Considerations for Automating *Salmonella* Serovar Identification within the Electronic Public Health Reporting Environment

Jeffry C. Alexander

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Biomedical and Veterinary Sciences

Jeffrey R. Wilcke, Chair

Julie M. Green

Michael K. Martin

Nammalwar Sriranganathan

July 22, 2015

Blacksburg, VA

Keywords: Public Health, *Salmonella*, serovar, terminology, ontology, SNOMED CT, Kauffmann-White

Considerations for Automating *Salmonella* Serovar Identification within an Electronic Public Health Reporting Environment

Jeffry C. Alexander

Abstract

CDC's requirements for *Salmonella* surveillance reporting include submission of serovars from the recognized naming scheme, Kauffmann-White (K-W), using identifiers curated by the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT®). Translating the serotype formula of a *Salmonella* isolate to the correct identifier has been a multistep manual process for users. Our goal was to determine whether a degree of automation could be achieved using an ontology based on K-W.

We investigated information artifacts presently available, namely K-W, SNOMED CT and CDC's Public Health Information Network – Vocabulary Access and Distribution System (PHIN-VADS). As SNOMED CT creates identifiers and associates them with serovar names, we performed detailed analysis on its coverage of K-W. An overall error rate of 13.1% included simple omissions and transcription errors. We limited our assessment of K-W and PHIN-VADS to the functional characteristics of the resources they distribute. K-W creates serovar names but does not provide identifiers. PHIN-VADS includes the identifiers but not antigenic formulae for most isolates. In summary, neither K-W nor PHIN-VADS contained all information users require.

Two different ontology prototypes were developed. Prototype I placed K-W serovars as terminal nodes in the hierarchy and these were given logic-based definitions. Prototype II added isolate classes as serovar subtypes. Only the isolate classes had complete logical definitions. Both prototypes were logically sound and functioned as expected. Prototype I paralleled existing SNOMED CT content but required more robust description logic than currently employed in SNOMED CT. Prototype II was more compatible with current functionality of SNOMED CT but created identifiers that would not meet current requirements for public health reporting.

Prototype I was fully populated as the *Salmonella* Serotype Designation Ontology (SSDO). As it stands, SSDO reliably places isolates in the appropriate classes, with few and predictable exceptions. Although SNOMED CT cannot accommodate its functionality at this time, SSDO can serve as the basis for a stand-alone application.

Ultimately whether by improving functionality of existing systems or providing a framework for an ancillary automated system, this work should facilitate real-time reporting and analysis of surveillance data that will prevent new or reduce severity of infectious disease outbreaks.

Acknowledgements

This work was supported in part by the Stamps Family Charitable Foundation, Inc. and was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

A sincere expression of gratitude goes out to the doctoral advisory committee. Without their guidance, encouragement, and helpful input, none of this work would have been possible.

Table of Contents

List of Figures

Figure X.11 (p. 121): Create a copy of the table, then delete the antigen phase column.

Figure X.12 (p. 122): In the duplicate table replace “Has component part” with “Has proper physical part” and delete the antigen column. Also remove any row that corresponds to an O antigen. The O antigens do not need a row that depicts the attribute “Has proper physical part”.

Figure X.13 (p. 122): Change column names to SNOMED relationship column names, create unique identifiers for each row, and substitute SCTID for terms where they exist (e.g. Salmonella Aachen = 114638004). Create them where they do not (some serovars and all attributes and values are not present in SNOMED). HCP = 90282300c, HPPP = 90282400c, Salmonella_O_Antigen = 90265600c, Salmonella_H_1 Antigen = 90265100c, O_17 = 80259500c, H_z35=90277400c, H_1=90265600c, H_6=90265900c. Note there is only one row for Role Group 1 (the O antigen row).

Figure X.14 (p. 123): Stated Relationships Table.

List of Tables

1. Introduction

Infection with *Salmonella* has been estimated to cause more than 1.2 million illnesses annually in the United States, with more than 23,000 hospitalizations and 450 deaths. (Scallan, et al., 2011) Salmonellosis is a reportable disease, and as such The Centers for Disease Control and Prevention (CDC) directs the National *Salmonella* Surveillance System, which houses and analyzes data collected by state and territorial public health laboratories (PHL). In these scenarios, physician's offices, health clinics, and clinical diagnostic laboratories submit the actual *Salmonella* isolates to a state or territorial PHL. The PHL then confirms the isolates as *Salmonella,* performs serotyping, identifies the isolates according to the Kauffmann-White (K-W) scheme, and reports the results to the CDC. Unusual or untypeable isolates are forwarded to the CDC National *Salmonella* Reference Laboratory at the Enteric Diseases Laboratory Branch (EDLB) for further characterization or confirmation; results are reported back to state and territorial PHLs. (CDC, 2011a)

1.1 Reporting and maintenance of data

State and territorial PHLs send reports electronically to the CDC through a variety of mechanisms. Initially, all surveillance data were transmitted through the Public Health Laboratory Information System (PHLIS), but other methods of data transmission have been implemented over time. Currently data are collected into the Laboratory-based Enteric Disease Surveillance (LEDS) system, which has subsequently replaced PHLIS. The Division of Foodborne, Waterborne, and Environmental Diseases (DFWED) in the National Center for Emerging and Zoonotic Infectious Diseases maintains the national *Salmonella* surveillance data in LEDS. The annual summaries of these data are the only regularly published national source of serotype information for *Salmonella.* (CDC, 2011b)

One of the stated goals of the 2009 American Recovery and Reinvestment Act (ARRA) is to increase the Meaningful Use (MU) of technology among the medical community, and an incentive program has been established to encourage adoption of enhanced technologies. CDC's involvement in MU includes using standard vocabulary to support lab result transactions to public health authorities. (Public Health Information Network, CDC, 2013)

The CDC has stipulated that laboratories are to report organism identities using the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT®) concept identifiers (SCTIDs). SNOMED CT, a product of the International Health Terminology Standards Development Organisation (IHTSDO), creates a unified and unique set of identifiers for a range of infectious disease agents (i.e., bacteria, viruses, fungi, protozoa, etc.). Where recognized authorities exist, the identifiers are assigned to official names maintained by those authorities. The WHO Collaborating Centre for Reference and Research on *Salmonella* (WHOCC-Salm) maintains K-W, the official list of *Salmonella* serovars.

SCTIDs are used in computer applications and information networks to facilitate accurate transmission and receipt of causative agents for disease outbreak tracking. The CDC's Public Health Information Network-Vocabulary Access and Distribution System (PHIN-VADS) extracts content from SNOMED CT and creates lists of terms and codes grouped by reporting function or project and makes them publicly available. *Salmonella* identifiers are found among all microorganisms contained in a CDC authored file that can be downloaded as a spreadsheet[1]. Each row in the spreadsheet includes the SCTID, fully specified name (FSN) and preferred name for each serotype in addition to several columns relating to the identity and version of the coding system. FSNs are intended to be unique, unambiguous descriptions for SNOMED CT concepts that include a hierarchy designator in the name (e.g. organism). For the *Salmonella* content in SNOMED CT the preferred terms are the FSNs with the hierarchy designator removed. Because PHIN-VADS is directly extracted from the SNOMED CT tables, this is also true for the PHIN-VADS preferred names. *Salmonella* serotypes are represented in PHIN-VADS as a term list populated from SNOMED CT's fields, which include K-W's definitive names.

[1] PHIN Vocabulary Access and Distribution System (VADS). PHVS_Microorganism_CDC accessed from: https://phinvads.cdc.gov/vads/ViewValueSet.action?id=06B09CEF-0E37-E111-A720-0050568D00F8

2. Background on *Salmonella* Nomenclature

The metalanguage that describes the classification of *Salmonella* is likely unfamiliar to many in the intended audience for this dissertation. This language is used both formally and informally by experts who are familiar with *Salmonella* serotyping and naming and so the specific meaning of terms may change with time and circumstance. Here in Table 1, as a reference and for clarity, we define several terms we intend to use consistently. It is our intent to use these terms in a formal way, eschewing informal usage for the sake of clarity.

Table 1. Definitions for common terms in this dissertation.

Salmonella antigen	Structural molecules of *Salmonella* that can be detected via serology (or other laboratory methods) and used to uniquely identify related isolates according to the K-W scheme. The K-W scheme is based on somatic (O), flagellar (H) and in special circumstances envelope (Vi) antigens.
Antigenic formula	List of *Salmonella* antigens arranged in a standard order and delimited by punctuation specified according to the K-W scheme.
Serotype	Formally, a name for a *Salmonella* isolate or serovar created by appending the antigenic formula to the *Salmonella* subspecies or species to create a unique serotype name. Informally, serotype is sometimes used as a synonym for antigenic formula.
Isolate serotype	The serotype of a clinical isolate. An instance of one of the serotype / serovar classes in K-W.
Serovar	A serologic variant formally recognized in K-W as a unique grouping of isolates. Informally, serovar is often used as a synonym for serotype and for antigenic formula.
K-W serovar	A name for a *Salmonella* isolate class formally defined in the K-W scheme. In certain cases more than one serotype is grouped under the same K-W serovar (e.g., *Salmonella* Typhimurium). This is a direct reference to the Title of the document released by WHOCC-Salm that creates the names for officially recognized *Salmonella* variants.
Serogroup	A collection of serovars that all bear a certain antigen (or group of antigens). For *Salmonella* in particular, the serogroups are formed as a set of K-W serovars that contain specific O antigen(s) for that group.

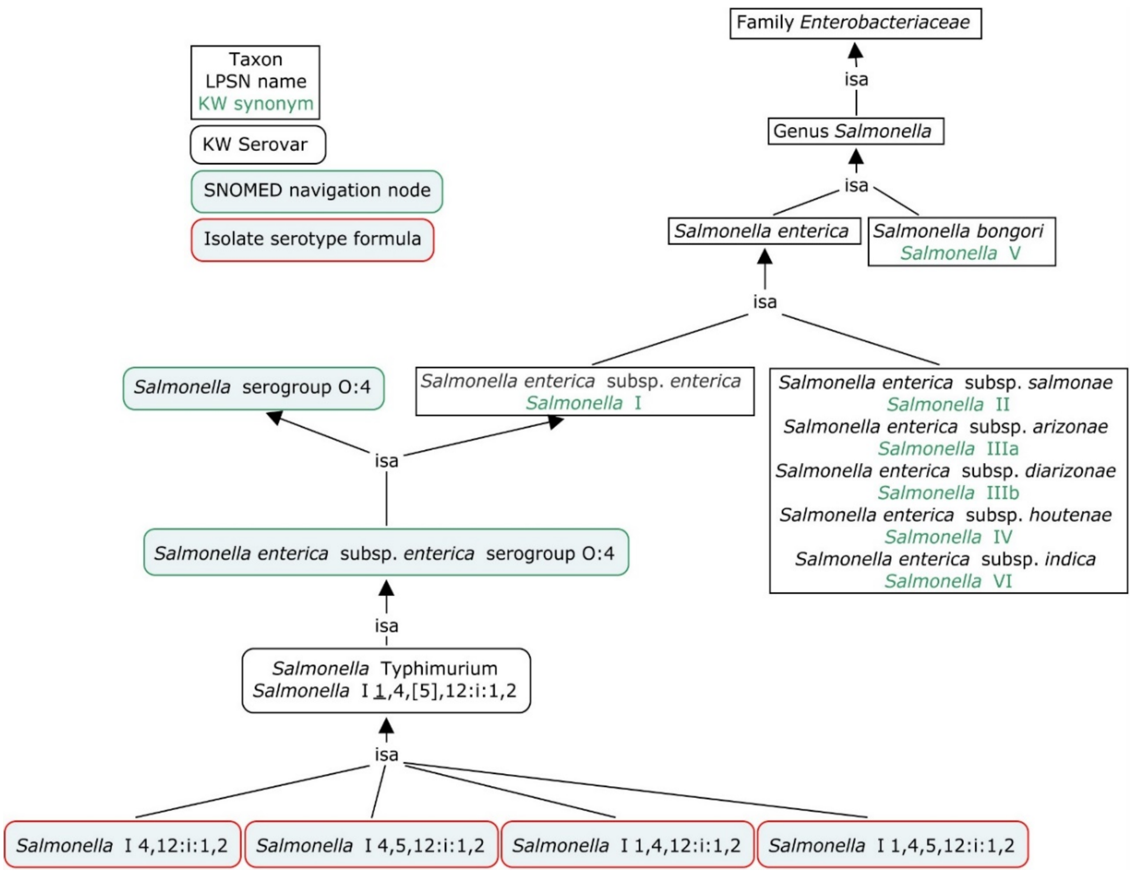
2.1 Taxonomy and Naming Conventions

Rules governing *Salmonella* taxonomy have evolved considerably with time, which now leads to significant complexity, ambiguity and debate for those who interact with the taxonomy. (Sneath, 2003) (Tindall, Grimont, Garrity, & Euzéby, 2005) Conventions related to the naming of living bacterial organisms are established in the International Code of Nomenclature of Bacteria. (Lapage, et al., 1992) The International Committee on Systematics of Prokaryotes (ICSP) developed this code, which describes the rules for denoting taxa of bacteria based on rank. The most recent edition of the Code is the 1990 Revision, published in 1992 by the American Society for Microbiology. Subsequent amendments have been made since, and these can be found in issues of the peer-reviewed *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) and are published online by the List of Prokaryotic names with Standing in Nomenclature (LPSN). (Euzéby, 1997)

The nomenclature of *Salmonella*, however, is taken further than the customary binomen and includes subspecies for the species *S. enterica*. Biologists commonly treat taxonomy in a very formal way and equate it with the system of names and categories first described by Linnaeus. Bacteriologists consider organism names ranging from Kingdom (e.g., Bacteria) to the subspecies level (e.g., *Salmonella enterica* subsp. *enterica*) to be “taxonomic”, a formal use of the term. The K-W scheme extends formal Linnaean taxonomy by organizing *Salmonella* subtypes according to serotyping. The scheme extends the hierarchy from subspecies to serotype. The resulting whole can be treated as a single taxonomy, a less formal application of the term.

K-W also added a variation to *Salmonella* naming in the formal Linnaean taxonomy in that it abbreviates species and subspecies with roman numerals which simplifies the lengthy names that result from the addition of their respective antigenic formulae. Figure 1 is a concept map of classes and relationships that position a single K-W serovar within these combined nomenclature schemes. The focus of the map is a single K-W serovar, *Salmonella* Typhimurium, which represents 4 isolate serotypes and is a taxonomic descendent of *Salmonella enterica* subsp. *enterica* from the Linnaean taxonomy.

Figure 1. A representative section of a unified *Salmonella* serotype taxonomy. K-W synonyms for *Salmonella* subspecies are included where they apply. Concepts referred to as SNOMED navigation nodes reference the K-W serogroups. Isolate serotype formulae are not included directly in SNOMED CT or in the K-W scheme.


Taxon
LPSN name
KW synonym
KW Serovar
SNOMED navigation node
Isolate serotype formula
Family Enterobacteriaceae
isa
Genus Salmonella
isa
Salmonella enterica
Salmonella bongori
Salmonella V
isa
Salmonella serogroup O:4
Salmonella enterica subsp. enterica
Salmonella I
Salmonella enterica subsp. salmonae
Salmonella II
Salmonella enterica subsp. arizonae
Salmonella IIIa
Salmonella enterica subsp. diarizonae
Salmonella IIIb
Salmonella enterica subsp. houtenae
Salmonella IV
Salmonella enterica subsp. indica
Salmonella VI
isa
Salmonella enterica subsp. enterica serogroup O:4
isa
Salmonella Typhimurium
Salmonella I 1,4,[5],12:i:1,2
isa
Salmonella I 4,12:i:1,2
Salmonella I 4,5,12:i:1,2
Salmonella I 1,4,12:i:1,2
Salmonella I 1,4,5,12:i:1,2

In the end, a full taxonomy (hierarchy) describing the genus *Salmonella* includes content defined by formal systematics and published in IJSEM as well as content defined by K-W that exists beyond the declared scope of formal systematics. This combination of naming conventions has led to substantial difficulty in laboratories reporting correct and timely information concerning *Salmonella* disease surveillance data.

As stated, the formal *Salmonella* taxonomy has undergone several revisions over time, and this project is focused on the current state of taxonomy, where the genus *Salmonella* is divided into two species, *Salmonella enterica* and *Salmonella bongori*. (Grimont & Weill, 2007) *Salmonella enterica* is further subdivided into six subspecies that are designated by taxonomic names ("latinized" trinomials) and denoted by the Roman numerals I, II, IIIa, IIIb, IV, and VI as noted in Table 2. *Salmonella bongori* was originally designated *S. enterica* subspecies V; it has since been determined to be a separate species of *Salmonella*. However, for simplicity and convenience, these strains are still sometimes referred to as "subspecies V". (McQuiston, et al., 2008)

Table 2. *Salmonella enterica* subspecies names and K-W synonyms.

Subspecies names	**K-W simplification for serotyping**
Salmonella enterica subsp. *enterica*	*Salmonella* I
Salmonella enterica subsp. *salamae*	*Salmonella* II
Salmonella enterica subsp. *arizonae*	*Salmonella* IIIa
Salmonella enterica subsp. *diarizonae*	*Salmonella* IIIb
Salmonella enterica subsp. *houtenae*	*Salmonella* IV
Salmonella enterica subsp. *indica*	*Salmonella* VI

2.2 K-W Scheme

Fritz Kauffmann and P. Bruce White proposed serotyping as the basis for a classification scheme for *Salmonella* in 1934. Known colloquially as the Kauffmann-White scheme and officially as the White- Kauffmann-Le Minor scheme, it has grown over time to include approximately 2600 individual serotypes. (Guibourdenche, et al., 2010)

As mentioned previously, the scheme is maintained by a division of the World Health Organization, the WHOCC-Salm. The historical role of WHOCC-Salm has been to maintain the comprehensive list of known *Salmonella* serotypes, but now duties include:

- Updating the *Salmonella* serotyping scheme (antigenic factors and serovar nomenclature).
- Technical support for *Salmonella* National Reference Centers (unusual antigenic structures or biochemical features).
- Updating protocols for the production of antisera.
- Research activities on *Salmonella* (molecular methods for the identification of serovars).
- Contribution to the WHO surveillance program (recommendations, training, participation to Global Salm-Surv).

Validation of new serotypes is done at WHOCC-Salm (Institut Pasteur) in collaboration with laboratories in Hamburg, Germany (Institut für Hygiene und Umwelt) and Atlanta (CDC). New serovars are approved when these three laboratories agree on their validation. (Grimont & Weill, 2007)

Per the K-W scheme, *Salmonella* serovars are identified on the basis of the immuno-reactivity of cell surface structures, predominantly the O and H antigens. After determining the species and/or subspecies, serological characterization of a *Salmonella* isolate begins by determining its complement of O antigens. An O antigen is a polysaccharide that is the outermost component of the cell surface layer of lipopolysaccharide (LPS). They are typically composed of four to six sugars, often called

O subunits. Differences between O antigens can result from 1) the sugar components of the O subunit, 2) the type of the covalent bonds between the sugars within the O subunit, or 3) the manner of the linkage between the O subunits that form the O antigen polymer. (CDC, 2011b)

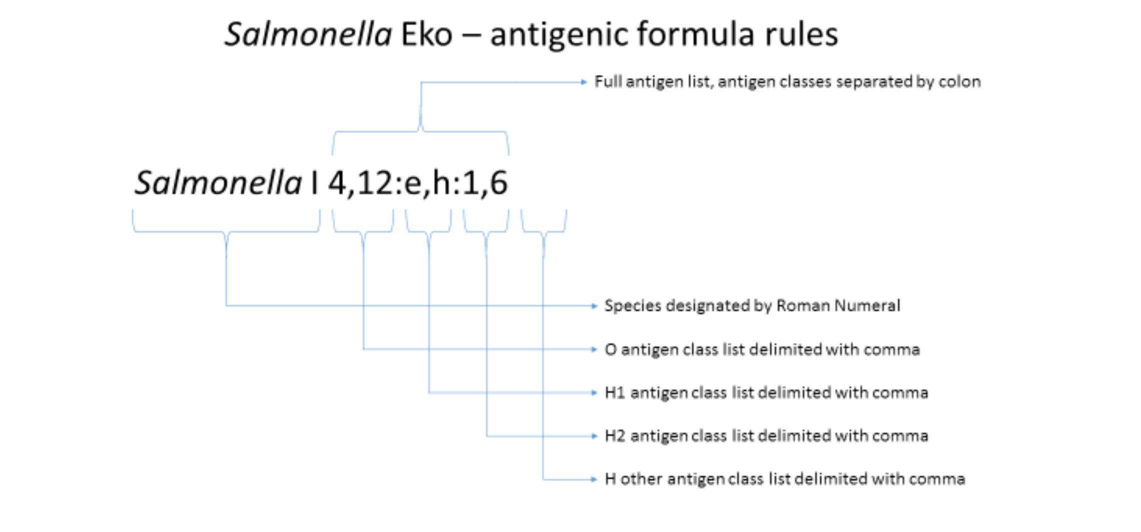
Within the K-W scheme, collections of serovars that contain a certain antigen or set of antigens are known as serogroups. These determining antigens are called O group antigens. Interestingly, individual isolates may also carry additional O antigens that do not define the serogroup. These additional polysaccharides are found on the core structure, and are referred to as ancillary O antigens. These ancillary antigens are associated with specific O serogroups and are often variably present as they are encoded by extra-chromosomal elements (e.g. bacteriophages, plasmids). (Fields, 2006) Regardless of ancillary antigens, K-W organizes the list of all serovars into tables according to their O group.

Serogroups were originally designated by alphabetic letters, and eventually it was necessary to continue with numbers. For consistency in the scheme, all serogroups were given a number designation; however, the most common O groups were originally labeled by letter and are still commonly referred to this way (e.g., serotype Enteritidis belongs to group O:9 or Group D1, serotype Typhimurium belongs to Group O:4 or Group B). When multiple O epitopes are present, they are listed sequentially and separated by commas. As of 2011, there are 46 described O groups in total. (Fitzgerald, et al., 2007) Appendix A shows the CDC's Table A of *Salmonella* O groups and associated O antigens.

H antigens are part of the filamentous portion of the bacterial flagellum, which is composed of protein subunits called flagellin. Each end of the flagellin is mostly conserved between serovars, while the center region is presumed to be the antigenically variable portion. Due to its structure, this is the part that is exposed on the surface of the flagellum. *Salmonella* is unique among enteric bacteria in that it can express two different flagellin antigens in separate phases, referred to as Phase 1 and Phase 2. Interestingly, only one phase is expressed at a time in a single bacterial cell. A few *Salmonella* isolates are monophasic and express only a single flagellin type. These occur naturally for some serotypes (e.g., serotypes Enteritidis and Typhi) or can occur through the loss or lack of expression of a flagellin gene. In rare instances, isolates can express a third flagellar antigen and are referred to as triphasic. The corresponding H antigen is listed in K-W as “other”. Currently, there are 114 recognized H antigens which may be expressed singly or in combination with other within a given phase. Antigen complexes within the same phase are separated by commas in the formula; for example, the second flagellar phase of serovar Typhimurium is composed of antigens 1 and 2, which is represented as “1,2”. (McQuiston, Waters, Dinsmore, Mikoleit, & Fields, 2011) Appendix B shows the CDC’s Table B of *Salmonella* H antigens.

