Scientific Ontologies, Digital Curation and the Learning Knowledge Ecosystem

Janna Hastings[1,2][0000−0002−3469−4923]

[1] Faculty of Computer Science,
Otto-von-Guericke Universität Magdeburg, Germany
hastings@ovgu.de
[2] Department of Clinical, Educational and Health Psychology,
University College London, UK

Abstract. The global coronavirus pandemic has brought another ongoing crisis into the spotlight: that of digital misinformation. While society at a global scale is facing challenges that demand scientific solutions as never before, trust in experts and scientific expertise is falling, and conspiracy theories abound.
At the same time, science itself is not without challenges, such as the reproducibility crisis across multiple domains. A contributing factor to misinformation is the way that scientific research is undertaken and reported in isolated and conflicting units, rather than as a holistic aggregate of information.
In this position paper, I will argue that scientific ontologies and digital curation will be essential tools for transforming how scientific research is conducted and reported to address the problem of misinformation.

Keywords: Ontologies · Scientific Research · Digital Curation · Machine Learning · Knowledge Ecosystems

1 Introduction

The ongoing coronavirus pandemic has brought many pre-existing societal problems into sharper focus. Among them is the pervasive challenge of 'fake news' [19, 30], in particular as it relates to misinformation – and disinformation – about science. In the terminology of a recent *Nature* report, the battle against coronavirus-related misinformation and conspiracy theories is 'epic' [3], necessitating coordinated action on all fronts. The World Health Organisation has repeatedly issued warnings about an 'infodemic' of misinformation. At a time when the need for scientific solutions has never been greater, the level of trust in science – and in 'experts' – is low.

Misinformation affects all disciplines, although it is particularly problematic for health-related information [31], with a case in point the engineered controversy surrounding vaccination and the resulting fall in vaccination uptake – and

resurgence of disease – throughout the developed world [7]. New media technologies are thought to play a key role as they enable rapid transmission of unfiltered information, and proposed solutions therefore emphasise fact-checking and 'inoculation' (e.g. [21, 28]). The importance of researchers themselves actively countering misinformation online has also been emphasized [17]. However, the problem of misinformation about science is not just a problem of new media technologies – nor of public awareness [33]. It is a problem that has grown against a background in which science itself is facing several transformative challenges including widespread reproducibility crises [16], and it is partly reflective of those challenges. Scientific research takes place against a backdrop of incentives, practices and cultures in which research career success and cumulative scientific progress are not always aligned [14], and isolated and implausible findings may be favoured over robust, cumulative and repeatable research. The resulting appearance of fragmentation across research outputs exacerbates the problem of misinformation.

Addressing these challenges requires action on multiple fronts, both societal and technological. Alongside relevant changes to incentives and practices, tools are needed that are able to show an *integrated view* across all existing findings, which is therefore able to contextualise new findings and media reports.

In this position paper, I set out a vision for an comprehensive suite of interacting technological components that bring together semantic technologies and digital curation with very large scale community-developed ontologies as the backbone for a comprehensive *learning knowledge ecosystem* that is robust against deliberate misuse and accidental misinterpretation.

2 Reproducibility and Fragmentation

The reproducibility crisis is a well-known methodological challenge facing scientific research: the results of many scientific studies have proven difficult to replicate on subsequent investigation. It affects multiple domains, including biomedicine [15], psychology [25], behavioural science [8] and neuroscience [27]. It is widely recognised that it will be necessary to harness multiple different strategies to improve the reproducibility of scientific research [24], including improved statistical and methodological procedures, mandatory replication of novel findings, shifting incentives and practices in scientific research, and appropriate use of theory to enable integration and aggregation.

When new scientific findings are published it is expected that they join, extend and accumulate with an existing body of knowledge. However, in practice, it is impractical to gain an overview of the full body of existing research or indeed to know to what extent different findings accumulate or supplant each other. This challenge is exacerbated by the way that different scientific results are reported on in isolation by the media, making it easier for results to be misinterpreted and misrepresented. However, isolated discoveries are being overturned or contradicted all the time, and this can affect whole programmes of research.

