Article:

Webb, T.L. orcid.org/0000-0001-9320-0068, Baird, H.M., Maikore, F.S. et al. (5 more authors) (2025) A method for evaluating the interoperability of ontology classes in the behavioural and social sciences [version 1; peer review: awaiting peer review]. Wellcome Open Research, 10. 525. ISSN: 2398-502X

https://doi.org/10.12688/wellcomeopenres.24206.1

Check for updates

RESEARCH ARTICLE

A Method for Evaluating the Interoperability of Ontology Classes in the Behavioural and Social Sciences

[version 1; peer review: awaiting peer review]

Thomas L. Webb[1], Harriet M. Baird[1], Fatima S. Maikore[2], Robert West[3], Janna Hastings[4-6], Suvodeep Mazumdar[7], Vitaveska Lanfranchi[2], Susan Michie[8]

[1]School of Psychology, The University of Sheffield, Sheffield, UK
[2]School of Computer Science, The University of Sheffield, Sheffield, UK
[3]University College London Department of Behavioural Science and Health, London, UK
[4]Institute for Implementation Science in Health Care, Faculty of Medicine, University of Zurich, Zurich, Switzerland
[5]University of St Gallen School of Medicine, St. Gallen, St. Gallen, Switzerland
[6]Swiss Institute of Bioinformatics, Lausanne, Vaud, Switzerland
[7]School of Information, Journalism and Communication, The University of Sheffield, Sheffield, UK
[8]University College London Centre for Behaviour Change, London, UK

v1 **First published:** 22 Sep 2025, **10**:525
https://doi.org/10.12688/wellcomeopenres.24206.1

Latest published: 22 Sep 2025, **10**:525
https://doi.org/10.12688/wellcomeopenres.24206.1

Abstract

Background

Ontologies are frameworks for representing information that promote clarity, consistency and coherence, reduce the fragmentation of knowledge, and allow datasets and knowledge to be linked across studies, disciplines and domains. To enable this, it is important to identify how concepts of interest ('classes') are represented in different ontologies and evaluate the extent to which such classes align (i.e., are 'interoperable'). This study aims to provide a method for doing this.

Methods

An automated tool using Meta's Llama 3 language model was developed and used to compare artificial intelligence (AI) and human approaches to matching ontology classes. The automated tool was then integrated into a hybrid method for identifying classes that appear to refer to the same thing across pairs of ontologies. The method was evaluated by three behavioural scientists who used it to identify similar classes in two ontologies and provided feedback on

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

their experience.

Results

The automated tool identified a larger number of potential matches than human-led review, so was used to generate a shortlist. The evaluation of the method produced mixed results. Users agreed which classes were identical or essentially the same across contexts, but none of the users identified similar classes that could be imported into an ontology without causing a contradiction or conflict. Users typically found using the method difficult, but many of the challenges related to using ontologies, rather than to the method specifically.

Conclusions

A combination of automated and human processes appears to be a feasible way to assess the interoperability of ontology classes. While further refinement is needed along with tools and resources that enable the use of ontologies by a broad range of researchers, the study provides a workable method for matching ontology classes in the behavioural and social sciences and offers a practical guide to support its implementation.

Plain language summary

Ontologies are frameworks that describe and organise knowledge and help researchers to connect knowledge across studies in a clear and structured way. In the behavioural and social sciences, multiple ontologies often describe related concepts (e.g., characteristics of behaviour or interventions designed to change behaviour), but these ontologies are not always compatible or easy to combine. This makes it difficult for researchers to use and share data effectively across different studies and disciplines.

This study developed an automated tool that uses a form of Artificial Intelligence (AI) to identify potentially similar things (e.g., types of behaviour or types of interventions). These are called ‘classes’ in ontologies. The classes identified by the automated tool were compared to the classes identified by experts and it was decided to use the automated tool as part of a method that helps researchers in behavioural and social sciences to:

1. Identify relevant ontologies for their work.

2. Determine whether these ontologies share a common structure.

3. Use the automated tool to find similar classes across ontologies.

4. Manually review the suggested matches to confirm whether they align.

5. Compare different researchers’ assessments to improve

accuracy.

We tested this method with behavioural scientists and found that users experienced challenges with some aspects of the method, especially when dealing with complex ontology structures. Based on their feedback, we improved the method and developed a step-by-step guide to help researchers integrate it into their work.

By improving how ontology classes are compared and linked, this method supports the integration of data across studies, disciplines and topic areas. This helps behavioural and social science researchers work more effectively.

Keywords
Ontology, interoperability, matching, classes

This article is included in the Human Behaviour-Change Project (including the APRICOT project) gateway.

Corresponding author: Thomas L. Webb (t.webb@sheffield.ac.uk)

Author roles: Webb TL: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Baird HM**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Maikore FS**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **West R**: Conceptualization, Methodology, Writing – Review & Editing; **Hastings J**: Conceptualization, Writing – Review & Editing; **Mazumdar S**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Lanfranchi V**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Michie S**: Conceptualization, Funding Acquisition, Writing – Review & Editing

Competing interests: RW and SM are unpaid directors of the Unlocking Behaviour Change Community Interest Company. RW undertakes training and consultancy the Freuds+ communications agency, Everyone Health, and Public Health Wales.

Grant information: This work was supported by Wellcome [201524] through a collaborative award to the Human Behaviour-Change Project, the Economic and Social Research Council under grant [ES/Y001044/1], and the APRICOT grant from the US National Institutes of Health [1U01CA291884-01].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Webb TL, Baird HM, Maikore FS *et al.* **A Method for Evaluating the Interoperability of Ontology Classes in the Behavioural and Social Sciences [version 1; peer review: awaiting peer review]** Wellcome Open Research 2025, **10**:525 https://doi.org/10.12688/wellcomeopenres.24206.1

First published: 22 Sep 2025, **10**:525 https://doi.org/10.12688/wellcomeopenres.24206.1

Introduction

Ontologies are frameworks for representing information that promote clarity, consistency, and coherence – things that are often lacking in the behavioural and social sciences (e.g., Krpan *et al.*, 2025). They do this by using a common language to describe things (referred to as 'classes') and organising information in a common structure. Ontologies can thus help to link information across data sets and study reports across academic disciplines and topics (Baird *et al.*, 2023; National Academies of Sciences, Engineering, and Medicine *et al.*, 2022).

It is not feasible for a single ontology to represent every class that is relevant to every topic in the behavioural and social sciences, and so different ontologies have been developed that focus on different topics. For example, the Behaviour Change Intervention Ontology (BCIO, Michie *et al.*, 2021) focuses on classes relevant to behaviour change interventions, while the Ontology for Modeling and Representation of Social Entities (OMRSE, Hicks *et al.*, 2016) focuses on classes that represent human social interactions, such as social acts, social roles, social groups, and organizations. However, while ontologies can focus on different topics, they will often share classes of entities (e.g., both the BCIO and OMRSE ontologies have classes reflecting people's identity) and so should ideally be able to be used together.