Currently, all *Salmonella* serovars can be designated by their subspecies Roman numeral and list of antigens. This formula name is comprised of the O and H antigens detected from serotyping. The antigen groups (i.e O, H1, H2, H other) are separated by colons; the individual antigen factors within a group are separated by commas (Figure 2). A simple example would be *Salmonella* I 4,12:e,h:1,6. In addition, serovars of *S. enterica* subspecies I are given non-formulaic names. In the previous example, *Salmonella* I 4,12:e,h:1,6 is reported as *Salmonella* Eko. In essence, all fully serotyped *Salmonellae* I are reported as binomial names, while all others (II-VI) are reported as an antigenic formulae, although some surveillance systems have the ability to capture both. (CDC, 2013)

Figure 2. Typical naming conventions for *Salmonella* serovars.


Salmonella Eko – antigenic formula rules
Full antigen list, antigen classes separated by colon
Salmonella I 4,12:e,h:1,6
Species designated by Roman Numeral
O antigen class list delimited with comma
H1 antigen class list delimited with comma
H2 antigen class list delimited with comma
H other antigen class list delimited with comma

2.2.1 Using K-W

After a laboratory has identified the *Salmonella* species and/or subspecies and the list of O and H antigens for a particular *Salmonella* isolate, the correct K-W serovar can be determined. The document distributed by WHOCC-Salm organizes serotype names in O serogroup tables. These tables are arranged to facilitate manual identification of serotype names when antigen type and value have been determined for an isolate.

A laboratory technician first locates the correct serogroup table based on the isolate's combination of O antigens. For most serovars this is a simple matter of matching the O antigens of the isolate with the name of the serogroup table. In some cases the O antigens of an isolate do not fully match the O antigens of the serogroup to which it belongs. As stated previously, K-W serovars may express O antigens in addition to those that are mandatory for the group, the so-called ancillary antigens. For rare K-W serovars, one or more of the serogroup antigens may be optional. When these exceptions are encountered, finding the correct K-W serovar based on the serotype of a particular isolate requires significant familiarity with the scheme.

Once the serogroup is properly identified, the table for that serogroup is traversed to complete the selection of the proper K-W serovar for the isolate. This process is initiated by the user matching the O antigens of the isolate, followed by the H1 antigens, the H2 antigens and finally the "other" H antigens. Once a satisfactory match is noted in all the columns the technician has successfully identified the K-W serovar.

2.3 Other attempts to facilitate assignment of serotype names

Partial automation of K-W was made available in 2012, when Luminex® released a Salmonella Analysis Tool that accompanied their xMAP® Salmonella Serotyping Assay. The analysis tool, distributed as a Hypertext Markup Language (HTML) document, allows the user to choose/input antigen data, and gives possible K-W serotypes as results. In essence, Luminex has created a graphical user interface (GUI) for K-W that allows users to utilize pull-down menus in which to enter specific antigens. Drawbacks to this application in the realm of PHL reporting are that the resulting serotype names do not have an associated SCTID, nor does Luminex fully account for the use of synonyms and optionality of antigens. In fact, Luminex alters several of the formatting and punctuation strategies that are used in K-W. This renders the matching of serotypes found in Luminex with those serotypes in K-W, SNOMED CT and PHIN-

VADS impossible. Moreover, Luminex does not provide all O antigens in the drop-downs, only those formerly designated by a letter.

3. Description of Problem

The CDC's expectation of reporting the correct SNOMED CT identifier (SCTID) for the *Salmonella* serotype associated with a particular submission is supported by three information artifacts that are maintained independently by separate organizations. As such, problems arise as each independent object, noted in Table 3, uses distinct logical schemes, and approaches access and distribution differently.

Table 3. *Salmonella* Serotype Information Artifacts.

Information artifact	**Distribution method/format**	**Role in *Salmonella* reporting**
Kauffmann-White (9th edition, 2007)	pdf	Establishes identities of *Salmonella* serovars, creates names and creates associations that group serovars based on required and optional antigens.
SNOMED CT	Relational database, xml/owl	Inserts the *Salmonella* serovars into a logical scheme and establishes (universally) unique identifiers for each.
PHIN-VADS (phvs_microorganism_cdc_v7)	Excel spreadsheet	Associates lists of concepts with specific program and reporting requirements.

Biochemical and serological tests identify *Salmonella* through presence or absence of particular enzymatic reactions and "anatomical" structures. The biochemical tests are performed mainly to differentiate species and subspecies, while serological results determine the O and H antigens, which distinguish individual serotypes. (Brenner & McWhorter-Murlin, 1998) (Mikoleit & Fields, 2006)

K-W is recognized internationally as the official source for *Salmonella* serovar naming and grouping and create organizing principles to determine whether particular serotype variations are important or inconsequential. K-W organizes serovars according to antigenic formulae, places them in serogroups, and assigns names to them. The CDC and Pasteur Institute evaluate isolates presented as part of epidemiologic surveillance for potential addition to, or incorporation within, the scheme. When a new unique antigenic formula is considered to represent a unique serovar it is added to the scheme. If however, the new unique antigenic formulae is considered to be a variation on the formula of an existing serovar, it is incorporated into the existing K-W serovar. The antigenic formula of the affected serovar may require adjustment to accommodate one or more antigens that are considered optional.

SNOMED CT incorporates the serovar and serogroup names from K-W and assigns unique identifiers. One functional advantage of SNOMED CT is that it is capable of managing synonymy, in this case allowing both a serovar name and the corresponding antigenic formula. The content is organized into a representative hierarchy, as individual serotypes are listed as subtypes of their respective O groups. SNOMED CT creates the identifiers and associates them with K-W serovar names. For K-W serovars of *Salmonella* I, each identifier is associated with at least two text strings, the K-W binomial and the antigenic formula. For K-W serovars of *Salmonella* II-VI, each identifier is associated only with the antigenic formula.

PHIN-VADS provides users with lists of identifiers for use in CDC messages. (CDC, 2011c) For *Salmonella*, the PHIN-VADS content associates the SNOMED CT identifiers with the SNOMED CT preferred term and the SNOMED CT FSN. PHIN-VADS does not specifically identify which concepts (table rows) are *Salmonella* serovars and does not include synonyms such as the antigenic formulae of *Salmonella* I serovars. In PHIN-VADS, SCTIDs for the *Salmonellae* I are only associated with their K-W binomial while *Salmonella* II-VI serovars are only associated with their K-W serovar names (which are antigenic formulae).

Taken together K-W, SNOMED CT and PHIN-VADS form a terminology information chain intended to inform the selection and submission of recognized names for *Salmonella* serovars in public health. The chain begins with the collection and serotyping of *Salmonella* isolates by PHLs. The information found within K-W determines the significance of antigenic variants amongst the isolates, assigning a specific serotype a serovar name and creating serogroups based on somatic antigen expression.

Once the antigens of an isolate have been determined, the K-W scheme can be used directly to identify the K-W serovar. In order to correctly name an isolate using the K-W scheme as distributed in pdf format, a lab technician would first locate the particular serogroup section, and then search alphabetically by the first flagellar antigen listing, followed by the second flagellar antigen until the specific serotype was found. In the case of a *Salmonella* I, the binomial name would be used to find the associated SCTID code (the value required by the public health network) by an alphabetic or string search of SNOMED CT or the PHIN-VADS spreadsheet. For a *Salmonella* II-VI, the actual serotype formula is used to find the serovar name in K-W, again by string searching. The challenge for these latter isolates is that the name and antigenic formula of the serovar may include optional antigens not present in the isolate.

SCTIDs may be listed in more than one place (e.g. SNOMED CT itself, and PHIN-VADS). Unfortunately, local installations of the entirety of SNOMED CT at the state public health lab level can be cumbersome and difficult to maintain, whereas PHIN-VADS supplies spreadsheet composed of a subset of all SNOMED microorganism content and associated SCTIDs. Problems arise here when 1) errors from SNOMED CT are carried through to PHIN-VADS, and 2) the identifiers used in PHIN-VADS are not associated with the text string that represents the antigenic formula of a laboratory isolate.

For any digital version of an artifact in the information chain described above, users assume that simple string matching is reliable[2]. Unfortunately, users are unable to locate correct identifiers based on either antigenic formulae or serovar binomial names via simple string matching for several reasons:

1) A direct representation of the antigenic formula of an isolate may not be present in any of the artifacts. This occurs when an isolate is one that is a subtype of a K-W serovar that includes optional antigens.
2) The K-W serovar name is not transcribed faithfully into SNOMED CT. As the PHIN-VADS organism list should be a faithful reproduction of SNOMED CT content, transcription errors will convey to PHIN-VADS.
3) The K-W serovar name is a K-W binomial. SNOMED CT generally associates the K-W binomial with the serotype formula but PHIN-VADS lists ONLY the K-W binomial.
4) The K-W serovar name includes formatting (e.g. underline in the case of *Salmonella* I 1,4,12:i:1,2) that will not match properly. SNOMED CT lists these antigens without underlining but this is not a direct representation of the actual K-W serovar name string.

[2] Simple text searching is available in most pdf readers (e.g. Adobe Acrobat for K-W), the VTSL online Browser (SNOMED CT) and Microsoft Excel (PHIN-VADS).

5) Search algorithms and indexing engines vary in their handling of string length and punctuation. Depending on the settings, antigenic formulae may be treated as lists of one and two character words. As a result they may simply be ignored by the search engine.

6) K-W itself does not represent the serovar formulae as simple strings, instead the antigens are listed in columns headed by O, H1, H2, and H other.

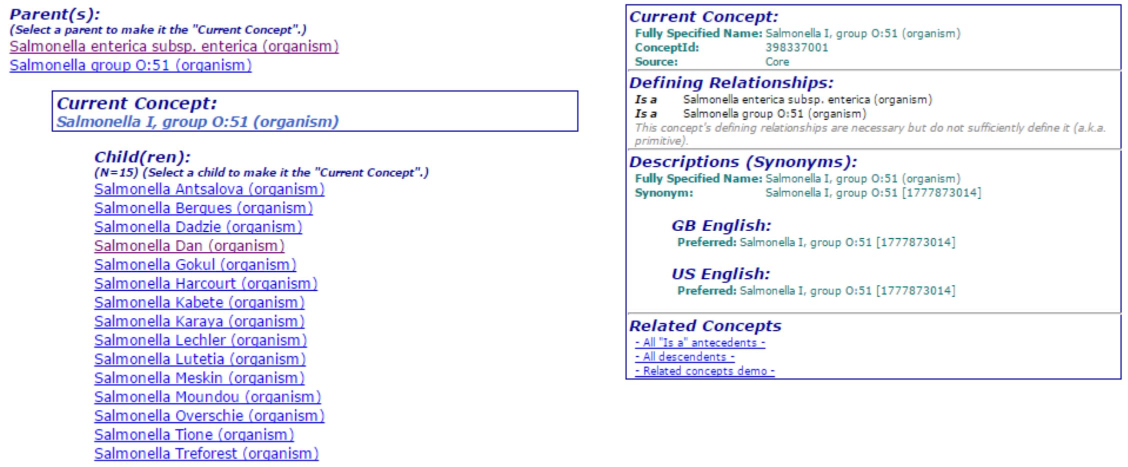
4. Dissertation Goal

The overarching goal of this investigation is to improve the ability of laboratory personnel to correctly assign the SCTID of a *Salmonella* serovar based on the antigenic formula of a laboratory isolate. Laboratory personnel presently rely on an information chain that begins with the experts at K-W and ends with the creation of PHIN-VADS content tables. We propose first to critically evaluate the current state of the information chain itself. Two issues in this regard will be considered. The first, and most important, is the fidelity of the information. Specifically, SNOMED CT must represent the K-W serovars faithfully when it creates the identifiers. We will also consider whether the desired utility can be achieved given structural and functional limitations of the artifacts. Finally, we will investigate whether or not the K-W serovars can be organized using description logics, and if an ontology based on these logics reliably associates antigenic formulae of laboratory isolates with the correct K-W serovar and its associated SCTID.

We were introduced to this subject when the Virginia Division of Consolidated Laboratory Services (VDCLS) asked if we could help them locate SCTIDs for *Salmonella* serovars using the online SNOMED CT Browser developed and maintained by Veterinary Terminology Services Laboratory (VTSL) at Virginia Tech. (VTSL, 2012) A specific issue from VDCLS was that SNOMED FSNs only contain the binomial designation for *Salmonella* I serotypes, not antigenic formulae, which is the natural output of serotyping from a laboratory. Staff were able to locate the lists of binomial names in the VTSL browser, as the lists contained the subtypes of the serogroup in which the particular isolate fell. We subsequently met with representatives of VDCLS to learn what difficulties they were experiencing.

Our conversation revealed that VDCLS staff were not using the text search feature of the browser. Instead they located the concept for the *Salmonella* subspecies they had isolated. The first level subtypes of any *Salmonella* subspecies in SNOMED CT is a list of serogroups that include members of that subspecies. Selecting the appropriate subspecies / serogroup concept generates the view seen in Figure 3a. Once this list was located, they systematically inspected the full entry for each serovar to view its antigenic formula. An example is shown in Figure 3b for *Salmonella* Dan where the antigenic formula is listed as a description.

Figure 3a. VTSL browser depiction of *Salmonella enterica* subsp. *enterica* serovars of Group O:51.



Parent(s):
(Select a parent to make it the "Current Concept".)
Salmonella enterica subsp. enterica (organism)
Salmonella group O:51 (organism)

Current Concept:
Salmonella I, group O:51 (organism)

Child(ren):
(N=15) (Select a child to make it the "Current Concept".)
Salmonella Antsalova (organism)
Salmonella Bergues (organism)
Salmonella Dadzie (organism)
Salmonella Dan (organism)
Salmonella Gokul (organism)
Salmonella Harcourt (organism)
Salmonella Kabete (organism)
Salmonella Karaya (organism)
Salmonella Lechler (organism)
Salmonella Lutetia (organism)
Salmonella Meskin (organism)
Salmonella Moundou (organism)
Salmonella Overschie (organism)
Salmonella Tione (organism)
Salmonella Treforest (organism)

Current Concept:
Fully Specified Name: Salmonella I, group O:51 (organism)
ConceptId: 398337001
Source: Core

Defining Relationships:
Is a Salmonella enterica subsp. enterica (organism)
Is a Salmonella group O:51 (organism)
This concept's defining relationships are necessary but do not sufficiently define it (a.k.a. primitive).

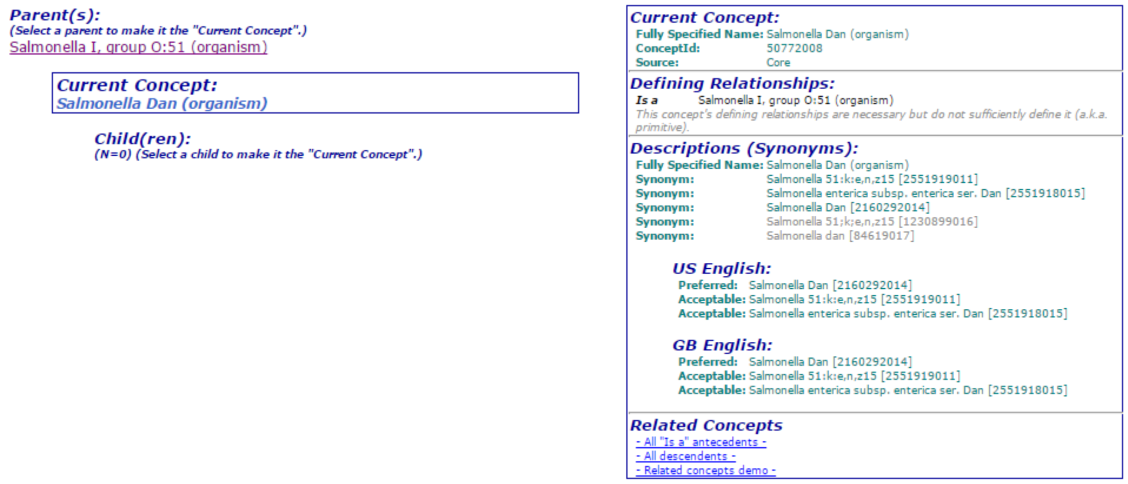
Descriptions (Synonyms):
Fully Specified Name: Salmonella I, group O:51 (organism)
Synonym: Salmonella I, group O:51 [1777873014]

GB English:
Preferred: Salmonella I, group O:51 [1777873014]

US English:
Preferred: Salmonella I, group O:51 [1777873014]

Related Concepts
- All "Is a" antecedents -
- All descendents -
- Related concepts demo -

Figure 3b. VTSL browser depiction of *Salmonella* Dan.



Parent(s):
(Select a parent to make it the "Current Concept".)
Salmonella I, group O:51 (organism)

Current Concept:
Salmonella Dan (organism)

Child(ren):
(N=0) (Select a child to make it the "Current Concept".)

Current Concept:

Fully Specified Name:	Salmonella Dan (organism)
ConceptId:	50772008
Source:	Core

Defining Relationships:

Is a Salmonella I, group O:51 (organism)
This concept's defining relationships are necessary but do not sufficiently define it (a.k.a. primitive).

Descriptions (Synonyms):

Fully Specified Name:	Salmonella Dan (organism)
Synonym:	Salmonella 51:k:e,n,z15 [2551919011]
Synonym:	Salmonella enterica subsp. enterica ser. Dan [2551918015]
Synonym:	Salmonella Dan [2160292014]
Synonym:	Salmonella 51;k;e,n,z15 [1230899016]
Synonym:	Salmonella dan [84619017]

US English:

Preferred:	Salmonella Dan [2160292014]
Acceptable:	Salmonella 51:k:e,n,z15 [2551919011]
Acceptable:	Salmonella enterica subsp. enterica ser. Dan [2551918015]

GB English:

Preferred:	Salmonella Dan [2160292014]
Acceptable:	Salmonella 51:k:e,n,z15 [2551919011]
Acceptable:	Salmonella enterica subsp. enterica ser. Dan [2551918015]

Related Concepts
- All "Is a" antecedents -
- All descendents -
- Related concepts demo -

As VDCLS staff were aware that SNOMED CT descriptions included the antigenic formulae for the serovars they had, in fact, attempted to search for them directly. Unfortunately the indexing engine used for the browser is set to ignore words of fewer than three characters. Because the antigenic formulae include punctuation (comma, colon, etc.) the search engine will not return descriptions that contain these strings. For example, it is necessary to search for the quoted string '51:k:e,n,z15' in order to locate *Salmonella* Dan based on its antigenic formula.

While this solved the immediate problem for VDCLS staff, it left us with concerns about ancillary problems. First, SNOMED CT does not include a fully correct entry for the antigenic formula based name as specified by K-W. In the case of *Salmonella* Dan, the correct formulaic name is '*Salmonella* I 51:k:e,n,z15'. In order to locate the concept one must know that it is incorrect to include "*Salmonella* I" in the search string. Further, if a formula based description in SNOMED CT does not represent a K-W serovar correctly, if the isolate is new and has not yet been represented in K-W, or if the laboratory results are incomplete or erroneous, the search will ultimately fail.

Our own attempts to use SNOMED CT and the PHIN-VADS microbiology tables to look up *Salmonella* serovar identifiers led us to conclude that the resources available to laboratory personnel were inadequate. Using the K-W pdf requires significant familiarity with the scheme and the tables themselves. String-based searching for serovar formulae was unreliable for a variety of reasons. There were errors in the SNOMED CT representation of K-W names and antigenic formula, and PHIN-VADS faithfully replicates these errors. As a flat table structure is unable to associate SCTIDs with SNOMED CT descriptions, PHIN-VADS does not include the antigenic formula for *Salmonella* I serovars. Finally, neither SNOMED CT nor PHIN-VADS included direct representations of the actual antigenic formula of the isolates that laboratories encounter on a regular basis.

This dissertation evolved from our desire to address the shortcomings found within these systems by identifying barriers and offering potential remediation strategies. The dissertation goal is to address limitations in Electronic Laboratory Reporting (ELR) of *Salmonella* serotypes and names by improving the ability of public health laboratories to select correct K-W serotype names and assign correct SCTIDs to real laboratory results. Specific objectives associated with specific tactics to accomplish them are listed in Table 4. These include: 1) determining the state of content fidelity between K-W and the rest of the information chain including SNOMED CT® (Jan. 2012) and the PHIN-VADS microbiology table, 2) reducing dependence on string matching and facilitate automation by representing K-W antigenic formulae in a logical framework; 3) determining whether SNOMED CT can accommodate the logical framework required.

Table 4. Dissertation Objectives and Tactics.

Objective	**Tactic**
Identification of obstacles to efficient reporting of *Salmonella* serotypes and names	Analysis of relevant terminology artifacts: • Kauffmann-White classification • SNOMED CT • Public Health Information Network (PHIN) – Vocabulary Access and Distribution System (VADS)
Propose and foster improvements	Develop a logical representation of K-W on which to base a reorganization of SNOMED CT, inform modifications to SNOMED extraction process employed by PHIN-VADS, and to serve as a foundation for automation of *Salmonella* serotype to SCTID conversion

5. Assessing the Information Chain

Following our meeting with VDCLS, we undertook a detailed analysis of the issues that hinder efficient reporting of *Salmonella* serotypes through PHIN by performing informal assessments of K-W and PHIN-VADS, and an in-depth analysis of *Salmonella* content found within SNOMED CT.

5.1 Assessment of Kauffmann-White

K-W uses two distinct naming conventions for *Salmonella* serotypes. *Salmonella enterica* subsp. *enterica* (*Salmonella* I) serotypes are all given a binomial serovar name (e.g. *Salmonella* Enteritidis also known as *Salmonella* I 1,9,12:g,m:-). These are the most frequently encountered *Salmonella* serovars, and have been known as binomials for so long that it would be unrealistic to suppress these names and substitute their antigenic formula. (Grimont & Weill, 2007) The remaining five *Salmonella enterica* subspecies, as well as *Salmonella bongori* are designated by a Roman numeral followed by their antigenic formulae (e.g. *Salmonella* II 9,12:a:1,5).

K-W is distributed in a digital form (pdf) that does not directly support automation. Historically, the pdf format has focused on human readability independent of software, hardware, or operating system in lieu of machine readability. Text searches are unreliable, and tables and figures are difficult to export. This has improved over time, however the pdf for K-W suffers the same fate of earlier pdfs, in that exporting is not a straight forward process. In the K-W document, a small table serving as the header demarcates each serogroup. This poses a problem when attempting simple export, as the header is treated as a table, not text. In addition, while the antigens are listed in separate columns, the differentiation between the columns cannot be determined. The antigenic formulae determined in the lab contain colons and a syntax created by K-W to demarcate the different classes of antigens. These do not readily lend themselves to exportation.

