Article:
West, R. orcid.org/0000-0001-6398-0921, Brown, J. orcid.org/0000-0002-2797-5428, Shahab, L. orcid.org/0000-0003-4033-442X et al. (10 more authors) (2025) Annotating datasets in behavioural and social sciences to promote interoperability: development of the schema for ontology-based dataset annotation (SODA) version 1.0. Wellcome Open Research, 10. p. 455. ISSN: 2398-502X

https://doi.org/10.12688/wellcomeopenres.24234.1

RESEARCH ARTICLE

Annotating datasets in behavioural and social sciences to promote interoperability: development of the Schema for Ontology-based Dataset Annotation (SODA) version 1.0

[version 1; peer review: awaiting peer review]

Robert West[1], Jamie Brown[1], Lion Shahab[1], Harriet Baird[2], Thomas Webb[2], Hazel Squires[3], Harry Tattan-Birch[1], Duncan Gillespie[3], Robin Purshouse[4], Alan Brennan[3], Suvodeep Mazumdar[5], Vitaveska Lamfranchi[6], Susan Michie[7]

[1]Department of Behavioural Science and Health, University College London, London, England, UK
[2]School of Psychology, University of Sheffield, Sheffield, UK
[3]School of Medicine and Population Health, University of Sheffield, Sheffield, UK
[4]School of Electrical and Electronic Engineering, University of Sheffield, Sheffield, UK
[5]School of Information, Communication and Journalism, University of Sheffield, Sheffield, UK
[6]School of Computer Science, University of Sheffield, Sheffield, UK
[7]Centre for Behaviour Change, University College London, London, England, UK

v1 **First published:** 20 Aug 2025, **10**:455
https://doi.org/10.12688/wellcomeopenres.24234.1
Latest published: 20 Aug 2025, **10**:455
https://doi.org/10.12688/wellcomeopenres.24234.1

Abstract

Background and aims

Ontologies are increasingly employed to help find, use and synthesise information, but methods for using them to annotate documents and datasets remain in their infancy in the behavioural and social sciences. The Behavioural Research UK DEMO-DATA project aimed to develop a prototype schema for annotating datasets in behavioural and social sciences.

Methods

A case-study dataset (the 'Smoking Toolkit Study'), used to inform an Agent-Based Model of trajectories in cigarette smoking and cessation in England, was chosen for annotation using two ontologies - The Behaviour Change Intervention Ontology (BCIO) and the Addiction Ontology (AddictO). The data set included 21 variables representing information about sociodemographic and tobacco and nicotine use

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.

attributes of the study population. A preliminary version of the schema for linking variables to ontology classes was developed as a basis for annotating each variable in the dataset. This was applied and revised iteratively until it was judged by an expert panel of domain experts and modellers to represent the variables sufficiently accurately to enable searching for and integration of data.

Results

The prototype Schema for Ontology-based Dataset Annotation (SODA) version 1.0 was developed over seven iterations. Variables were represented by an 'object property'|'ontology class' expression (e.g., 'has characteristic'|'extent of social smoking') together with information about the data types (e.g., numbers, ontology subclasses, or Boolean values), measurement source, unit of measurement, any coding or data transformations and whether or not the variable was fully characterised by the annotation. The prototype schema was applied successfully to the smoking dataset with 15 new ontology classes being created as required.

Conclusions

A prototype schema for annotating behavioural and social science datasets was developed and successfully applied to a dataset on smoking in England using ontology relations and classes. The next step is to further develop and evaluate the schema by application to case studies with a range of users and other datasets.

Plain language summary

This study focused on creating a standardised framework or 'schema' to organise and label datasets in behavioural and social sciences to make them easier to find and use. The study team created a prototype system called SODA (Schema for Ontology-based Dataset Annotation) to label datasets using specialised classification systems called 'ontologies'. The approach was tested on a dataset about smoking in England, using two classification systems: the Behaviour Change Intervention Ontology and the Addiction Ontology. The dataset contained 21 variables about people's demographics and tobacco use. The schema went through seven versions, improving their labelling system until the team of experts agreed it accurately represented the information in the dataset.

The final labelling system successfully organised the smoking dataset by creating expressions that connect properties to classifications (for example, 'has characteristic | extent of social smoking'). Each label included information about the type of data, how it was measured, and other important details. During this process, 15 new classification categories were created to fully describe everything in the dataset.

This new approach could make it much easier for researchers to find,

understand, and combine information across different studies in behavioural and social sciences. The next step is to test the system with different types of datasets and users to improve it.

Keywords
Ontologies, datasets, annotation, metadata, behavioural science, social science, FAIR principles

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

Corresponding author: Robert West (robert.west@ucl.ac.uk)

Author roles: West R: Conceptualization, Formal Analysis, Methodology, Resources, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Brown J**: Formal Analysis, Validation, Writing – Review & Editing; **Shahab L**: Formal Analysis, Validation, Writing – Review & Editing; **Baird H**: Conceptualization, Validation, Writing – Review & Editing; **Webb T**: Conceptualization, Funding Acquisition, Project Administration, Validation, Writing – Review & Editing; **Squires H**: Methodology, Validation, Writing – Review & Editing; **Tattan-Birch H**: Methodology, Validation, Writing – Review & Editing; **Gillespie D**: Methodology, Validation, Writing – Review & Editing; **Purshouse R**: Methodology, Validation, Writing – Review & Editing; **Brennan A**: Methodology, Validation, Writing – Review & Editing; **Mazumdar S**: Conceptualization, Validation, Writing – Review & Editing; **Lamfranchi V**: Conceptualization, Validation, Writing – Review & Editing; **Michie S**: Conceptualization, Validation, Writing – Review & Editing

Competing interests: RW has undertaken research and consultancy for companies that develop and manufacture smoking cessation medicines (Pfizer, GSK, Qnovia, Lilly). He is an unpaid Director of the Unlocking Behaviour Change Community Interest Company. JB has received (most recently in 2018) unrestricted research funding to study smoking cessation from Pfizer and J&J, who manufacture medicinally licensed smoking cessation treatments. LS has received honoraria for talks, unrestricted research grants and travel expenses to attend meetings and workshops from manufactures of smoking cessation medications (Pfizer; J&J) and has acted as paid reviewer for grant awarding bodies and as a paid consultant for health care companies.