The extent to which ontologies – and ontology classes – can be used together in the same activity without causing an inconsistency or conflict, is referred to as interoperability. In practice, this means that they are either identical (i.e., have the same Internationalized Resource Identifier or IRI), contextually exchangeable, or useably consistent. Contextually exchangeable means that an ontology class is judged to be interchangeable with a class in another ontology for a given set of purposes. For example, a researcher interested in identifying the beliefs that are associated with a health behaviour may view BCIO:006154: 'self-efficacy belief for a behaviour' from the BCIO as contextually exchangeable with SCTID:405100003: 'Health belief: perceived control' from the SNOMED CT (International Edition). Note that unless two classes are identical in every respect, then there will be some subjectivity in what is considered exchangeable, and this will depend on the purpose or purposes for which the two classes are being used. For example, a researcher interested in how different theoretical frameworks characterise control beliefs might not deem the two classes referenced in the example above to be contextually exchangeable. The property 'Is contextually exchangeable with' therefore falls between two existing properties (from the Simple Knowledge Organization System [SKOS] vocabulary): (i) skos:closematch (i.e., indicating that two classes share some meaning but may have some differences in scope, connotation, or application) and (ii) skos: exactMatch (i.e., indicating that two classes mean exactly the same thing and can be used interchangeably).

'Useably consistent' means that an ontology class could be imported into an ontology without causing a contradiction or conflict. In practice, this means that two classes from different ontologies might be considered children and / or parent classes of one another, and neither they nor their ancestor classes should use the same label to represent different concepts across the ontologies. For example, ENVO:00000469: 'research facility' is not useably consistent with OMRSE:00000102: 'healthcare facility' because the former has ENVO:03501288: 'facility' as a parent while the latter has OMRSE:00000062: 'facility' as a parent, and these two parent classes are defined differently and have different ancestor classes in the wider ontology. In contrast, an entity from the Contextualised and Personalised Physical Activity and Exercise Recommendations (COPPER) ontology (Braun *et al.*, 2024b), COPPER:4003: 'affirm commitment despite barriers BCT' could be imported into the BCIO because COPPER uses the same structure to define Behaviour Change Techniques (BCTs) as the BCIO and COPPER:4003 is a child of a parent class from the BCIO (BCIO:007015: 'Affirm commitment BCT').

Existing efforts to match ontologies

Given the scale and complexity of ontologies, manually comparing two ontologies, especially those comprising hundreds or thousands of classes, is not only time consuming and labour intensive, but also prone to errors and inconsistencies. Large language models (LLMs) are revolutionising ontology matching by leveraging the power of natural language processing to enhance both the accuracy and efficiency of the matching process. For example, the LLMs4OM tool (https://github.com/HamedBabaei/LLMs4OM) employs retrieval and matching algorithms like BERT and RAG to identify semantically similar classes based on textual attributes. Key attributes extracted from each ontology class include IRI, label, subclasses, parent classes, synonyms, and comment. LLMs4OM follows a staged approach to generate matches between ontologies:

1. Retrieval: Potential matches are identified by encoding class attributes using models like BERT.
2. Matching: Further analysis is conducted, often with RAG, which combines retrieval and generation to refine matches.
3. Post-Processing: Matches are refined based on thresholds for confidence and similarity.
4. Evaluation: The tool compares generated matches against a reference file containing expected matches, providing detailed evaluation results such as match counts and performance scores.

Evaluation of the LLMs4OM tool using 20 datasets has shown promise (Giglou *et al.*, 2024), but adapting LLMs4OM for other ontologies (i.e., outside the 20 ontology matching datasets tested with the tool), poses challenges due to its reliance on built-in evaluation and matching structures. Specifically, the tool requires a reference file for evaluation, which may not be available for some projects, necessitating code modifications to bypass this step. Without evaluation, LLMs4OM's default output focuses on evaluation metrics rather than directly providing matched pairs, requiring further adjustments to display or save the matches.

Additionally, LLMs4OM does not prioritise attributes like the 'definition' (IRI: http://purl.obolibrary.org/obo/IAO_0000115). Several ontologies in the behavioural and social sciences, including BCIO and OMRSE, use this attribute to provide textual definitions of classes. Integrating this attribute into the matching process requires additional coding. Taken together, the lack of reference files, the need for specific outputs, and the emphasis on certain attributes would require significant alterations to the code to make the LLMs4OM tool work effectively for ontologies in the behavioural and social sciences.

The OLaLa system (Hertling & Paulheim, 2023) employs LLMs to facilitate ontology matching by effectively processing natural language information. This method seeks to improve the accuracy and efficiency of ontology matching tasks through a series of steps, including the creation of optimised prompts, the integration of knowledge graph information, the selection of appropriate LLMs, the provision of pre-existing correspondences, and the generation of candidate matches.

Agent-OM (Qiang *et al.*, 2023) presents a proof-of-concept framework for ontology matching based on LLM agents. The architecture comprises two Siamese agents for retrieval and matching, augmented by prompt-based ontology matching tools. Evaluation results from various Ontology Alignment Evaluation Initiative tracks demonstrate that Agent-OM achieves competitive performance on standard ontology matching tasks and significantly improves performance in complex and few-shot settings.

There are however disadvantages in employing LLMs for ontology matching, including that LLMs can be computationally expensive to train and deploy, requiring significant energy, water and infrastructure (Garg *et al.*, 2025). The performance of LLMs also depends on the quality and coverage of their training data. If the training data is incomplete or biased, then the LLM may struggle to generalise. In addition, the decision-making processes of LLMs are not transparent or easily explained, making it difficult to understand why a particular correspondence was identified. Furthermore, while LLMs can capture complex linguistic patterns, they may lack human reasoning abilities, leading to errors in matching concepts that require real-world knowledge, understanding and/or contextualisation. Taken together, it is possible that a combination of automated matching using LLMs and human review may provide an appropriate balance between efficiency and accuracy when evaluating the interoperability of ontology classes, enabling researchers to focus on validating results rather than manually searching for potential matches.

The present research

The present study aimed to develop guidance for researchers in the behavioural and social sciences who want to identify how concepts of interest are represented in different ontologies and evaluate the extent to which such classes are interoperable. This involved developing an automated tool that uses LLMs to compare classes between pairs of ontologies, producing a short list (the 'common entities list') for subsequent human review. The automated tool was developed iteratively and compared to manual (human-led) methods using the BCIO and OMRSE as an initial test case. An initial guide for implementing a method for evaluating the interoperability of ontology classes that incorporates the automated tool was then produced and a team of behavioural scientists evaluated the guide. Their outputs were assessed for accuracy, consistency, and the ease with which they were able to implement the method. Finally, the guide was improved based on the feedback.

