *Wikidata* itself, although the most extensive collaborative LOD platform in the world, does not store its data as RDF but serves OWL-compliant RDF that meets the criteria for all five stars.[44] All this is to underscore the notion that, technically as well as generically speaking, there exists a spectrum from web publication to LOD.[45]

Notwithstanding this spectrum, however, there is an important distinction, semantically, between a thing and a representation of it; from a description logics perspective, conflating the two is wrong. This matters less if what matters is access, but ambiguous semantics pose a serious problem for reasoning. In one sense, grabbing both authorities' identifiers and sources about those identified for external-ids is just fine: it falls within the scope of *Wikidata*'s definition. But, at the same time, the P1343 'described by source' property lies fallow, even though the data would be more useful if external-ids were used for authority URIs only and not also for representational sources.

The semantic limitations outlined here reflect *Wikidata*'s collaboratively produced ontology, 'loosely defined by the relationships between the Items in the graph'.[46] They are exacerbated by a recognized Semantic Web design flaw, the 'HTTP Range 14 problem' ensuing from the lack of means of distinguishing informational and non-informational resources in URLs.[47] The 'strong tradition of two-dimensional thinking derived from the paper-world'[48] in the cultural heritage and digital humanities communities also contributes, and, in so doing, undermines Berners-Lee's claim that: 'Linked data is essential to actually connect the semantic web. It is quite easy to do with a little thought and becomes second nature. Various common-sense considerations determine when to make a link and when not to.'[49] For a *Wikidata* contributor without ontology expertise, guided by their community of practice, it is commonsensical to think of an *Orlando* entry about a person and an OCLC page about a person as the same, although they do semantically distinct things in a Linked Data context.

Wikidata is the leading experiment in using LOD to navigate differences of language, domain and culture using an open contributory model. As such it constitutes a rich site for evaluating approaches to diversity. *Wikidata*'s collaborative, bottom-up approach means that sometimes even fundamental – but by no means commonsensical – semantic distinctions, such as between *P31 (instance of)* and *P279 (subclass of)*, that is, between Peirce's notion of tokens and types, are not necessarily understood. The result is that the *Wikidata* ontology is 'large and messy' and not easily amenable to inference or reasoning without the application of external methods or structures.[50] *Wikidata* has developed a knowledge graph with fairly 'free-form semantics' as opposed to a consistent and consistently applied ontology, and the slippage outlined above speaks to the challenges of rigorous semantic structures, which were criticized early in the Semantic Web project and which remain a barrier to participation.[51] *Wikidata* illustrates the extent to which the accessibility essential to fostering data creation from marginalized groups sits in tension with a desire for greater semantic precision.

5. SAMEAS IDENTITIES

The 'sameAs problem' stemming from the ubiquitous use of the *owl:sameAs* property is summarized by Halpin et al.: 'Much of the supposed "crisis" over the proliferation of *sameAs* in Linked Data can be traced to the fact that many mutually incompatible intuitions motivate the use of *owl:sameAs* in Linked Data. These intuitions almost always violate the rather strict logical semantics of identity demanded by *owl:sameAs* as officially defined.'[52] To highlight their implications, I here give examples particularly pertinent to Cultural Data of several forms of *sameAs* misuse identified by Halpin et al.

'Identical but Referentially Opaque' applies when the specificity of names matters for contextual reasons. For instance, underground railroad agent and conductor Daniel Hughes lived in what was known as N— Hollow, later renamed Freedom Road, in Muncy, Pennsylvania.[53] The two names refer to the same place, but the racial slur embedded in the earlier one demands contextualization and carries with it properties that may not carry over to the current name. Referential opacity also occurs when there is a level of identification between entities, but the properties are not interchangeable because of differences in conceptualization. The late-Victorian author 'Michael Field' is an example that tests the boundaries of referential opacity. This string might be classed in one dataset as a pseudonym, in another as a persona, in yet another as an author. In one sense all refer to the same thing, but the properties associated with each conceptualization would be very different, including whether 'Michael Field' was a person.[54]