Grant information: This work was supported by Wellcome [201524]; Cancer Research UK; the Economic and Social Research Council [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: West R, Brown J, Shahab L *et al.* **Annotating datasets in behavioural and social sciences to promote interoperability: development of the Schema for Ontology-based Dataset Annotation (SODA) version 1.0 [version 1; peer review: awaiting peer review]** Wellcome Open Research 2025, **10**:455 https://doi.org/10.12688/wellcomeopenres.24234.1

First published: 20 Aug 2025, **10**:455 https://doi.org/10.12688/wellcomeopenres.24234.1

Introduction

A substantial amount of research effort in behavioural and social sciences is wasted because of an inability to find and integrate data from different sources[1]. As a consequence, there is a movement to improve 'Findability, Accessibility, Interoperability and Re-usability' of data (FAIR data principles)[2] and apply these to behavioural and social sciences. 'Interoperability' means that different collections of data can work together, be combined, compared, or analysed as a unified whole, even when they come from different sources. Ontologies are structured representations of knowledge within a specific domain, defining concepts and their relationships in a way that promotes discoverability and interoperability[3–5]. There is a need to develop systematic frameworks (here called 'schemas') for using ontologies to annotate datasets in behavioural and social sciences to promote FAIR principles. This paper reports the initial development of a prototype schema.

Progress in adopting FAIR data principles in the behavioural and social sciences has been slow, but collaborations are growing to set standards for data documentation in order to accelerate progress (e.g., the Data Documentation Initiative (DDI))[6]. A key area for development is in documentation of data and harmonisation of variables and metadata[7,8]. Ontologies can be helpful in achieving this by providing structured vocabularies that can standardise how variables and metadata are defined across datasets. They consist of defined classes located in a coherent semantic hierarchy with unique identifiers '(Internationalised Resource Identifiers' or IRIs)[9] and properties that are defined by their relations with other classes. By mapping diverse terminologies to common conceptual frameworks, ontologies enable researchers to precisely communicate what their data represent, facilitating accurate interpretation, comparison, and integration of findings across studies.

A 2022 US National Academies of Science Engineering and Medicine report suggested that ontologies are likely to play a central role in advancing behavioural and social research in the coming decade[4]. On this basis, the US National Institutes of Health has invested more than 10 million dollars in a 5-year programme to develop and evaluate tools to help ontologies to become embedded in behavioural and social sciences[10]. One of the component projects is building tools for enabling the use of the Behaviour Change Intervention Ontology and associated ontologies such as the Addiction Ontology[11,12]. In the UK, the Economic and Social Science Research Council (ESRC) has set up a 5-year programme (Behavioural Research UK – BRUK) to accelerate advances in behavioural and social sciences, and part of that funding is being used to promote the development and enable the use of ontologies[13]. The present study forms part of both of these initiatives.

The literature already contains annotation tools, methods and resources (see Table 1 for examples) to facilitate annotation of datasets. These range from data standards, (e.g.,[14]) through to fully fledged data ecosystems (e.g.,[15]). What is required, however, is a standard framework or schema that sets out what information needs to be captured and precisely how the information is to be represented using ontologies.

In order to be suitable for widespread adoption in behavioural and social sciences, we propose that the schema needs to work with readily available software, be accessible to researchers without expertise in data science or ontologies, and tailored to the kinds of datasets created in this area. (See also the Open Data Institute's open standards for data[16] and the UK Government's open standards principles[17].) None of the existing resources currently meet these requirements. The schema can take the form of a template for a machine-readable annotation file that can accompany the data file in a commonly used format that requires no specialist software. (See Table 2.)

This study focuses on annotation of variables and their values within datasets, rather than providing metadata about the dataset itself (i.e., it excludes details of the methods used to collect the data, the study sample, the purpose for which the data were collected, the owners of the data, rules regarding data access, etc.). There are existing checklists to promote comprehensive and transparent reporting on complex surveys (e.g., the PRICSSA checklist)[18] to support the creation of appropriate variables to include in datasets and accompanying codebooks.

Thus, the aim of this study was to develop and undertake a preliminary evaluation of an ontology-based annotation schema for datasets in the behavioural and social sciences. The primary research question was: Is it feasible to develop schema that can meet the requirements set out in Table 2?

Methods

This was an exploratory study with methods that evolved over the course of the study in response to experiences and findings. An initial protocol outlining the aims of the research and proposed approach was registered on the Open Science Framework in March 2024 (https://osf.io/284gw/files/osfstorage/65f445ae719c061250d49a33).

The dataset

A strategically important dataset was chosen, containing variables of multiple types and familiar to the research team: the Smoking Toolkit Study (STS)[19,20]. The STS is one of the world's most comprehensive datasets on smoking, a research topic of high public health significance. The STS is a long-running research project conducted by University College London (UCL) that monitors patterns of smoking and tobacco use in England, Wales and Scotland. Established in 2006, it collects data through household or telephone surveys of a fresh sample of approximately 1,700 adults living in England each month.