Method

Protocols for the project as a whole and evaluation of the guide were pre-registered on the Open Science Framework (https://doi.org/10.17605/OSF.IO/4T65G) and ethical approval was sought and provided by the research ethics committee in the School of Psychology at the University of Sheffield (Reference Number 059430) prior to starting the study (approved 11th April 2024).

Development of an automated tool to compare classes

We developed an automated tool consisting of three Python scripts that handle the three different steps of the class matching process, as described below.

STEP 1. Class IRI matching

The first step involves identifying classes in two ontologies that share the same IRI. An IRI is a unique identifier assigned to each class in an ontology and, if two ontologies use the same IRI for a class, then it indicates that they refer to the same concept and indicates where the two ontologies already align. To automate this process, we developed a Python script that compares the IRIs in two ontologies and outputs the results into a CSV file. This file lists the matching IRIs, along with the corresponding labels and definitions from both ontologies.

STEP 2. Class label and synonym matching

While IRI matching identifies exact overlaps, it does not identify where two ontologies represent the same concept using different IRIs or terminology. For example, the class "walking" is represented in BCIO with the IRI BCIO:036108, while in COPPER, it has the IRI COPPER:1020. To address this, we developed a second Python script that identifies potentially equivalent classes based on their labels and synonyms.

For this step, the automated tool leverages ontology properties from the Information Artifact Ontology (IAO) which defines the properties "alternative label" (IAO:0000118), "has exact synonym" (oio:hasExactSynonym), "has broad synonym" (oio:hasBroadSynonym), "has narrow synonym" (oio:hasNarrowSynonym), and "has related synonym" (oio:hasRelatedSynonym). By examining these properties, the automated tool can detect cases where one ontology uses a different term or phrasing to describe a concept that appears elsewhere under another name.

The process begins by extracting all class labels and synonyms from both ontologies, excluding classes that share the same IRI. The automated tool then compares the class labels, flagging exact matches. For classes where the labels do not match, it checks whether a term in one ontology appears as a synonym in the other. This allows for the detection of meaningful connections even when different terms are used. Once potential matches are identified, the tool records the corresponding class IRIs, labels and synonym properties in a structured format as a CSV file.

STEP 3. Class definition matching

The IRI and label-based matching conducted in Steps 1 and 2 do not account for the underlying meanings of the classes, which can result in semantically similar concepts being overlooked. For example, “Abusive behaviour” (OMRSE:00000243) and “Harmful behaviour to others” (BCIO:050398) describe potentially similar concepts but use different labels and IRIs and so would be overlooked in Steps 1 and 2. Label-based matching may also lead to incorrect matches when two classes share a similar label but have different meanings. For instance, “Cold” in the Medical Dictionary for Regulatory Activities Terminology (MedDRA) refers to Nasopharyngitis; an upper respiratory tract infection (see MEDDRA:10009851), while in the National Cancer Institute Thesaurus (NCIT), “Cold” means “having less heat energy than the object against which it is compared; the absence of heat” (see NCIT:C62180). In such cases, relying solely on label similarity could produce false positives.

To provide a richer semantic description of classes, ontology developers typically include textual definitions using properties such as “definition” (IAO:0000115 or rdfs:definition), following best practices outlined in principles such as those set out by the OBO Foundry (https://obofoundry.org/principles/fp-000-summary.html). These principles emphasize the importance of well-defined class meanings to ensure clarity, interoperability, and reuse across ontologies. However, despite these best practices, variations in how definitions are phrased can still create problems when trying to align ontologies.

To address these limitations, we designed a third, more advanced Python script, implementing a quantized[1] version of the Llama 3 language model to compare textual definitions of classes across the two ontologies. This step analyses the semantic content of definitions, allowing for more precise identification of equivalent or closely related classes.

The initial implementation of the language model-based approach was designed to compare every class definition in one ontology (e.g., the BCIO) against every class definition in a second ontology (e.g., OMRSE). However, while this approach ensured thorough coverage, it quickly became computationally impractical when trying to compare large ontologies. For example, comparing 1,000 classes from one ontology against 1,000 classes in another, results in 1,000,000 pairwise comparisons, requiring significant computational resources and time. On a MacBook Pro with an M3 Max chip (36GB memory), comparing one class “education process” (SDGIO:00010001) from the BCIO against 669 classes from OMRSE took 369 seconds, meaning that a full comparison of 1,000 x 1,000 classes would take around 551,550 seconds (i.e., over 6 days of continuous processing). This made the approach unsuitable for practical use.

One potential solution to this challenge is to use high-performance computing (HPC) systems or cloud-based multi-GPU clusters. Such systems, equipped with AI accelerators like NVIDIA H100 GPUs or TPU clusters, could significantly reduce processing time, potentially completing the full-scale comparison in under an hour. However, access to HPC resources is often limited, expensive, carbon intensive and not feasible for the average user of behavioural research. Therefore, to improve efficiency, we refined the automated tool to focus on comparing a single target class from one ontology against all classes in another. This targeted approach significantly reduced the number of comparisons while maintaining accuracy. It also aligns more closely with real-world use cases, where researchers are likely to compare ontologies to find matching classes for a specific concept that, for example, they aim to annotate in a dataset or paper. By focusing on one class at a time, the automated tool became significantly faster, reducing processing time from days to minutes per comparison task, and making it more accessible for users without HPC resources.

Evaluation of the automated tool

The automated tool was evaluated by an empirical comparison with human approaches to matching ontology classes. We focused the evaluation of the automated tool on the language model-based approach for matching class definitions (i.e., Step 3), as the other two components—IRI matching and label and synonym matching—were straightforward and did not require extensive testing.

TEST 1: Identifying matches in the OMRSE for three (broad) classes from the BCIO

In the first test a member of the research team (TLW) manually reviewed the OMRSE ontology to identify potential matches for three target BCIO classes: BCIO:050437: ‘health-related behaviour’, BCIO:006017: ‘beliefs about barriers to behaviour’, and ADDICTO:0000381: ‘identity’. We then investigated whether the automated tool could identify the same and/or additional similar classes. Table 1 summarises the number of matches identified. The full list of matched classes from the manual method and automated tool are provided in Table SX on the Open Science Framework (see the Data Availability statement).

The manual approach identified two potential matches for ADDICTO:0000381: ‘identity’ in OMRSE — OMRSE:00000204: ‘social identity information content entity’ and OMRSE:00000278: ‘social identity’ —but found no matches for the other two BCIO classes (BCIO:050437: ‘health-related behaviour’ and BCIO:006017: ‘beliefs about barriers to behaviour’). By contrast, the automated tool identified more matches overall,

[1] ‘Quantization’ is a technique to reduce the size of large language models by converting high precision values of model weights and activation functions to types of data with lower precision.