'Identity as Claims' involves questions about truth status, because an *owl:sameAs* relationship means that all properties of the linked URIs apply to each other. This kind of equivalence might lead to assertions about 'Michael Field' that would be incorrect: for instance, Robert Browning describes visiting Field; however, a pseudonym might be incapable of social relationships according to an ontology governing it. Property transference can also be problematic when entities change over time. For instance, the individual George Routledge founded in 1836 a publishing house that became George Routledge & Sons and went through many name permutations until absorbed by Taylor & Francis. An event-based ontology could tease out Routledge as a person and as a series of organizations by type, but few have world enough and time for such detail and precision. Meanwhile, a careless match between a URI for the person and one for the current publisher could result in ludicrous assertions. *owl:sameAs* relationships can thus lead to false statements and ontological errors.

'Matching' and 'Similar' identity errors involve close likeness or association, such as the versions of Routledge: matching is contingent upon particular contexts or purposes, while 'similar' entails some but not all properties being shared. 'Related' misuse of *owl:sameAs* links discrete entities that are connected in some way even though they do not share properties, as in the conflation of URIs with representations. The Simple Knowledge Organization System (SKOS) addresses all three of these cases with terms such as 'skos:closeMatch', 'skos:broader' and 'skos:narrower', providing a middle ground between complete sameness and difference, but unfortunately not supporting logical operations.

The sameAs problem thus relates to the semantically erroneous declaration of a relationship of equivalence or identity between two entities, either through misconception of the implications of doing so, as a recourse given the lack of alternatives to *owl:sameAs*, or both. Carrying over properties in such cases, including typing or categorizing entities in ways that might be incommensurable across disparate ontologies, introduces semantic ambiguity into the larger LOD graph. Halpin et al. mapped out a similarity ontology in recognition of the fact that 'there is a nuanced heterogeneous structure of similarity' at play in the use of sameAs, but concluded it could not be 'reliably deployed'. Instead, the modest proposal of an RDFS 'seeAlso' property 'to indicate a resource that might provide additional information about the subject resource' was added to the specification.[55] Fully tackling the *owl:sameAs* challenge, they observe, 'may require a certain refactoring of some core constructs of RDF'.[56]

Considerations of the *owl:sameAs* problem increasingly flag the importance of background and context,[57] confirming the importance of these factors to dealing with the challenges posed by difference and diversity. Halpin et al. indeed argue that 'much of the variance between identity is due to domain-specific contextual uses of identity' and propose an extension to RDF to recognize context.[58] However, this proposal works against the potential for cross-fertilization, serendipity and inference across different domains and sectors of knowledge production that would be hugely useful for cultural analysis: it could have the unfortunate effect of reinstating the data segregation that Linked Data was meant to overcome.

6. THE CHALLENGES OF DIFFERENCE

The slippages between, and conflations of, entities surveyed here demonstrate merely one way in which the ambiguity in human expression and thought are manifest in LOD. Differences that coexist with similarity rather than being absolute are being incorporated into assertions of identity in ways that work against nuanced representations of cultural complexity by absorbing difference into sameness or silence, the latter being an absence of links. The examples provided here related to named entities are relatively straightforward; challenges only multiply when one turns to cultural identities and other dynamic social constructions, or moves into aspects of ontology structure beyond those considered here.[59] The flattening of difference brings to mind the colloquial expression 'same difference', the linguistic equivalent of a shrug that overwrites some kind of distinction or contrast through an assertion of sameness.[60] The challenges are not trivial, given that they stem in large part from the logics that ground LOD. Unless a means of formalizing and attaching logical operations to relationships that sit on the spectrum between identity and difference can be devised, the ability to leverage the semantics of Linked Data to get at cultural complexity will be seriously hampered. Yet that is no reason for a collective shrug on the part of researchers or cultural institutions interested in using LOD with a regard for nuance and diversity. Several approaches, especially if widely adopted across sectors, could help to mitigate the problems of over-identification.