The extract was a subset of variables being used to characterise adults in England to build an Agent-Based Model of smoking to inform policy making (See Table 3)[21]. The variables were those selected at an early stage in the model development process.

Table 1. Examples of existing resources and methodologies for annotating datasets.

Tool Name	URL	Description
Aridhia DaSH	https://www.aridhia.com/	Data management platform with variable-level annotation capabilities for health and social science data
CEDAR Workbench	https://metadatacenter.org/	Template-based metadata creation tool that supports ontology-based variable annotations
CLOSER Variable Harmonisation Tools	https://discovery.closer.ac.uk/	Tools for harmonising and annotating variables across multiple longitudinal studies
Colectica	https://www.colectica.com/	Metadata management tool that works directly with SPSS files and allows variable-level annotations using DDI standards and ontology terms
DataSchema Framework	https://schema.org/Dataset	Ontology classes that can be used to describe features of datasets
DataVerse Metadata Blocks	https://dataverse.org/	Repository system with customisable metadata schemas for dataset and variable annotation
DDI-CDI	https://ddialliance.org/ddi-cdi	Cross-domain integration vocabulary for mapping variables to ontology terms
ELSST	https://elsst.ukdataservice.ac.uk/	European Language Social Science Thesaurus tools for annotating social science variables
EML Tools (Morpho)	https://knb.ecoinformatics.org/tools	Ecological Metadata Language tools that support variable annotation with ontology terms
Iqvoc	https://github.com/innoq/iqvoc	Vocabulary management system that can be used to maintain controlled terms for dataset variables
ISA-Tools	https://isa-tools.org/	Suite of tools for managing experimental metadata including variable annotations
Apache Jena	https://jena.apache.org/	Framework for annotating statistical variables with ontology terms, generating RDF/XML linkage files
Ontotext Refine	https://www.ontotext.com/products/ontotext-refine/	Data cleaning tool with reconciliation services to match dataset variables to ontology terms
Opal	https://www.obiba.org/pages/products/opal/	Open-source epidemiological data management software with ontology annotation features
Questionnaire Variable Ontologiser	https://www.phenxtoolkit.org/	Tool specifically for annotating questionnaire variables in health and behavioral research
RDF Data Cube Tools	https://www.w3.org/TR/vocab-data-cube/	Implementation of W3C standard for representing statistical data variables with semantic annotations
SPSS Dimension Management	https://www.ibm.com/products/spss-statistics	Native SPSS functionality for adding metadata to variables, with some ontology integration capabilities
StatWiki	No public URL available (typically institutional)	Collaborative platform for annotating statistical data variables with ontological classifications
Varanto	https://www.maelstrom-research.org/technology/software	Variable harmonisation tool that links dataset variables to standardised ontology terms

Table 2. Proposed requirements for a prototype schema for annotating variables in behavioural and social science datasets using ontologies.

Requirement	Description
Functionality	Enables variables in databases, and the values of those variables, to be represented by linking cases to the most appropriate ontology classes, together with enough information to be discoverable in automated search processes and match with variables in other relevant datasets.
Accessibility	Uses a data format that is usable by researchers who are unfamiliar with data science principles, without specialist software.
Applicability	Applicable to datasets and ontologies in behavioural and social sciences.

Table 3. Variables used for annotation.

Variable label	Variable name in the dataset	Variable description
Age	actage	Age in years
Gender	sexz	Gender
Occupational Social Grade	sgz	Occupational Social Grade as defined in the UK's National Readership Survey
Educational level	qual	Highest educational qualification
Housing tenure	tenure	Ownership or rental status of primary residence
Region	gore	One of nine official regions in England
Enjoyment of smoking	q632x4	Level of enjoyment of smoking
Desire to stop smoking	qmotiv4	Any desire to stop smoking
Intention to stop smoking soon	qmotiv3	Intention to quit in the next 3 months
Interaction with smokers in social network	qimw891	Proportion of smoking that occurs with other smokers
Diagnosis of mental health disorder	mdiag	Combination of 12 variables. Since age 16 years, any diagnosis of one or more of: depression, anxiety, obsessive compulsive disorder, panic disorder or phobia, post-traumatic stress disorder, psychosis, personality disorder, attention deficit hyperactivity disorder, eating disorder, alcohol use disorder, drug use disorder, problem gambling
Number of recent quit attempts	q632b7_1	Number of serious quit attempts made in the past year.
Cigarettes per day	basecpd3	Current average daily cigarette consumption with non-smokers assigned 0
AUDIT-C score	auditc	AUDIT-C score (alcohol consumption)
NRT use	allnrt	Current Nicotine Replacement Therapy use
E-cigarette use	allecig	Current e-cigarette use (nicotine vaping)
Smoker self-identity	q632e9	Whether think of self as smoker or non-smoker
Strength of cigarette addiction	sturge	Strength of Urges To Smoke (SUTS) scale
Varenicline use in a recent quit attempt	chac	Past-year use of varenicline in a quit attempt
Use of behavioural support in a recent quit attempt	behc	Past-year use of behavioural support in a quit attempt
Spending on cigarettes	qspend_1	Average weekly spend on cigarettes or tobacco

The ontologies

The ontologies used to annotate the dataset had to comply with good practice regarding the development and maintenance of ontologies, be interoperable with each other, and cover the domain of interest. We are at an early stage in the development of ontologies in the behavioural and social sciences, and there are few ontologies that meet these requirements[22,23]. The two main ones identified as likely to cover the variables in the dataset were: The Behaviour Change Intervention Ontology (BCIO)[24] and the Addiction Ontology (AddictO)[25].