Table 1. Number of classes in OMRSE identified as potential matches to target classes from the BCIO by manual and automated approaches.

Target BCIO Class ID	Target BCIO Class Label	Matches in OMRSE identified by manual method	Matches in OMRSE identified by automated tool
BCIO:050437	Health-related behaviour	0	2
BCIO:006017	Belief about barriers to a behaviour	0	0
ADDICTO:0000381	Identity	2	28

identifying (i) two matches for BCIO:050437: 'health-related behaviour' - OMRSE:00000039: 'smoker role' and OMRSE:00000161: 'leaving a health care facility after receiving care', and (ii) 28 matches for ADDICTO:0000381: 'identity'. However, some of these were from the upper-level ontology IAO (e.g., IAO:0000708: 'ORCID identifier'), prompting discussion about whether such broad conceptual matches should be included in ontology alignment. Additionally, several of the 28 matches were subclasses of already identified classes. For example, OMRSE:00000205: "Ethnic identity information content entity" and OMRSE:00000209: "Gender identity information content entity" are both subclasses of OMRSE:00000204: "Social identity information content entity". This indicates that the tool was not only identifying direct matches but also more granular, specific instances of the concept. This raised further questions about whether subclass relationships should be treated as valid matches, or if the automated tool should be adjusted to prioritize only direct class-level alignments. The automated tool did not identify any matches for BCIO:006017: 'beliefs about barriers to behaviour', consistent with the manual method.

The findings of the first test highlight both the strengths and limitations of the automated tool. While the tool successfully identified all manual matches, it also identified a broader set of related but potentially less relevant matches, particularly by identifying subclasses or indirectly related concepts and including upper-level ontology classes.

TEST 2: Testing the automated tool on three pairs of classes from the BCIO and OMRSE that have similar meanings, but are described in different terms

To further assess the automated tool, we tested whether it could identify class matches that had already been established through manual review. Three pairs of classes from the BCIO and OMRSE deemed by the research team to be conceptually similar were selected (see Table 2). Each pair of classes described the same idea but using slightly different terminology and so may present a challenge for an automated tool. Each target class from OMRSE was entered into the automated tool, which then searched for potential matches within BCIO. The process was also conducted in the opposite direction, where target classes from BCIO were checked for matches in OMRSE. The results from both directions were then compared against the matches identified manually. Table 2 summarises the number of matches identified by the automated tool for each of the pairs of classes. Details of each matched class are provided in Table SY on the Open Science Framework (see the Data Availability statement).

For SDGIO:00010001: 'Education process', the tool identified 12 matches, including OMRSE:00002032: 'Completing a program of education', which was deemed a close match by the research team. For BCIO:050398: 'Harmful behaviour to others', the tool found three matches: OMRSE:00000243: 'Abusive behavior' and two of its subclasses (OMRSE:00000245: 'Physically abusive behavior' and OMRSE:00000244: 'Psychologically abusive behavior'). However, OMRSE:00000246: 'Sexually abusive behavior' was not retrieved, suggesting that the tool does not always capture all subclasses of a matched class as initially expected. For BCIO:010013: 'Nursing professional', the tool unexpectedly matched OMRSE:00002059: 'Caregiver role' rather than the anticipated OMRSE:00000014: 'Nurse role'. For OMRSE:00002032: 'Completing a program of education', the tool found 42 matches, including SDGIO:00010001: 'Education process'. However, OBI:0000011: 'Planned process' appeared multiple times due to having multiple labels and definitions, though it was counted only once in the final analysis. For OMRSE:00000243: 'Abusive behaviour', the tool found 13 matches, including BCIO:050398: 'Harmful behaviour to others'. For OMRSE:00000014: 'Nurse role', the tool identified 9 matches, including BCIO:010013: 'Nursing professional'.

One key finding was that identifying a match in one direction did not always guarantee that the automated tool would identify the same match in the reverse direction. While the tool successfully identified some of the expected matches, it also retrieved additional matches, often subclasses or conceptually related terms. This suggests that the automated tool captures broader relationships but may not always align with human expectations.

Another difference was how definitions were handled. Manual review using the Ontology Lookup Service (OLS) included both class definitions and associated comments (from 'rdfs:comment' property) as these were not differentiated in OLS, whereas the automated tool considered only the IAO:0000115:

Table 2. Number of classes in OMRSE identified as potential matches to target classes.

Target class ID	Target class label	Matching direction	Matches identified by automated tool
SDGIO:00010001	Education process	BCIO → OMRSE	12
BCIO:050398	Harmful behaviour to others	BCIO → OMRSE	3
BCIO:010013	Nursing professional	BCIO → OMRSE	1
OMRSE:00002032	Completing a program of education	OMRSE → BCIO	42
OMRSE:00000243	Abusive behaviour	OMRSE → BCIO	13
OMRSE:00000014	Nurse role	OMRSE → BCIO	9

'definition' property. To assess the impact of only considering the 'definition' property, we tested the script with both the 'definition' and 'comment' properties. As expected, processing longer definitions took significantly more time, and fewer matches were identified—always a subset of those found using definitions alone. For BCIO:050398: 'Harmful behaviour to others', including comments had no effect. Based on these findings, we decided to exclude comments from both manual and automated matching. Comments often provide context for ontology users and developers, such as modelling decisions, rather than defining the class itself.

TEST 3: Identifying matches in the BCIO for three sub-classes from the OMRSE

For this test, two members of the research team (TLW and HMB) manually reviewed the list of classes in the BCIO to identify potential matches to three target classes from the OMRSE. Their results were then compared to the potential matches identified by the automated tool. The three target classes were OMRSE:00000012: 'Healthcare provider role', OMRSE:00000011: 'Patient role', and OMRSE:00002068: 'Communication'. TLW and HMB identified the same classes in the BCIO as matches (the findings are presented in Table SZ on the Open Science Framework – see the Data Availability statement)).

The automated tool identified 29 matches for OMRSE:00000012: 'Health care provider role' but failed to identify BCIO:010008: 'Health professional' as one of the matches. It identified 9 matches for OMRSE:00000011: 'Patient role' and 16 matches for OMRSE:00002068: 'Communication', including the matches identified by TLW and HMB.

These results suggest that the automated tool successfully identified many relevant matches, particularly by capturing variations in terminology and broader semantic relationships. However, the failure to match OMRSE:00000012: 'Health care provider role' with BCIO:010008: 'Health professional' highlights a limitation—likely due to differences in how roles and professions are defined within the respective ontologies. The definition of 'health care provider role' within OMRSE focuses on the role itself, describing it as something that inheres in an individual or organization and is realized through the act of providing healthcare services. In contrast, the definition of 'Health professional' provided by the BCIO describes a professional category, highlighting the development and application of medical and health-related knowledge rather than the act of providing care. These conceptual differences in how the ontologies frame roles versus professional identities may have caused the automated tool to perceive them as distinct rather than equivalent.