To address fragmentation, there is a need to move beyond isolated research findings towards a comprehensive and integrated body of evidence. In neuroscience, for example, differences in analytical workflows have been shown to lead to differences in results even on the same dataset [6], however, meta-analyses across the different results converged on a consensus. It is imperative that we find ways to systematically integrate all findings and evidence. The ability to meaningfully aggregate across studies and to grow the background against which surprising findings are tested is absolutely key to the progress of science as a whole.

It is necessary, but not sufficient, to make data available and open, because a flood of unintegrated and uninterpreted data overwhelms consumers unless they already have the expertise to process and integrate such datasets. Intelligent and continuous integration is needed in order to link data to theory and conclusions, provide overviews and summaries that represent the scientific findings as a whole in a way that is accessible for all consumers including those who are not experts in that field.

Given the volumes of research involved, sophisticated computational support is essential for all aspects of addressing this challenge.

3 The Role of Scientific Ontologies

Scientific ontologies are standardised computable representations of the entities that are the subject matter of scientific investigations in a domain, built on semantic technologies [10]. Scientific ontologies have been used in multiple disciplines, such as biology [1], chemistry [12], behavioural science [23, 13] and medicine [5]. They serve many different purposes, including as indexes on large-scale data resources such as databases and the semantic web, to integrate and compare data across different studies, and to aggregate individual findings into meaningful categories [9].

An ontology standardises the terminology and the categorisation of entities in a domain. Therefore, it needs to be flexible enough to represent the full breadth of research in the domain, as well as remaining extensible, as new entities are suggested in the course of ongoing scientific research. Annotations are associations between an ontology and some data, linking the ontology to what is known, what has been discovered, and, often, *how* it has been discovered. When it serves as a hub for a whole community, an ontology-organised knowledge base provides a view across the whole of what is known in a given field, an integrated synthesis of the available evidence.

A key feature of scientific ontologies when used to facilitate scientific integration is that they include not just a representation of *schematic types* or broad groupings of kinds of entity (such as molecule, gene, behaviour, emotion) but also a detailed representation of, and hierarchical arrangement of, the entities at the level of detail that features in scientific investigations (e.g. *L-dopamine*, *BRCA1*, *hand-washing* and *happiness*). Having a semantic, defined, annotated and hierarchically arranged index for the entities that feature in scientific inves-

tigations enables data about such entities to be integrated across studies and aggregated flexibly and dynamically.

To further address fragmentation and in particular the gaps that develop between different disciplines and theoretical perspectives, theoretical integration and translation within and between disciplines is needed, which requires both explicit formalisation of theories and the mapping of the elements of theories to the elements of ontologies in order to systematically link between theory and evidence [11]. In addition to theoretical integration, it is also important to be able to connect predictive mathematical and computational models using the same ontologies as indexes.

4 The Role of Digital Curation

Scientific ontologies and their association with datasets across databases and the semantic web have to a large extent been created by the careful and meticulous work of human experts. They formulate domain knowledge in computable form, read the literature, and associate ontology terms with relevant results and findings in databases and resources.

Advances in the comprehensiveness and scope of the research results available via open data resources will directly advance scientific research, results and practice as well as reduce the opportunities for media reports to stand in isolation, as such databases provide a background into which novel findings can be integrated and synthesised.

Human resources for digital curation are always limited, thus, innovation in the ways in which curation takes place have the potential to have significant downstream cumulative effects. Such innovations have been proposed in several different directions: involvement of the scientific research community directly in (co) curation of their data; enhancements in tool support; and the development of 'human in the loop' semi-automated artificial intelligence systems.