The ontology annotations needed to capture enough information about each dataset variable to establish correspondence with variables in other databases. The correspondence might be exact or approximate. If it was approximate, then information needed to be available about where the imprecision lay.

Development of the annotation schema

The prototype schema was developed iteratively by the lead author, in consultation initially with JB and LS (see author list) and then with the other members of the team (all authors on this paper). With each iteration an attempt was made to apply the schema to the dataset and issues with its comprehensiveness and accuracy were identified. Then the schema was revised and the process repeated until it was judged that the prototype schema captured all the information required for users to be able to annotate the database using relevant ontologies. In the following paragraphs, readers can refer to Table 4 to see the fields that were added in each iteration.

In the *first iteration* of the annotation schema, the BCIO and AddictO were searched to identify ontology classes that might correspond to variables in the database. (See 'Entity label' and 'Entity IRI' in Table 4.) This revealed that, in most cases, the ontology classes would be too general to represent the variables and so more specific classes would have to be developed. Therefore new proposed ontology classes were developed with the teams managing the ontologies using their established procedures[26]. (See 'Entity added' in Table 4.)

However, it was also noted that database variables could not be represented by ontology classes alone because variables represent *relations* between individuals in the database and ontology classes. (See 'Relation to ontology class' and 'Relation IRI' in Table 4.) For example, the variable 'cigarettes per day' represents the idea that the person in the database currently smokes a given number of cigarettes per day; they *enact* a behaviour pattern. Therefore, in a *second iteration*, the annotation needed to include a relation of the person to the ontology class as well as the ontology class itself. Relations were drawn from established ontologies that could be used for this purpose (e.g., the Relations Ontology)[27].

On further testing, it was noted that measurement processes used to generate data may materially influence the interpretation of the data points. This may be because they involve different levels or types of bias or because they are associated with somewhat different constructs. For example, a construct such as 'strength of addiction to cigarette smoking' can be construed and measured in different ways that affect how it relates to other constructs such as 'relapse to smoking following a quit attempt'[28]. Therefore, when annotating variables it can be important to represent the measurement process, at least insofar as it is likely to influence the interpretation of the data. Therefore, in the *third iteration* we added that option of representing measurement processes to the schema. (See 'Measure' and 'Measure IRI' in Table 4.) There could be cases where it can be assumed that the measurement process does not make a material difference, for example, when recording a person's age in a population survey. In those cases, a judgement may be made by the annotator that the measured value is the true value, and how it was measured need not be annotated.

For some variables, the values in the database would be simple presence or absence; in other cases they would be subclasses of the class representing the variable; in other cases they could be alphanumeric 'strings'; while in other cases they would be numbers representing a quantity. In the last case it was found to be important to have a field in the schema to represent the unit of measurement where there was one. (See 'Database value data type' and 'Unit of measurement' in Table 4.)

Variables often use numeric codes for categorical data or to denote presence or absence of a feature (represented as a Boolean value). For example, the nine Government Office Regions were coded as numbers from 1 to 9. These mappings are typically captured in the dataset by the use of value labels. Therefore, in the *fourth iteration* an additional item of information was added to the schema to match some variable values in datasets onto ontology classes. (See 'Transformation' in Table 4.) For example, in the social grade variable we specify that the numeric value 1 maps on to the ontology class 'SOCIAL GRADE AB' which has the unique ID in the Addiction Ontology 'ADDICTO:0001389'.

In addition, variables may be derived from other variables in the dataset computationally or by manually recording values. For example, an ordinal scale with seven categories may be dichotomised in different ways to extract different types of information. Therefore, in the *fifth iteration* fields were added to the schema to represent the derivation and coding of values in variables. (See 'Derivation' in Table 4.)

It was noted that there would be occasions when the variables were so specific to a given dataset that a precisely ontologised annotation would create ontology classes that would be unlikely to be used in other datasets. This could happen, for example, when a very specific classification of educational level was used in the dataset (see Table 5). In those cases, it would make more sense to annotate at a higher level of generality and to include an additional annotation to indicate that this had been done so that attempts to link datasets could take

Table 4. Fields in the annotation data schema.