7. CONCLUSION: TOWARDS LINKED OPEN DATA FOR DIFFERENCE

The lack of a ready solution to representing diversity in Linked Open Data is unsurprising. Grounded in the attempt to formalize natural language, Linked Data is reliant on the very logic of sameness and difference that is undermined by *différance*, the deferral of determinate meaning. For Derrida, attempts to sidestep illusory notions of determinacy are inevitably complicit with those structures, at the same time that they verge on the unthinkable.[61] Humanities researchers and cultural organizations working with Linked Data are likewise inevitably reliant on structures of formal logic that they will at times be forced to supplement, exceed and transgress in the manner of Derrida's *bricolage*. Parallel problems, of course, plague natural language representations of diversity.

Attempts to address the tension between sameness and difference within the constraints of formal logic have yielded little to date. An immediate strategy for researchers and memory institutions is to push on the capacity of carefully formalized ontologies such as CIDOC-CRM to see what can be achieved by representing provenance and context.[62] Logical operations on the structural representation of context my help to elucidate differences and distinctions among related entities.[63] Linked Data properties devised specifically to support diversity are also being explored, although they cannot solve the core problems with logic. The Canadian Writing Research Collaboratory, for instance, uses a 'label' property (different from a skos:label) for dealing with interrelated social identities in ways not unlike the 'pollarding' technique proposed by Halpin et al. to avoid conflating entities.[64] A complementary approach is to investigate what purchase SKOS can provide on nuance, by experimenting with highly granular structured vocabularies such as *Homosaurus*.[65] In this context, too, it would be worth evaluating how well CIDOC-CRM, which uses domain-specific vocabularies for typing as well as for entities, can leverage those specificities in logical operations.[66] By designing for difference,[67] user interfaces might leverage highly granular taxonomic relationships for prioritization, filtering and faceting in ways that promote the representation and understanding of diversity.

Community building and cross-sectoral collaboration may be most important of all to address diversity effectively, given the substantial resources and infrastructure required to work with LOD. Humanities scholars have a stake both in the Web's cultural content and in analysing the construction of meaning. They are, therefore, ideally positioned to partner with data stewards who think deeply about context and material culture. If these complementary perspectives could inform shared experimentation, iteration and evaluation of LOD, that might generate mutually acceptable strategies and a broad community of practice that could support the representation of difference in LOD in broadly intelligible ways.

Ensuring that LOD represents diversity well is crucial to more than the study and dissemination of cultural heritage. As Wendy Chun argues, a digitally driven passion for sameness, 'homophily (the notion that similarity breeds connection)', is one of the enemies of social justice in our current technological landscape.[68] For Leif Weatherby, 'data's dual aspect as both representation and infrastructure' creates a pressing need to engage critically with the 'semiotic metaphysics' of the data-driven systems that order our world. As part of that engagement, those working in LOD must grapple with the in-between of sameness and difference to create cultural datasets and technical systems that will allow us to 'mine the mathesis of difference and similarity to explore unexpected formations, trends, and linkages influencing what we think identities have been and can be'.[69] Linked Data standards and technologies are designed to work with heterogeneous data sources. A sophisticated, modular and dynamic approach of the kind envisioned by Sowa may rely heavily on description logics, but it will almost certainly involve *bricolage*. Notwithstanding the significant challenges outlined here, cultural researchers and institutions have compelling reasons to pursue strategies for the formal representation of knowledge that incorporates diversity meaningfully.

ORCID

Susan Brown https://orcid.org/0000-0002-0267-7344

ACKNOWLEDGEMENTS

I would like to acknowledge my co-authors on the Canadian Writing Research Collaboratory ontology (https://sparql.cwrc.ca/) and the Linked Infrastructure for Networked Cultural Scholarship team (https://lincsproject.ca/team/); Kelly Hughes for assistance with manuscript preparation; and funding from the Social Sciences and Humanities Research Council and the Canada Foundation for Innovation.