Models of researcher-involving co-curation enable joint efforts between ontology and database experts and the researchers who publish primary findings to annotate novel research reports. Such approaches have been adopted by e.g. the PomBase database [22] and by Reactome [18].

Typically, the tools that support this work are custom-built for each database and group. Although a few cross-domain tools do exist, they are not yet widely adopted. Tools also exist that are allow researchers themselves to formulate metadata associated with a publication in computable form, for example the ISA suite of metadata editing tools [29] and the Addiction Paper Authoring Tool [32].

The potential for artificial intelligence approaches to support the work of digital curation is enormous. Although results for fully automated curation pipelines are not yet sufficiently reliable and generalisable to be able to replace human expertise in most domains, approaches which support human curation by aiding in filtering, retrieval and organisation ensure that the effort of the human

is used as efficiently as possible, as well as ensuring robust and comprehensive information flows between different producers and platforms.

5 Towards a Learning Knowledge Ecosystem

In the past decade there has been a shift in the medical domain towards a *learning healthcare system*, which aims for a continual interchange between research and practice in medicine, underscored by widespread data integration and sharing between electronic health record systems and clinical research [26].

I suggest that scientific research needs a similar revolution and shift in thinking towards the creation of a systems-wide integrated and comprehensive *ecosystem* for discovery science across domains and disciplines. This ecosystem must be oriented around *knowledge* rather than mere data, which means that it must simultaneously support multiple levels of detail analogous to 'zooming': from a broad overview of all the research on a given topic down to the detailed evidence supporting each finding. It further needs to be actively *learning* in the sense that new findings need to rapidly feed into it, which will be possible at scale only if it is built around artificial intelligence technologies that are able to integrate semantic content with powerful machine learning approaches.

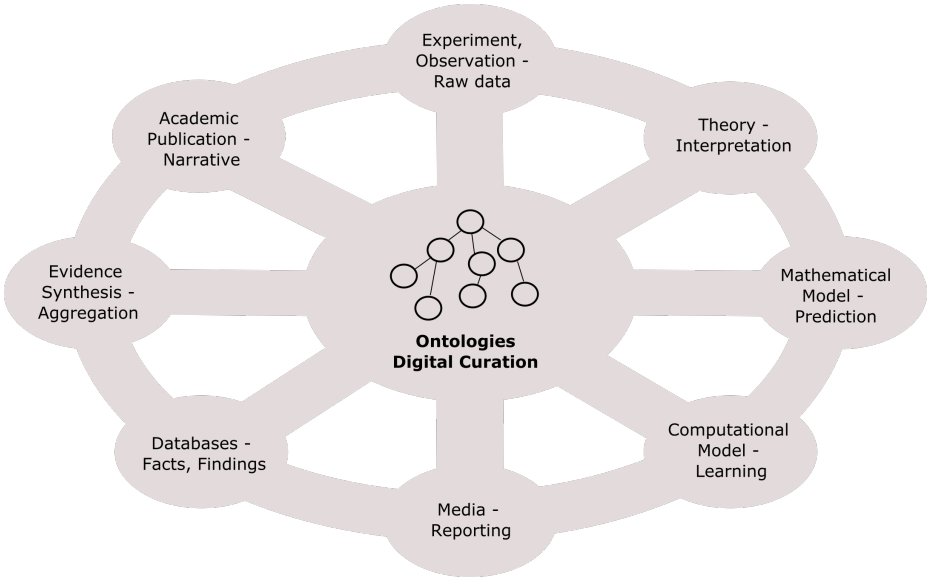

Fig. 1. Towards a Learning Knowledge Ecosystem. The components of the learning knowledge ecosystem are illustrated together with their information flows.

The components of a learning knowledge ecosystem are illustrated in Figure 1, with ontologies and digital curation at the heart of a more robust way of

doing scientific research with digital support. The core role of digitally curated scientific ontologies within this ecosystem is to provide unambiguous semantic shared identifiers as well as to provide a framework for the representation of consensus elements of the domain. They also serve as hubs around which communities can organise consensus-building and participatory activities.