Annotation field	Description
Variable label	The label field typically included in the data file being annotated (e.g., Social grade).
Variable name	The variable name in the data file being annotated (e.g., sgz).
Variable description	A brief description of the variable in the data file, giving sufficient detail to serve as the basis for the remainder of the annotation (e.g., Occupational Social Grade as defined in the UK's National Readership Survey).
Relation to ontology class	The label of the ontology relation linking the person to the ontology class (e.g., 'has characteristic').
Relation IRI	The International Resource Identifier of the relation linking the person to the ontology class, consisting of the namespace signifying the ontology containing the relation and the unique ID within that ontology (e.g., RO:0000053 which is the IRI for 'has characteristic' in the Relations Ontology).
Entity label	The label of the ontology class used to annotate the variable (e.g., 'social grade' which is a class in the Addiction Ontology).
Entity IRI	The International Resource Identifier of the ontology class used to annotate the variable, consisting of the namespace of the ontology and the unique ID of the class within that ontology (e.g., ADDICTO:0001324 which is the IRI for the class 'social grade' in the Addictions Ontology).
Entity added	Indication of whether a new ontology class had to be created for the variable ('Y') or not ('N') (e.g., 'N' because 'social grade' was already in the ontology).
Ontology expression	An expression derived automatically by concatenating the relation label and the class label using a Manchester Web Ontology Language (OWL) expression. Manchester OWL is a version of OWL that allows entities to be represented by combinations of relations, classes and data, e.g., ('has characteristic') ('social grade').
Measure	The form of measure used to obtain the data, expressed either as a generic type of measure (e.g., 'self-report questionnaire') or, if one is used, a standardised measurement instrument or scale (e.g., Motivation to Stop Smoking scale).
Measure IRI	If a standardised instrument or scale is used, an IRI for that scale, where possible taking the form of a web address to locate the scale online (e.g., https://doi.org/10.1016/j.drugalcdep.2012.07.012 as the IRI of the paper reporting the Motivation To Stop Smoking scale).
Coding	Where appropriate, a description of how values in the database correspond to values or classes *to attempt to be* mapped to the ontology. This corresponds to the 'Value Label' command in SPSS. Mappings for specific values by semicolons with the database value specified first, followed by an equals sign followed by annotation value. Quotations are used to signify strings (e.g., 1='Male';2='Female';3='Other'). This field is left blank there is no coding used.
Derivation	If the variable is derived from another variable in the dataset, this field specifies the syntax for the derivation in the language of the software that was used to derive the data (e.g., SPSS).
Transformation	Where appropriate, how values in the database have been transformed to ontology values such as ontology classes or Boolean values.
Database value data type	The type of data being represented by the annotation: number, string, subclass, Boolean. 'Subclass' indicates that the values in the cells are subclasses of the ontology class being used to annotate the variable (e.g., the values in the variable 'social grade' can be any of its subclasses in the Addiction Ontology such as 'social grade AB').
Unit of measurement	Where appropriate, the unit of measurement for the data being represented by the variable (e.g., 'years' for the class 'human age in years').
Ontology class direct match to variable	Indication of whether the annotation exactly matches the variable being annotated ('Y') or is an approximate match ('N') (e.g. In the case of the variable 'Gender', the annotated class 'gender identity' does not map exactly on to the values in the variable because it has specific subclasses for gender identities other than 'male' and 'female' while the dataset uses the value 'Other'. In addition, there is no specification of the precise question formulation and, in this case, it might make a difference to the values in the database).
Reason for no precise match	Where appropriate, a description of the reason why the annotation is not an exact match (e.g., 'Ad-hoc question and response options').
Notes	Explanations for decisions made in the annotation.

Table 5. Five example annotations.

Variable label	Example 1: Age	Example 2: Gender	Example 3: Social Grade	Example 4: Educational level	Example 5: Desire to stop smoking
Variable name	actage	sexz	sgz	qual	qmotiv4
Variable description	Age in years	Gender	Occupational Social Grade	Highest educational qualification	Any desire to stop smoking
Relation to ontology class	has datum	has characteristic	has characteristic	has characteristic	has disposition
Relation IRI	HSO:0000067	RO:0000053	RO:0000053	RO:0000053	RO:0000091
Entity label	human age in years	gender identity	social grade	highest level of formal educational qualification achieved	disposition to want to stop cigarette smoking to some degree
Entity IRI	ADDICTO:0001370	BCIO:015098	ADDICTO:0001324	BCIO:015043	ADDICTO:0001415
Entity added	N	N	Y	N	Y
Ontology expression	('has datum') ('human age in years')	('has characteristic') ('gender identity')	('has characteristic') ('social grade')	('has characteristic') ('highest level of formal educational qualification achieved')	('has disposition') ('disposition to want to stop cigarette smoking to some degree')
Measure	Self-report questionnaire	Self-report questionnaire	Self-report questionnaire	Self-report questionnaire	Motivation To Stop Smoking scale
Measure IRI	ADDICTO:0000155	ADDICTO:0000155	ADDICTO:0000155	ADDICTO:0000155	https://doi.org/10.1016/j.drugalcdep.2012.07.012
Coding		1='MEN'; 2='WOMEN'; 3='IN ANOTHER WAY'	1='AB'; 2='C1'; 3='C2'; 4='D'; 5='E'	1='GCSE/O-LEVEL/CSE'; 2='VOCATIONAL QUALIFICATIONS (=NVQ1+2)'; 3='A-LEVEL OR EQUIVALENT (=NVQ3)'; 4='BACHELOR DEGREE OR EQUIVALENT (=NVQ4)'; 5='MASTERS/PHD OR EQUIVALENT'; 6='OTHER'; 7='NO FORMAL QUALIFICATIONS'; 8='STILL STUDYING'; 9='DON'T KNOW'	
Derivation					RECODE qmotiv (1 THRU 5=1) (6 THRU 9=0) (-1=0) INTO qmotiv4.

Wellcome Open Research 2025, 10:455 Last updated: 21 AUG 2025

Variable label	Example 1: Age	Example 2: Gender	Example 3: Social Grade	Example 4: Educational level	Example 5: Desire to stop smoking
Transformation		1='FEMALE GENDER' [BCIO:010111]; 2='MALE GENDER' [BCIO:010112];3=NOT MAPPED	1='SOCIAL GRADE AB' [ADDICTO:0001389]; 2='SOCIAL GRADE C1' [ADDICTO:0001390]; 3='SOCIAL GRADE C2' [ADDICTO:0001391]; 4='SOCIAL GRADE D' [ADDICTO:0001392]; 5='SOCIAL GRADE E' [ADDICTO:0001393]	1='ACHIEVED LOWER SECONDARY EDUCATION' [BCIO:015047]; 2='ACHIEVED LOWER SECONDARY EDUCATION' [BCIO:015047]; 3='ACHIEVED UPPER SECONDARY EDUCATION' [BCIO:015048]; 4='BACHEOLOR DEGREE OR EQUIVALENT' [BCIO:015049]; 5='ACHIEVED MASTERS OR EQUIVALENT LEVEL' OR 'ACHIEVED DOCTORAL OR EQUIVALENT LEVEL EDUCATION' [BCIO:015050 OR BCIO:015051]; 6 THRU 7='ACHIEVED PRIMARY EDUCATION' [BCIO:015046]	0='FALSE'; 1='TRUE'
Database value data type	Number	Number	Number	Number	Boolean
Ontology value data type	Number	Subclass	Subclass	Subclass	Boolean
Unit of measurement	Years				
Ontology class direct match to variable	Y	Y	Y	N	Y
Reason for no precise match				The subclasses do not fully correspond to the response options.	
Notes		'IN ANTOHER WAY' is not mapped because it could refer to many different classes. The variable is mapped to gender identity rather than what may be considered biological sex because it is assessed by self-report.	In this usage there is no upper age limit for the person characterised, but in the primary usage the upper age limit is 64.	The values are mapped to classes in BCIO that are aimed at being internationally generalisable. However, this means that the mapping is not precise in this instance.	This is a derived variable from a self-report scale, qmotiv in the dataset.