END NOTES

[1] M. Douglas, *How institutions think* (London, 1986). Cited here at 55.

[2] 'diversity, *n.*' *Oxford English dictionary (Online)* (Oxford, June 2021).

[3] 'identity, *n.*' *Oxford English dictionary*.

[4] d. boyd and K. Crawford, 'Six provocations for Big Data', *Proceedings of the Oxford Internet Institute's A decade in internet time: symposium on the dynamics of the Internet and society, Oxford 2011* (Oxford, 2011), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431, last accessed 19 July 2021. Cited here at 3.

[5] S. U. Noble, *Algorithms of oppression* (New York, 2018); B. Rieder, 'Big Data and the paradox of diversity', *Digital Culture & Society,* 2.2 (2016), 39–54.

[6] R. J. Deibert, *Reset: reclaiming the Internet for civil society* (Toronto, 2020); S. Zuboff, *The age of surveillance capitalism: the fight for a human future at the new frontier of power* (London, 2019).

[7] M. H. Drosou, H. V. Jagadish, E. Pitoura and J. Stoyanovich, 'Diversity in Big Data: a review', *Big Data*, 5.2 (2017), 73–84. Cited here at 74.

[8] T. Berners-Lee, J. Hendler and O. Lassila, 'The Semantic Web', *Scientific American,* 284.5 (2001), 34–43; World Wide Web Consortium [W3C], 'Linked Data Cookbook', *W3C EGov Wiki*, https://www.w3.org/egov/wiki/Linked_Open_Data; W3C, 'Linked Data Cookbook', *W3C*, https://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook; F. Bauer and M. Kaltenböck, *Linked Open Data: the essentials. The climate knowledge brokering edition* (Vienna, 2016). https://www.reeep.org/sites/default/files/LOD-TheEssentials2016.pdf. All last accessed 19 July 2021.

[9] R. Kitchin, *The data revolution: Big data, open data, data infrastructures and their consequences* (London, 2014). Cited here at 66.

[10] On the Linked Data cloud: culture is subsumed within 'cross domain' datasets, a category that is itself shrinking in relation to other domains. J. P. McCrae et al., *The Linked Open Data cloud*, https://lod-cloud.net/. Last accessed 19 July 2021.

[11] Kitchin, *The data revolution*, 48–66; J. E. Simpson, 'Inference and linking on the humanist's Semantic Web', *Scholarly and Research Communication*, 5.4 (2014). Cited here at 4.

[12] Digital Public Library of America, https://dp.la/; 'Linked Open Data', *Documentation de la TGIR Huma-Num*, https://documentation.huma-num.fr/humanum-en/; Europeana Pro, https://pro.europeana.eu/page/linked-open-data; CultureSampo, https://seco.cs.aalto.fi/applications/kulttuurisampo/. All last accessed 19 July 2021.

[13] R. Sanderson, 'Shout it out: LOUD', EuropeanaTech conference, *YouTube*, https://www.youtube.com/watch?v=r4afi8mGVAY; R. Sanderson, 'The illusion of grandeur: trust and belief in cultural heritage Linked Open Data', https://www.birmingham.ac.uk/schools/historycultures/departments/ironbridge/news/2021/illusions.aspx, last accessed 19 July 2021.

[14] T. Berners-Lee, W3C, *Linked Data*, https://www.w3.org/DesignIssues/LinkedData.html; W3C, *RDF*, https://www.w3.org/RDF/; W3C, *Standards*, https://www.w3.org/standards/; M. Hausenblas et al., 5 Star Data, *5 star Open Data,* 2021 https://5stardata.info/en/. All last accessed 19 July 2021.