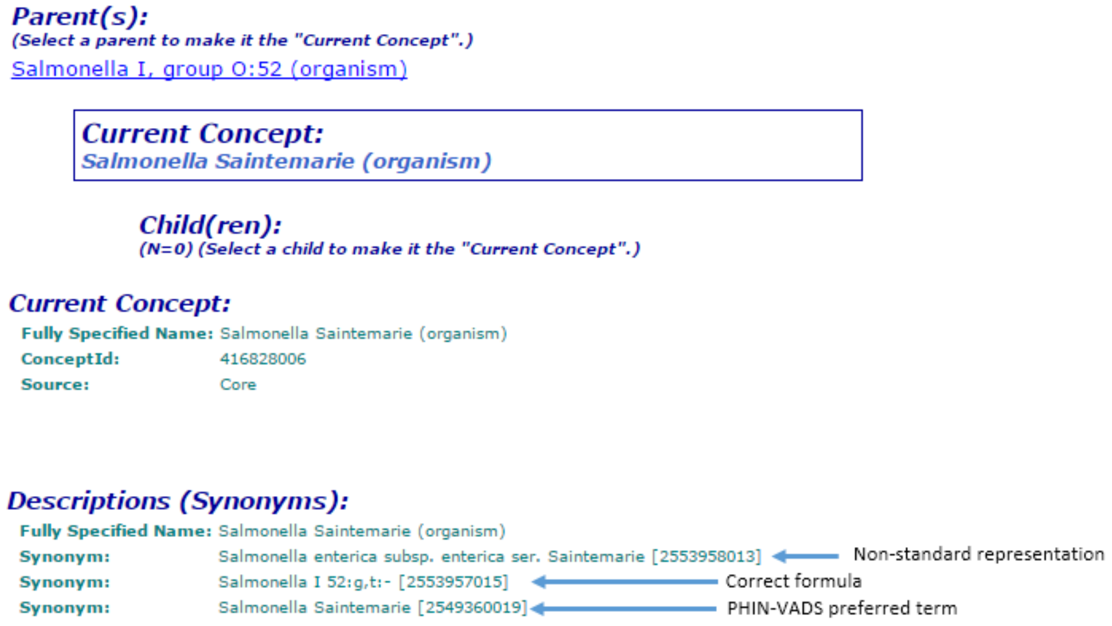
Some of the underlying logic of K-W, related especially to the presence of optional antigens in a formula, is conveyed to human users through punctuation (e.g. [ ]) and formatting (e.g. underline). Processing of text by computers, as is required for string matching, generally will recognize punctuation characters in a string but formatting is processed separately from the string. For this reason, non-textual representations (e.g. underlining) in K-W antigenic formulae are not accessible via any search function of SNOMED CT tables or the K-W pdf itself. The resulting antigenic formula of a *Salmonella* isolate relates directly to some specific row in the table that can be located by a human reader, but it cannot be matched using the relatively simple search capabilities built into pdf readers.

5.2 Assessment of SNOMED CT

From a logical perspective, SNOMED CT appears to correctly place individual serotypes within serogroups as specified in the K-W document. Furthermore, SNOMED CT replicates the K-W naming convention. Fully specified names are binomial (e.g., *Salmonella* Typhimurium) for *Salmonella* I and serotype formulae *Salmonella* for II-VI (e.g. *Salmonella* II 9,46:z:z39). In addition, the list of SNOMED CT descriptions for each K-W serovar includes at least one representation of its antigenic formula. Generally, the serotypes are correctly associated with the binomial names among the *Salmonella* I serotypes. Figure 4a depicts an example of a K-W serovar that has been faithfully captured in SNOMED CT, while Figure 4b shows an example where the information was not accurately represented. Infidelity such as this produced intermittent negative results often enough to be noticed when repeated queries of known serotype formulas were processed through the VTSL browser interface.

Figure 4a. Representation of *Salmonella* Saintemarie in SNOMED CT. The image is a faithful representation of K-W information in SNOMED CT. The full antigenic formula is represented correctly.

Salmonella Saintemarie



Parent(s):
(Select a parent to make it the "Current Concept".)
Salmonella I, group O:52 (organism)

Current Concept:
Salmonella Saintemarie (organism)

Child(ren):
(N=0) (Select a child to make it the "Current Concept".)

Current Concept:

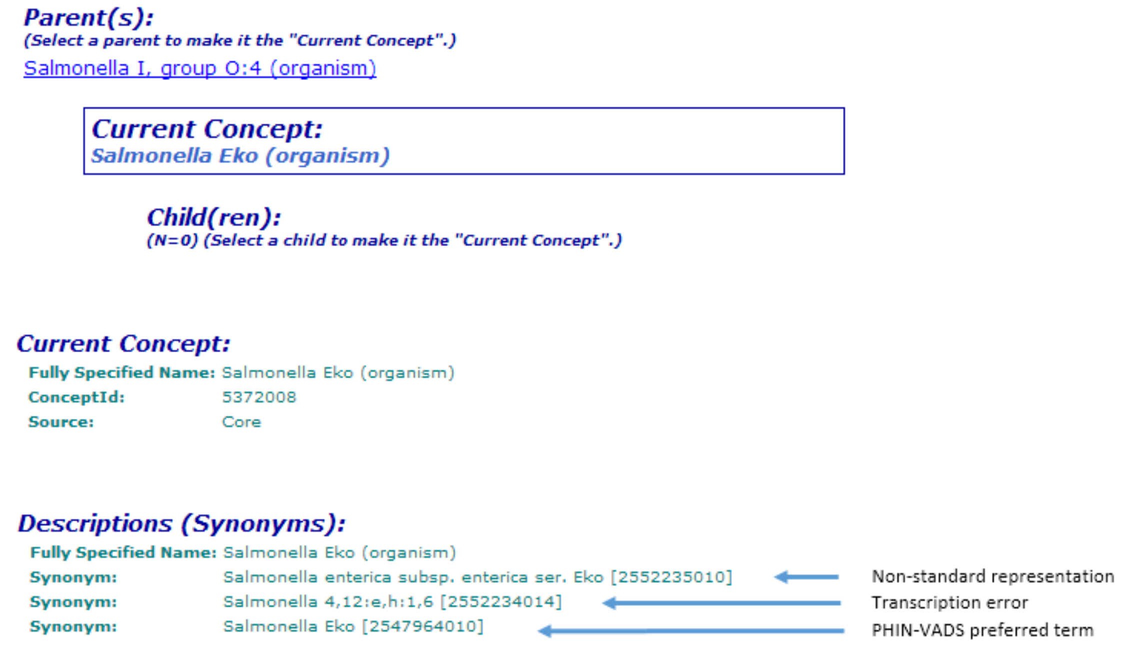
Fully Specified Name:	Salmonella Saintemarie (organism)
ConceptId:	416828006
Source:	Core

Descriptions (Synonyms):

Fully Specified Name:	Salmonella Saintemarie (organism)	
Synonym:	Salmonella enterica subsp. enterica ser. Saintemarie [2553958013]	Non-standard representation
Synonym:	Salmonella I 52:g,t:- [2553957015]	Correct formula
Synonym:	Salmonella Saintemarie [2549360019]	PHIN-VADS preferred term

Figure 4b. Representation of *Salmonella* Eko in SNOMED CT. The image is a faithful representation of SNOMED CT content. The antigenic formula is incorrect in that it does not include the subspecies designation (Roman numeral I).

Salmonella Eko



Parent(s):
(Select a parent to make it the "Current Concept".)
Salmonella I, group O:4 (organism)

Current Concept:
Salmonella Eko (organism)

Child(ren):
(N=0) (Select a child to make it the "Current Concept".)

Current Concept:

Fully Specified Name:	Salmonella Eko (organism)
ConceptId:	5372008
Source:	Core

Descriptions (Synonyms):

Fully Specified Name:	Salmonella Eko (organism)	
Synonym:	Salmonella enterica subsp. enterica ser. Eko [2552235010]	Non-standard representation
Synonym:	Salmonella 4,12:e,h:1,6 [2552234014]	Transcription error
Synonym:	Salmonella Eko [2547964010]	PHIN-VADS preferred term

After K-W authors complete and release an official version or update, mapping K-W serogroups and serotype names into the SNOMED CT organism hierarchy is the only link in the chain that requires human editorial intervention.

For public health reporting purposes, it is imperative that the fidelity between K-W, SNOMED CT, and PHIN-VADS be maintained. As SNOMED CT's fidelity with the expert source has such an important influence on a lab's ability to report the correct organism identification, we performed a comparative analysis of K-W and SNOMED CT content.

5.2.1 Methods

We established two primary goals for this assessment. We wanted to determine the number of K-W serovars that were not correctly represented in SNOMED CT and we wanted to know whether SNOMED CT included serovar content that was not in K-W. A bi-directional comparison of K-W serovars and the corresponding content in SNOMED CT was conducted using system query language (SQL) outer joins of relational database tables. This required creation of tables that accurately represented K-W and tables from SNOMED CT limited to the *Salmonella* serovars.

Antigenic formulae of the Salmonella serovars, 9th edn (2007) (Grimont & Weill, 2007) was obtained from the Institut Pasteur website. New serovars and editorial changes were incorporated from the 2010 supplement. (Guibourdenche, et al., 2010) Two spreadsheets, the first for *Salmonella* I, and the second containing *Salmonella* II-VI, were created by extracting selected tables from these sources using commercial pdf conversion software. (Weblite Solutions Corp., 2013) K-W *Salmonella* serovar names in the two spreadsheets were normalized to match corresponding FSN representations in the January 2012 release of SNOMED CT. This normalization was accomplished in several steps using Microsoft Excel (Microsoft Corp., 2013). First, the correct punctuation was added (i.e. commas between individual antigens, colons between antigen groups such as O and H1). Next the term “Salmonella” was added to the beginning and “(organism)” was appended to the end of each entry. Finally, the spreadsheets were manually checked and corrected for conversion errors (e.g. inclusion of page numbers from the K-W pdf, superscript and subscript notations within the K-W pdf, etc.). These spreadsheets were then imported into a MySQL relational database. (Oracle Corp., 2013) The Concepts table from the January 2012 release provided the corresponding SNOMED CT content.

As discussed earlier, due to the different naming conventions found within K-W, entries from *Salmonella* I were matched on the common name (e.g., *Salmonella* Dublin), while the *Salmonella* II-VI were matched based on the proper serovar designation (e.g., *Salmonella* II 9,46:z:z39). Entries that had an exact match were noted, as well as those that only had an entry in K-W, and those that were only found in SNOMED.

As previously stated, there could be errors in either of the two tables from where matching was attempted, therefore we performed a bi-directional analysis. The first analysis discovered K-W serovars not present in SNOMED CT (Category I errors). In this case the join produced a two column table. Column 1 contained the entry from the K-W table. Column 2 contained either an exact match from the SNOMED CT table or was blank (indicating absence from SNOMED CT). The second analysis looked for serovar concepts in SNOMED CT that had no matching term (Category II errors). This latter analysis required significant processing of SNOMED CT tables to limit content to only the serovars.

The *Salmonella* serovar-specific SNOMED CT tables were constructed by retrieving all concepts that contained the phrase "Salmonella" from concepts contained in the January 2012 release of SNOMED CT. A search for concepts that included common misspellings of "Salmonella" was also conducted and none were found. Additional processing removed concepts that were not of the organism hierarchy or that were not serovars. (e.g., Genus *Salmonella*, Salmonella serogroup O:4, *Salmonella enterica* subsp.

enterica), and those that did not fall within the organism hierarchy (e.g. Salmonella food poisoning (disorder)). Queries were developed to join the K-W tables with IHTSDO-released SNOMED CT Concept tables based on matching the FSN fields.

Join tables were manually examined to identify affected content and to determine the cause of discordance. Category I error rates were calculated as (errors/total K-W serotypes * 100). Category II error rates were calculated as (errors/total SNOMED subtypes * 100).

5.2.2 Results

The error rates for known K-W serovars not found in SNOMED CT (Category 1) were 9.3% for *Salmonella* I serotypes and a mean of 18.7% for *Salmonella* II-VI (Table 5). Error rates for the individual subspecies of *Salmonella* II-VI were similar and ranged from 17.3% to 23.1%. The majority of the errors for *Salmonella* I subspecies were simple omissions from SNOMED CT. Non-textual errors (colons, commas, etc.) account for a large percentage of discrepancies in the II-VI group.

Table 5. *Salmonella* serotype names not present in SNOMED CT® (Category I errors).

	Salmonella I		*Salmonella* II-VI	
Discrepancy	**n**	**% of errors**	**n**	**% of errors**
Omission (prior to 2010)	126	87.5	85	43.1
Omission (2010 supplement)	11	7.6	7	3.6
Editing error	6*	4.2	76	38.6
SNOMED error	1†	0.7	29‡	14.7
Total discrepancies	**144**		**197**	
Total K-W serovars	1542		1053	
Error rate	**9.3%**		**18.7%**	

* Transcription errors (i.e. punctuation substitutions, misspellings, antigens listed out of order, etc.) with or without altered meaning.

† *Salmonella* Butantan is listed in SNOMED as "*Salmonella enterica* subspecies *enterica* serovar Butantan". While technically correct, this is the only *Salmonella* serotype in SNOMED named in this manner and it does not conform to the K-W naming scheme.

‡ All errors in this subcategory appear to be an incorrect subspecies assignment for an existing serotype.

Simple omissions accounted for 126 (87.5%) of the Category I errors for *Salmonella* I serotypes and 85 (43.1%) of Category I errors for *Salmonella* II-VI. Additional omissions were associated with the 2010 K-W supplement and considered separately. Eleven (7.6%) *Salmonella* I omissions and 7 (3.6%) *Salmonella* II omissions were newly introduced serovars in the 2010 supplement. The remainder were transcription errors (i.e. punctuation substitutions, misspellings, antigens listed out of order, etc.). Most of these are accompanied by SNOMED CT descriptions (synonyms) with the proper subspecies assignment. For example the unmatched SNOMED CT FSN '*Salmonella* IV 40:z4,z24:- ' , is associated with the description '*Salmonella* III arizonae 40:z4,z24:- '. A serovar with this formula does not exist in the K-W scheme. Although one cannot be completely certain, the SNOMED CT concept appears to be a misrepresentation of the K-W serovar '*Salmonella* IIIa 40:z4,z24:- ' as *Salmonella* *enterica* subsp. *arizonae* is known as *Salmonella* IIIa. Transcription errors such as this were much less frequent for *Salmonella* I than for *Salmonella* II-VI.

The error rates for *Salmonella* serovar names found within SNOMED CT, but not recognized by K-W (Category II) were 4.2% for *Salmonella* I and 12.7% for *Salmonella* II-VI (Table 6). One group of errors was unique to *Salmonella* I subtypes. Specifically, 41 (66.1%) of the Category II (*Salmonella* I) errors resulted from the use "variant X" in the fully specified name to refer to the optionality of antigen X. We believe these represent a systematic error created by SNOMED CT editors misinterpreting the intent of specific entries in the section of K-W called "Alphabetic List of Serovar Names Withdrawn from the Scheme." An example is SNOMED CT's FSN '*Salmonella* London var 15+ (organism)'. Previously, this variation of *Salmonella* London was known as *Salmonella* Portsmouth. However, K-W has decided to withdraw the specific variants and have them grouped with the "parent" serovar denoted through non-textual coding (i.e. square brackets around optional antigens in the serovar formula). In the above referenced section, K-W used the phrase, '*Salmonella* London var. 15^{+}' to make specific reference to the affected serotype. However, their intent was that the serovar formerly named *Salmonella* Portsmouth should become *Salmonella* London. A second group of errors was also unique to *Salmonella* II subtypes. These 75 (60.0%) of the Category II (*Salmonella* II-VI) errors were produced by variations of inappropriate punctuation applied to *Salmonella* II fully specified names. Editing or transcription errors accounted for the existence of Category II errors for 21 (33.9%) of the *Salmonella* I and 50 (40.0%) of the *Salmonella* II-VI SNOMED CT subtypes. Details of transcription errors are shown in Table 7.

Table 6. Erroneous SNOMED CT *Salmonella* serotype names not present in Kauffmann-White (Category II errors).

	Salmonella I		*Salmonella* II-VI	
Discrepancy	**n**	**%**	**n**	**%**
Use of "variant" to address K-W formatting in fully specified name	41	66.1		
Punctuation error in formula-based fully specified name			75	60.0
Editing /transcription error†	21	33.9	50	40.0
Total discrepancies	**62**		**125**	
Total SNOMED CT subtypes	1493		983	
Error rate	**4.2%**		**12.7%**	

† Misspelling, lack of updating, incorrect or missing parts of antigenic formula. See Table 7 for examples of these errors.

Table 7. Examples of errors in subcategories used for Table 6.

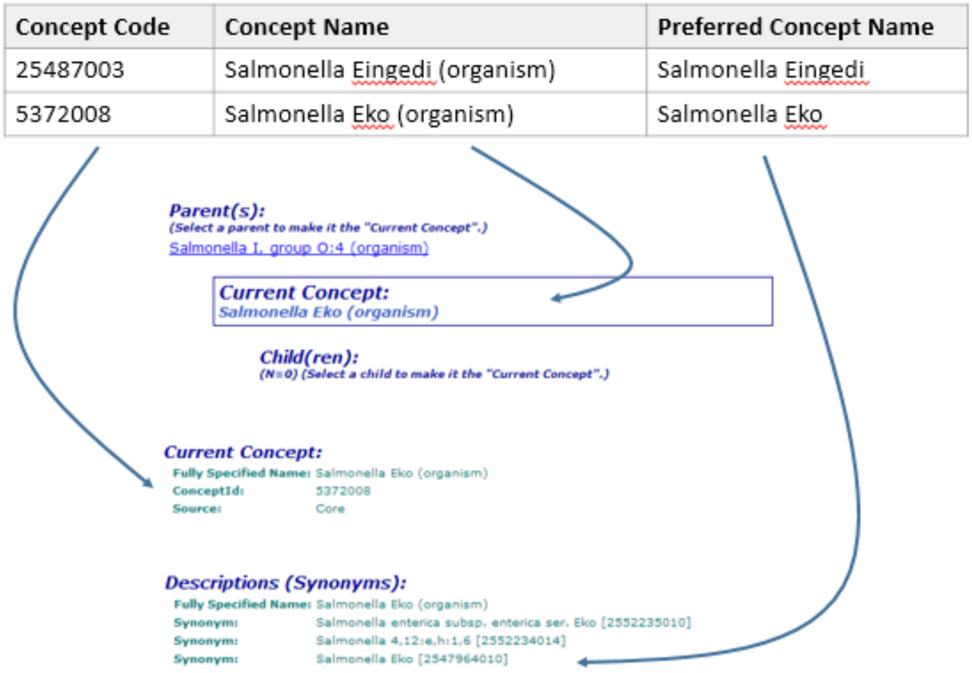
Error category	**SNOMED CT concept**	**Error**
Editing errors or transcription errors	Salmonella II 11:a:d:e,n,z15 (organism) 114314008	"d" is in wrong location (listed as Other H in K-W). Correct is Salmonella II 11:a:e,n,z15:d
	Salmonella 1,13,23:g,m,s,t:1,5 (organism) 302702005	No roman numeral, concept is listed as a child of Salmonella II
	Salmonella IIIa 42:r:- (organism) 404444000	K-W lists as IIIb
	Salmonella 3,10:R1,z40:1,7 (organism) 398537000	Roman numeral absent and formula wholly ambiguous
Punctuation error	Salmonella II 4,12,:z:z39 (organism) 114489008	Added comma after "12"

5.3 Assessment of PHIN-VADS

The spreadsheet provided by PHIN-VADS appears to be a direct extract of SNOMED CT content. In the manner PHIN-VADS faithfully represents the text strings and identifiers in SNOMED, PHIN-VADS also faithfully maintains the transcription errors in SNOMED CT. Other than preferred concept names associated with FSNs, PHIN-VADS provides no access to synonyms present in either K-W or SNOMED CT.

Figure 5 depicts the representation of an example serovar in both PHIN-VADS and SNOMED. The values for PHIN-VADS' Concept Code and SNOMED CT's concept identifier are in alignment, as well as the representation for Concept Name and FSN. Because PHIN-VADS does not represent synonyms, the link between binomial names and formulae is not present.

Figure 5. Representation of *Salmonella* Eko in PHIN-VADS and SNOMED CT.



Concept Code	Concept Name	Preferred Concept Name
25487003	Salmonella Eingedi (organism)	Salmonella Eingedi
5372008	Salmonella Eko (organism)	Salmonella Eko

Parent(s):
(Select a parent to make it the "Current Concept".)
Salmonella I, group O:4 (organism)

Current Concept:
Salmonella Eko (organism)

Child(ren):
(N=0) (Select a child to make it the "Current Concept".)

Current Concept:

Fully Specified Name:	Salmonella Eko (organism)
ConceptId:	5372008
Source:	Core

Descriptions (Synonyms):

Fully Specified Name:	Salmonella Eko (organism)
Synonym:	Salmonella enterica subsp. enterica ser. Eko [2552235010]
Synonym:	Salmonella 4,12:e,h:1,6 [2552234014]
Synonym:	Salmonella Eko [2547964010]

Under the present organizational scheme, it is not possible to locate SCTIDs for *Salmonella* I serotypes using PHIN-VADS unless the binomial name is known by other means. Given the logical arrangement of *Salmonella* binomial names and formulae, the antigenic formulae are held as synonyms and PHIN-VADS does not include them. SNOMED CT and serotype synonyms are necessary intermediaries, and utility depends in turn, on accurate transcription of serotype strings and logical fidelity with the K-W organization scheme. In short, the microbiology spreadsheet provided by PHIN-VADS cannot be used to “look up” SCTIDs for all K-W serovars.

5.4 Current state of the chain: SNOMED and PHIN-VADS representation of K-W

As discussed, K-W, SNOMED CT and PHIN-VADS form a terminology information chain to assist with the selection and reporting of recognized identifiers for *Salmonella* serotypes. One of the primary outcomes of the deliberations represented by K-W is the determination that a particular serotype is either distinct or belongs to a group of *Salmonella* listed under a single serovar. Table 8 lists the four primary arrangements that appear to exist.

Table 8. *Salmonella* formulae and names from K-W through SNOMED CT. Example 1 depicts an antigenic formula and a serovar name that are identical throughout the information chain. Example 2 represents a serovar name (and formula) with optional antigens present. Example 3 shows a serovar name that is a binomial. Lastly, example 4 represents a complex case, in which both optional antigens and a binomial need to be addressed.

	Antigenic Formula	**K-W Serovar (Formula)**	**K-W Serovar (Name)**	**SNOMED/PHIN (As listed)**
1.	II 4,12:g,z_{62}:-	II 4,12:g,z_{62}:-	II 4,12:g,z_{62}:-	II 4,12:g,z_{62}:-
2.a.	II 6,7:g,t:e,n,x:z_{42}	II 6,7:g,t:[e,n,x]:z_{42}	II 6,7:g,t:[e,n,x]:z_{42}	II 6,7:g,t:e,n,x:z_{42}*
b.	II 6,7:g,t:-:z_{42}			
3.	I 4,12:e,h:1,6	I 4,12:e,h:1,6	Eko	Eko
4.a.	I 1,4,5,12:i:1,2	I $\underline{1}$,4,[5],12:i:1,2	Typhimurium	Typhimurium
b.	I 1,4,12:i:1,2			Typhimurium var. Copenhagen
c.	I 4,5,12:i:1,2			Typhimurium
d.	I 4,12:i:1,2			Typhimurium var. Copenhagen

*This entry is erroneous in that it is missing square brackets around the 2nd H antigen. It should be listed as II 6,7:g,t:[e,n,x]:z_{42}

In example 1, the antigen formula and the serovar name are identical throughout the information chain. This is the simplest example, as the antigenic formula and the K-W name both share 1:1 identity with SNOMED CT and PHIN listings.