Additionally, some of the matches found by the automated tool were conceptually related but not necessarily directly equivalent. This aligns with findings from previous tests, where the tool retrieved a mix of exact, broader, and narrower matches, sometimes missing key domain-specific relationships. Taken together the findings of Test 3 underscore the importance of human oversight to carefully evaluate the matches, particularly when the ontology structures differ significantly.

Further refinements to the automated tool

Based on the results of the tests, we refined the automated tool to address inefficiencies and false positives, improving accuracy. Some changes were driven by testing, while others enhanced relevance by focusing on key ontology classes. These refinements reduced computational load and ensured that the language model was applied where it added the most value. The following refinements were made:

Filtering upper-level ontology classes

We excluded upper-level ontology classes by adding a filter to ensure that classes from the Basic Formal Ontology (BFO) and the Information Artifact Ontology (IAO) were not included in the comparisons. These foundational ontologies provide high-level categories rather than domain-specific concepts, making them unsuitable for direct class matching. Removing them ensured that the comparison focused on meaningful, domain-level classes.

Avoiding redundant comparisons

Since Steps 1 and 2 had already identified classes with identical IRIs, matching labels, and synonyms, we excluded these cases from the comparisons made by the automated tool using the language model. This ensured that Step 3 focused solely on comparing textual definitions, rather than repeating matches already found through previous steps.

Pre-filtering definitions before applying the language model

To further improve efficiency, we introduced a pre-filtering process before applying the language model to the automated tool. Initially, we experimented with SequenceMatcher, a method that compares definitions based on shared character sequences. However, this method produced many false positives, as it does not consider actual meaning—it simply detects overlapping words. This meant that sentences with similar wording, but different meanings, were mistakenly marked as matches. For example, BCIO:006081: 'personal role' and OMRSE:00000069: 'burn patient role' were marked as similar. While both definitions contain the phrase *"a role that inheres in"*, their actual meanings differ significantly. Recognizing the limitations of SequenceMatcher, we replaced it with a cosine similarity check, which measures how semantically related two definitions are. This method better captures conceptual similarity rather than just textual overlap, ensuring that only definitions with meaningful relationships were passed onto the language model for final comparison.

Focusing on native ontology classes

Another refinement was the exclusion of imported classes, ensuring that comparisons were only made between native ontology classes—that is, classes originally defined within the ontology rather than those imported from external sources. This step helped to eliminate redundancy and ensured that we evaluated only the core content of each ontology rather than duplicating comparisons involving reused classes.

Integrating the automated tool for comparing class definitions into a guide for researchers

The automated tool for comparing class definitions was integrated into a guide designed to help researchers identify how concepts of interest are represented in different ontologies and evaluate the extent to which such classes are interoperable. The guide detailed a series of stages including (i) identifying relevant ontologies, (ii) establishing that the ontologies share the same upper-level ontology, (iii) creating a list of classes in pairs of ontologies that appear to refer to the same thing, (iv) identifying classes that are (a) identical (i.e., classes that have the same IRI), (b) contextually exchangeable and (c) useably consistent, (v) comparing assessments between users, and (vi) identifying cross-references. The version of the guide used for the evaluation reported below (Version 3.0) can be viewed on the Open Science Framework (see the data availability statement).

Piloting the guide for evaluating the interoperability of ontology classes

Two members of the research team (TLW and HMB) piloted version 3.0 of the implementation guide by using it to identify identical, contextually exchangeable, and useably consistent classes representing behaviour in the Human Behaviour Ontology (HBO, Schenk *et al.*, 2024), and the Contextualised and Personalised Physical Activity and Exercise Recommendations (COPPER) Ontology (Braun *et al.*, 2024b). The pilot suggested that it was suitable for evaluation by researchers outside the project.

Evaluation of the guide for evaluating the interoperability of ontology classes in the behavioural and social sciences

The formal evaluation involved asking three behavioural scientists who had collaborated with the research team on a previous project (Scott *et al.*, 2022) and were familiar with ontologies but had not worked on interoperability to apply the method to the HBO and COPPER ontologies. Participants were provided with information about the research prior to taking part and provided written informed consent via email. We compared the classes that each user identified in each category (i.e., identical, contextually exchangeable, and useably consistent classes) and measured users' experiences of the process via a short questionnaire. The experiences of the three users were assessed in the following ways:

Self-reported ability and confidence in using ontologies

Users' perceived ability and confidence in using ontologies was measured by asking how proficient they felt in the use of ontologies before they used the guide. The options were: (i) I had not used or come across ontologies at all, (ii) I had heard of ontologies but not used them in my work, (iii) I had used them to a small extent, (iv) I had used them a fair amount but had not contributed to any, and (v) I had contributed to or developed one or more ontologies.

Usability of the guide

The usability of the guide was measured using a modified version of the System Usability Scale (SUS, Brooke, 1996). This included 10-items that participants were asked to respond to on 5-point scales ranging from 'strongly disagree' to 'strongly agree': (i) I think I would like to use the guide frequently, (ii) I found the guide unnecessarily complex, (iii) I thought the guide was easy to use, (iv) I think that I would need the support of a technical person to be able to use the guide, (v) I found the various steps in the guide were well integrated, (vi) I thought there was too much inconsistency in the guide, (vii) I would imagine that most people would learn to use the guide very quickly, (viii) I found the guide cumbersome to use, (ix) I felt very confident using the guide, and (x) I needed to learn a lot of things before I could get going with the guide.

Difficulty completing the steps outline in the guide

Users were asked how difficult they found it to: (i) identify whether ontologies shared the same upper-level ontology, (ii) create a list of classes from the two candidate ontologies that have identical IRIs, (iii) create a list of classes from the two candidate ontologies that referred to essentially the same thing (what the guide refers to as 'contextually exchangeable'), (iv) create a list of classes from the two candidate ontologies that were useably consistent, (v) identify whether they agreed with other users, (vi) resolve disagreements with other users, (vi) identify whether similar classes were cross-referenced in the respective ontologies, (vii) provide reasons for their assessment of (i) to (vi), and (viii) suggest any ways in which the guide could be improved.

Results

Classes identified by users before and following discussion

The classes identified by users in each of the categories are shown in Table 3. All three users identified and agreed that three classes (BCIO:036042: 'Physical performance behaviour', BCIO:050300: 'Personal attribute', and BCIO:006099: 'Social influence behaviour') had identical IRIs in the HBO and the COPPER Ontology, and identified one class that had the same label and a similar definition but a different IRI (BCIO:036108: 'Walking' and COPPER:1020: 'Walking') and so was deemed to be contextually exchangeable. One of the users did not identify any similar classes in HBO and COPPER that could be deemed useably consistent; the other two users identified one (different) pair of classes each that they believed were similar and might be deemed useably consistent, although they were both unsure.

