CONCEPTUAL MODELS OF THE SOCIOTECHNICAL
ASIS&T Workshop, 22-23 October 2020

Recursive Ontologization: Standardization, Semantic Violence, and Equifinal Closure in Semantic Interoperability Work

Andrew S. Hoffman
a.hoffman@ftr.ru.nl
iHub - Interdisciplinary Hub for Security, Privacy and Data Governance, Radboud University
Nijmegen, the Netherlands

KEYWORDS
computational ontologies, data models, standardization, interoperability, data science ethnography

EXTENDED ABSTRACT

In the vernacular of contemporary data science, ‘interoperability’ has been posited as a guiding principle for the conduct of data-intensive work [20], and its adjectival rendering – ‘interoperable’ – the quality ascribed to entities that can be more or less seamlessly fused together despite initially appearing in divergent formats and/or emanating from different domains. Meanwhile, in biomedical research parlance, the term ‘translation’ has traditionally been used to describe the set of processes through which insights and discoveries made in basic or fundamental research are transformed – or ‘translated,’ as it were – into clinically actionable interventions, such as the development or repurposing of pharmaceuticals, diagnostics, and behavioral interventions [1]. A closer look at the policies and practices driving contemporary translational science reveals that there is a process of synonymization underway whereby ‘translation’ and ‘interoperation’ are becoming increasingly indistinguishable terms [2].

There is perhaps no better example of this convergence than is found in the Biomedical Data Translator consortium, a recent team science initiative funded by the National Center for Advancing

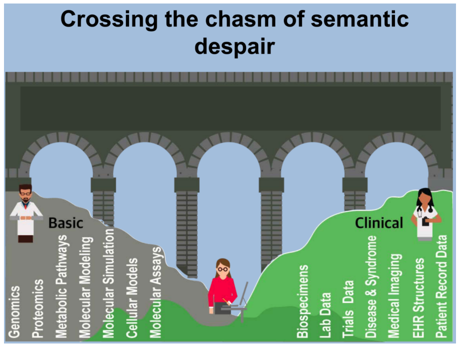

Figure 1: The Chasm of Semantic Despair [4]

Translational Science (NCATS), one of the US National Institutes of Health. The Translator consortium aims to build a next-generation question answering system (henceforth, the Translator) – part smart assistant, part expert system – that can mutually translate the diversity of 'languages' spoken by different biomedical research stakeholders in an attempt to cross the 'chasm of semantic despair' (Figure 1). The technical work to accomplish this feat of translation is focused on two main task-areas: (1) interoperating data and knowledge from heterogeneous domains, and onboarding them into the Translator ecosystem; and (2) developing and deploying multiple instances of 'Reasoning Agent' software that can make sense of this information at scale, where the Reasoning Agents are also interoperated to allow for collaboration in answering specific (parts of) user queries [5, 6].

Envisioned and existing use-cases range from speeding the development of research hypotheses to be tested in laboratory experiments, to more clinically-based applications such as the discovery of therapies to treat rare diseases in the setting of precision medicine [9]. However, as the Translator is not yet a publicly-facing system, a great deal of interpretive flexibility [16] hangs over the project as a whole, and the individuals and teams contributing to its construction operate upon sometimes-conflicting socio-technical imaginaries [12] about how such a tool might ultimately be taken up in – and indeed, change the very fabric of – translational research. Importantly, the convergence of translation and interoperation is not simply a rhetorical accomplishment, but rather relies on a myriad of existing and novel socio-technical infrastructures, human and material alike.

Based on ongoing ethnographic fieldwork with the Biomedical Data Translator consortium, this paper zooms in on aforementioned two task-areas mentioned above. More specifically, we focus on the relationship between a set of extant ontologies that have been onboarded into the Translator ecosystem, on the one hand – and in particular, a set of edge labels, or 'predicates,' which make statements about how entities within a knowledge graph-based infrastructure relate to each other – and, on the other hand, a data model called the Biolink Model, which has been developed as a means of harmonizing those extant predicates such that relationships culled from a myriad of domain-specific databases and knowledge sources can be semantically rendered as the 'same thing' to facilitate automated reasoning and the production of coherent results by the Translator [17].