The second example represents a serovar that contains optional antigens. Familiarity with K-W's use of non-textual identifiers is needed to recognize that the antigenic formula depicted in 2.b. maps to the same serovar as 2.a. In this case, the secondary H antigen (e,n,x) may or may not be present, which will lead to a string mismatch between the antigenic formula in 2.b. (II 6,7:g,t:-:z_{42}) and the proper K-W serovar designation (II 6,7:g,t:[e,n,x]:z_{42}). The antigenic formula can be located using the VTSL SNOMED CT Browser provided the antigenic formula is surrounded by quotation marks. However, difficulties can arise, as the input string must match exactly what is listed in SNOMED CT's description field. Due to indexing capabilities within MySQL, this string search capability ignores brackets and parenthesis, where string searching in PHIN-VADS (Excel spreadsheet) does not. For example, the string in 2.a. would be found in SNOMED CT, as the browser would return hits that contain square brackets. As an aside, both SNOMED CT and PHIN-VADS contain an error in this example, in that no brackets are around the second H antigen and if it were corrected, using Excel's search function of PHIN-VADS for 'II 6,7:g,t:e,n,x:z_{42}' would not return a result, as the correct entry should be 'II 6,7:g,t:[e,n,x]:z_{42}'.

Examples 3 and 4 are concerned with *Salmonellae* I, those given a binomial serovar designation. Again, a simple example is shown in 3. It contains no optional antigens and maps directly to serovar Eko. In this case, K-W is essential to translate the antigenic formula to the serovar name, as the antigenic formulae for these serovars is not found in PHIN-VADS. By searching for the serovar name (Eko, in this case) in PHIN-VADS, one would be led directly to the SCTID. In theory, the structure of SNOMED CT would prove beneficial in comparison to using only K-W with PHIN-VADS. In SNOMED CT, concepts are associated with descriptions (or synonyms). In this example, the concept is listed as *Salmonella* Eko, with the antigenic formula '4,12:e,h:1,6' given as a description. Unfortunately, an error exists in this description, as the Roman numeral designator is absent. Given this inaccuracy, one must search the VTSL Browser for the exact antigenic formula included in the description '4,12:e,h:1,6' to find *Salmonella* Eko and its associated SCTID.

Example 4 revisits the challenge presented by a serovar such as *Salmonella* Typhimurium. The antigenic formula of the serovar '1,4,[5],12:i:1,2' represents O_1 and O_5 as optional. When optional antigens are present, no individual isolate is represented by an antigenic formula that matches that of its serovar. SNOMED CT could add the antigenic formulae of all possible isolates to the list of descriptions of a serovar. For *Salmonella* Typhimurium this would only be four additional descriptions. Unfortunately, the antigenic formula for serovar Senftenberg contains eight optional antigens. If all combinatorial possibilities were considered, the *Salmonella* Senftenberg would be comprised of 256 individual serotypes. In addition to the fact that all 256 variations may not actually exist, the need to enumerate all possibilities makes this solution undesirable. PHIN-VADS, the next link in the chain, lists only SNOMED CT fully specified names (concepts) and does not include synonyms (descriptions). Even if fully executed, this solution would not convey to the PHIN-VADS tables.

An additional issue not shown in Table 8 is related to the fact that previous versions of K-W sub-classified certain serotypes and gave them their own unique designations. Serotypes that did not contain the optional antigens were given the designation var. (variant). An example is shown in rows 4.b. and 4.d. where the serotype does not contain O_5 and is named Typhimurium var. Copenhagen. K-W has since abandoned this naming convention, however this error persists due to a lack of maintaining current updates on SNOMED CT's behalf. This legacy error, when corrected, will result in only one serovar name (Typhimurium) in SNOMED CT and PHIN-VADS instead of two.

5.5 Discussion

5.5.1 K-W

One of the aims of K-W is to group *Salmonella* isolates with similar antigenic formulae under single serovar names. For example, a K-W serovar is represented by a binomial name such as *Salmonella* Typhimurium which is also represented by the serovar formula is '1,4,[5],12:i:1,2'. The formula includes two antigens that may or may not be present (1 and [5]). Depending on the presence of the optional O_1 and O_5, the K-W serovar (and SNOMED CT FSN) "*Salmonella* Typhimurium" is the binomial name for multiple *Salmonella* isolates with different serotype formulae ('4,12:i:1,2', '4,5,12:i:1,2', '1,4,12:i:1,2' and '1,4,5,12:i:1,2'). In the end, there is a relationship between the antigenic formula of the isolates and the serovar (class) of which they are members that may not be 1:1. The result is a lexical disconnect between a serotyping result (4,12:i:1,2) and the formulaic name that represents the serovar (1,4,[5],12:i:1,2). Neither the K-W pdf nor SNOMED CT includes strings that match the actual isolate formulae.

Serotype formulae such as ‘1,4,[5],12:i:1,2’ present additional difficulties for the conversion of the K-W scheme to SNOMED CT content. In general, the *SNOMED CT Editorial Guide* (IHTSDO, 2012) indicates that the use of punctuation in terms should be minimized, and there are no examples to suggest that punctuation should convey meaning beyond conventional English uses. In this *Salmonella* Typimurium example, commas are used in a conventional way (list separation) while colons are used to convey a change in the type of antigen (O, H1, H2) being identified. Further, while text formatting (e.g underscore, italics, bolding, etc.) can be included in descriptions, it is not used to convey meaning in the SNOMED CT organism hierarchy. K-W uses non-text representation as an important method for conveying meaning. Square brackets (e.g., [5]) are placed around antigens that are expressed variably, without cause to meaning. Curly braces are used to indicate that particular antigens are present to the exclusion of others (e.g. {10}{15} indicates that if 10 is present, 15 is not and vice versa). In addition, the use of an underscore conveys that an antigen in particular was detected through conversion by a corresponding bacteriophage, and also considered optional.

5.5.2 SNOMED CT

This analysis considered bi-directional content fidelity between an official naming convention for *Salmonella* serotypes and a reference terminology that is used to facilitate application of the naming convention in public health information networks. Correct and consistent serotype naming in public health messages requires that the content of K-W (all serovars) be present in SNOMED CT. The analysis presented in Table 5 represents this perspective and indicates that approximately 13% of the serotypes in the K-W scheme are not represented in SNOMED CT$^{®}$ (Jan. 2012). These Category I errors are problematic for public health labs, as some number of the isolates they may need to report cannot and will not appear in the terminology extract that they are required to use, as PHIN-VADS is compiled directly from SNOMED CT. Moreover, SNOMED CT contains serotype names that are not represented in K-W (Category II errors). These are a quality control problem in SNOMED CT as these are not actually K-W serovars. While correction requires editorial effort, they probably do not interfere with public health case reporting in any meaningful way.

The error rate we detected is not surprising given factors that make for a difficult conversion from K-W to SNOMED CT. These factors include:

1) K-W specifies two distinct naming conventions for *Salmonella* serotypes. *Salmonella enterica* subsp. *enterica* (*Salmonella* I) serotypes are all given a binomial serovar name (e.g. *Salmonella* Enteritidis and a formulaic name '*Salmonella* I 1,9,12:[f],g,m,[p]:[1,7]'). The editing error rates for *Salmonella* I serotypes (n = 1542) and for the combined *Salmonella* II-VI serotypes (n = 1049) were 4.17% and 38.58%, respectively.
2) *Salmonella* I K-W serovar names may be more easily transcribed, as there is no requirement for attention to serotype syntax (e.g. colons). It may also be that it is easier to maintain accuracy for labels that are reasonably familiar words and not "strings of serotype codes."
3) K-W is not readily available in a useful database or spreadsheet form. Databases and spreadsheets are traditional digital artifacts that facilitate automation which should reduce errors that occur during manual transcription and editing.

4) K-W's primary electronic means for distributing its list of serotype names is in the form of pdf files. Although some automation of the conversion can be performed as was done in this analysis, considerable manual inspection was still required to transfer the names.

5) SNOMED CT can only serve as an adequate resource inasmuch as its content is current. Table 5 shows that the 11 *Salmonella* I serovars and the 7 *Salmonella* II-VI serovars from the 2010 supplement were not present in SNOMED CT. This reflects on the general difficulty associated with terminology maintenance especially when the terminology attempts to reflect the content of a recognized external authority on a separate update schedule. SNOMED CT provides biannual updates, however new versions of K-W are released at multi-year intervals, with periodic updates.

SNOMED CT maintains the K-W scheme insofar as having the FSN or preferred names be binomial for *Salmonella* I and antigenic formulae for *Salmonella* II-VI. Correcting the content gaps identified in this analysis is relatively straightforward. Essentially, it is a matter of applying adequate editorial resources to the problem. Our analysis could be used to reconcile SNOMED CT with the most current edition of K-W and errors in SNOMED CT would be corrected. Going forward, editorial resources can be dedicated to evaluating new versions of K-W to maintain the proper current status of SNOMED CT. However, the difficulties associated with using SNOMED CT to determine the proper names of *Salmonella* serotype does not end with the content problems this analysis identifies.

At present, SNOMED CT includes the binomial name *Salmonella* Typhimurium (organism) as the FSN. An incomplete but useful representation of the formulaic name ‘1,4,[5],12:i:1,2’ is included as a synonym. As mentioned previously, SNOMED CT does not include terms that match the individual isolate formula. To do so presents an interesting conundrum. SNOMED CT could include isolate formulae as synonyms of the serovar. Ignoring the fact that they are not truly synonyms, full representation would require that SNOMED CT include all synonyms. Recalling the example made by *Salmonella* Senftenberg, 256 isolate synonyms would require creation. Alternately, isolate concepts classes could be created and placed as subtypes of the serovars giving each isolate an SCTID. Unfortunately using the formula of an isolate to find a serovar would return the SCTID of the isolate rather than the SCTID of its serovar. This creates a different automation challenge in that means would have to be devised to determine the “correct” SCTID for submission to CDC.

5.5.3 PHIN-VADS

PHIN-VADS lists only SNOMED CT fully specified names and does not include synonyms. As PHIN-VADS is distributed as a spreadsheet, there is no ability to add synonymous descriptions without adding new entries. This causes inherent reporting difficulty for public health laboratories that, by protocol, identify an isolate by a serotype formula. This especially rings true for classifying the binomial names of *Salmonella* I isolates, in addition to any serovars that contain optional antigens.

5.5.4 K-W to SNOMED CT to PHIN-VADS

Neither K-W nor PHIN-VADS includes direct representations of the antigenic formula of all isolates, nor is either distributed in a format that is suitable for their inclusion. SNOMED CT can be forced to accommodate the isolate formulae but this creates the problems alluded to previously. The use of one naming scheme for *Salmonella* I and another for *Salmonella* II–VI and subsequent effects on information recording is a matter for the editors of K-W to consider. In summary, the chain of terminological content for *Salmonella* serovars does not provide laboratory personnel with a reliable and efficient means to identify SCTIDs for the *Salmonella* isolates they encounter. Introducing automation to improve efficiency will require application of alternative information technologies.

6. Ontologies

The K-W scheme is ostensibly a naming convention driven by intrinsic characteristics of the *Salmonella* serovars. Finding the serovar name of an isolate by using the pdf made available by WHOCC-Salm is a matter of matching the characteristics of the isolate to those of a known serovar. This aligns with the functional features of logical frameworks called ontologies. The technical infrastructure of SNOMED CT allows one to consider its organism hierarchy to be the primitive precursor of a functional ontology. At present, SNOMED CT organism classes are not defined in this way, but a logical model based on functional characteristics is being developed. (Wilcke, 2014) As the end goal of the laboratory task is to find an identifier from SNOMED CT, we propose to organize the serovars as an ontology based on serotyping results. The ontology we develop will inform the SNOMED CT model and may facilitate automation of serovar name identification, in addition to identification of the representative SCTID.

6.1 Background

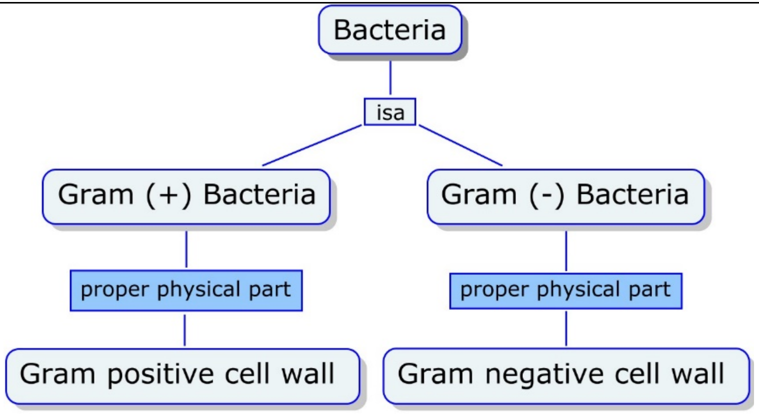
Ontologies are structural frameworks for organizing information and were first defined for the computer science environment by Tom Gruber in 1993 as “specifications of a conceptualization”. (Gruber, 1993) In essence, an ontology is a method of representing items of knowledge (ideas, facts, things) in a way that defines the relationships and classifications of concepts within a specified domain of knowledge.

Ontologies are useful in that they define classes (or concepts) in a hierarchy of types and subtypes. Once created, ontologies can be used to infer that a particular instance is a member of the defined concept class or classes to which it belongs. Ontology creators assign properties to those relationships that define the way the inferences are made. Ontologies are especially helpful for research in areas with vast amounts of available data, and where the relationships to be investigated do not fit neatly into a simple hierarchy, such as biomedical research.

For all their eventual complexity, formal ontologies are built using very few logical structures and functions. The Pizza Ontology developed by the University of Manchester (Horridge, 2011) was created to teach fundamental principles of ontologies. The basic structures of the Pizza Ontology include concepts such as "Thin Crust Pepperoni Pizza," attributes such as "*has_topping*" and "*has_crust*" and values such as "pepperoni" and "thin crust" to describe specific Pizza (concepts) and allow for automated classification. In this way, a Thin Crust Pepperoni Pizza that "*has_crust*" = "thin crust" can be automatically distinguished from a Hand Tossed Pepperoni Pizza that "*has_crust*" = "hand tossed." The number and nature of the attributes and values depends entirely on the domain that the ontology attempts to describe.

Figure 6 depicts a simple example regarding a bacterial relationship with a super-type (Bacteria) and subtypes (Gram (+) Bacteria and Gram (-) Bacteria). These relationships are based on attribute-value pairs. In this case, the attribute is "proper physical part", and the values are kinds of cell walls. Arranged formally in an ontology, a computer can determine that gram negative bacteria and gram positive bacteria are different. This depiction is representative of how ontologies ultimately enable automated information sharing.

Figure 6. Simple computable classification system in bacteria that shows attribute-value pairs.



Ontologies are ways to organize and share knowledge across domains. In the past, biomedical domains have developed discipline-specific and even project-specific terminologies. This in turn leads to the phenomenon referred to as "information silos", wherein two domains share an interest in a subject but cannot communicate as they do not share a language about the subject. These silos of disparate data arise as very low-level domain terms often lack precise and unambiguous definitions. (Mirhaji, 2009) This lack of interoperability leads to inconsistencies, fragmentation and overlap both within and in-between terminologies of different domains.

6.2 Related Biomedical Ontologies

Among the first applications of ontology in biomedicine, the Gene Ontology (GO) began in 1998 as a collaborative effort among three model organism databases—FlyBase, the Saccharomyces Genome Database (SGD), and the Mouse Genome Database (MGD)—to merge the information contained in separate databases for fruit fly, yeast, and mouse genomes, respectively. GO was developed to represent and process information about gene products and functions and has since grown to include more than 17 individual databases of genetic information. It has been described by the authors as a "controlled vocabulary" directed toward providing a practically useful framework for maintaining the biological annotations that are applied to gene products in a variety of contexts. (Gene Ontology Consortium, 2001)

The Gene Ontology is, in essence, three separate ontologies that describe gene products in terms of associated 1) biological processes, 2) cellular components and 3) molecular functions in a species-independent manner. In addition to maintaining the ontologies, the project's researchers provide gene annotation and develop tools to facilitate access and search.

GO makes extensive use of some of the fundamental characteristics of ontologies. Because single genetic entities may be known by multiple names, the ontologies use synonyms to broaden or narrow the scope of inquiry as necessary. For example, child terms (subclasses) might be related to multiple parent terms (superclasses). The ontologies use five basic relationships: *is_a*, *part_of*, *regulates*, *positively_reglates*, and *negatively_regulates*. (Gene Ontology Consortium, 2014)

The developers of GO have received criticism that they have not focused on software implementations or on logical expression. Their efforts have been directed, rather, toward the ease of populating and maintaining the vocabulary in lieu of the ability to provide automated reasoning. (Smith, Kumar, & Bittner, 2005) In response to this criticism, the Open Biological and Biomedical Ontologies (OBO) Foundry was created. The OBO Foundry has established a set of principles for ontology development with the goal of creating a collection of interoperable reference ontologies in the biomedical domain. One of the key requirements to achieve this goal is to ensure that ontology developers reuse concepts and definitions that others have already created rather than create their own definitions, thereby making the ontologies orthogonal by lacking redundancy. If these classes are used in subsequent ontologies, the class in the new ontology refers back to the original instead of creating a duplicate class. (Open Biological and Biomedical Ontologies Foundry, 2011)

In order to connect biomedical ontologies, they must be developed in such a way that links them to upper level ontologies, such as Basic Formal Ontology (BFO) and OBO's Relation Ontology (RO). BFO is a true upper ontology designed to promote data interoperability in scientific as well as other domains. It does not contain specialized terms that would properly fall within the coverage domains of the individual sciences. (Smith, Kumar, & Bittner, 2005) In addition, RO is a compilation of relationships intended for standardization across biomedical ontologies. It incorporates core upper-level relations such as *part_of* as well as biology-specific relationship types such as *develops_from*. (Smith, et al., 2005)

SNOMED CT has been identified as a biomedical ontology, and descriptions of its potential alignment with the OBO Foundry and BFO are the subject of other investigations. (National Center for Biomedical Ontology, 2015) (Hogan, 2008) For purposes of public health reporting of *Salmonella* serotypes, SNOMED CT provides the required digital identifiers for reporting *Salmonella* isolates to CDC. The *Salmonella* serovar names have been placed in a primitive hierarchy, but no formal logical model has been described and no definitions are provided. We will attempt to show that adding logical definitions can enhance existing SNOMED CT content, and that these logical definitions can support *Salmonella* isolate identification and naming.

Previously, the Animals in Context Ontology (ACO), (Santamaria, Fallon, Green, Schulz, & Wilcke, 2012) was created from content within SNOMED CT using the principles of the OBO Foundry. ACO includes development stages and physiologic animal classes in addition to animal classes where humans have assigned the animal's role in agricultural production or use. ACO in includes classes from and integrates with other OBO Foundry ontologies, such as the Gene Ontology (GO). In summary, ACO classes are of interest to science, medicine and agriculture, and can connect previously poorly integrated information between animal and human systems.

6.3 *Salmonella* Serotype Designation Ontology

ACO demonstrated that SNOMED CT content could be created in a way that attended to the principles of the OBO Foundry. ACO has been made available in two forms. It can be downloaded as an OBO compliant ontology that includes SCTIDs as alternate (external) identifiers. In addition, the same content is also incorporated in the Veterinary Extension to SNOMED CT. At the time of its creation, ACO was entirely novel content, and new SCTIDs were created when it was incorporated in the veterinary extension. The challenge for organizing the *Salmonella* serovar names as an ontology is slightly different. As demonstrated by our previous evaluations, most of the *Salmonella* serovar concept classes are represented correctly in SNOMED CT. The ontology development goal for the present work is to create a functional ontology that minimally disrupts existing SNOMED CT content, if at all. We plan to develop a domain level, OBO compliant ontology to assist with the reporting of *Salmonella* isolates to the CDC from PHLs. Specifically, it will enable the PHLs to determine the correct serotype designation (name and/or formula) and associated SCTID with minimal effort.

7. Ontology Development Methods

In creating its list of serovars the editors of K-W not only create names for isolates with unique antigenic formulae, they also group serotypes that share significant portions of their antigenic formulae under a single name. In essence, K-W creates a kind of synonymy between serotypes and serovar names. For example, *Salmonella* Canada is given the antigenic formula designation '*Salmonella* I 4,12,[27]:b:1,6'. Because the O_27 antigen is designated as [optional] the two distinct serotypes '*Salmonella* I 4,12:b:1,6' and '*Salmonella* I 4,12,27:b:1,6' will both be called *Salmonella* Canada. While human beings reading K-W can easily understand this near synonymy, the two serotypes are, in essence, logical subtypes of the serovar name.

K-W serovars can be viewed as classes in an ontology and the isolates as subclasses. As stated previously, ontologies are created using very few logical devices. As a result, there are a limited number of ways that an ontology of K-W serovars and isolate classes can be constructed. The most elegant solution would be to create logical definitions of the K-W serovars and place them as the terminal classes in the hierarchy. The logical definitions would then be used to classify laboratory instances. The difficulty with this approach is created by K-W's use of options and their grouping of several serotypes into a single serovar class. Logical definitions for these classes are challenging to create and can only be assessed by software that can evaluate them based on the principles of description logic. This problem can also be addressed using an approach that is computationally simpler, namely establishing the individual serotypes that make up serovar groups as classes in their own right. As the antigens of an individual serotype do not have options (which is also true for most of the K-W serovar classes) the logical definitions are much simpler and can be “computed” using much simpler tools.

It would seem reasonable that the antigenic formulas of isolate classes could be incorporated as synonyms of each serovar class rather than as subtypes. The system would not rely on logical classification, but would employ text matching of antigenic formulae associated with the correct K-W serovar class (and associated SCTID). Unfortunately this approach requires enumeration of all possible options. As stated previously, *Salmonella* Senftenberg has 8 of its antigens listed as optional. The resulting possible number of combinations is 2^8 or 256 synonyms. Rather than reducing the maintenance burden and possible problems associated with text matching, this approach would seem to increase them. Further, this solution would not be technically correct as the isolate classes are not actually synonyms of the serovar class as they are subtype members of the class.