Wellcome Open Research 2025, 10:455 Last updated: 21 AUG 2025

account of this. In the *sixth iteration*, this annotation field was added. (See 'Ontology class direct match to variable' and 'Reason for no precise match' in Table 4.)

Finally, in the *seventh* iteration an annotation option for the 'Transformation' field (dealing with mapping values of a variable to ontology subclasses) had to be created for cases where a given value could not be mapped on to an ontology class (for example, when there was a value corresponding to 'Other'). For this the annotation 'NOT MAPPED' was chosen (see Table 5 for the variable 'gender'). A field was also added for the annotator to provide notes on the annotation to help users to understand decisions made during the process.

Results

The annotation schema

To maximise usability and access without the need for bespoke software, and without requiring programming, this first prototype data schema for the annotation was specified as an Excel spreadsheet. Table 4 shows the spreadsheet fields with descriptions. Table 5 shows five example annotations.

The supplementary file contains the final completed annotation spreadsheet.

A total of 17 out of 21 (80.9%) variables could be annotated sufficiently precisely using the schema to permit direct mapping to similar variables in other datasets. In the case of the four variables where this was judged not to be the case, the dominant reasons were the use of highly specific questions or response options in the survey that may not be used in other surveys and therefore it made sense to map them at a more general level. In those cases, mapping to similar variables in other datasets would have to be at the level of the construct rather than the data.

For 15 (71.4%) variables, new ontology classes had to be added to ADDICTO to represent the associated construct. This reflected the fact that ontologies in the domain of behavioural and social sciences are at an early stage of development and classes relating to constructs in specific sub-domains such as tobacco and alcohol use need to be elaborated.

In the case of four (19.0%) of the variables, a standardised measure was used. In none of those cases could an ontology class be identified that was dedicated to describing the measure in a structured format, so the identifier pointed to a digital object identifier (DOI) of a journal article describing the measure and its use. In cases where a non-standardised measure was used, the type of measurement instrument was identified (e.g., self-report questionnaire) and the measure IRI annotated referred to an ontology class for this type of instrument.

In the case of the 'Government Office Region' variable, with each region having a numeric code, it was decided to represent the names of the regions as strings (i.e., text), but they could have been added to a relevant ontology as 'instances' or 'individuals'.

Discussion

This study developed a schema for annotating behavioural science datasets using ontology classes and relations. The schema was designed to support the FAIR principles (Findable, Accessible, Interoperable, and Reusable) and was tested on a subset of variables from the Smoking Toolkit Study used to inform an Agent-Based Model (ABM) of smoking behaviours. Our findings show it is feasible to create machine-readable annotations for variables in behavioural and social science datasets using existing ontologies, supplemented where necessary with new ontology classes.

The annotation schema evolved through multiple iterations to address increasing complexity in representation requirements. We found that variables in behavioural science datasets could not be adequately represented by ontology classes alone, but required relation-class pairs (e.g., 'has characteristic'|'social grade') together with additional metadata about measurement processes, data transformations, coding systems, and units of measurement. This approach enabled approximately 80% of the variables to be annotated with sufficient precision to permit direct matching with similar variables in other datasets.

Given that current ontologies in the behavioural and social sciences domain are still at an early stage of development, for 15 of the 21 variables examined, new ontology classes had to be added to the Addiction Ontology to represent the associated constructs. This highlights both the nascent state of ontology development in this field and the vital role that annotation efforts can play in expanding and refining these ontologies. The process of annotation can thus serve a dual purpose: enhancing dataset interoperability while simultaneously contributing to ontology development.

The Excel-based format of the annotation schema proved to be a practical solution that met our requirement for accessibility to researchers unfamiliar with data science principles, without requiring specialist software. The schema could equally be contained in a simpler delimited data file (e.g., csv) file given that no special formatting was required, which is important as future updates of Excel-specific formats could lead to compatibility problems. This approach addresses a key barrier to wider adoption of ontology-based annotation in behavioural and social sciences. The barrier to using ontology editing tools is significant, where even ontology engineers with experience in ontology development and tooling support encounter challenges like limited scope, integration problems, identifying and selecting appropriate tools, bugs, difficulty in access and learning time.

The dataset chosen for this project was one used to inform an ABM. Theoretical models of behaviour exhibit the same lack of systematisation, ambiguity and coherence as we see in datasets[29,30]. These problems extend to computational models such as ABMs[31]. Ontologisation of existing models and the use of ontologies in new models would go a long way to addressing this problem. A start on this has been made with the development of an ontology-based

modelling system for expressing theories of behaviour change[29,32]; this is an active area for further work.