Critically, over the course of a number of meetings which we attended as participant-observers – including weekly remote calls of the Translator Data Modeling Working Group as well as remote and in-person 'hackathons' and 'relay meetings,' where all members of the consortium gather on a bi-annual for several days' worth of intensive collaborative work – project participants engaged in several conversations pertaining to how the Biolink Model was being used to standardize predicates. One major source of tension that arose, and which persisted over several months of Working Group meetings, relates to the granularity of knowledge representation and thus expressiveness of Biolink Model predicate terms. That is, certain members of the Translator consortium began to raise concerns about the possibility that the level of granularity at which the Biolink Model standardizes edge

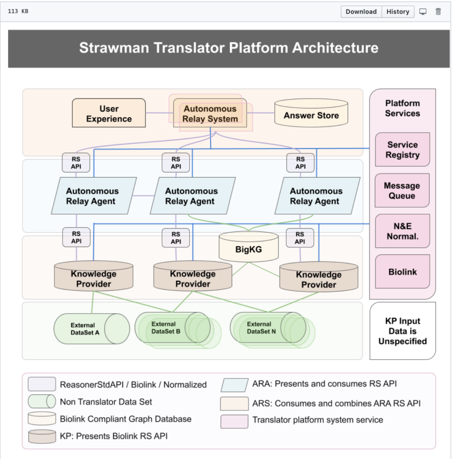

Figure 2: Strawman Architecture c. April 2020

labels would lead to information loss, thereby having a marked impact on the overall utility of the information exposed through the system and possibly edging out certain categories of (prospective) Translator end-users and use-cases.

The focus of this paper is thus on the sociotechnical work oriented towards facilitating semantic interoperability [10]; the frictions [3, 8] encountered amidst efforts to semantically harmonize knowledge graph edge labels; and the moves taken by members of the Translator consortium to problem-solve their way out of the conundra encountered therein. In the first section, we give a more detailed overview of the (envisioned) Translator architecture as well as a brief history of the Biolink Model and its role in facilitating interoperability within the Translator ecosystem to contextualize the ensuing analysis (Figure 2).

The second section then turns to our empirical analysis, where we introduce three vignettes focused on three different moments of standardization:

- The first moment recounts the project-wide effort to account for all extant edge labels being used by different components of the system and, in turn, to map these to existing edge labels within the Biolink model;
- The second moment finds Consortium members confronting the possible forms of exclusion and overflow – what one participant referred to as *semantic violence* – that the work of predicate harmonization precipitates [11, 13, 14], reasoning about scenarios in which (prospective) end-users and use-cases may be excluded as a result of standardizing at a given level of granularity/expressivity, and debating a proposal to abandon the Biolink Model altogether in favor of some other mode of standardization;
- And in the third moment, participants arrive at a kind of *equifinal closure* [7, 15] to these debates, ultimately coming to a consensus on keeping the Biolink Model as the mapping standard, but also putting forth the idea of creating a new 'strike team' of Consortium members who would occupy a kind of intermediary role, liaising with the many teams building different components of the Translator system to aid them in mapping efforts as well as in adding new (and possibly more granular) predicates to the core data model, where necessary.

In the final two sections of the paper, we reflect on these three moments – and, more broadly, on the relationships between (extant) ontologies and (internal) data models in interoperability work – as signaling a process we call *recursive ontologization*. By introducing this notion, we aim to refocus CSCW, HCI, and critical information studies scholarship on ontologies and data models that has been largely preoccupied with the development and implementation of single schemas [15, 18, 19], to one that takes on wider ecologies of ontologies and data models, and to the ongoing, practical work practices deployed to make these tools useful for scientific work. In doing so, we hope to be better positioned to account not only for the affordances and critical occlusions that such a process portends, but also for the ways in which it leads to new categories of data work, data workers, and redistributions of labor within large-scale collaborations in the data-intensive (bio)sciences.