7.1 Prototypes

Based on our understanding of the underlying logic of K-W, we developed two distinct *Salmonella* Serotype Designation Ontology (SSDO) prototypes. Prototype I aligned with the existing content in SNOMED CT in that the most granular hierarchy nodes align fully with K-W serovar names. Isolate classes were not created. Classification depended on stated subtypes for the *Salmonella* subspecies and the logical definitions of serovars. Definitions were not created to classify the *Salmonella* species or subspecies and these classes remained primitive. Classification of the species would be based on biochemical characteristics of these organism classes that occurs “before” the K-W scheme classifies the serovars. This occurs above serovar classification in the hierarchy and is beyond the scope of this particular project. The definitions of K-W serovars that represent groups included appropriate representations of the optionality of various antigens. Prototype II was similar to Prototype I, except that isolate serotypes were created as subclasses of the K-W serovar classes in the ontology. Definition of K-W serovar classes is simplified considerably as relationships for optional antigens were moved to define the isolate serotype concepts. There was no requirement to create complicated logical definitions that incorporated the optional antigens in the K-W serovar classes. Figure 7 is a side-by-side representation of *Salmonella* Typhimurium rendered in Prototype I and Prototype II.

Figure 7. Comparison of the structure of Prototype I vs. Prototype II. Serovar classes were the most distal in Prototype I. Isolate classes were the most distal in Prototype II.

Prototype I	**Prototype II**
Genus *Salmonella* *Salmonella* enterica *Salmonella* enterica subsp. enterica *Salmonella* Typhimurium (1,4,[5],12:i:1,2)	Genus *Salmonella* *Salmonella* enterica *Salmonella* enterica subsp. enterica *Salmonella* Typhimurium (serovar) '4,12:i:1,2' (isolate) '1,4,12:i:1,2' (isolate) '4,5,12:i:1,2' (isolate) '1,4,5,12:i:1,2' (isolate)

In addition to creating logical definitions and classes based on K-W, we also considered the eventual need to adapt our ontology to the information chain, specifically how our prototypes would support SNOMED CT and PHIN-VADS. Table 9 describes the three optional approaches and the fit of various possible prototypes with functional capabilities of parts of the information chain derived from K-W.

Table 9. Three approaches to describe the K-W information chain in an ontology.

Approach	**Facilitates lab lookup of serovars via SNOMED**	**Can be incorporated into existing PHIN-VADS table**
K-W serovars are classes. Antigenic formulae of isolates are synonyms. K-W serovars with many optional antigens result in the creation of many serotype synonyms.	**Yes** Could reliably find K-W serovar based on isolate antigenic formula.	**No** PHIN-VADS does not include synonyms from SNOMED and relationships to K-W serovars cannot be represented via spreadsheet.
K-W serovars are classes, subtypes not included. Classification based on logical definition of the K-W serovar. (Prototype I)	**Yes**	**Yes**
K-W serovars are classes, Isolates created as subtype classes. K-W serovars with many optional antigens result in the creation of dozens of serotype classes. (Prototype II)	**No** K-W serovars may be terminal or a supertype of a terminal node. One would have to control traversing the hierarchy. It would be difficult to select K-W serovar based on matching laboratory serotypes.	**Yes** Isolate serotypes would be associated with SCTIDs. Matching these serotypes to an antigenic formula would not return the K-W serovar SCTID.

The *Salmonella* serotype designation ontology was managed using Protégé. (Stanford Center for Biomedical Informatics Research, 2014) Protégé is open source software designed specifically to develop domain ontologies. Created at Stanford University in collaboration with the University of Manchester, it has more than 200,000 registered users and has been called the “leading ontological engineering tool.” (Gašević, Djurić, & Devedžić, 2009) Protégé simplifies the aspect of creating ontology classes, adding object and data properties, and running queries through its intuitive interface.

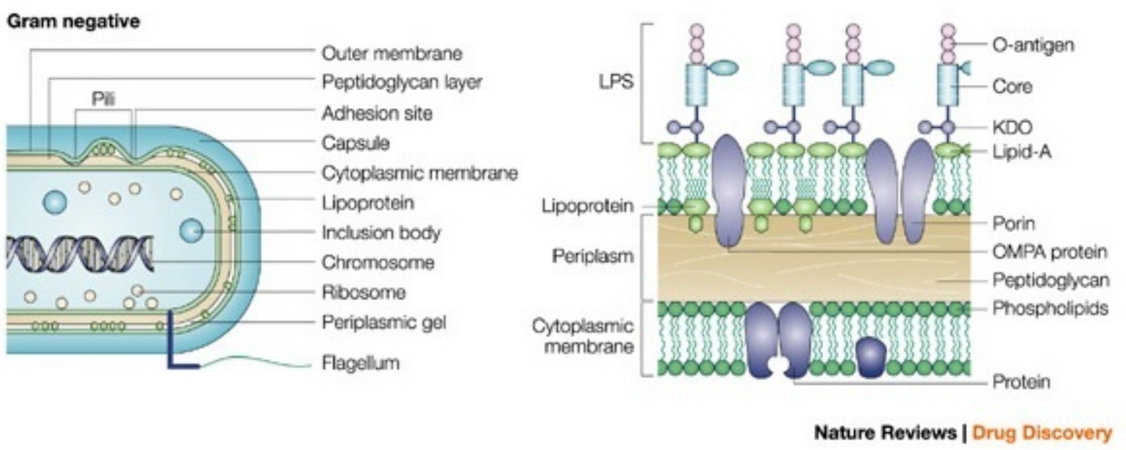
For both ontologies a limited set of defining attributes were created and antigen classes were added to serve as values for defining relationships. Attributes and attribute values were identical for the two prototypes.

7.2 Defining Attributes

The selection and naming of attributes for any ontology is somewhat arbitrary. Consistent use of the attributes, and assignment of values to them, actually produce the desired functionality. An ontology can be entirely self-contained with its attributes created for "internal use only." However, we intend that our *Salmonella* serotype designation ontology function within the larger world of biomedical ontologies and believe that it can be used to improve functional characteristics of SNOMED CT. We worked carefully to choose attributes based on existing OBO ontologies. (Open Biological and Biomedical Ontologies Foundry, 2011) OBO provides a consistent framework for incorporating domain knowledge such as this, in the form of domain ontologies, into the body of biomedical knowledge. Links to OBO ontologies provide an ability to cross-walk between domains. For example, a *Salmonella* genomics researcher may appreciate the ability to simultaneously leverage GO and this SSDO because both are based on principles of OBO. BioTop Lite is the ontology we selected as the bridge between the upper level OBO ontologies and our serotype designation ontology. BioTop Lite includes definitions for the foundational concepts of biomedicine as represented in OBO. In particular, it provides text-based definitions for attributes that adhere to OBO's formal design policies. Moreover, it includes object properties that are mapped to the relations in RO. (Beißwanger, Schulz, Stenzhorn, & Hahn, 2008)

In addition to aligning with OBO, we intend that the attributes align with the actual physical nature of the organism. Figure 8 is a graphic representation of a Gram negative bacteria. The antigens used to determine the identity of *Salmonella* isolates are either incorporated in the cell wall (O antigens) or into flagella (H antigens). In *Salmonellae*, the O antigens are present and embedded in the cell wall and do not change as identity testing progresses. H antigens, on the other hand, may be present in one of several phases that represent the presentation of a particular class of flagella (H1, H2, H other) during various stages of testing. In either case, the antigens are constituent molecular parts of larger macromolecular structures.

Figure 8. Graphic representation of a Gram negative bacterial cell.



[Adapted by permission from Macmillan Publishers Ltd: *Nature Reviews Drug Discovery* Lolis & Bucala, Copyright 2003.]

The BioTop attribute *hasComponentPart* was selected to link both the O and H antigens to the logical definitions of the serovars. BioTop Lite specifies that *hasComponentPart*:

> "relates components with a compound. Components strictly partition the compound, and the compound is the mereological sum of its components. A loss of some component affects the integrity of the compound, and possibly the type it

> instantiates, e.g. a complete vs. a defective organism. Components should be (at least) partly *bona fide* parts. The use of this relation also requires the commitment to an underlying granularity level." (Beißwanger, Schulz, Stenzhorn, & Hahn, 2008)

The attribute *hasComponentPart* applies to both O and H antigens as the antigens are molecular components of the outer membrane or a flagellum. To create defining relationships, specific antigens were linked to each *Salmonella* serotype class using *hasComponentPart* to form the logical representation of various parts of the antigenic formula of that class.

Antigens are incorporated in flagella produced by the organism during one or more of the three flagellar phases (H1, H2, H other) of the serotype test procedure. At each stage of testing, the bacteria is evaluated for the presence of flagella, each with its own compliment of antigens. Each flagellar phase, except the first, is produced by blocking the flagellar antigen(s) determined in the previous phase. If a specific *Salmonella* isolate has the ability to produce more than one type of H antigen, blocking induces the bacteria to produce those flagella with different antigens than those blocked. There is only a single set of *Salmonella* H antigens, and these may be expressed in any of the three phases. In order to differentiate between the flagellar phases, we chose *hasProperPhysicalPart* from BioTop Lite. (Beißwanger, Schulz, Stenzhorn, & Hahn, 2008) Defining relationships were created in the ontology by linking *hasProperPhysicalPart* to values for each of the three flagella.

Although it would not be inappropriate to name the cell membrane as a proper physical part and associate it with O antigens, doing so would not contribute to classification in this ontology. The *hasComponentPart* attribute stands alone for defining O antigens because they are all incorporated in the same structure that does not vary during testing. In other words, O antigens are always integral to a single cell wall that has only one relationship to the serotype definition.

7.3 Value Sets

O and H antigen classes are included in the ontology to serve as values that complete defining relationships using *hasComponentPart*. The O antigens are numeric, while the H antigens are alphanumeric. Although the two categories share some labels (e.g., there is an O antigen 5 and an H antigen 5), the designations are applied to different antigens (i.e., O_5 is biochemically distinct from H_5).

One of our stated goals was alignment with OBO principles. This implies that our ontology will not recreate values that are present in another. Thus, appropriate domain ontologies were searched for pertinent value concepts. A search of the domain ontologies in National Centers for Biomedical Ontology's (NCBO) BioPortal for the phrase "Salmonella Antigen" revealed related but dissimilar content in SNOMED CT, the National Cancer Institute Thesaurus (NCIT), and National Library of Medicine's Medical Subject Headings (MeSH). Unfortunately, none of these are "true" ontologies built on OBO principles and none included a comprehensive list of *Salmonella* antigens. Another search revealed listings for "Salmonella" in the Host Pathogen Interaction Ontology (HPIO) with subclasses consistent with Linnaean taxonomy of *Salmonella* (e.g. *Salmonella enterica* subsp. *arizonae*). The Infectious Disease Ontology (IDO), which is a set of interoperable ontologies under development, attempts to provide coverage for the infectious disease domain. Currently, IDO has coverage for Brucellosis, Influenza and *Staphylococcus aureus*, among others. However, we were unable to locate existing *Salmonella* content within IDO. Based on these negative results, we developed the necessary and specific lists of somatic and flagellar antigens.

Although the individual antigens were not represented in any of the ontologies that we investigated, SNOMED CT does include classes for “Salmonella O (somatic) antigen” and “Salmonella H (flagellar) antigen.” Reference to these SNOMED CT classes were made by referring to their SCTIDs in the ontology and they served as superclasses for the individual O and H antigens respectively.

7.4 Description logic

Formal reasoning functionality of OBO ontologies is based on description logics (DL), a family of formal knowledge representation languages. (Krötzsch, Simančík, & Horrocks, 2012) Analogous to the order of operations that influence solutions to algebraic equations, DL provide operators (e.g., or, not, and) and rules for evaluating logical statements in an ontology. Different ontologies have been built using several different levels of description logic that convey more or less flexibility and power. Protégé employs DL to compute internal logical consistency using software tools called reasoners. Protégé is compatible with a number of reasoners that differ in their ability to process DL of varying levels of complexity and capability. We employed the FaCT++ (Fast Classification of Terminologies) reasoner for use with this project. (University of Manchester, 2012) The FaCT system includes two reasoners, one pertaining to transitive roles, functional roles and role hierarchy, and the other for inverse roles and qualified number restrictions. (Tsarkov & Horrocks, 2006) During development of the prototype ontologies, restrictions were not imposed in terms of description logic functions and abilities. In other words, we created the ontology using a DL sufficient to our needs without concern as to whether SNOMED CT used a DL of equal functionality.

7.5 Prototype class selection

The prototype ontologies were created from specific known serovar classes within K-W. These classes were strategically selected so that all variations in optionality were considered, expanding on those shown previously in Table 8. Examples of each type of singular "optionality" as well as combinations were represented. There were 8 predetermined serovars (one with no optional antigens, one for each of the three singular options, three pair wise combinations, and one with all three optional antigens) for each *Salmonella* I and *Salmonella* II-VI for a total of 16.

7.6 Prototype selection

Once we were satisfied that the logical variations of the *Salmonella* serovars were included in both prototypes, one was selected for the final ontology of all the K-W serovars. The first obvious criterion was that the prototype could correctly classify the prototype test set of 16 isolates. Both prototype designs met this criterion. Selection was ultimately based on factors that reflected the application of the fully populated ontology within the K-W to SNOMED CT to PHIN-VADS information chain. These factors included 1) adherence to the principles of the OBO Foundry, 2) re-use of existing RO relationships, 3) whether the ontology could be recreated in SNOMED CT given its functionalities and limitations, 4) the ability to represent serotypes of other organisms in SNOMED (e.g. *E. coli*), 5) avoidance of disrupting or adding to the PHIN-VADS microorganism table and 6) complexity and size of tasks related to creation, editing and long term maintenance of the ontology.

7.6.1 Alignment with SNOMED CT

Alignment with SNOMED CT was considered in the context of the emerging organism model under development and not the current state of SNOMED CT. First, we considered whether or not the description logic subset SNOMED CT will be using can correctly assess the logic of our final ontology. Next was whether an undue maintenance burden would be imposed on SNOMED CT in incorporating our ontology. The maintenance burden includes the creation of new concept classes, new descriptions and alignment with the current hierarchy, all of which should be kept to a minimum. Of lesser but still important concern was whether or not changes in SNOMED CT would affect the next link in the chain, namely PHIN-VADS.

7.6.2 Alignment with PHIN-VADS

CDC states that PHIN-VADS exists for accessing, searching, and distributing standards-based vocabularies used within PHIN to local, state and national PHIN partners. (CDC, 2011d) In addition, PHIN-VADS tables can be used to validate the identifiers submitted by laboratories reporting in CDC systems are in fact legitimate SCTIDs. The specific question is whether using one prototype versus the other would alter selection and incorporation of concepts that make up the PHIN-VADS tables. The preferred outcome of this project relative to PHIN-VADS is that, when implemented in SNOMED CT, the result would not add superfluous or misleading content to PHIN-VADS tables nor would it disrupt CDCs extraction (processes) of SNOMED CT content.

7.6.3 Selection of Prototype I

We developed two structurally distinct prototypes and evaluated each as a basis for an ontology representing all K-W names. Prototype I was pursued for several reasons. The terminal nodes of Prototype I are more easily selected and align more fully with SNOMED CT and CDC reporting requirements. Although Prototype II could be executed in SNOMED CT without altering SNOMED CT's logical configuration, Prototype I produced better alignment with CDC's terminology systems and does not require creation of SNOMED CT concept classes that are not reportable. A stand-alone tool (app) based on Prototype I could facilitate identification of the proper SCTID a lab should report for a given isolate.

As designed, Prototype I exposes limits that SNOMED CT has placed on its internal logic or DL. The current DL used for SNOMED CT does not permit the use of the logical disjunction "or" and we are unaware of plans to include it. An example is the need to have antigens that are mutually exclusive, as is the case with serogroup O:3,10, where factors O_15 or O_15 and O_34 are substituted for O_10 when they are present. As mentioned previously, Prototype I required more robust description logic than did Prototype II.

Prototype I aligns well with the PHIN-VADS microbiology table because the terminal serovar classes represent the concepts as required for laboratory reporting. As stated above, elaboration of Prototype II in SNOMED CT would create classes distal to those required for reporting purposes. As the PHIN-VADS microbiology table appears to be a list of SNOMED CT concept classes represented by SCTIDs, SNOMED CT FSNs and preferred terms (PT), these new classes would be incorporated in these tables. For users of the tables the concept classes would appear to overlap as a given serovar such as *Salmonella* Typhimurium would be represented by five rows in the table only one of which would be correct for reporting.

7.7 Populating SSDO

Prototypes I and II were created manually in Protégé through the keyboard interface. As the numbers of concepts and definitions were relatively small, this task was not particularly burdensome. On the other hand, the full ontology includes 2595 K-W serovars and their definitions. For this reason we determined that some automation of the ontology creation process was desirable. The automation approach was developed using some of the same database tables used to assess the *Salmonella* content in SNOMED CT. We also used a Perl script (IHTDSO, 2015) designed to generate OWL files from SNOMED CT tables that can be imported by Protégé. A complete description of the ontology creation process can be found in Appendix C. Briefly, the prior SNOMED CT assessment had identified the specific rows in SNOMED CT tables that represented the concept classes for the *Salmonella* serotypes. This information was used to generate three text files, one each for concepts, descriptions and relationships. These files were converted by the Perl script to a Web Ontology Language (OWL) file that represented the taxonomy of Genus *Salmonella*, the K-W serovars, and the antigens and relationships required to create logical definitions for the K-W serovars.

7.7.1 Manual entry of definitions

Once the ontology had been created as an OWL file, it was imported into Protégé. Definitions for the K-W serovars were derived from rows created in SNOMED CT relationship tables as described above. The three major antigen formula styles, shown in

Table 10 required three different types of logical definitions. Antigenic formulae with no optional antigens required no additional processing. Simple optional antigens (e.g., [5]), and phage induced (e.g., 1) were removed as part of the table creation process, as our assessment of Prototype I indicated that they did not influence classification. As such, no further editing of these types was required. Definitions for the K-W serovars containing exclusive optional antigens were refined manually in Protégé.

Table 10. *Salmonella* antigen formula styles and their logical representations in Protégé.

Antigen Optionality	**K-W serovar name**	**Antigenic formula**	**Logical definition as rendered in Protégé**
none	*Salmonella* Aachen	*Salmonella* I 17:z_{35}:1,6	'Salmonella enterica subsp. enterica' and (hCP some O_17) and (hPPP some (SH1 and hCP some H_z_{35})) and (hPPP some (SH2 and hCP some H_1)) and (hPPP some (SH2 and hCP some H_6)) Each antigen is represented by an axiom.
[simple]	*Salmonella* Typhimurium	*Salmonella* I 1,4,[5],12:i:1,2	'Salmonella enterica subsp. enterica' and (hCP some O_4) and (hCP some O_12) and (hPPP some (SH1 and hCP some H_i)) and (hPPP some (SH2 and hCP some H_1)) and (hPPP some (SH2 and hCP some H_2)) Axioms for O_1 and O_5 are not included.
{exclusive}	*Salmonella* Stockholm	*Salmonella* I 3,{10}{15}:y:z_6	'Salmonella enterica subsp. enterica' and (hCP some O_3) and ((hCP some O_10) **or** (hCP some O_15)) and (hPPP some (SH1 and hCP some H_y)) and (hPPP some (SH2 and hCP some H_z_6)) Universal restriction '**or**' creates exclusive optionality.

hCP = hasComponentPart
hPPP = hasProperPhysicalPart
			SH1 = Salmonella_phase_H1
			SH2 = Salmonella_phase H2

7.8 Initial refinement

The full ontology was tested and refined iteratively. Each version of the ontology was inspected for logical consistency and for correct classification of both K-W serovar classes. Logical consistency among the serovars was judged on the basis of both the desired and correct subsumption of serovar classes.

7.8.1 Class membership

Final evaluation of the ontology was conducted by creating individuals and noting placement of individuals in serovar classes. Two distinct approaches were taken to the creation of individuals. Initial testing was performed using the serovar classes shown in Table 11a. These classes were strategically selected in order to demonstrate correct classification based on each major logical construction used in SSDO. Individuals were created directly in Protégé (Individuals tab) and evaluated as to whether they were assigned membership in the expected serovar classes. Examples were created for each logical variation among the serovars and situations where unexpected subtyping and equivalence occurred. Table 11a lists the serovars tested, their respective formulae and the purpose of each test.

Table 11a. Serovars (SSDO classes) against which individual serotypes were strategically tested.

Serovar	**Formula**	**Classification tested**
Salmonella Eko	*Salmonella* I 4,12:e,h:1,6	*Salmonella* I with no options Serovar includes antigens for O, H_1, H_2. No antigens are optional.
Salmonella II 9,46:z_{10}:z_6	*Salmonella* II 9,46:z_{10}:z_6	*Salmonella* II with no options (antigens same as a *Salmonella* I class) Logical definition includes its subtype relationship to *Salmonella* II but is otherwise identical to *Salmonella* Louisiana.
Salmonella Texas	*Salmonella* I 4,[5],12:k:e,n,z_{15}	Simple optionality test Logical definition does not include an axiom for O_5 antigen.
Salmonella Treforest	*Salmonella* I 1,51:z:1,6	Simple optionality test Logical definition does not include an axiom for O_1 antigen.
Salmonella Paratyphi C	*Salmonella* I 6,7,[Vi]:c:1,5	Simple optionality test Logical definition does not include an axiom for cell surface (envelope) antigen Vi.
Salmonella Abortusequi	*Salmonella* I 4,12:-:e,n,x	Assertion of absence Logical definition includes "and not (H_1 antigen phase…)".
Salmonella Anatum	*Salmonella* I 3,{10}{15}{15,34}:e,h:1,6:[z_{64}]	Exclusive optionality Logical definition includes the union construction 'or' to specify that the O antigens must be one of a series but not more than one of a series.

Salmonella Uno	*Salmonella* I 6,8:z_{29}:[e,n,z_{15}]	Assertion of absence can be included any time H antigen has no values. Demonstrates that including 'not' axiom does not cause misclassification when entire H antigen set is optional.
Salmonella Miami *Salmonella* Sendai	*Salmonella* I 1,9,12:a:1,5	K-W Defined equivalence These two serovars are equivalent in the ontology as they share antigenic formulae (and there are no optionals).
Salmonella II 6,7:l,z_{28}:e,n,x	*Salmonella* II 6,7:l,z_{28}:e,n,x	† Equivalence between H_x and H_z_{16} Demonstrates that the assertion of equivalence between two *Salmonella* H antigens (H_x ≡ H_z_{16}) results in proper classification.
Salmonella Inchpark	*Salmonella* I 6,8:y:1,7	Serovar is a subtype of another serovar (*Salmonella* Alagbon).

† K-W substitutes H_x for H_ z_{16} or omits H_z_{16} from serovar formulae when H_z_{16} is present in isolate antigenic formulae. See section 9.3.5 of this document for expanded discussion of this equivalence.

Once the initial evaluation was complete, a separate list of serovars was created from several sources. The list shown in table 11b included the top 10 serovars reported to 1) CDC for human clinical isolates, 2) National Veterinary Services Laboratory (NVSL) for non-human clinical isolates, and 3) NVSL for non-human non-clinical isolates (CDC, 2014). As serovars appeared in each of the three reports, 20 unique serovars captured the 3 top 10 lists. When there were no optional antigens in the serovar formula, a single individual was created. While it seemed appropriate to create individuals representing all combinations of optional antigens, a complete list of individuals based on the eight optional antigens of *Salmonella* Senftenberg would include 256 different individuals. We assert that it was unnecessary to test each combination based on strategic testing we had conducted and the fact that simple optional antigens do not influence classification in SSDO. When serovar class definitions included optional antigens, we tested definitions of one antigenic formula that included only the required antigens and a definition that included all listed antigens. When the serovar definitions included exclusive optionality or the assertion of absence, additional individuals were created to account for the more complicated definitions. Table 11b lists the serovars tested, their respective formulae and the source of the isolate.