[15] W3C, *RDF.*

[16] W3C, *owl:sameAs*, https://www.w3.org/TR/owl-ref/#sameAs-def. *owl:equivalentClass* and *owl:equivalentProperty* do not necessarily indicate conceptual identity like *owl:sameAs* (https://www.w3.org/TR/owl-ref/#equivalentClass-def; https://www.w3.org/TR/owl-ref/#equivalentProperty-def). OWL Full permits *owl:sameAs* relationships between classes and properties. All last accessed 19 July 2021.

[17] Properties require their own discussion due to their complexity and the greater complexity of sameness in relation to ontologies. In addition, Halpin et al. find that *owl:sameAs* is much more used than *owl:equivalentClass* or *owl:equivalentProperty* statements, Halpin et al., 'When sameAs isn't the same'. The argument here about ambiguity also applies generally to properties. One experiment in reasoning on cultural heritage data in *Wikidata* investigated property alignment but focused on technical challenges of processing rather than the quality of the inferred statements. N. Freire and D. Proença, 'RDF reasoning on large ontologies: a study on cultural heritage and Wikidata', *Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations* (New York, 2020), 381–93; H. Halpin, P. J. Hayes and H. S. Thompson, 'When owl:sameAs isn't the same redux: towards a theory of identity, context, and inference on the Semantic Web', in H. Christiansen, I. Stojanovic and

G. Papadopoulos, eds, *Modeling and using context: lecture notes in computer science*, 9405 (New York, 2015), 47–60. Cited here at 52.

[18] W3C, *owl:differentFrom*, https://www.w3.org/TR/owl-ref/#differentFrom-def; C3C, *owl:disjointWith*, https://www.w3.org/TR/owl-ref/#disjointWith-def. All last accessed 19 July 2021.

[19] M. Krötzsch, F. Simančík and I. Horrocks, 'Description logics', *IEEE Intelligent Systems,* 29.1 (2013), 12–19. Cited here at 12.

[20] S. T. Piantadosi, H. Tily and E. Gibson, 'The communicative function of ambiguity in language', *Cognition,*122.3 (2012), 280–91; T. Wasow, 'Ambiguity avoidance is overrated', in S. Winkler, ed., *Ambiguity: language and communication* (Berlin, 2015), 23–31.

[21] A. Sennet, 'Ambiguity', in E. N. Zalta, ed., *The Stanford encyclopedia of philosophy*, https://plato.stanford.edu/entries/ambiguity/. Last accessed 9 July 2021.

[22] 'Ambiguity', *Oxford English dictionary.*

[23] An ontology is 'a formal, explicit specification of a shared conceptualization that is characterized by high semantic expressiveness required for increased complexity'. C. Feilmayr and W. Wöß, 'An analysis of ontologies and their success factors for application to business', *Data & Knowledge Engineering*, 101 (2016), 1–23. Cited here at 3, 4, drawing on Gruber et al.

[24] J. Derrida, 'Structure, sign, and play in the discourse of the human sciences', *Writing and difference,* trans. A. Bass, 1967 (London, 2001), 351–70. Cited here at 352, 354, 351.

[25] Derrida, 'Structure, sign, and play', 356.

[26] J. Sowa, 'Signs and reality', *Applied Ontology*, 10.3–4 (2015), 273–84. Cited here at 277.

[27] C. S. Peirce, *The Essential Peirce, volume 2: Selected philosophical writings (1893–1913)* (Bloomington, 1998). Cited here at 375.

[28] Sowa, 'Signs and reality', 275.

[29] J. Barnouw, 'Peirce and Derrida: "Natural Signs" empiricism versus "Originary Trace" deconstruction', *Poetics Today*, 7.1 (1986), 73–94. Cited here at 73, 79; Derrida, *Of grammatology*, 1967 (Baltimore, 2016). Cited here at 49, 50.