The first pair of classes that was identified by one user as similar and useably consistent was BCIO:07002: 'Goal setting BCT and COPPER:0000: 'Behaviour change plan'. The definitions of these classes suggests some overlap – both refer to setting goals or plans to facilitate a change in behaviour within an individual and the classes are useably consistent as BCIO and COPPER position the classes in different semantic hierarchies – the BCIO positions 'Goal setting BCT' within the parent class BCIO:033000: 'Behaviour change technique' while COPPER positions 'Behaviour change plan' within the parent class 'Plan' from the Ontology for Biomedical Investigations (OBI:0000260). However, the conceptual differences between goal setting and planning (Heckhausen & Gollwitzer, 1987), suggest that the two classes are not contextually exchangeable and so should be used separately.

The second pair of classes that was identified by one (different) user as potentially useably consistent was BCIO:006148: 'Evaluative belief about the consequences of behaviour' and BCIO:050828: 'Positive behavioural consequence' within COPPER. However, as these classes are differentiated within the BCIO (on the basis that, while both classes describe the consequences of behaviour, one refers to people's beliefs about this consequence while the other refers to the consequence itself and so is a continuant rather than behavioural consequences), they would not be deemed useably consistent – or if they were then users would need to propose that the BCIO be revised to combine the classes.

Users' experiences of the process

The three users' responses to the System Usability Scale provided scores of 85 (deemed 'excellent' usability), 43 (deemed 'awful'), and 30 (deemed 'awful'; Brooke, 1996). When asked about specific tasks, the users reported difficulty (i) identifying classes with identical IRIs (2 out of 3 participants), (ii) identifying contextually exchangeable classes (1 out of 3 participants), (iii) identifying useably consistent classes (3 out of 3 participants), and (iv) determining whether similar classes were cross-referenced (2 out of 3 participants). However, participants did not find it difficult to (i) determine agreement with other users (0 out of 3 participants) or (ii) resolve disagreements with other users (0 out of 3 participants).

To better understand these challenges, we examined participants' qualitative responses (available in the file "BR-UK Evaluation (Responses) + TLW" on the Open Science Framework) and categorised the difficulties into three main areas: (i) issues understanding and using ontologies (e.g., "*I don't*

Table 3. Classes in HBO and COPPER Identified by users as Identical, Contextually Exchangeable, and Useably Consistent.

Human Behaviour Ontology (HBO)	Contextualised and Personalised Physical Activity and Exercise Recommendations (COPPER) Ontology	Users identifying match
Classes with identical IRIs		
BCIO:036042: 'Physical performance behaviour'	BCIO:036042: 'Physical performance behaviour'	1, 2, 3
BCIO:050300: 'Personal attribute'	BCIO:050300: 'Personal attribute'	1, 2, 3
BCIO:006099: 'Social influence behaviour'	BCIO:006099: 'Social influence behaviour'	1, 2, 3
Classes deemed contextually exchangeable		
BCIO:036108: 'Walking'	COPPER:1020: 'Walking'	1, 2, 3
Classes deemed useably consistent		
BCIO:07002: 'Goal setting BCT'	COPPER:0000: 'Behaviour Change Plan'	1
BCIO:006148: 'Evaluative belief about the consequences of behaviour'	BCIO:050828: 'Positive behavioural consequence'	3

fully understand what exactly an upper-level ontology means", "*I had some difficulties importing the HBO owl file into web protege*"), (ii) issues with the ontologies under consideration (e.g., "*The selected ontologies were highly generic, which was a bit of an obstacle to working out what interoperability I was specifically interested in searching for*"), and (iii) issues with the guide (e.g., "*At present it is dense, technical and difficult to follow*"). While we acknowledge that some difficulties may arise from a combination of these factors, our primary focus was on challenges related to the method for evaluating the interoperability of ontology classes. This aligns with the core objective of our research, as we assume that those wanting to evaluate the interoperability of ontology classes are likely to have some prior experience working with ontologies.

Comments on using the guide and suggestions for improvement

All three participants found it difficult to determine whether the two ontologies shared the same upper-level ontology. Several key challenges emerged:

(i) **Lack of clarity on upper-level ontologies.** Participants noted uncertainty about what "upper-level ontology" meant in this context: *"I don't fully understand what exactly an upper-level ontology means, although I understand that there are different levels in an ontology and that certain levels from different ontologies are likely to overlap."*

(ii) **Difficulty in providing evidence.** Another challenge related to proving that the two ontologies share the same upper-level ontology: *"I am not sure how to evidence that the two ontologies share the same upper level. HBO references BFO a couple of times in the .pdf..."*

(iii) **Structural differences between ontologies.** A third difficulty occurred because the ontologies were structured differently. For example, one participant mentioned "*the way the ontologies are structured - the COPPER ontology did share some BFO entities with the BCIO behaviour, but these were very scattered*" and explained that the different levels of classes in the two different ontologies also made it difficult to identify links to an upper level ontology: "*Sometimes the BFO entity was not the highest level in the COPPER ontology (e.g. BFO class role was a child of COPPER class user characteristic), which seemed to go against the purpose of an upper ontology*".

(iv) **Assumptions based on the framing of the question.** Participants felt that the framing of the question made the task of identifying whether ontologies share the same upper-level ontology difficult. For example, one participant assumed that the ontologies shared the same upper-level ontology: "*the 2 chosen ontologies for the exercise meet this criterion or they wouldn't have been chosen.*"

The next task in the implementation guide required participants to identify classes in the two ontologies that had identical IRIs. Participants approached this task differently, leading to varied experiences:

(i) **Straightforward for those using the automated output.** One of the participants found this task obvious as they explored the output from the automated tool (an Excel file) that was intended to provide an initial set of suggestions for participants to inspect in greater detail: *"It's very obvious from the excel file/output which classes have identical IRIs"*.

(ii) **Difficulties in selecting candidate classes.** The second participant found it difficult to select candidate entities for the automated tool and explored the hierarchy of the two ontologies to identify classes with potentially identical IRIs, following specific topics of interest and respective labels to be used as candidates for matching in the other ontology: *"At first I tried listing labels which might generate matches but was doing this intuitively, so my approach seemed opportunistic and potentially random / biased"*.

(iii) **Confusion over the role of automation.** A third participant found the mix of manual and automated steps within the guide disorganized: *"It felt disorganized to come up with some suggestions for potential matches manually and then have the matches identified automatically."* The purpose of the automated approach was to provide participants with an initial set of candidate classes for them to manually inspect and match. Based on this participant's response, we believe that this purpose was misunderstood, and thus that the guide could better set out how the automated process could and should be used.