Ontologisation also supports integration, re-use, adaptation and verification of computational models used to appraise public health policy. This would help to reduce the long development times for such models and improve their robustness. The Agent-Based Model that was developed as a companion to the present study integrates the example ontology–all entities in the model have IRIs present in the design and that can be queried at run-time[21]. Ontologisation supports recent ambitions in the Agent-Based Modelling community to develop reusable building blocks and enables model discovery processes[31], i.e., the automated construction and evaluation of candidate computational models[33].

Limitations and future directions

Several limitations should be noted when interpreting the findings. First, this was an exploratory study using a limited subset of variables from a single dataset. The generalisability of the schema to other datasets and domains within behavioural and social sciences requires further investigation. While the dataset covered core constructs in population surveys of smoking, variables related to addiction, particularly smoking and e-cigarette use, may have unique characteristics that are not representative of behavioural science datasets more broadly. In addition, in long-term surveillance data series of this kind, variables can involve different question wording or be coded differently in different waves. This will need to be addressed in future schemas that can link datasets corresponding to different waves of essentially the same study.

Secondly, while the schema could be used to annotate the variables in our test dataset, we did not conduct formal usability testing with researchers outside the development team. This is the next phase for this work. The acceptability, accessibility and usability of the schema to researchers without expertise in ontologies or data science remains to be demonstrated empirically and the schema is likely to require additional development to be widely adopted.

Thirdly, we found that standardised measures in behavioural and social sciences often lack dedicated IRIs for structured description, requiring the use of journal article DOIs as proxies. This highlights a gap in the infrastructure supporting FAIR data principles in behavioural sciences that will need to be addressed.

Fourthly, the schema focused on annotating variables within datasets rather than providing comprehensive metadata about the datasets themselves. For full implementation of FAIR principles, this variable-level annotation would need to be integrated with dataset-level metadata systems building on the methods set out in Table 1.

Finally, our evaluation did not test the practical utility of the annotations for data discovery or integration tasks. This will require ongoing evaluation as usage develops.

Building on this initial prototype, several key steps are necessary to advance the development and adoption of ontology-based annotation in behavioural and social sciences:

1. **Validation across diverse datasets**: The schema should be tested with a broader range of datasets spanning different sub-domains of behavioural and social sciences to assess its generalisability and identify domain-specific requirements.

2. **User testing**: Formal usability studies should be conducted with researchers who have varying levels of expertise in data management and ontologies to evaluate the acceptability, accessibility and usability of the schema and identify barriers to adoption.

3. **Integration with existing systems**: The annotation schema should be integrated with existing metadata systems and repositories to create a more comprehensive approach to implementing FAIR principles for behavioural and social science data.

4. **Development of supporting tools**: To facilitate wider adoption, supporting tools and standard operating procedures should be developed or adapted, such as ontology browsers tailored to behavioural science concepts, suggestion systems for matching variables to ontology classes, and validation tools for annotation quality.

5. **Evaluation of utility**: Studies should be conducted to evaluate whether datasets annotated using the schema are indeed more findable, accessible, interoperable, and reusable in practice, using concrete use cases for data discovery and integration.

6. **Community engagement**: Broader engagement with the behavioural and social science research community is needed to build consensus around annotation practices and to encourage contribution to ontology development. It is likely that training will be required for capacity building. Our hope is that initiatives like the establishment of the Behavioural and Social Sciences Ontology (BSSO) Foundry will prove useful in this regard[3].

7. **Standardisation of measures**: Work with the research community to develop structured descriptions with dedicated IRIs for commonly used standardised measures in behavioural sciences.

8. **Expansion of relevant ontologies**: Continued development of ontologies covering the diverse domains of behavioural and social sciences, with particular attention to concepts frequently used in dataset variables.

9. **Extension to data-driven models:** Ontologisation of data-based models of behaviour will be an important next step. We have begun this process when building a systems map underpinning the Agent-Based Model described earlier. We will report on this process in a separate paper.

This work represents an initial step in addressing the challenge of data fragmentation in behavioural and social sciences through ontology-based annotation. The schema has been developed for annotation of existing datasets but could also provide a basis for developing new datasets that would promote interoperability from the start. The iterative development of annotation schemas, alongside the growth of domain-specific ontologies, has the potential to significantly enhance data interoperability and hence data integration and analysis, thereby reducing research waste in these fields. While considerable challenges remain, this prototype appears to demonstrate a pragmatic approach that balances the complexities of semantic representation with the practical needs of researchers in behavioural and social sciences.

Conclusions

It is feasible to develop a schema for annotating behavioural and social science datasets using ontology classes and relations. The schema captured the complexity of variables in these datasets by representing them as relation-class pairs accompanied by essential metadata about measurement processes, coding systems, and data transformations. While current ontologies in behavioural and social sciences remain at an early stage of development, requiring expansion to adequately represent domain-specific constructs, our approach shows promise for enhancing data interoperability without requiring researchers to master complex data science principles or use specialised software. The next steps are to validate the schema across diverse datasets, conduct formal usability testing with researchers, develop supporting tools, and engage the broader research community in building consensus around annotation practices. By addressing these challenges, ontology-based annotation has the potential to reduce research waste in behavioural and social sciences through improved data findability, accessibility, interoperability, and reusability, ultimately accelerating scientific progress in these fields.

Ethics and consent

Ethical approval was not required for this study.

Data availability

The data used to develop the schema are from the Smoking Toolkit Study available on request. Interested parties can access the data by completing the access request form: https://smokinginengland.info/resources/sts-documents.