[30] D. E. Pettigrew, 'Peirce and Derrida: from sign to sign', in V. M. Colapietro and T. M. Olshewsky, eds, *Peirce's doctrine of signs* (Brussels, 2011), 365–78.

[31] F. Merrell, *Peirce, signs, and meaning* (Toronto, 1997). Cited here at ix.

[32] Sowa, 'Signs and reality', 382.

[33] D. F. Almeida and G. G. Joseph., 'Eurocentrism in the history of mathematics: the case of the Kerala School', *Race & Class,* 45.4 (2004): 45–59.

[34] J. Sowa, 'A dynamic theory of ontology', in B. Bennett and C. Fellbaum, eds, *Formal ontology in information systems* (Amsterdam, 2004), 204–13. Cited here at 213.

[35] Sowa, 'A dynamic theory', 204.

[36] 'External identifiers', *Wikidata*, https://www.wikidata.org/wiki/Wikidata:External_identifiers, last accessed 19 July 2021.

[37] 'Orlando author ID', *Wikidata*, https://www.wikidata.org/wiki/Property:P6745; 'Wikidata property for authority control for writers', *Wikidata*, https://www.wikidata.org/wiki/Q55650689. Both last accessed 19 July 2021.

[38] 'Radclyffe Hall', in S. Brown, P. Clements and I. Grundy, eds, *Orlando: women's writing in the British Isles from the beginnings to the present* (Cambridge, 2006–21). Last accessed 19 July 2021.

[39] 'Radclyffe Hall', *Wikidata*, https://www.wikidata.org/wiki/Q237639, last accessed 19 July 2021.

[40] While this is true of the textbase, its XML has been used to produce a Linked Open Dataset accessible at https://sparql.cwrc.ca/. S. Brown et al., 'The *Orlando* British women's writing

dataset release 1: Biography and bibliography V.1', *Scholars Portal Dataverse*, 2019, https://doi.org/10.5683/SP2/EOB9S6, last accessed 19 July 2021.

[41] 'Resource description and access toolkit'. *Resource Description and Access*. https://www.rdatoolkit.org/, last accessed 19 July 2021.

[42] E. Purvis, 'Collections information and access policy', *National Portrait Gallery*, https://www.npg.org.uk/about/corporate/gallery-policies/collections-information-and-access-policy, last accessed 19 July 2021.

[43] 'APIs', *OCLC Developer Network*, https://www.oclc.org/developer/api.en.html. 'Shared entity management infrastructure', *OCLC*. https://www.oclc.org/en/worldcat/oclc-and-linked-data/shared-entity-management-infrastructure.html/. Both last accessed 19 July 2021.

[44] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez and D. Vrandečić, 'Introducing Wikidata to the Linked Data Web', *Proceedings of the international Semantic Web conference* (New York, 2014), 50–65.

[45] B. Cope, M. Kalantzis and L. Magee, *Towards a Semantic Web: connecting knowledge in academic research* (Oxford, 2011). Cited here at 226–7.

[46] A. Piscopo and E. Simperl, 'Who models the world?: Collaborative ontology creation and user roles in Wikidata', *Proceedings of the ACM on human–computer interaction,* 2, issue CSCW (2018), 1–18. Cited here at 4.

[47] 'HTTPRange-14', *Wikipedia*, https://en.wikipedia.org/wiki/HTTPRange-14, last accessed 19 July 2021.

[48] M. Barbera, "Linked (Open) Data at web scale: research, social and engineering challenges in the digital humanities", *JLIS*, 4.1 (2013), 91–104. Cited here at 96.

[49] T. Berners-Lee, 'Linked Data', *W3C* (2009), https://www.w3.org/DesignIssues/LinkedData.html, last accessed 19 July 2021.

[50] T. Hanika, M. Marx and G. Stumme, 'Discovering implicational knowledge in Wikidata', *Proceedings of the international conference on formal concept analysis* (New York, 2019), 315–23.