Table 11b. Serovars (SSDO classes) reported to CDC and NVSL against which individual serotypes were tested.

Serovar	**Formula**	**Isolate source**
Salmonella Agona	*Salmonella* I 1,4,[5],12:f,g,s:[1,2]:[z_{27}],[z_{45}]	Clinical (NVSL)
Salmonella Bareilly	*Salmonella* I 6,7,14:y:1,5	Clinical (CDC)
Salmonella Braenderup	*Salmonella* I 6,7,14:e,h:e,n,z_{15}	Non-clinical (NVSL)
Salmonella Cerro	*Salmonella* I 6,14,18:z_4,z_{23}:[1,5]:[z_{45}],[z_{82}]	Clinical (NVSL)
Salmonella Derby	*Salmonella* I 1,4,[5],12:f,g:[1,2]	Clinical (NVSL)
Salmonella Dublin	*Salmonella* I 1,9,12,[Vi]:g,p:–	Clinical (NVSL)
Salmonella Enteritidis	*Salmonella* I 1,9,12:g,m:–	Clinical (CDC) Clinical (NVSL) Non-clinical (NVSL)
Salmonella Heidelberg	*Salmonella* I 1,4,[5],12:r:1,2	Clinical (CDC) Clinical (NVSL) Non-clinical (NVSL)
Salmonella I 4,[5],12:i:-	*Salmonella* I 4,[5],12:i:-	Clinical (CDC) Clinical (NVSL)
Salmonella Infantis	*Salmonella* I 6,7,14:r:1,5:[R1…],[z_{37}],[z_{45}],[z_{49}]	Clinical (CDC) Clinical (NVSL)
Salmonella Javiana	*Salmonella* I 1,9,12:l,z_{28}:1,5:[R1…]	Clinical (CDC)
Salmonella Kentucky	*Salmonella* I 8,20:i:z_6	Non-clinical (NVSL)
Salmonella Mbandaka	*Salmonella* I 6,7,14:z_{10}:e,n,z_{15}:[z_{37}],[z_{45}]	Non-clinical (NVSL)
Salmonella Montevideo	*Salmonella* I {6,7,14}{54}:g,m,[p],s:[1,2,7]	Clinical (CDC) Clinical (NVSL) Non-clinical (NVSL)
Salmonella Muenchen	*Salmonella* I 6,8:d:1,2:[z_{67}]	Clinical (CDC)
Salmonella Muenster	*Salmonella* I 3,{10}{15}{15,34}:e,h:1,5:[z_{48}]	Non-clinical (NVSL)
Salmonella Newport	*Salmonella* I 6,8,20:e,h:1,2:[z_{67}],[z_{78}]	Clinical (CDC) Clinical (NVSL)
Salmonella Senftenberg	*Salmonella* I 1,3,19:g,[s],t:– :[z_{27}],[z_{34}],[z_{37}],[z_{43}],[z_{45}],[z_{46}],[z_{82}]	Non-clinical (NVSL)
Salmonella Thompson	*Salmonella* I 6,7,14:k:1,5:[R1…]	Non-clinical (NVSL)
Salmonella Typhimurium	*Salmonella* I 1,4,[5],12:i:1,2	Clinical (CDC) Clinical (NVSL) Non-clinical (NVSL)

8. Results

8.1 Classification

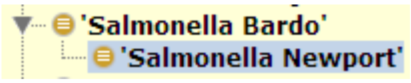
We were able to create logically sound definitions for all K-W serovars, including those added via the K-W supplement of 2010. Table 12 provides descriptive statistics of SSDO, listing serovars by sub-species, the first (immediate) proximal primitive classes. Once a reasoner (FaCT++) was applied to SSDO, a limited amount of logical re-arrangement occurred. One hundred forty-two became level one supertypes of other classes. Level one supertypes remained direct descendants of the primitive *Salmonella* classes. Conversely, 175 classes became subtypes of other serovars. There was a limited amount of subtype nesting in that some level 1 supertypes acquired multiple direct descendants, and a very few were involved in subtyping that spanned multiple generations. Nineteen serovars became logical subtypes of more than one supertype and there were 5 instances of class equivalence involving a total of 12 K-W serovars. Figures 9a, 9b and 9c give examples of the supertype-subtype relationships as depicted in the graphical user interface of Protégé.

Table 12. Descriptive statistics of SSDO. The 2595 SSDO serovars (classes) were distributed across the seven *Salmonella* subspecies. Following classification, logical relationships were identified between and among the serovars.

Subspecies[†]	**SSDO serovars**	**Level 1 supertypes**	**Subtypes of Level 1**	**Multiple supertypes**	**Instances of Equivalence**
Salmonella I	1542	99	129	9	3
Salmonella II	510	30	37	5	1
Salmonella IIIa	100	12	7	5	0
Salmonella IIIb	334	1	2	0	1
Salmonella IV	73	0	0	0	0
Salmonella V	23	0	0	0	0
Salmonella VI	13	0	0	0	0
Total	**2595**	**142**	**175**	**19**	**5**

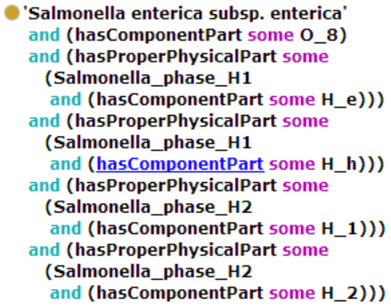
† *Salmonella* V is not actually a subspecies, but *Salmonella bongori,* the other *Salmonella* species.

Figure 9a. Level 1 Supertype (*Salmonella* Bardo) and its Subtype (*Salmonella* Newport). The definition of *Salmonella* Newport includes all axioms of *Salmonella* Bardo plus the axiom for the O_6 antigen.



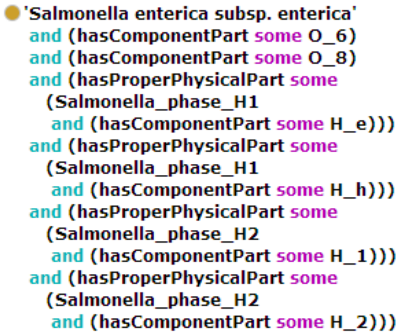
'Salmonella Bardo'
'Salmonella Newport'

***Salmonella* Bardo**



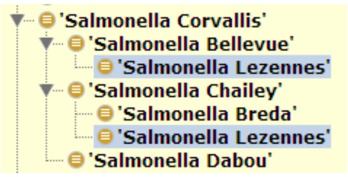
```
'Salmonella enterica subsp. enterica'
 and (hasComponentPart some O_8)
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_e)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_h)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_1)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_2)))
```

***Salmonella* Newport**



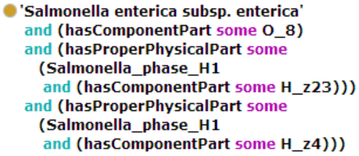
```
'Salmonella enterica subsp. enterica'
 and (hasComponentPart some O_6)
 and (hasComponentPart some O_8)
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_e)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_h)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_1)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_2)))
```

Figure 9b. Multiple supertypes. *Salmonella* Lezennes is a subtype of both *Salmonella* Bellevue and *Salmonella* Chailey which are, in turn, subtypes of *Salmonella* Corvallis. *Salmonella* Bellevue adds axioms for phase H2 H_1 and H_7 antigens to the definition of *Salmonella* Corvallis while *Salmonella* Chailey adds an axiom for O_6. *Salmonella* Lezennes adds both the H2 axioms and the O_6 axiom.



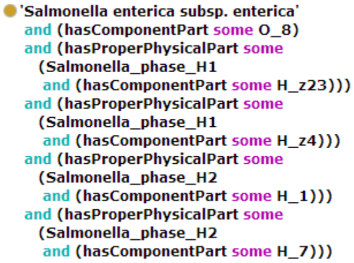
```
'Salmonella Corvallis'
    'Salmonella Bellevue'
        'Salmonella Lezennes'
    'Salmonella Chailey'
        'Salmonella Breda'
        'Salmonella Lezennes'
    'Salmonella Dabou'
```

***Salmonella* Corvallis**



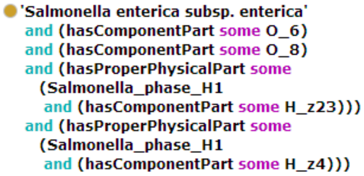
```
'Salmonella enterica subsp. enterica'
 and (hasComponentPart some O_8)
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_z23)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_z4)))
```

***Salmonella* Bellevue**



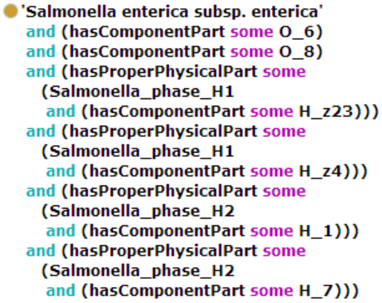
```
'Salmonella enterica subsp. enterica'
 and (hasComponentPart some O_8)
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_z23)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_z4)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_1)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_7)))
```

***Salmonella* Chailey**



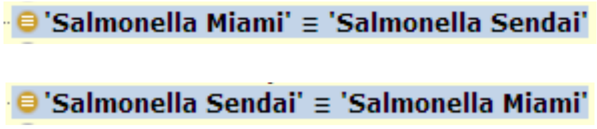
```
'Salmonella enterica subsp. enterica'
 and (hasComponentPart some O_6)
 and (hasComponentPart some O_8)
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_z23)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_z4)))
```

***Salmonella* Lezennes**



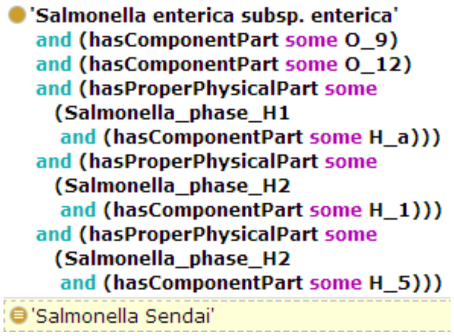
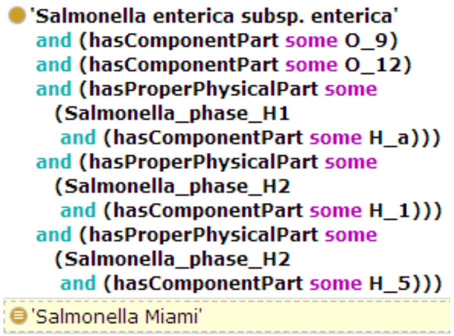
```
'Salmonella enterica subsp. enterica'
 and (hasComponentPart some O_6)
 and (hasComponentPart some O_8)
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_z23)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_z4)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_1)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_7)))
```

Figure 9c. Serovar Equivalence. *Salmonella* Miami and *Salmonella* Sendai share antigenic formulae and definitions. Each has an entry in the hierarchy and are shown to be equivalent through the symbol ≡.



'Salmonella Miami' ≡ 'Salmonella Sendai'

'Salmonella Sendai' ≡ 'Salmonella Miami'

Salmonella Miami	*Salmonella* Sendai
 'Salmonella enterica subsp. enterica' and (hasComponentPart some O_9) and (hasComponentPart some O_12) and (hasProperPhysicalPart some (Salmonella_phase_H1 and (hasComponentPart some H_a))) and (hasProperPhysicalPart some (Salmonella_phase_H2 and (hasComponentPart some H_1))) and (hasProperPhysicalPart some (Salmonella_phase_H2 and (hasComponentPart some H_5))) 'Salmonella Sendai'	 'Salmonella enterica subsp. enterica' and (hasComponentPart some O_9) and (hasComponentPart some O_12) and (hasProperPhysicalPart some (Salmonella_phase_H1 and (hasComponentPart some H_a))) and (hasProperPhysicalPart some (Salmonella_phase_H2 and (hasComponentPart some H_1))) and (hasProperPhysicalPart some (Salmonella_phase_H2 and (hasComponentPart some H_5))) 'Salmonella Miami'

8.2 Strategic Testing

Development of SSDO proceeded through a series of iterative steps that lead us to create a limited number of definition types to account for the rules of the K-W scheme. Table 13a lists the tests conducted, the individual antigenic formulas tested as individuals and the classification result obtained.

Table 13a. Results of SSDO strategic testing after classification. Antigenic formulae of the isolates represent the complexity associated with the antigenic formula of the serovar.

Test Serovar	Individual formula	Result
No options *Salmonella* Eko	*Salmonella* I 4,12:e,h:1,6	Classified as expected
No options *Salmonella* II 9,46:z_{10}:z_6	*Salmonella* II 9,46:z_{10}:z_6	Classified as expected
Simple optionality *Salmonella* Texas	*Salmonella* I 4.12:k:e,n,z_{15} *Salmonella* I 4,5,12:k:e,n,z_{15}	Classified as expected Classified as expected
Simple optionality *Salmonella* Treforest	*Salmonella* I 51:z:1,6 *Salmonella* I 1,51:z:1,6	Classified as expected Classified as expected
Simple optionality *Salmonella* Paratyphi C	*Salmonella* I 6,7:c:1,5 *Salmonella* I 6,7,Vi:c:1,5	Classified as expected Classified as expected
Assertion of absence [*] *Salmonella* Abortusequi	*Salmonella* I 4,12:-:e,n,x	Classified as expected
Exclusive optionality *Salmonella* Anatum	*Salmonella* I 3,10:e,h:1,6:z_{64} *Salmonella* I 3,15:e,h:1,6:z_{64} *Salmonella* I 3,15,34:e,h:1,6:z_{64}	Classified as expected Classified as expected Classified as expected
Assertion of absence [†] *Salmonella* Uno	*Salmonella* I 6,8:z29:- *Salmonella* I 6,8:z29:e,n,x,z_{15}	Classified as expected Classified as expected
K-W Defined equivalence *Salmonella* Miami *Salmonella* Sendai	*Salmonella* I 1,9,12:a:1,5	Classified as expected
H_x and H_z_{16} equivalence *Salmonella* II 6,7:l,z_{28}:e,n,x	*Salmonella* II 6,7:l,z28:e,n,z_{16}	Classified as expected
Serovar is subtype [‡] *Salmonella* Inchpark	*Salmonella* I 6,8:y:1,7	Classified as expected

[*] - definitions of serovar and individual include the axiom "and not (hasProperPhysicalPart some Salmonella_phase_H1)"

[†] - definition of the serovar class does not include an axiom for the H_2 phase (as the e,n,x,z_{15}) complex is listed as optional in the antigenic formula. The definition of the individual isolate includes the axiom "and not (hasProperPhysicalPart some Salmonella_phase_H2)"

[‡] - the individual is a member of the *Salmonella* Inchpark class but as *Salmonella* Inchpark is a subtype of *Salmonella* Alagbon, the individual is also listed as a member of *Salmonella* Alagbon.

8.3 Classification of CDC and NVSL isolates.

Once the strategic testing was completed, we proceeded to evaluate the list of serovars reported in the incidence lists provided by CDC and NVSL. Table 13b lists the CDC and NVSL serovars, the antigenic formulae of the individuals tested and the classification outcome. In most cases, the individuals were identified as members of the correct serovar class and the relationship was 1:1. When the correct serovar class was a logical subtype of another serovar class, the individuals were also members of the supertype class, which is logically correct as members of any class are, by definition, members of associated superclasses. For example, all isolates are subtypes of Genus *Salmonella* as well as some particular serovar. In several cases, the individual was identified as being a member of more than one serovar class not involved in subtype-supertype relationships.

Table 13b. Results of classification based on the antigenic formula of isolates. Serovars were selected from CDC and NVSL incidence lists. Antigenic formulae of the two isolates represented the minimum and maximum complexity possible given the antigenic formula of the serovar.

Surveillance serovar	**Individual formula**	**Classification Result**
Salmonella Agona	*Salmonella* I 4,12:f,g,s:– *Salmonella* I 1,4,5,12:f,g,s:1,2:z_{27},z_{45}	As expected As expected
Salmonella Bareilly	*Salmonella* I 6,7:y:1,5 *Salmonella* I 6,7,14:y:1,5	As expected As expected
Salmonella Braenderup	*Salmonella* I 6,7:e,h:e,n,z_{15} *Salmonella* I 6,7,14:e,h:e,n,z_{15}	As expected As expected
Salmonella Cerro	*Salmonella* I 18:z_4,z_{23}:– *Salmonella* I 6,14,18:z_4,z_{23}:1,5:z_{45},z_{82}	As expected **Dual**
Salmonella Derby	*Salmonella* I 4,12:f,g:– *Salmonella* I 1,4,5,12:f,g:1,2	As expected As expected
Salmonella Dublin	*Salmonella* I 9,12:g,p:– *Salmonella* I 1,9,12,Vi:g,p:–	As expected As expected
Salmonella Enteritidis	*Salmonella* I 9,12:g,m:– *Salmonella* I 1,9,12:g,m:1,7	**Dual** As expected
Salmonella Heidelberg	*Salmonella* I 4,12:r:1,2 *Salmonella* I 1,4,5,12:r:1,2	As expected As expected
Salmonella I 4,[5],12:i:-	*Salmonella* I 4,12:i:– *Salmonella* I 4,5,12:i:–	As expected As expected
Salmonella Infantis	*Salmonella* I 6,7:r:1,5 *Salmonella* I 6,7,14:r:1,5:R1…,z_{37},z_{45},z_{49}	As expected As expected
Salmonella Javiana	*Salmonella* I 9,12:l,z_{28}:1,5 *Salmonella* I 1,9,12:l,z_{28}:1,5:R1…	As expected As expected
Salmonella Kentucky	*Salmonella* I 8:i:z_6 *Salmonella* I 8,20:i:z_6	As expected As expected
Salmonella Mbandaka	*Salmonella* I 6,7:z_{10}:e,n,z_{15} *Salmonella* I 6,7,14:z_{10}:e,n,z_{15}:z_{37},z_{45}	As expected As expected

Salmonella Montevideo	*Salmonella* I 6,7:g,m,s:–	**Triple**
	Salmonella I 6,7,14:g,m,p,s:1,2,7	**Dual**
	Salmonella I 6,7,14:g,m,s:–	**Quadruple**
	Salmonella I 54:g,m,,s:–	As expected
Salmonella Muenchen	*Salmonella* I 6,8:d:1,2	As expected
	Salmonella I 6,8:d:1,2:[z_{67}]	As expected
Salmonella Muenster	*Salmonella* I 3,10:e,h:1,5	As expected
	Salmonella I 3,10:e,h:1,5:z_{48}	As expected
	Salmonella I 3,15:e,h:1,5	As expected
	Salmonella I 3,15:e,h:1,5:z_{48}	As expected
	Salmonella I 3,15,34:e,h:1,5	As expected
	Salmonella I 3,15,34:e,h:1,5:z_{48}	As expected
Salmonella Newport	*Salmonella* I 6,8:e,h:1,2	As expected
	Salmonella I 6,8,20:e,h:1,2:z_{67},z_{78}	As expected
Salmonella Senftenberg	*Salmonella* I 1,3,19:g,t:–	As expected
	Salmonella I 1,3,19:g,s,t:–:z_{27},z_{34},z_{37},z_{43},z_{45},z_{46},z_{82}	As expected
Salmonella Thompson	*Salmonella* I 6,7:k:1,5	As expected
	Salmonella I 6,7,14:k:1,5:R1…	**Dual**
Salmonella Typhimurium	*Salmonella* I 4,12:i:1,2	As expected
	Salmonella I 1,4,5,12:i:1,2	As expected

9. Discussion

9.1 General discussion

In general, we were successful in our effort to produce an ontology that correctly classifies the K-W serovars. If a logical definition is created for a laboratory isolate and the certain conditions are met, the serovar of the isolate is consistently and correctly determined. Conditions required for 100% fidelity in this regard are that the *Salmonella* subspecies is correctly identified, there are no false negative test results (all antigens are properly detected), there are no false positive test results (no antigens are detected that are not present) and that the isolate is, in fact, a member of a serovar that has been previously described and incorporated into K-W.

The SSDO is a non-unique solution to classification of the K-W serovars. As we considered various options for logical definitions and the inclusion or exclusion of possible ontologic classes, we were attempting to provide a solution that was compatible with external information systems used to communicate organism identity associated with cases of salmonellosis. Limitations arise from a need to integrate the serovars into SNOMED CT, by the structure and content of the PHIN-VADS microbiology table and by the K-W scheme itself. During our investigation and the development of SSDO, we made specific choices in the design that attempt to provide a best fit.

9.1.1 Logical definitions

Logical definitions are created from at least one root primitive class, a set of attributes (characteristics) and a set of values. A logical definition is essentially a list of axioms or statements that use attributes to describe the association between the concept class being defined and the values that differentiate it from the other classes in the ontology. Description logic then interprets the combinations of attributes and values to place classes in the ontology.

9.1.1.1 Attribute selection

The fact that SSDO can correctly classify serotype individuals (instances) does not depend on any particular meaning of the attributes and values on which the definitions are based. As long as the definitions create a logic that treats somatic and flagellar antigens separately and distinguishes between the three flagellar phases, the K-W serovars will classify correctly. Being certain that SSDO is properly integrated with other ontologies in OBO and that the attributes selected from RO represent the true nature of the relationship between the classes and values in the definition is more challenging.

Classification in SSDO depends entirely on correct stated supertypes of the K-W serovars and the application of two attributes namely *hasComponentPart* and *hasProperPhysicalPart*, both of which are contained in RO. In our opinion, the *hasComponentPart* attribute is straightforward and correctly represents the relationship between the K-W serovar class and the antigens that make up its antigenic formula. In other words, antigens are a good fit for the information that the attribute is supposed to convey. On the other hand the *hasProperPhysicalPart* relationship, while functional, may not be the only option.

We came to select *hasProperPhysicalPart* for two fairly straightforward reasons. First, the logic would not work if a reasoner could not distinguish between the three flagellar phases. An attribute distinct from the one used for antigen values was needed. Second, it seems very reasonable to assert that flagella are proper physical parts of bacteria. However, reexamination of the associated value set (H1 flagellum, H2 flagellum, H other flagellum) suggests that this assertion is not the only option. It can be argued that H1, H2 and H other are not really three different kinds of flagella but rather just the flagella that are present during three distinct phases of laboratory testing. This would mean that "phases" are in fact information about conduct of the test and not information about the flagella (or even the bacteria) themselves. If this is the case, an alternative attribute may well be *hasAbstractPart*, also from RO. This change will not disrupt SSDO in any logical way, as a global substitution of one attribute for another will not alter the way the reasoned treats the affected definitions. This idea will be reassessed before SSDO is submitted as a candidate ontology for OBO.