The digitalisation and interconnection of these components - observation, theory, prediction, learning, reporting, aggregation, narrative and databases - is the objective of much of the open science agenda, and is now in place for some topics or subject areas in some domains, but there are many gaps to fill. Moreover, many of the cross-connecting information flows are not yet in place. Thus, researchers or consumers wishing to connect different components have a difficult task at present. This is particularly severe for those who need to work on questions that cross multiple topic areas from multiple disciplines, as different discipline-specific approaches may have idiosyncratic infrastructures that may not be easy to apply in combination.

Figure 1 illustrates interrelationships and flows relating to a given entity or group of entities within the process of knowledge creation and construction, from experiment through interpretation to publication and media reporting. Use of common identifiers and shared semantic representation across these different aspects allows scientific findings, outcomes and aggregate bodies of evidence to be presented *as a whole*, consistently against the same shared background. In turn, the adoption of a shared background facilitates a more grounded media presentation, less susceptible to sensationalism. Stabilising the evidence for accumulation of knowledge via centralisation and exchange allows pockets of uncertainty and contradictions in the evidence base to become more apparent, which paradoxically may serve to increase trust in science overall [4]. It is furthermore important for public trust that the learning knowledge ecosystem be maintained and managed by a plurality of cooperating public, not-for-profit institutions and that a large degree of international cooperation be evident – risk of bias should be actively managed for all participants. No one institute should dominate, nor one country or language.

6 Conclusions

The immediate societal crisis engendered by the coronavirus pandemic may well be solved - in due course - by the progress of science, but the deeper challenges that the pandemic has highlighted will take longer to address.

Scientific ontologies and the informatics technologies that support data curation, storage, exchange and discovery are already transforming research processes and practices. There are exciting new developments in technologies for widespread interlinked scientific data and knowledge representation, such as the Open Research Knowledge Graph [2]. However, even as such efforts gain traction, the need to close the gaps and reduce systemic redundancy and wasted effort becomes more urgent.

One pressing question is how to bring the scientific research community itself directly and actively into the process of curating its own findings into an aggregated whole. To achieve this will involve more than good intentions: powerful incentives are needed to change embedded practices. One possible direction this might take is given by considering one of the drivers of the Gene Ontology's wide-ranging success: scientists will be motivated to contribute to shared knowledge resources if those resources enable them to answer *scientific* questions that would not otherwise be answerable [20].

The vision of a learning knowledge ecosystem, in which the data science of novel discovery is interfaced seamlessly with what is already known, points towards a new era of synthesis after an era of increasing fragmentation. We might call this *knowledge science*.

