DO LOD CONVENTIONS IMPEDE THE REPRESENTATION OF DIVERSITY? THE CASE OF DISABLED ACTORS IN DBPEDIA AND WIKIDATA

Bettina Berendt
TU Berlin & Weizenbaum Institute, Germany,
KU Leuven, Belgium

Jan Kabus
TU Berlin,
Germany

Abstract

Do social knowledge bases depict the diversity of society? We investigated the representation of disability in film and of disabled actors in Wikidata and DBpedia with stakeholder interviews and queries and found underrepresentation in film and in the knowledge bases, as well as structural impediments to inclusion in the film industry – and in the knowledge bases: heterogeneous conceptualisations, heterogeneous terms, relationality, and individuation/autonomy. These challenges of Linked Open Data (LOD) knowledge representation and governance can contribute to the enduring underrepresentation of socially marginalized groups and pose challenges for adequate representations of people in general.

Keywords: diversity, representation, DBpedia, Wikidata, critical data studies

It appears desirable for social knowledge bases such as Wikidata (WD) and Wikipedia (WP, and with it DBpedia, DBp) to show and enhance diversity. A key mechanism is to make socially marginalized groups and individuals (MGI) more visible, by counteracting the frequently occurring underrepresentation of the MGI.

We combined queries to WD and DBp with qualitative stakeholder interviews (all details in Kabus, 2024) and found under-representation of "disability" in movies and television, confirming findings from other sources.[1] We found a (compounding) underrepresentation of disabled actors in WD and DBp, in the sense of much lower proportion of entries on individuals from the MGI in the knowledge sources when compared to their proportion in the film media and in society at large. At first sight, what is needed seems to be more editors and/or focussing their attention on creating more entries on MGI, and encouraging/simplifying imports from existing data sources. But the interviews also highlighted, in addition to the lack of knowledge about WP/DBp/WD, structural factors (such as poor or lacking accessibility of film sets and acting schools, or "cripping up", the casting of non-disabled actors to play disabled roles) that, together with enduring social prejudices that also get perpetuated in movie scripts, impede the inclusion of disabled actors. This perspective helped us to identify, through the query results, possible reasons beyond a scarcity of editors or data: four types of *structural impediments* to inclusion that are inherent in, or at least favoured by, features of the knowledge representation and governance structures of WP/DBp/WD.

1. Heterogeneous conceptualisations. Wikidata and DBpedia refer to different official category systems (such as ICD-10) and further categories, whose origin is not always clear. Modelling decisions made by editors are not uniform. Similar challenges are well-known and probably unavoidable in a collectively and often bottom-up constructed ontology.

Frequently, disability is modelled similarly to age (implicit in birth date), nationality, and other information about a person that is represented as an RDF property: a person *has a* birth date and may *have a* disability (e.g., wdt:P1050 & Q12131). Guidelines such as those by NHS England[2] or the US CDC[3] correspond to this, suggesting *people-/person-first language (PFL)* such as "a person who has a communication disorder", "a person who is hard of hearing", or "an actor with a disability". Person-first language is meant to be respectful and to avoid the reduction a person to their medical condition, which labels such as "a deaf person" or "a disabled actor" seem to do. So are subclass/is-a representations (e.g., https://dbpedia.org/page/Category:Deaf_actors) a remnant of an ableist past? This may or may not be the case – in fact, many disability activists today advocate the use of *identity-first / disability-first language*[4] because "we do not separate gender, race, sexuality or religion from our identity in [this] way. I am a woman, not a 'person with womanliness' "[5] (see 2. below), and "[to voice] that our identities are affected by the way society treats us, not by our condition" (ibid., see 3.).

This heterogeneity poses a challenge for WP/DBp/WD analysis: it requires matching and mappings to aggregate towards statements such as "Wikidata contains *n* actors (in a

[1] e.g., https://www.nielsen.com/insights/2022/closing-the-inclusion-gap-for-people-with-disabilities/
[2] https://service-manual.nhs.uk/content/inclusive-content/disabilities-and-conditions
[3] https://www.cdc.gov/ncbddd/disabilityandhealth/pdf/disabilityposter_visual_alt.pdf
[4] https://radicalcopyeditor.com/2017/07/03/person-centered-language/
[5] https://nowthenmagazine.com/articles/crip-a-story-of-reclamation

specified group).", which are needed to derive findings such as "Actors (in the specified group) are underrepresented in the sense of their proportion in Wikidata compared with the proportion in movies / in the population at large." How should these matchings and mappings be done?