9.1.1.2 Handling optional antigens

K-W serovars are representations of *Salmonella* organisms that can cause disease and are identified by public health laboratories. The actual organisms that are members of a serovar may have identical antigenic formulae. If this is the case, the antigenic formula is fixed and none of the antigens are optional. A list of simple existential statements is adequate to allow individual isolates to be properly classified. For other serovars, the organism members share at least some antigens but other antigens may be optionally expressed. For these, it was necessary to omit existential statements that named the optional antigens. Because the operator is existential ("some") the definition does not preclude that a member may possess antigens that are not listed in the definition.

There are two exceptions to the pattern described in the previous paragraph. As was described previously, antigen optionality is created by whole substitution of antigens between isolates rather than a simple presence or absence of a single antigen. An example would be *Salmonella* Stockholm (*Salmonella* I 3,{10}{$\underline{15}$}:y:z_6). In this serovar, the isolates will all bear the O_3 antigen and either O_10 or O_15, but not both. For proper classification to occur, the logical definition of this and similar serovars included the union operator as shown in Table 14. The other exception is our use of "assertions of absence" to clearly define organisms that were not diphasic. This exception is discussed in detail in the following section "Subtyping of K-W Serovars"

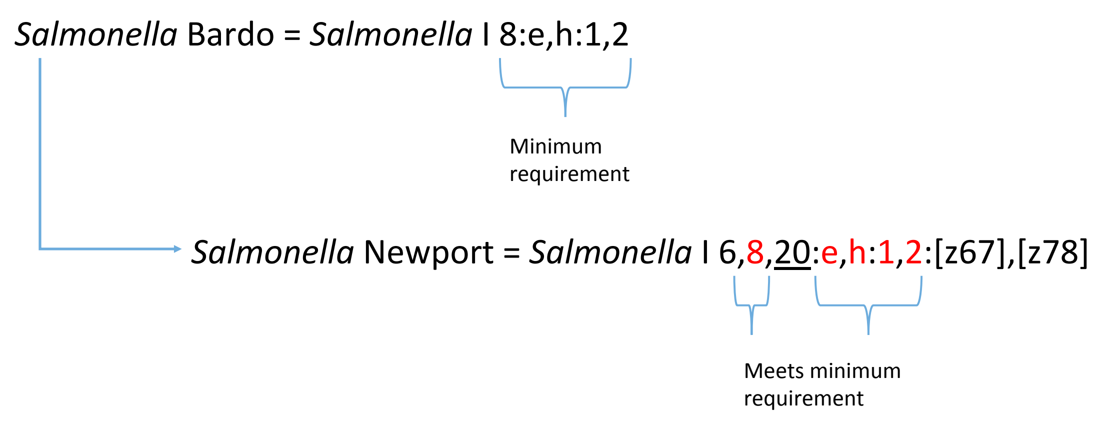
Table 14. Construction of logical definitions for serovars containing union sets ('or' operator).

{exclusive}	*Salmonella* Stockholm	*Salmonella* I 3,{10}{$\underline{15}$}:y:z_6	'Salmonella enterica subsp. enterica' and (hCP some O_3) and ((hCP some O_10) **or** (hCP some O_15)) and (hPPP some (SH1 and hCP some H_y)) and (hPPP some (SH2 and hCP some H_z6))

hCP = hasComponentPart
hPPP = hasProperPhysicalPart
SH1 = Salmonella_phase_H1
SH2 = Salmonella_phase H2

Although proper classification occurs for serovars that include optional antigens and those that do not, a subtle logical side effect occurs namely, some K-W serovar classes become subtypes of others after reasoning. This organization into a hierarchy is based on the logical definitions of the serovars themselves. An example is shown in Figure 10. In this way, SSDO is not in perfect alignment with the K-W scheme even though it correctly represents the relationship between serovars and isolates.

Figure 10. Antigenic formulae of two serovars. *Salmonella* Newport is a subtype of *Salmonella* Bardo because serovar Newport contains the minimum required (stated) antigens.

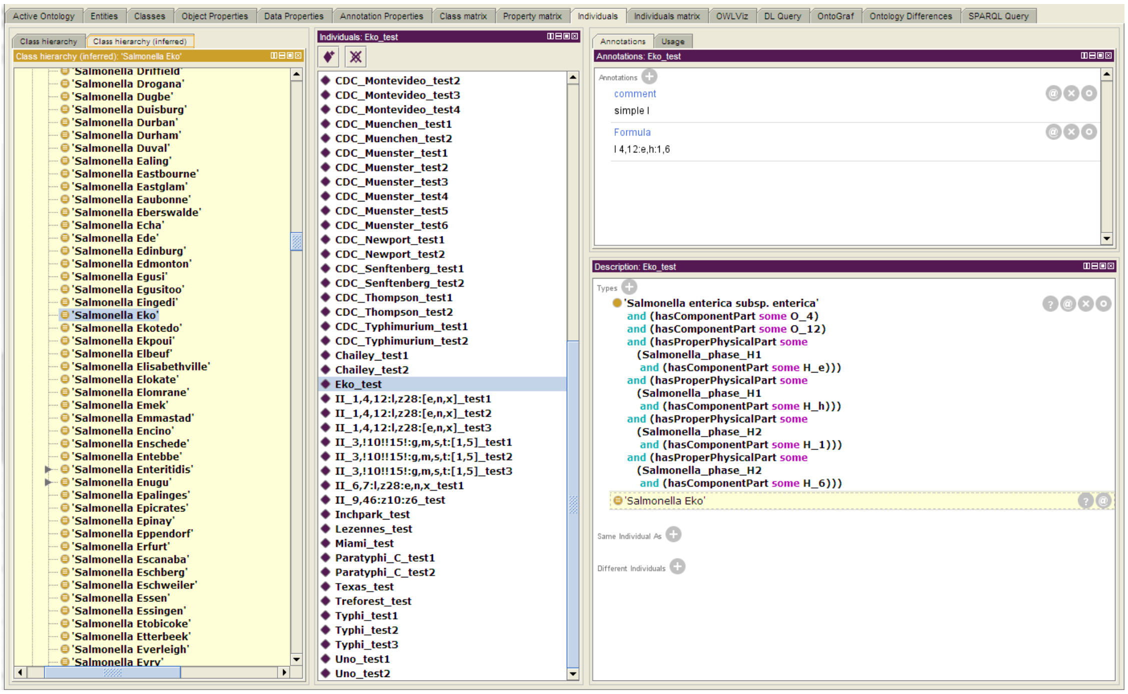

Salmonella Bardo = *Salmonella* I 8:e,h:1,2
Minimum requirement
Salmonella Newport = *Salmonella* I 6,8,20:e,h:1,2:[z67],[z78]
Meets minimum requirement

9.1.2 Classification of isolates

The primary focus of these investigations has been to facilitate the assignment of proper serovar names to laboratory isolates. Within an ontology, this assignment is made by determining the "membership" of the isolate within a serovar class. An isolate is deemed to be a member of a class if its stated characteristics match those of a class. The simplest example is a 1:1 match. *Salmonella* Eko is a serovar class defined by a list of required antigens. All laboratory isolates that are members of the class will possess all of the required antigens. When an antigen is optional, which is the case for O_1 in the serovar *Salmonella* Treforest (*Salmonella* I 1,51:z:1,6), laboratory isolates may either be '*Salmonella* I 1,51:z:1,6' or '*Salmonella* I 51:z:1,6'. The more optional antigens present in the formula of a serovar, the longer the list of distinct antigenic formulae among its isolate members.

Functional testing of SSDO proceeded by determining that isolates of fixed antigenic formulae became members of the serovar class to which they should be assigned. To meet the needs of laboratories attempting to assign serovar names to isolates, each isolate should be correctly assigned to one and only one serovar class as shown in Figures 11a and 11b. This criteria is met by SSDO in all but two cases. First, when serovar class subtyping occurs, the members of the most dependent class are automatically members of that class plus any supertype class. This result is anticipated and presents no obstacle to correct name assignment provided one can determine which class is the most dependent. Handling of individuals by Protégé demonstrates this appropriately as shown in Figures 12a, 12b and 12c. From the individuals tab, an individual is identified by the dependent class of which it is a member. From the class tab, supertype classes will show members that accrue to it from its subtypes. In addition to the effect of class supertyping on membership, there are certain logical combinations of requirements and optionality that can produce a situation where an isolate is classified as a member of two or more serovars that are not in a subtype/supertype relationship. The logical constructions that lead to these exceptions are discussed further in the next two sections.

Figure 11a. Individuals tab view of functional testing of SSDO with *Salmonella* Eko. The left pane highlights the serovar class *Salmonella* Eko. The center pane highlights the test individual, and the lower right pane shows that test isolate Eko_test is a member of the serovar class *Salmonella* Eko.



Class hierarchy (inferred): 'Salmonella Eko'

'Salmonella Driffield'
'Salmonella Drogana'
'Salmonella Dugbe'
'Salmonella Duisburg'
'Salmonella Durban'
'Salmonella Durham'
'Salmonella Duval'
'Salmonella Ealing'
'Salmonella Eastbourne'
'Salmonella Eastglam'
'Salmonella Eaubonne'
'Salmonella Eberswalde'
'Salmonella Echa'
'Salmonella Ede'
'Salmonella Edinburg'
'Salmonella Edmonton'
'Salmonella Egusi'
'Salmonella Egusitoo'
'Salmonella Eingedi'
'Salmonella Eko'
'Salmonella Ekotedo'
'Salmonella Ekpoui'
'Salmonella Elbeuf'
'Salmonella Elisabethville'
'Salmonella Elokate'
'Salmonella Elomrane'
'Salmonella Emek'
'Salmonella Emmastad'
'Salmonella Encino'
'Salmonella Enschede'
'Salmonella Entebbe'
'Salmonella Enteritidis'
'Salmonella Enugu'
'Salmonella Epalinges'
'Salmonella Epicrates'
'Salmonella Epinay'
'Salmonella Eppendorf'
'Salmonella Erfurt'
'Salmonella Escanaba'
'Salmonella Eschberg'
'Salmonella Eschweiler'
'Salmonella Essen'
'Salmonella Essingen'
'Salmonella Etobicoke'
'Salmonella Etterbeek'
'Salmonella Everleigh'
'Salmonella Evry'

Individuals: Eko_test

CDC_Montevideo_test2
CDC_Montevideo_test3
CDC_Montevideo_test4
CDC_Muenchen_test1
CDC_Muenchen_test2
CDC_Muenster_test1
CDC_Muenster_test2
CDC_Muenster_test3
CDC_Muenster_test4
CDC_Muenster_test5
CDC_Muenster_test6
CDC_Newport_test1
CDC_Newport_test2
CDC_Senftenberg_test1
CDC_Senftenberg_test2
CDC_Thompson_test1
CDC_Thompson_test2
CDC_Typhimurium_test1
CDC_Typhimurium_test2
Chailey_test1
Chailey_test2
Eko_test
II_1,4,12:l,z28:[e,n,x]_test1
II_1,4,12:l,z28:[e,n,x]_test2
II_1,4,12:l,z28:[e,n,x]_test3
II_3,!10!!15!:g,m,s,t:[1,5]_test1
II_3,!10!!15!:g,m,s,t:[1,5]_test2
II_3,!10!!15!:g,m,s,t:[1,5]_test3
II_6,7:l,z28:e,n,x_test1
II_9,46:z10:z6_test
Inchpark_test
Lezennes_test
Miami_test
Paratyphi_C_test1
Paratyphi_C_test2
Texas_test
Treforest_test
Typhi_test1
Typhi_test2
Typhi_test3
Uno_test1
Uno_test2

Annotations: Eko_test

Annotations

comment
simple I

Formula
I 4,12:e,h:1,6

Description: Eko_test

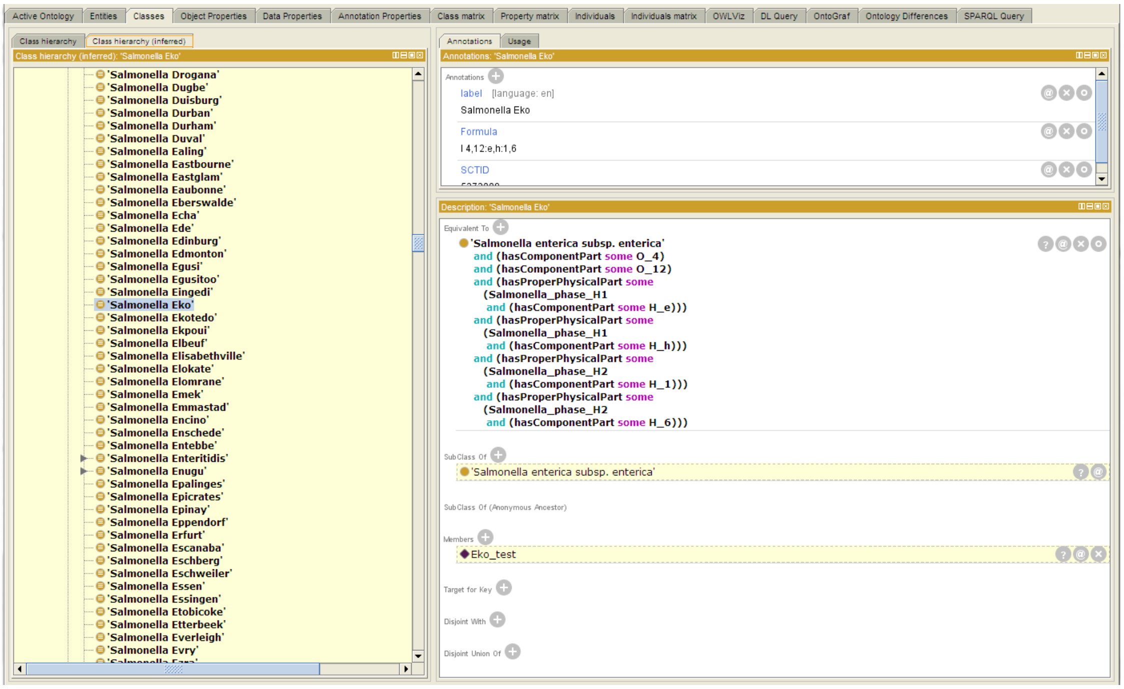
Types

```
'Salmonella enterica subsp. enterica'
 and (hasComponentPart some O_4)
 and (hasComponentPart some O_12)
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_e)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_h)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_1)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_6)))
'Salmonella Eko'
```

Same Individual As

Different Individuals

Figure 11b. Class tab view of functional testing of SSDO with *Salmonella* Eko. The left pane highlights the serovar class *Salmonella* Eko. The lower right pane shows that test isolate Eko_test is a member of the serovar class *Salmonella* Eko.



Class hierarchy (inferred): 'Salmonella Eko'

'Salmonella Drogana'
'Salmonella Dugbe'
'Salmonella Duisburg'
'Salmonella Durban'
'Salmonella Durham'
'Salmonella Duval'
'Salmonella Ealing'
'Salmonella Eastbourne'
'Salmonella Eastglam'
'Salmonella Eaubonne'
'Salmonella Eberswalde'
'Salmonella Echa'
'Salmonella Ede'
'Salmonella Edinburg'
'Salmonella Edmonton'
'Salmonella Egusi'
'Salmonella Egusitoo'
'Salmonella Eingedi'
'Salmonella Eko'
'Salmonella Ekotedo'
'Salmonella Ekpoui'
'Salmonella Elbeuf'
'Salmonella Elisabethville'
'Salmonella Elokate'
'Salmonella Elomrane'
'Salmonella Emek'
'Salmonella Emmastad'
'Salmonella Encino'
'Salmonella Enschede'
'Salmonella Entebbe'
'Salmonella Enteritidis'
'Salmonella Enugu'
'Salmonella Epalinges'
'Salmonella Epicrates'
'Salmonella Epinay'
'Salmonella Eppendorf'
'Salmonella Erfurt'
'Salmonella Escanaba'
'Salmonella Eschberg'
'Salmonella Eschweiler'
'Salmonella Essen'
'Salmonella Essingen'
'Salmonella Etobicoke'
'Salmonella Etterbeek'
'Salmonella Everleigh'
'Salmonella Evry'
'Salmonella Ezra'

Annotations: 'Salmonella Eko'

Annotations

label [language: en]
Salmonella Eko

Formula
I 4,12:e,h:1,6

SCTID

Description: 'Salmonella Eko'

Equivalent To

```
'Salmonella enterica subsp. enterica'
 and (hasComponentPart some O_4)
 and (hasComponentPart some O_12)
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_e)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H1
    and (hasComponentPart some H_h)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_1)))
 and (hasProperPhysicalPart some
   (Salmonella_phase_H2
    and (hasComponentPart some H_6)))
```

SubClass Of
'Salmonella enterica subsp. enterica'

SubClass Of (Anonymous Ancestor)

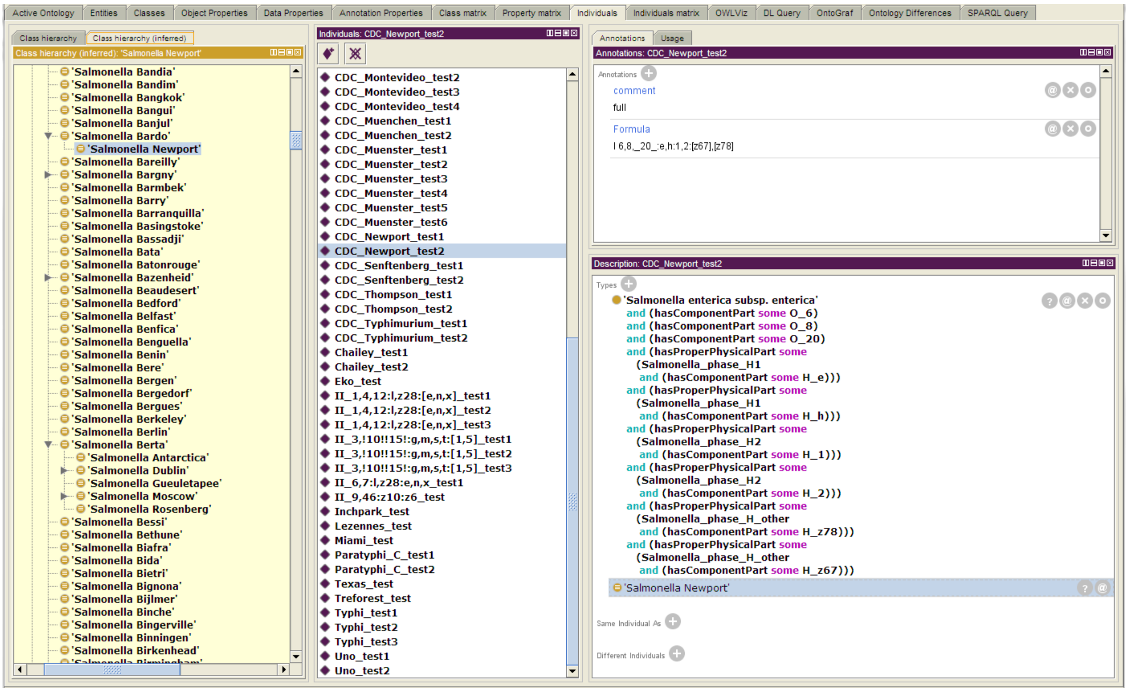
Members
Eko_test

Target for Key

Disjoint With

Disjoint Union Of

Figure 12a. Individuals tab view of functional testing of SSDO with *Salmonella* Newport. The left pane highlights the serovar class *Salmonella* Newport. The center pane highlights the test individual, and the lower right pane shows that test isolate CDC_Newport_test2 is a member of the serovar class *Salmonella* Newport.



Class hierarchy (inferred): 'Salmonella Newport'

'Salmonella Bandia'
'Salmonella Bandim'
'Salmonella Bangkok'
'Salmonella Bangui'
'Salmonella Banjul'
'Salmonella Bardo'
'Salmonella Newport'
'Salmonella Bareilly'
'Salmonella Bargny'
'Salmonella Barmbek'
'Salmonella Barry'
'Salmonella Barranquilla'
'Salmonella Basingstoke'
'Salmonella Bassadji'
'Salmonella Bata'
'Salmonella Batonrouge'
'Salmonella Bazenheid'
'Salmonella Beaudesert'
'Salmonella Bedford'
'Salmonella Belfast'
'Salmonella Benfica'
'Salmonella Benguella'
'Salmonella Benin'
'Salmonella Bere'
'Salmonella Bergen'
'Salmonella Bergedorf'
'Salmonella Bergues'
'Salmonella Berkeley'
'Salmonella Berlin'
'Salmonella Berta'
'Salmonella Antarctica'
'Salmonella Dublin'
'Salmonella Gueuletapee'
'Salmonella Moscow'
'Salmonella Rosenberg'
'Salmonella Bessi'
'Salmonella Bethune'
'Salmonella Biafra'
'Salmonella Bida'
'Salmonella Bietri'
'Salmonella Bignona'
'Salmonella Bijlmer'
'Salmonella Binche'
'Salmonella Bingerville'
'Salmonella Binningen'
'Salmonella Birkenhead'
'Salmonella Birmingham'

Individuals: CDC_Newport_test2

CDC_Montevideo_test2
CDC_Montevideo_test3
CDC_Montevideo_test4
CDC_Muenchen_test1
CDC_Muenchen_test2
CDC_Muenster_test1
CDC_Muenster_test2
CDC_Muenster_test3
CDC_Muenster_test4
CDC_Muenster_test5
CDC_Muenster_test6
CDC_Newport_test1
CDC_Newport_test2
CDC_Senftenberg_test1
CDC_Senftenberg_test2
CDC_Thompson_test1
CDC_Thompson_test2
CDC_Typhimurium_test1
CDC_Typhimurium_test2
Chailey_test1
Chailey_test2
Eko_test
II_1,4,12:l,z28:[e,n,x]_test1
II_1,4,12:l,z28:[e,n,x]_test2
II_1,4,12:l,z28:[e,n,x]_test3
II_3,!10!!15!:g,m,s,t:[1,5]_test1
II_3,!10!!15!:g,m,s,t:[1,5]_test2
II_3,!10!!15!:g,m,s,t:[1,5]_test3
II_6,7:l,z28:e,n,x_test1
II_9,46:z10:z6_test
Inchpark_test
Lezennes_test
Miami_test
Paratyphi_C_test1
Paratyphi_C_test2
Texas_test
Treforest_test
Typhi_test1
Typhi_test2
Typhi_test3
Uno_test1
Uno_test2

Annotations: CDC_Newport_test2

Annotations

comment
full

Formula
I 6,8,_20_:e,h:1,2:[z67],[z78]

Description: CDC_Newport_test2

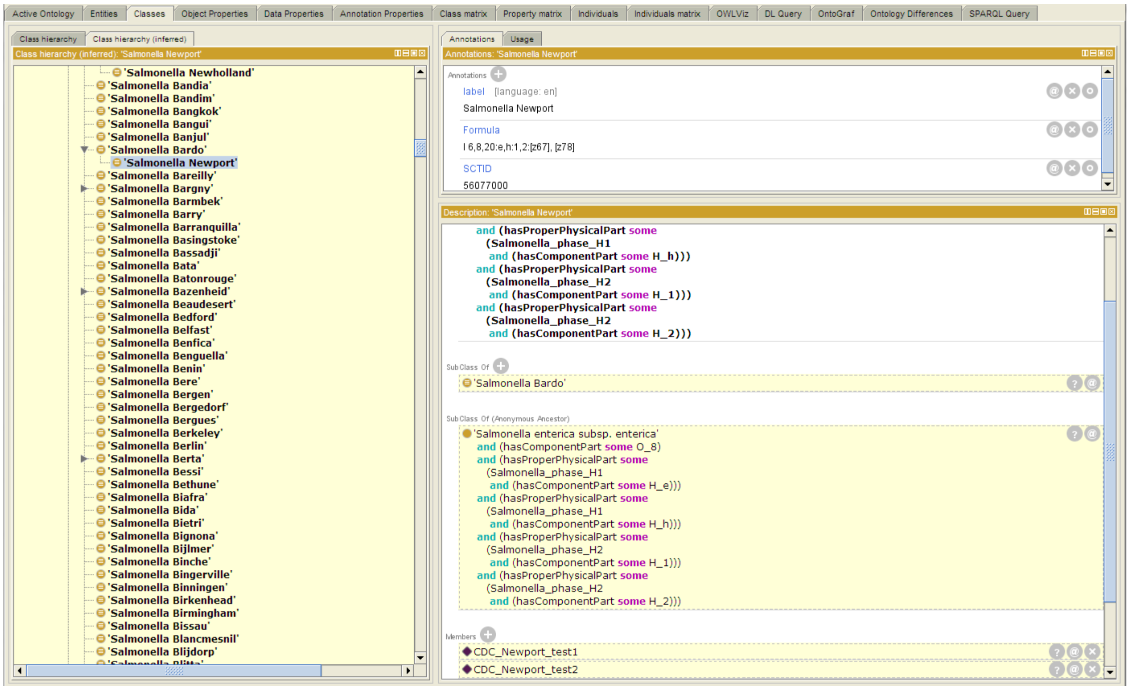
Types

```
'Salmonella enterica subsp. enterica'
  and (hasComponentPart some O_6)
  and (hasComponentPart some O_8)
  and (hasComponentPart some O_20)
  and (hasProperPhysicalPart some
    (Salmonella_phase_H1
     and (hasComponentPart some H_e)))
  and (hasProperPhysicalPart some
    (Salmonella_phase_H1
     and (hasComponentPart some H_h)))
  and (hasProperPhysicalPart some
    (Salmonella_phase_H2
     and (hasComponentPart some H_1)))
  and (hasProperPhysicalPart some
    (Salmonella_phase_H2
     and (hasComponentPart some H_2)))
  and (hasProperPhysicalPart some
    (Salmonella_phase_H_other
     and (hasComponentPart some H_z78)))
  and (hasProperPhysicalPart some
    (Salmonella_phase_H_other
     and (hasComponentPart some H_z67)))
'Salmonella Newport'
```

Same Individual As

Different Individuals

Figure 12b. Class tab view of functional testing of SSDO with *Salmonella* Newport. The left pane highlights the serovar class *Salmonella* Newport. The lower right pane shows that test isolate CDC_Newport_test2 is a member of the serovar class *Salmonella* Newport. Also in the lower right pane, note that *Salmonella* Newport is a subclass of *Salmonella* Bardo.