References

1. Ashburner, M., Ball, C.A., Blake, J.A., et al.: Gene ontology: tool for the unification of biology. the Gene Ontology Consortium. Nature Genetics **25**, 25–29 (May 2000). https://doi.org/10.1038/75556
2. Auer, S.: Towards an Open Research Knowledge Graph. Tech. rep., Zenodo (Jan 2018). https://doi.org/10.5281/zenodo.1157185, `https://zenodo.org/record/1157185#.YASlUVlrzaY`
3. Ball, P., Maxmen, A.: The epic battle against coronavirus misinformation and conspiracy theories. Nature **581**, 371–374 (2020)
4. van der Bles, A.M., van der Linden, S., Freeman, A.L.J., Spiegelhalter, D.J.: The effects of communicating uncertainty on public trust in facts and numbers. Proceedings of the National Academy of Sciences **117**(14), 7672–7683 (2020). https://doi.org/10.1073/pnas.1913678117
5. Bodenreider, O., Cornet, R., Vreeman, D.J.: Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. Yearbook of Medical Informatics **27**(1), 129–139 (Aug 2018). https://doi.org/10.1055/s-0038-1667077
6. Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., et al.: Variability in the analysis of a single neuroimaging dataset by many teams. Nature (2020). https://doi.org/10.1038/s41586-020-2314-9
7. Browne, M.: Epistemic divides and ontological confusions: The psychology of vaccine scepticism. Human vaccines & immunotherapeutics **14**, 2540–2542 (2018). https://doi.org/10.1080/21645515.2018.1480244
8. Camerer, C.F., Dreber, A., Holzmeister, F., et al.: Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour **2**(9), 637–644 (Sep 2018). https://doi.org/10.1038/s41562-018-0399-z, number: 9 Publisher: Nature Publishing Group
9. Haendel, M.A., Chute, C.G., Robinson, P.N.: Classification, Ontology, and Precision Medicine. New England Journal of Medicine **379**(15), 1452–1462 (Oct 2018). https://doi.org/10.1056/NEJMra1615014, `http://www.nejm.org/doi/10.1056/NEJMra1615014`
10. Hastings, J.: Primer on Ontologies. In: Dessimoz, C., Škunca, N. (eds.) The Gene Ontology Handbook, vol. 1446, pp. 3–13. Springer New York, New York, NY (2017), series Title: Methods in Molecular Biology

11. Hastings, J., Michie, S., Johnston, M.: Theory and Ontology in Building Cumulative Behavioural Science (2019), https://osf.io/9te3x/
12. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., Steinbeck, C.: ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research **44**, D1214–D1219 (Jan 2016). https://doi.org/10.1093/nar/gkv1031
13. Hastings, J., Schulz, S.: Ontologies for Human Behavior Analysis and Their Application to Clinical Data. In: International Review of Neurobiology, vol. 103, pp. 89–107. Elsevier (2012). https://doi.org/10.1016/B978-0-12-388408-4.00005-8, https://linkinghub.elsevier.com/retrieve/pii/B9780123884084000058
14. Higginson, A.D., Munafò, M.R.: Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. PLOS Biology **14**(11), e2000995 (Nov 2016). https://doi.org/10.1371/journal.pbio.2000995, https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2000995, publisher: Public Library of Science
15. Ioannidis, J.P.A.: Why Most Published Research Findings Are False. PLOS Medicine **2**(8), e124 (Aug 2005). https://doi.org/10.1371/journal.pmed.0020124, https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124, publisher: Public Library of Science
16. Ioannidis, J.P.A.: Why Science Is Not Necessarily Self-Correcting. Perspectives on Psychological Science **7**(6), 645–654 (Nov 2012). https://doi.org/10.1177/1745691612464056, http://journals.sagepub.com/doi/10.1177/1745691612464056
17. Iyengar, S., Massey, D.S.: Scientific communication in a post-truth society. Proceedings of the National Academy of Sciences **116**(16), 7656–7661 (2019). https://doi.org/10.1073/pnas.1805868115, https://www.pnas.org/content/116/16/7656
18. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P.: The reactome pathway knowledgebase. Nucleic Acids Research **48**(D1), D498–D503 (2020). https://doi.org/10.1093/nar/gkz1031
19. Lazer, D.M.J., Baum, M.A., Benkler, Y., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018). https://doi.org/10.1126/science.aao2998, https://science.sciencemag.org/content/359/6380/1094
20. Lewis, S.E.: The Vision and Challenges of the Gene Ontology. In: Dessimoz, C., Škunca, N. (eds.) The Gene Ontology Handbook, pp. 291–302. Methods in Molecular Biology, Springer, New York, NY (2017)
21. van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., Lewandowsky, S.: Inoculating against misinformation. Science **358**(6367), 1141–1142 (2017). https://doi.org/10.1126/science.aar4533, https://science.sciencemag.org/content/358/6367/1141.2
22. Lock, A., Harris, M.A., Rutherford, K., Hayles, J., Wood, V.: Community curation in PomBase: enabling fission yeast experts to provide detailed, standardized, sharable annotation from research publications. Database **2020** (04 2020). https://doi.org/10.1093/database/baaa028, https://doi.org/10.1093/database/baaa028, baaa028
23. Michie, S., West, R., Finnerty, A.N., Norris, E., Wright, A.J., Marques, M.M., Johnston, M., Kelly, M.P., Thomas, J., Hastings, J.: Representation of behaviour change interventions and their evaluation: Development of the Upper