2. Concepts and terms: shared, preferred, reclaimed? Who should define the preferred type of modelling and nomenclature? If self-ascription is prioritised over external ascription: who may speak for the "selves" – are there "communities", and what happens when different stakeholder communities and/or individuals have conflicting views and language? Some linguistic conventions mirror those regarding other dimensions of intersectionality (e.g., "Deaf person" or "Little People"), and it may be the case that a formerly derogatory term gets *reclaimed* by affected persons, including in ways that challenge existing classifications beyond terminology,[6] which may or may not be meant for usage also by non-disabled others. Resulting heterogeneities are more than 'mere data quality problems'.

Different conceptualisations and terminologies challenge DBp/WD representations on a fundamental level: if we understand ontology in the classic sense of "an explicit specification of a shared conceptualization" (Gruber, 1993), what happens if some things are not or must not be shared?

3. Relationality: How to avoid essentialism in LOD? PFL aims to not equate a person with their medical condition, i.e. to avoid a disrespectful essentialism. However, it follows the *medical model of disability* in which the medical condition is named and emphasized in a way that suggests that the condition could be separated from the person and vice versa (which appears doubtful when viewed as "identity", see 1.) and that, as disability activists have argued,[7] the person would be better off without it. The *social model of disability*, which today is espoused by many definitions such as that by the WHO[8] or in labour laws, in contrast emphasizes that in many cases, there is nothing inherently negative or deficient in the person, but that they *get disabled by* the way society 'does things' – the disability only arises/exists in relations and interactions. For example, Deaf persons experience deficits only in a world where information is conveyed by sound, and Little People in a world built to the scale of taller individuals.

The social model poses a challenge for LOD representation: A person-centric property triple, ascribing a unary property, cannot capture it. A class instance statement may appear to be a way out, but it shifts the burden of definition to the class (and fails at individuation, see 4.). Statements to adequately describe disabling interactions and their results may require a stronger expressiveness than that of RDF and highly skilled, well-collaborating editors with domain knowledge.

4. Individuation and autonomy: Who defines whether a category applies? Given the complexities outlined above, it becomes clear why at least in legal social practice, the decision whether a concrete person is "disabled" is highly context-dependent and individual. Many social security and labour law employ lists of impairments and their extents, which are assessed individually and, if above a threshold, yield a status of "disabled". The person *may* use this status (or not). For example, in labour contexts, it is not uncommon that employees forgo the opportunity.

Our findings mirrored this observation, with interviewees referencing colleagues explicitly not considering themselves "disabled". Interviewees described and lauded their casting agencies' actor databases providing a field "inclusion in film" with a free-form textual value that the actor themselves can fill (if they want). The *individual* autonomy to decide whether and if so, how, one is disabled, was emphasized and valued highly by all our interviewees.

Could this be a useful design choice for WP/DBp/WD? The importance of individual self-ascription poses a challenge for LOD policy rules: In the interest of objectivity, WP discourages that persons write about themselves, and in the interest of plurality (and provenance documentation as a basis for user-led conflict resolution), WD encourages that references be provided for statements. The first rule 'taints' a self-ascription, and the second requires that one provide a reference for describing one's identity regarding a matter that for many can be highly intimate. Thus, these rules, while sensible and useful in many other contexts, may become privacy-invasive, unfeasible, or plain absurd in this one.

Conclusion. These impediments can all contribute to underrepresentation, poor visibility, and lack of diversity regarding not only disability and disabled actors, but also other MGI by fragmentation and misrepresentation. Also, while self-ascription should be preferred, given that WD/WP practically rely on editors who enter information about others, discouraging external ascription may compound underrepresentation. These factors thus constitute general issues for the adequate representation of people in collectively authored LOD and should inform future design.

References

Gruber, T.R. (1993). Towards principles for the design of ontologies used for knowledge sharing. In *Formal Ontology in Conceptual Analysis and Knowledge Representation.* Kluwer Academic Publishers.

Kabus, J. (2024). *Unterrepräsentation und Probleme für Schauspieler:Innen mit Behinderungen in sozialen Datenbanken.* Bachelor Thesis, TU Berlin. www.berendt.de/DD

[6] e.g. https://en.wikipedia.org/wiki/Crip_(disability_term)

[7] autisticadvocacy.org/about-asan/identity-first-language/

[8] https://www.who.int/health-topics/disability