REFERENCES

[1] Christopher P. Austin. 2018. Translating Translation. *Nature reviews. Drug discovery* 17, 7 (Jul 2018), 455–456. https://doi.org/10.1038/nrd.2018.27

[2] Christopher P. Austin, Christine M. Colvis, and Noel T. Southall. 2019. Deconstructing the Translational Tower of Babel. *Clinical and Translational Science* 12, 2 (Mar 2019), 85. https://doi.org/10.1111/cts.12595

[3] Morten Bonde, Claus Bossen, and Peter Danholt. 2019. Data-work and friction: Investigating the practices of repurposing healthcare data. *Health Informatics Journal* 25, 3 (Sep 2019), 558–566. https://doi.org/10.1177/1460458219856462

[4] Christopher G. Chute. 2018. Crossing the chasm of semantic despair: integrating knowledge and data from science and clinical practice. In *Clinical Research Forum–IT Roundtable, Washington DC. https://cdn.ymaws.com/sites/crforum.site-ym.com/resource/resmgr/docs/*$IT_{R}oundtable/CRF_{I}T_{R}oundtable_{3}Nov2017re.pdf(November2, 2017)., Vol.$ 21.

[5] Biomedical Data Translator Consortium et al. 2019. The Biomedical Data Translator program: conception, culture, and community. *Clinical and translational science* 12, 2 (2019), 91.

[6] Biomedical Data Translator Consortium et al. 2019. Toward a universal biomedical data translator. *Clinical and translational science* 12, 2 (2019), 86.

[7] Anne Donnellon, Barbara Gray, and Michel G. Bougon. 1986. Communication, meaning, and organized action. *Administrative Science Quarterly* 31, 1 (1986), 43–55. https://doi.org/10.2307/2392765

[8] Paul N. Edwards, Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41, 5 (Oct 2011), 667–690. https://doi.org/10.1177/0306312711413314

[9] Ruth Hailu. 2019. NIH-funded project aims to build a "Google" for biomedical data. https://www.statnews.com/2019/07/31/nih-funded-project-aims-to-build-a-google-for-biomedical-data/

[10] Sandra Heiler. 1995. Semantic interoperability. *Comput. Surveys* 27, 2 (Jun 1995), 271–273. https://doi.org/10.1145/210376.210392

[11] Fidelia Ibekwe-SanJuan and Geoffrey C. Bowker. 2017. Implications of Big Data for Knowledge Organization. *KNOWLEDGE ORGANIZATION* 44, 3 (2017), 187–198. https://doi.org/10.5771/0943-7444-2017-3-187

[12] Sheila Jasanoff and Sang-Hyun Kim. 2015. *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power.* University of Chicago Press. Google-Books-ID: 5XxTCgAAQBAJ.

[13] Cory Philip Knobel. 2010. *Ontic Occlusion and Exposure in Sociotechnical Systems.* Ph.D. Dissertation. University of Michigan.

[14] Matthew T. Mccarthy. 2017. The Semantic Web and Its Entanglements. *Science, Technology and Society* 22, 1 (Mar 2017), 21–37. https://doi.org/10.1177/0971721816682796

[15] Elena Parmiggiani and Vidar Hepsø. 2013. Pragmatic information management for environmental monitoring in oil and gas. *ECIS 2013 Completed Research, Paper 65* (2013).

[16] Trevor J. Pinch and Wiebe E. Bijker. 1984. The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science* 14, 3 (Aug 1984), 399–441. https://doi.org/10.1177/030631284014003004

[17] Lindsay Poirier. 2015. The Stickiness of Difference in the Semantic Web. https://thesocietypages.org/cyborgology/2015/07/20/the-stickiness-of-difference-in-the-semantic-web/

[18] Dave Randall, Rob Procter, Yuwei Lin, Meik Poschen, Wes Sharrock, and Robert Stevens. 2011. Distributed ontology building as practical work. *International Journal of Human-Computer Studies* 69, 4 (Apr 2011), 220–233. https://doi.org/10.1016/j.ijhcs.2010.12.011

[19] David Ribes and Geoffrey C. Bowker. 2009. Between meaning and machine: Learning to represent the knowledge of communities. *Information and Organization* 19, 4 (Oct 2009), 199–217. https://doi.org/10.1016/j.infoandorg.2009.04.001

[20] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.