The automated first step to matching classes in the two ontologies was appreciated, as highlighted by two participants who noted the ease of the process - "*I sent a list of candidate classes for the tool to generate candidates and it produced matches that made sense to me*" and "*It was fairly easy to compare the meaning of the labels from different ontologies*". It was, however, noted that the definition of classes in the two ontologies could also introduce some difficulty. For example, one participant noted that "*sometimes the [definition] would differ even when the labels were similar*". Another participant suggested that a fully automated process could be more valuable, stating that "*...the manual part feels haphazard. An entirely automated first step would be better*".

A broader point was also raised with respect to the purpose of determining matches and overlaps – namely, that it may be helpful if the guide introduced a specific context or task for which matches could be either consistent or inconsistent. For example, one participant mentioned that "*there were a few possible matches where one class [from] one ontology might be considered a parent class of another class from another ontology (e.g., locomotive behaviour and running). Whether or not these are deemed to be useably consistent will depend on the purpose of the task – if I'm interested in different types of physical activity then they are not useably consistent, whereas if*

I'm interested in all types of physical activity then they are useably consistent". In a sense, this comment reflects the slightly artificial nature of the evaluation, with participants being asked to use the guide as part of a research project, rather than because they had a task that required evaluating the interoperability of ontology classes (e.g., annotating a dataset or paper).

Our overall conclusions from the users' evaluation and feedback on the guide for evaluating the interoperability of ontology classes are that: (i) users may have different understandings of upper-level ontologies and how they relate to the ontologies being considered. If the ontologies are differently structured, with a different hierarchical organisation, then the potential links with an upper-level ontology may not be obvious; (ii) while the automated process for shortlisting potential matches between classes in the two ontologies offered some convenience, it could be better explained and streamlined within the workflow for identifying related classes; and (iii) the experience of matching two ontologies and identifying similar classes depends on the users' experience with, and knowledge of, ontologies, as well as the disciplinary perspectives and interests that they approach the task with.

Revising the guide

After participants had completed the exercise above, we invited suggestions on how the guide might be improved. They made several suggestions, covering cosmetic changes, structural changes and level of detail. For example, one suggestion was to improve the fonts, illustrations and text layouts ("*consider using larger font, double spaced text, flow diagrams*"). Another was that it would be helpful to provide more practical information on accessing ontologies, loading them to WebProtege and using the automated tool. A third proposed structural changes ("*...guide would be improved by a clearer introduction on the purpose. I would put the glossary at the end*)". Two of the three participants mentioned the need to improve the automated tool so that it can be seamlessly integrated within a workflow.

We made several improvements to the guide based on this feedback and our understanding of users' experiences using the guide, as outlined below.

(i) **Intended audience and application.** The revised guide (v4.0) explicitly states the intended audience of the guide, helping potential users to determine if the guide is relevant to them and their intended goals. Although we specify that the guide assumes a working knowledge of ontologies, we now provide links to additional resources for users who may benefit from further information or training on using ontologies. These have been specifically designed for social and behavioural scientists and as such align with potential users of the guide. We also included an example use case for how the guide may be applied in practice.

(ii) **Background and conceptual clarity**. The introduction to the guide was expanded to better explain the background and purpose of the guide. In particular, we now draw a distinction between class matching and ontology interoperability, explaining how matching ontology classes enhances broader interoperability. This helps to highlight possible use cases for when users may want to apply such a method in their work.

(iii) **Improvements to the method**. To enhance usability, we restructured the method into five distinct steps, making the guide easier to follow. Where certain steps require specific knowledge, we provide brief background explanations along with links to further information. For example, we expanded and clarified key concepts, such as "upper-level ontology." The revised guide also includes a detailed explanation of the BFO to make this step more accessible.

Additionally, we broke down complex steps into more manageable actions. For example, we provide explicit, step-by-step instructions for using OLS to find relevant ontologies, including an example on searching for concepts, reviewing results, and navigating ontology structures. We also provide an example table to illustrate how researchers can compare assessments, resolve discrepancies, and document their decisions systematically. To enhance the overall workflow, particularly with respect to integrating the automated tool, we now include a direct link to the tool on GitHub along with detailed instructions to facilitate accessibility and ease of use.

(iv) **Cosmetic and structural changes.** We made cosmetic and structural changes to improve the readability and usability of the guide. Additional subheadings now clearly distinguish each step of the process, while the use of color-coded and bold-type fonts differentiates background information from the actions needed to execute the method. To enhance logical flow, the order of sections presented in the guide was reorganized: The guide now begins with a clear outline of the intended audience, purpose, aims, and example use cases, followed by the steps comprising the method, and concluding with a glossary of key terms. Additionally, we provide links to training and support materials, including online workshops run by the team, GitHub resources, and pages on the Open Science Framework that provide researchers with more learning opportunities.

Discussion

Ontologies confer a number of benefits for scientists, including the ability to articulate precise understanding of concepts and ideas that can be shared and linked to datasets and research reports (Baird *et al.*, 2023; National Academies of Sciences, Engineering, and Medicine *et al.*, 2022). As researchers start to appreciate the benefits of and use ontologies, the number of ontologies proliferates and inevitably there is overlap in the domains and concepts specified by ontologies. This overlap affords an opportunity to connect ontologies and facilitate discussion around similarities and differences between concepts and ideas (e.g., Braun *et al.*, 2024a). Identifying and aligning overlapping classes can reduce the unnecessary duplication of concepts (classes), leading to a more efficient and precise representation of knowledge. It also makes it easier for researchers to reuse and build upon existing frameworks, reducing the time and effort required to develop new ontologies. The present research developed and evaluated a method for researchers who want to evaluate the interoperability of ontology classes in the social and behavioural sciences.

The findings of this study highlight the usefulness and limitations of automated and manual approaches to matching ontology classes, as well as the challenges faced by potential users in this domain. For example, while automated tools can make identifying potential ontology class matches more efficient – something that is essential given the size of many ontologies – our automated approach to matching also produced a high number of false positives (i.e., classes that were not potentially similar). This highlighted the need for a hybrid approach that utilises automated matching alongside manual review to enable both efficient and accurate identification of potentially similar ontology classes.

This hybrid approach was integrated into a step-by-step guide for evaluating the interoperability of ontology classes. The guide prompted users to use the automated script to create a list of classes in pairs of ontologies that appear to refer to the same thing and review this list to identify classes that are (a) identical (i.e., that have the same IRI), (b) contextually exchangeable (i.e., that have distinct IRIs, but are sufficiently similar to be deemed exchangeable) and (c) useably consistent (i.e., that could be imported into an ontology without causing a contradiction or conflict). If this process is conducted by several users or members of a research team, then they can compare assessments.