[51] 'Knowledge graph', *Wikipedia*, https://en.wikipedia.org/wiki/Knowledge_graph, last accessed 19 July 2021; Cope et al., *Towards a Semantic Web*, 229.

[52] H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness and H. S. Thompson, 'When owl:sameAs isn't the same: an analysis of identity in Linked Data', *Proceedings of the international Semantic Web conference* (Springer, 2010), 305–20. Cited here at 306. Halpin et al., 'When owl:sameAs isn't the same redux'.

[53] B. DeRamus, *Forbidden fruit: Love stories from the Underground Railroad* (New York, 2005). Cited at 219.

[54] S. Brown and J. Simpson. 'The curious identity of Michael Field and its implications for humanities research with the Semantic Web', *Proceedings of the 2013 IEEE international conference on Big Data*, 77–85. IEEE, 2013.

[55] C3C, '*RDF Schema 1.1*', https://www.w3.org/TR/rdf-schema/, last accessed 19 July 2021.

[56] Halpin et al., 'When owl:sameAs isn't the same', 319.

[57] Halpin et al., 'When owl:sameAs isn't the same', 317.

[58] Halpin et al., 'When owl:sameAs isn't the same redux'.

[59] S. Brown, J. Cummings, A. Lemak, J. Drudge-Willson, K. Martin, A. Mo and D. Stacey, 'Cultural formations: structuring a Linked Data ontology for intersectional identities', *Cultural Analytics*, forthcoming.

[60] Fittingly, there are subtle differences in definitions of this phrase, which is generally uttered in the wake of some kind of contrast or distinction having been made. The *Oxford English dictionary* defines it as simply 'the same thing, no difference', whereas the *Cambridge dictionary* definition pulled from the *Advanced learner's dictionary*

and thesaurus gets closer: it is 'said when you agree that what you said was not exactly correct, but you think the difference is not important'. 'same', *Oxford English dictionary* (Oxford, June 2021). 'same difference', *Cambridge dictionary* (Cambridge, n.d.), https://dictionary.cambridge.org/dictionary/english/same-difference. Both last accessed 19 July 2021.

[61] Derrida, 'Structure, sign, and play', 352, 370.

[62] W3C, *Web annotation data model*, https://www.w3.org/TR/annotation-model/; *CIDOC-CRM* [Conceptual Reference Model of the International Committee for Documentation of the International Council of Museums], http://www.cidoc-crm.org/. Both last accessed 19 July 2021.

[63] I. Kyvernitou and A. Bikakis, 'An ontology for gendered content representation of cultural heritage artefacts', *Digital Humanities Quarterly*, 11.3 (2017), http://www.digitalhumanities.org/dhq/vol/11/3/000316/000316.html, last accessed 19 July 2021.

[64] Halpin et al., 'When sameAs isn't the same redux', 50; *CWRC ontology preamble*, https://sparql.cwrc.ca/ontologies/cwrc-preamble-EN.html; '*The CWRC ontology specification 0.99.6*', https://sparql.cwrc.ca/ontologies/cwrc-2018-02-22.html, DOI: https://doi.org/10.5683/SP2/HXMS24, last accessed 19 July 2021.

[65] 'Homosaurus vocabulary', *Digital Transgender Archive*, https://homosaurus.org/v3, last accessed 19 July 2021.

[66] Vocabulary reuse has been articulated as key to the next phase of Semantic Web integration. K. Janowicz, P. Hitzler, B. Adams, D. Kolas and C. Vardeman II, 'Five stars of Linked Data vocabulary use', *Semantic Web*, 5.3 (2014), 173–6.

[67] T. McPherson, 'Designing for difference', *differences*, 25.1 (2014), 177–88.

[68] W. H. K. Chun, *Discriminating data* (Cambridge, MA, 2021). Cited at 36.

[69] A. Liu, 'Toward a diversity stack: digital humanities and diversity as technical problem', *PMLA*, 135.1 (2020), 130–51. Cited here at 145.