Level of the Behaviour Change Intervention Ontology. Wellcome Open Research **5**, 123 (Jun 2020). https://doi.org/10.12688/wellcomeopenres.15902.1, https://wellcomeopenresearch.org/articles/5-123/v1

24. Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.J., Ware, J.J., Ioannidis, J.P.A.: A manifesto for reproducible science. Nature Human Behaviour **1**(1), 0021 (Jan 2017). https://doi.org/10.1038/s41562-016-0021, http://www.nature.com/articles/s41562-016-0021
25. Open Science Collaboration: Estimating the reproducibility of psychological science. Science **349**(6251), aac4716–aac4716 (Aug 2015). https://doi.org/10.1126/science.aac4716, https://www.sciencemag.org/lookup/doi/10.1126/science.aac4716
26. Platt, J.E., Raj, M., Wienroth, M.: An Analysis of the Learning Health System in Its First Decade in Practice: Scoping Review. Journal of Medical Internet Research **22**(3), e17026 (2020). https://doi.org/10.2196/17026, https://www.jmir.org/2020/3/e17026/, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada
27. Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.B., Vul, E., Yarkoni, T.: Scanning the horizon: towards transparent and reproducible neuroimaging research. Nature Reviews Neuroscience **18**(2), 115–126 (Feb 2017). https://doi.org/10.1038/nrn.2016.167, https://www.nature.com/articles/nrn.2016.167, number: 2 Publisher: Nature Publishing Group
28. Roozenbeek, J., van der Linden, S.: Fake news game confers psychological resistance against online misinformation. Palgrave Communications **5**(1), 65 (2019). https://doi.org/10.1057/s41599-019-0279-9, https://doi.org/10.1057/s41599-019-0279-9
29. Sansone, S.A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.A., Copeland, J., Das, S., Daruvar, A.d., Matos, P.d., Dix, I., Edmunds, S., Evelo, C.T., Forster, M.J., Gaudet, P., Gilbert, J., Goble, C., Griffin, J.L., Jacob, D., Kleinjans, J., Harland, L., Haug, K., Hermjakob, H., Sui, S.J.H., Laederach, A., Liang, S., Marshall, S., McGrath, A., Merrill, E., Reilly, D., Roux, M., Shamu, C.E., Shang, C.A., Steinbeck, C., Trefethen, A., Williams-Jones, B., Wolstencroft, K., Xenarios, I., Hide, W.: Toward interoperable bioscience data. Nature Genetics **44**, 121–126 (2012)
30. Scheufele, D.A., Krause, N.M.: Science audiences, misinformation, and fake news. Proceedings of the National Academy of Sciences of the United States of America **116**, 7662–7669 (Apr 2019). https://doi.org/10.1073/pnas.1805871115
31. Wang, Y., McKee, M., Torbica, A., Stuckler, D.: Systematic literature review on the spread of health-related misinformation on social media. Social science & medicine (1982) **240**, 112552 (Sep 2019). https://doi.org/10.1016/j.socscimed.2019.112552
32. West, R.: An online Paper Authoring Tool (PAT) to improve reporting of, and synthesis of evidence from, trials in behavioral sciences. Health Psychology **39**(9), 846–850 (2020). https://doi.org/10.1037/hea0000927, place: US Publisher: American Psychological Association
33. Wynne, B.: Public engagement as a means of restoring public trust in science - hitting the notes, but missing the music? Community Genet **9**, 211–220 (2006)