The evaluation of the guide suggested that it had potential in the sense that users were able to identify classes in two ontologies that were identical or contextually exchangeable. However, it was apparent that users needed to have some experience working with ontologies (e.g., understand what is meant by an ‘upper-level ontology’ and how to assess if two ontologies use the same upper-level ontology) and found using the guide difficult. Given the challenges that users had understanding ontology structures, future work should focus on developing tools and training resources that make working with ontologies easier (Michie *et al.*, 2024), providing the basis for assisting researchers in aligning ontology classes (and ontologies more broadly) more effectively. A broader evaluation with a diverse set of behavioural and social science researchers could provide further insights into improving the method and refining the implementation guide.

To improve accessibility and usability, it may be helpful to deploy the automated tool as a web-based application. This would allow researchers to access the tool through a user-friendly interface, eliminating the need to install software or run scripts manually. A web-based application could also embed the implementation guide into its workflow, offering step-by-step support and automating relevant parts of the class alignment process. On the technical side, future iterations of the tool could explore more advanced semantic matching techniques. For example, by incorporating larger context windows or using fine-tuned, domain-specific language models. These enhancements could help the tool better capture subtle differences in meaning and improve the accuracy of class matching. Finally, the evaluation also indicated that some users would have preferred a fully automated process. This preference appeared to stem from the hybrid approach which involved switching between manual review and the automated tool, potentially disrupting the workflow and making the process feel more complex.

Conclusions

The present research developed a hybrid method that combined automated and manual approaches to identifying and evaluating the interoperability of ontology classes in the behavioural and social sciences. While challenges remain and further refinement is needed, we provide a step-by-step guide that enables users to implement the method to identify potentially similar classes across pairs of ontologies. This guide is available on the Open Science Framework (see Data Availability statement) but in time will join other tools and guides for working with ontologies in the Behavioural and Social Sciences Ontology Foundry (Hastings *et al.*, 2024). Continued advancements in LLM-based tools for ontology matching, alongside structured workflows and user training, will be essential for promoting the usability and interoperability of ontologies in the behavioural and social sciences.

Data availability

The data (Webb *et al.*, 2025) and supplementary materials supporting this article, including v4.0 of the guide for implementing the resulting method for evaluating the interoperability of ontology classes, are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/4T65G) under a CC-By Attribution 4.0 International license.

Software availability

Information related to the automated tools, including code and documentation, can be accessed via GitHub (https://github.com/fatibaba/DEMO-INTER-AutomatedTool), archived on Zenodo (https://doi.org/10.5281/zenodo.16738179). The code is distributed under an Apache License 2.0.

Acknowledgements

We are grateful to Vyv Huddy, Paul Norman, and Richard Rowe for helping to evaluate the guide and Colbie Reed and Bill Hogan for discussions around evaluating the interoperability of ontology classes.

References

Baird HM, Hastings J, Johnston M, *et al.*: **Ontologies of behaviour: current perspectives and future potential in health psychology.** *European Health Psychologist.* 2023; **23**(2): 1011–1016.
Reference Source

Braun M, Baird HM, Peters GJ, *et al.*: **Protecting pluralism or committing to consensus? Risks and opportunities of ontologies in behavioral sciences**. *Roundtable discussion at the European Health Psychology Conference*. Cascais, Portugal, 2024a.
Publisher Full Text

Braun M, Carlier S, De Paepe A, *et al.*: **Development and evaluation of the Contextualised and Personalised Physical Activity and Exercise Recommendations (COPPER) ontology**. 2024b.
Publisher Full Text

Brooke J: **SUS: a "quick and dirty" usability scale.** In: P. W. Jordan; B. Thomas; B. A. Weerdmeester; A. L. McClelland (eds.). *Usability Evaluation in Industry.* London: Taylor and Francis, 1996.
Reference Source

Garg A, Kitsara I, Bérubé S: **The hidden cost of AI: unpacking its energy and water footprint**. 25 February, 2025.
Reference Source

Giglou HB, D'Souza J, Engel F, *et al.*: **LLMs4OM: matching ontologies with Large Language Models.** *ArXiv.* 2024.
Publisher Full Text

Hastings J, Zhang L, Schenk P, *et al.:* **The BSSO foundry: a community of practice for ontologies in the behavioural and social sciences [version 1; peer review: 1 approved, 3 approved with reservations].** *Wellcome Open Res.* 2024; **9**: 656.
PubMed Abstract | Publisher Full Text | Free Full Text

Heckhausen H, Gollwitzer PM: **Thought contents and cognitive functioning in motivational versus volitional states of mind.** *Motiv Emot.* 1987; **11**(2): 101–120.
Publisher Full Text

Hertling S, Paulheim H: **OLaLa: ontology matching with Large Language Models.** *Proceedings of the 12th Knowledge Capture Conference.* 2023; 131–139.
Publisher Full Text

Hicks A, Hanna J, Welch D, *et al.*: **The ontology of medically related social entities: recent developments.** *J Biomed Semantics.* 2016; **7**(1): 47.
PubMed Abstract | Publisher Full Text | Free Full Text

Krpan D, Fasolo B, Schneider L: **A call for precision in the study of behaviour and decision.** *Nat Hum Behav.* 2025; **9**(3): 433–436.
PubMed Abstract | Publisher Full Text

Michie S, West R, Finnerty AN, *et al.*: **Representation of Behaviour Change Intervention and their evaluation: development of the upper level of the Behaviour Change Intervention Ontology [version 2; peer review: 2 approved].** *Wellcome Open Res.* 2021; **5**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

Michie S, West R, Hastings J, *et al.*: **The human behaviour-change project phase 2: advancing behavioural and social sciences through ontology tools [version 1; peer review: not peer reviewed].** *Wellcome Open Res.* 2024; **9**: 730.
Publisher Full Text

National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education, Board on Behavioral, Cognitive, and Sensory Sciences, *et al.:* **Ontologies in the behavioral sciences: accelerating research and the spread of knowledge**. Washington, DC: The National Academies Press, 2022.
PubMed Abstract | Publisher Full Text

Qiang Z, Wang W, Taylor K: **Agent-om: leveraging LLM agents for ontology matching**. *arXiv preprint arXiv: 2312.00326*, 2023.
Publisher Full Text

Schenk PM, Hastings J, Michie S: **Developing the mental health ontology: protocol for a step-wise method to develop an ontology for the mental health domain as part of the GALENOS Project [version 1; peer review: 1 approved with reservations, 1 not approved].** *Wellcome Open Res.* 2024; **9**: 40.
Publisher Full Text

Scott A, Webb TL, Norman P, *et al.*: **A new resource for behavioural science - developing tools for understanding the relationship between behaviours**. *Poster presented at the 36th Conference of the European Health Psychology Society*. Bratislava, Slovakia, August, 2022.

Webb TL, Baird HM, Michie S, *et al.*: **BRUK Evaluation_Responses.** [Data set]. Open Science Framework. 2025.
https://osf.io/4t65g/files/osfstorage/68777388942e72b47029d7d8