The data schema is open source and available to be used, with acknowledgement from the lead author. It is available on the Open Science Framework at this link: https://osf.io/j43wm/[34] DOI 10.17605/OSF.IO/J43WM[34]

Extended data

This Open Science Framework component (https://osf.io/zbe72) contains the supplementary file associated with the paper. https://doi.org/10.17605/OSF.IO/J43WM[34]

References

1. Riley WT: **Behavioral and social sciences at the National Institutes of Health: methods, measures, and data infrastructures as a scientific priority.** *Health Psychol.* 2017; **36**(1): 5–7.
PubMed Abstract | Publisher Full Text

2. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR guiding principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | Publisher Full Text | Free Full Text

3. 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

4. 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**. National Academies Press, Washington, D.C., 2022.
PubMed Abstract | Publisher Full Text

5. Sharp C, Kaplan RM, Strauman TJ: **The use of ontologies to accelerate the behavioral sciences: promises and challenges.** *Curr Dir Psychol Sci.* 2023; **32**(5): 418–426.
Publisher Full Text

6. DDI Alliance: **About the DDI Alliance**. 2025.
Reference Source

7. Gorman L, Browne WJ, Woods CJ, *et al.*: **What's stopping knowledge synthesis? A systematic review of recent practices in research on smallholder diversity.** *Front Sustain Food Syst.* 2021; **5**.
Publisher Full Text

8. Arslan RC: **How to automatically document data with the codebook package to facilitate data reuse.** *Adv Methods Pract Psychol Sci.* 2019; **2**: 169–187.
Publisher Full Text

9. **Internationalized Resource Identifier**. *Wikipedia*. 2024.

10. **Ontology development and its use in the behavioral and social sciences**. Office of Behavioral and Social Sciences Research.
Reference Source

11. 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

12. Hastings J, Cox S, West R, *et al.*: **Addiction ontology: applying basic formal ontology in the addiction domain.** *Qeios.* 2020.
Publisher Full Text

13. **Behavioural Research UK (BR-UK).** The university of edinburgh. 2024.
Reference Source

14. PhenX Toolkit: **About the PhenX Toolkit**.
Reference Source

15. Dataverse: **About the Dataverse Project**. 2024.
Reference Source

16. Open Data Institute: **What are open standards for data?** *Open Standards for Data Guidebook.* 2025.
Reference Source

17. **Open standards principles**. *GOV.UK*.
Reference Source

18. Seidenberg AB, Moser RP, West BT: **Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA)**. *Journal of Survey Statistics and Methodology.*

Oxford Academic, 2023; **11**(4): 743–757.
Publisher Full Text

19. Fidler JA, Shahab L, West O, *et al.*: **'The smoking toolkit study': a national study of smoking and smoking cessation in England.** *BMC Public Health.* 2011; **11**: 479.
PubMed Abstract | Publisher Full Text | Free Full Text

20. UCL Smoking and Alclhol Research Group: **The smoking toolkit study**.
Reference Source

21. Tian D, Squires HY, Buckley C, *et al.*: **Incorporating the COM-B Model for behavior change into an agent-based model of smoking behaviors: an object-oriented design**. In: *2024 Winter Simulation Conference (WSC)*. 2024; 252–263.
Publisher Full Text

22. Norris E, Finnerty AN, Hastings J, *et al.*: **A scoping review of ontologies related to human behaviour change.** *Nat Hum Behav.* 2019; **3**(2): 164–172.
PubMed Abstract | Publisher Full Text

23. **Accelerating social and behavioral science through ontology development and use**. National Academies.
Reference Source

24. Michie S, West R, Finnerty AN, *et al.*: **Representation of behaviour change interventions and their evaluation: development of the upper level of the behaviour change intervention ontology [version 2; peer review: 2 approved].** *Wellcome Open Res.* 2020; **5**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

25. **AddictO Vocab**.
Reference Source

26. Wright AJ, Norris E, Finnerty AN, *et al.*: **Ontologies relevant to behaviour change interventions: a method for their development [version 3; peer review: 2 approved, 1 approved with reservations].** *Wellcome Open Res.* 2020; **5**: 126.
PubMed Abstract | Publisher Full Text | Free Full Text

27. **The Relations Ontology**. 2025.
Reference Source

28. Fidler JA, Shahab L, West R: **Strength of urges to smoke as a measure of severity of cigarette dependence: comparison with the Fagerström test for Nicotine Dependence and its components.** *Addiction.* 2011; **106**(3): 631–638.
PubMed Abstract | Publisher Full Text

29. West R, Godinho CA, Bohlen LC, *et al.*: **Development of a formal system for representing behaviour-change theories.** *Nat Hum Behav.* 2019; **3**(5): 526–536.
PubMed Abstract | Publisher Full Text

30. Hale J, Hastings J, West R, *et al.*: **An Ontology-Based Modelling System (OBMS) for representing behaviour change theories applied to 76 theories [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2020; **5**: 177.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Berger U, Bell A, Michael Barton C, *et al.*: **Towards reusable building blocks for agent-based modelling and theory development.** *Environ Model Softw.* 2024; **175**: 106003.
Publisher Full Text

32. Bolock AE, Abdennadher S, Herbert C: **An ontology-based framework for psychological monitoring in education during the COVID-19 pandemic.** *Front Psychol.* 2021; **12**: 2879.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Vu TM, Buckley C, Bai H, *et al.*: **Multiobjective genetic programming can improve the explanatory capabilities of mechanism-based models of social systems.** *Complexity.* 2020; **2020**: 8923197.
PubMed Abstract | Publisher Full Text | Free Full Text

34. West R, *et al.*: **Tools and resources for annotating datasets using ontologies.** *Open Science Framework.* 2025.