Class hierarchy (inferred): 'Salmonella Newport'

```
        'Salmonella Newholland'
    'Salmonella Bandia'
    'Salmonella Bandim'
    'Salmonella Bangkok'
    'Salmonella Bangui'
    'Salmonella Banjul'
  ▼ 'Salmonella Bardo'
        'Salmonella Newport'
    'Salmonella Bareilly'
  ▶ 'Salmonella Bargny'
    'Salmonella Barmbek'
    'Salmonella Barry'
    'Salmonella Barranquilla'
    'Salmonella Basingstoke'
    'Salmonella Bassadji'
    'Salmonella Bata'
    'Salmonella Batonrouge'
  ▶ 'Salmonella Bazenheid'
    'Salmonella Beaudesert'
    'Salmonella Bedford'
    'Salmonella Belfast'
    'Salmonella Benfica'
    'Salmonella Benguella'
    'Salmonella Benin'
    'Salmonella Bere'
    'Salmonella Bergen'
    'Salmonella Bergedorf'
    'Salmonella Bergues'
    'Salmonella Berkeley'
    'Salmonella Berlin'
  ▶ 'Salmonella Berta'
    'Salmonella Bessi'
    'Salmonella Bethune'
    'Salmonella Biafra'
    'Salmonella Bida'
    'Salmonella Bietri'
    'Salmonella Bignona'
    'Salmonella Bijlmer'
    'Salmonella Binche'
    'Salmonella Bingerville'
    'Salmonella Binningen'
    'Salmonella Birkenhead'
    'Salmonella Birmingham'
    'Salmonella Bissau'
    'Salmonella Blancmesnil'
    'Salmonella Blijdorp'
```

Annotations: 'Salmonella Newport'

Annotations

label [language: en]
Salmonella Newport

Formula
I 6,8,20:e,h:1,2:[z67], [z78]

SCTID
56077000

Description: 'Salmonella Newport'

```
    and (hasProperPhysicalPart some
      (Salmonella_phase_H1
       and (hasComponentPart some H_h)))
    and (hasProperPhysicalPart some
      (Salmonella_phase_H2
       and (hasComponentPart some H_1)))
    and (hasProperPhysicalPart some
      (Salmonella_phase_H2
       and (hasComponentPart some H_2)))
```

SubClass Of

'Salmonella Bardo'

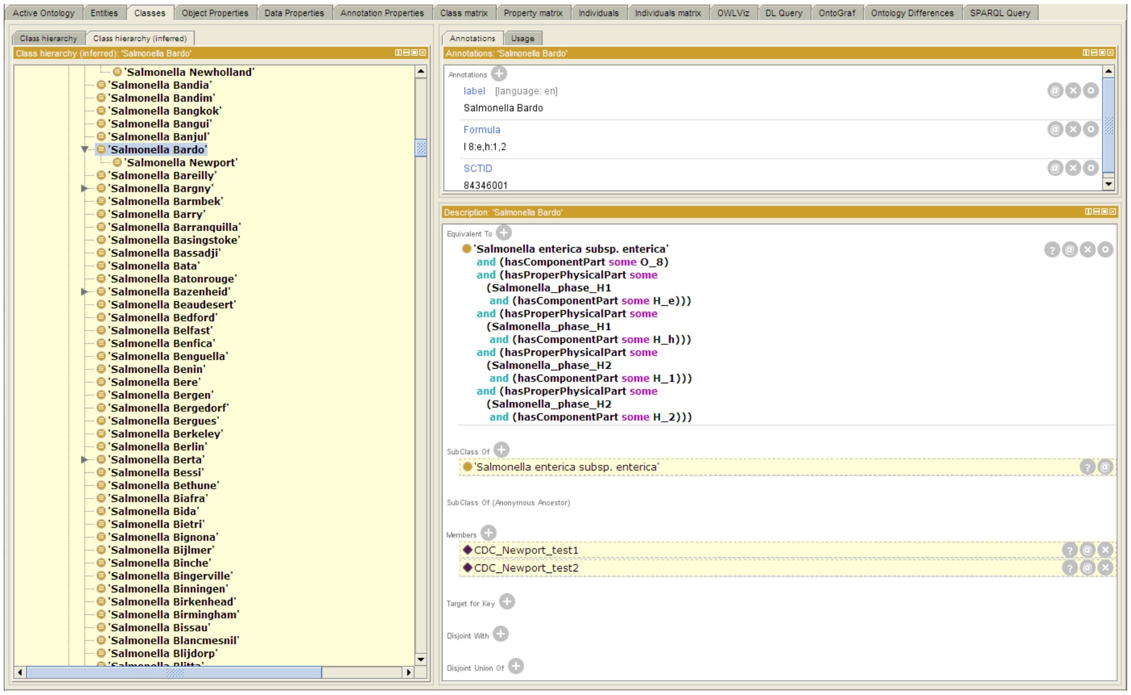
SubClass Of (Anonymous Ancestor)

```
'Salmonella enterica subsp. enterica'
  and (hasComponentPart some O_8)
  and (hasProperPhysicalPart some
    (Salmonella_phase_H1
     and (hasComponentPart some H_e)))
  and (hasProperPhysicalPart some
    (Salmonella_phase_H1
     and (hasComponentPart some H_h)))
  and (hasProperPhysicalPart some
    (Salmonella_phase_H2
     and (hasComponentPart some H_1)))
  and (hasProperPhysicalPart some
    (Salmonella_phase_H2
     and (hasComponentPart some H_2)))
```

Members

◆ CDC_Newport_test1
◆ CDC_Newport_test2

Figure 12c. Class tab view of functional testing of SSDO with *Salmonella* Newport. The left pane highlights the serovar class *Salmonella* Bardo. The lower right pane shows that test isolate CDC_Newport_test2 is a member of the serovar class *Salmonella* Bardo. This is an expected result as membership in a subtype class (*Salmonella* Newport) confers membership in the supertype class (*Salmonella* Bardo).



```
'Salmonella Newholland'
'Salmonella Bandia'
'Salmonella Bandim'
'Salmonella Bangkok'
'Salmonella Bangui'
'Salmonella Banjul'
'Salmonella Bardo'
    'Salmonella Newport'
'Salmonella Bareilly'
'Salmonella Bargny'
'Salmonella Barmbek'
'Salmonella Barry'
'Salmonella Barranquilla'
'Salmonella Basingstoke'
'Salmonella Bassadji'
'Salmonella Bata'
'Salmonella Batonrouge'
'Salmonella Bazenheid'
'Salmonella Beaudesert'
'Salmonella Bedford'
'Salmonella Belfast'
'Salmonella Benfica'
'Salmonella Benguella'
'Salmonella Benin'
'Salmonella Bere'
'Salmonella Bergen'
'Salmonella Bergedorf'
'Salmonella Bergues'
'Salmonella Berkeley'
'Salmonella Berlin'
'Salmonella Berta'
'Salmonella Bessi'
'Salmonella Bethune'
'Salmonella Biafra'
'Salmonella Bida'
'Salmonella Bietri'
'Salmonella Bignona'
'Salmonella Bijlmer'
'Salmonella Binche'
'Salmonella Bingerville'
'Salmonella Binningen'
'Salmonella Birkenhead'
'Salmonella Birmingham'
'Salmonella Bissau'
'Salmonella Blancmesnil'
'Salmonella Blijdorp'
'Salmonella Blitta'
```

```
Annotations
  label  [language: en]
  Salmonella Bardo
  Formula
  I 8:e,h:1,2
  SCTID
  84346001
```

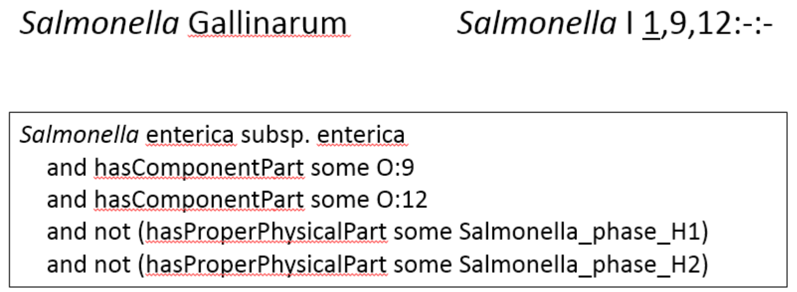
```
Equivalent To
  'Salmonella enterica subsp. enterica'
   and (hasComponentPart some O_8)
   and (hasProperPhysicalPart some
     (Salmonella_phase_H1
      and (hasComponentPart some H_e)))
   and (hasProperPhysicalPart some
     (Salmonella_phase_H1
      and (hasComponentPart some H_h)))
   and (hasProperPhysicalPart some
     (Salmonella_phase_H2
      and (hasComponentPart some H_1)))
   and (hasProperPhysicalPart some
     (Salmonella_phase_H2
      and (hasComponentPart some H_2)))

SubClass Of
  'Salmonella enterica subsp. enterica'

SubClass Of (Anonymous Ancestor)

Members
  CDC_Newport_test1
  CDC_Newport_test2

Target for Key

Disjoint With

Disjoint Union Of
```

9.1.3 Subtyping of K-W serovar classes

One goal for this ontology was to keep the logical definitions relatively simple while still facilitating the correct classification of individual isolates (instances). Specifically, we determined early on that optional antigens (underscore 1, bracket [5], etc.) had no effect on the subsequent matches between serovar classes and serotype instances. Interestingly, the authors of K-W make the following assertion: "*Presence or absence of accessory O factors (underlined or in square brackets) does not interfere with serovar identification.*" Therefore our decision not to include optional antigens in definitions is consistent with the K-W scheme.

When the resulting ontology is classified, a side-effect of omitting optional antigens is revealed. As mentioned previously, serovars form natural subtype hierarchies that K-W does not represent. As we demonstrated correct classification of instances, we did not force alignment between the ontology and K-W with one notable exception. In the case of monophasic serovars, such as *Salmonella* Abortusequi (*Salmonella* I 4,12:-:e,n,x), asserting that the value of the H1 antigen is optional is simply not correct. We suggest that the ontology is in better alignment with the biology of the serovars if we assert that the serovar does NOT possess a particular flagellar phase. The most extreme case is *Salmonella* Gallinarum (Figure 13) that does not bear any flagellar antigens (*Salmonella* I 1,9,12:-:-). Treating H1 and H2 antigens as optional in this serovar resulted in the accumulation of large numbers of serovar subtypes that were actually motile (having one or more H phase flagella). In order to avoid this inappropriate subtyping, we deliberately added assertions concerning the absence of flagellar phases when the absence defined the serovar(s). The remaining subtyping that occurs is consistent, as the formula of a serovar subtype always matches the formula of a supertype, but with the simple addition of one or more antigens in any position (O, H1, H2, H_other).

Figure 13. Logical definition of the non-motie serovar *Salmonella* Gallinarum.



Salmonella Gallinarum *Salmonella* I 1,9,12:-:-

```
Salmonella enterica subsp. enterica
   and hasComponentPart some O:9
   and hasComponentPart some O:12
   and not (hasProperPhysicalPart some Salmonella_phase_H1)
   and not (hasProperPhysicalPart some Salmonella_phase_H2)
```

One side effect of asserting that particular flagellar phases do not exist becomes apparent when one considers the actual information available to laboratory personnel when they attempt to classify any isolate that is monophasic. When the absence of a flagellar phase is asserted in the definition of a serovar class, the definition of individuals that should be identified as members of that serovar must also contain the absence assertion “and not (hasProperPhysicalPart some Salmonella_phase_H*x*).” A system designed to classify isolates according to the ontology could easily add the “not” axiom to the definition of isolates that did not possess antigens of a particular phase. Initially we questioned the correctness of this for situation where a serovar class includes isolates such as *Salmonella* Paratyphi A (*Salmonella* I 1,2,12:a:[1,5]) for which an entire flagellar phase is optional. Although it seems paradoxical, if the system creates the absence assertion in such a case, the logic we used will still classify the isolate correctly. This is not only consistent with our application of description logic, but it is also likely to be true in a biological sense for that isolate.

While we can prove that subtyping within the ontology does not adversely affect classification of individual instances, we also postulate that there could be some value to its existence. The presence of optional antigens suggests that K-W must decide whether similar antigenic formulae with and without a particular antigen should be separate serovars or combined into one using optionality. Historically, *Salmonella* Typhimurium was presented as two serovars, namely *Salmonella* Typhimurium (with O:5) and *Salmonella* Typhimurium var. Copenhagen (without O:5). In subsequent revisions of the scheme, these two were reconciled to a single serovar and O:5 was declared to be optional.

When we were considering the subtyping issue in depth, we noticed one rather complicated example. Specifically *Salmonella* Hessarek (*Salmonella* I 4,12,[27]:a:1,5) classified as a subtype of *Salmonella* Fulica (*Salmonella* I 4,[5],12:a:[1,5]). The similarity between these two formulae may not be apparent to the casual observer. Because the logical definitions do not include optional antigens, these two differ only by the H2 antigens. The phase H2_1,5 antigens are optional in *Salmonella* Fulica (and therefore absent from the definition) and mandatory for *Salmonella* Hessarek (and therefore present in the definition). The potential importance of this finding is supported by a brief discussion in K-W concerning these two serovars.

> "Serovars Hessarek (4,12,[27]:a:1,5) and Fulica (4,[5],12:a:[1,5]), which formula could be similar, are not combined because they differ by biochemical characters. Rhamnose, gas production from glucose, dulcitol, trehalose, Simmons citrate, L(+) tartrate (= d-tartrate), mucate, H2S, and tetrathionate-reductase are positive for Hessarek and negative for Fulica. This latter serovar is very rare."

K-W considered these two serovars for unification. The fact that these serovars differ biochemically may have brought them to the attention of the K-W editors. However this may not be mere coincidence. Attempting to assess similarities such as these among nearly 2600 serovars would be a daunting manual task. Our ontology does nothing to determine whether classes should be combined or maintained separately, however viewing our ontology after classification makes the existence of antigenically related serovars much more obvious.

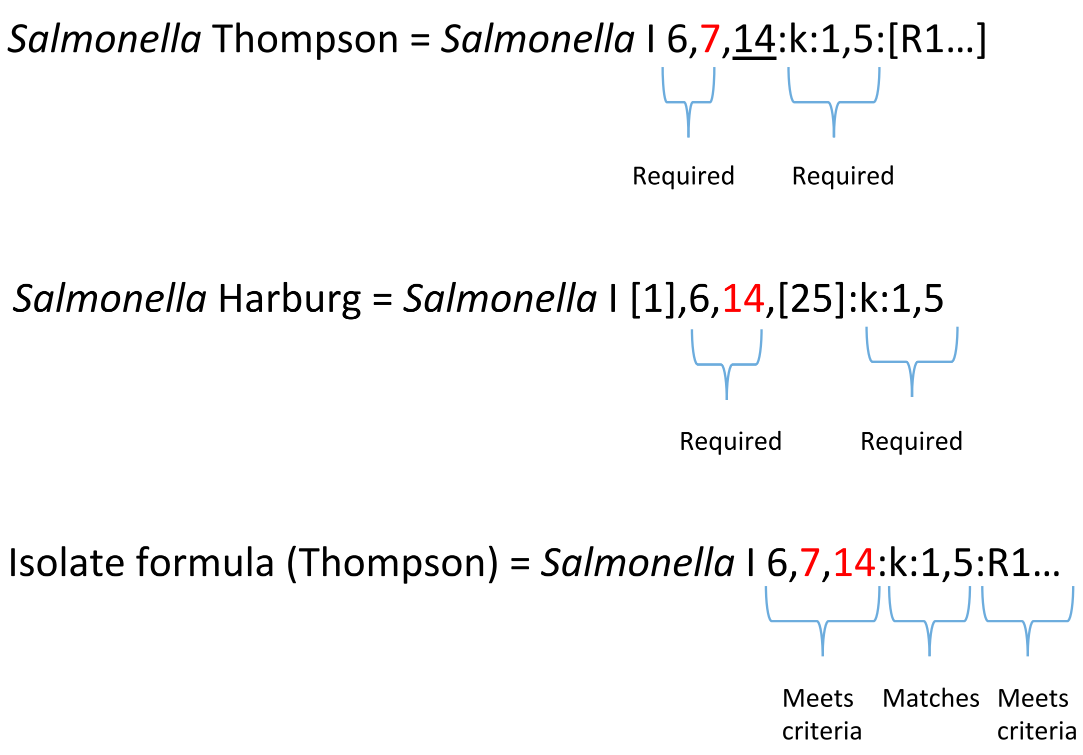
9.1.4 Multiple classification of isolates without subtyping

There are instances in K-W of serovar classes that have antigenic formulae related in such a way that individuals may become members of multiple classes even though the classes are not subtypes of each other. In this case, the ontology assigns the isolate to multiple serovar classes, only one of which is correct. Specifically, this can occur when the following conditions are met:

1) The stated characteristics of the isolate match the stated characteristics of the serovar (this is true for ALL memberships).
2) Optionality of serovar must allow for presence of additional antigens (this is true for ALL memberships).
3) Required (stated) characteristics of serovars precludes subtyping.

Table 13b lists classification results for serovars extracted from CDC and NVSL incidence lists. Individual antigenic formulae generated for four serovar classes were identified as members of more than one serovar, and those serovars were not related by subtyping. Two instances appeared to be a natural result of coincidental similarities in the antigenic formulae of the serovars. At least one isolate formula that matches *Salmonella* Thompson also meets the minimum criteria for *Salmonella* Harburg. This example is shown in Figure 14. Similarly, at least one isolate formula that matches *Salmonella* Cerro also meets the minimum criteria for *Salmonella* Arapahoe. Any system that leverages this ontology for classification of isolates will need to evaluate the dual classification on the basis of a best fit (matching) algorithm. Best fit in this case simply matches the list of antigens of the isolate. The correct serovar will always include a reference to all antigens present in the isolate.

Figure 14. Antigenic formulae of two serovars and a test isolate. *Salmonella* Thompson and *Salmonella* Harburg are not related by subtyping because required (stated) antigens do not match. The isolate classifies as *Salmonella* Thompson because its formula matches. The isolate classifies as serovar Harburg because its definition does not preclude the existence of O_7 or H_R1... among its members.


Salmonella Thompson = *Salmonella* I 6,7,14:k:1,5:[R1…]
Required
Required
Salmonella Harburg = *Salmonella* I [1],6,14,[25]:k:1,5
Required
Required
Isolate formula (Thompson) = *Salmonella* I 6,7,14:k:1,5:R1…
Meets criteria
Matches
Meets criteria

An additional two serovars, *Salmonella* Enteritidis and *Salmonella* Montevideo also resulted in multiple classifications of test isolates. For these two serovars, K-W retains a certain level of ambiguity in the antigenic formulae. In the case of *Salmonella* Enteritidis, the isolate tested classifies as both *Salmonella* Enteritidis and *Salmonella* Berta. This is unavoidable, but depending on the antigenic formula specified for *Salmonella* Enteritidis, it is either unrelated to *Salmonella* Berta or is its subtype.

The specific issues surrounding serovar Montevideo are more complex but revolve around genetic substitution of O_54 for O_6,7. In the absence of this substitution and corresponding changes in the logical definition of the class, *Salmonella* Montevideo would be involved in subtype relationships with one or more serovars. The end result is that multiple different test isolate formulae classify as members of different combinations of these related serovars. Additional details concerning these exceptions to the K-W scheme are described in subsequent sections of this discussion.

9.1.5 Multiple supertypes

As we described in the previous section, serovar subtyping occurs either because a subtype has one or more extra antigens than its supertype or because the antigenic formula of a supertype includes optional antigens or even optional H antigen phases. It is possible that a serovar will align with more than one supertype when it meets multiple criteria independently. An example within SSDO occurs for two serovars, *Salmonella* Lezennes and *Salmonella* Blegdam, which each classify under two supertypes. Table 15 includes the antigenic formulas for these two serovars as well as the formulae of the supertypes and the specific reasons for multiple subtyping.

These instances of multiple supertypes do not represent errors as long as misclassification of isolates does not occur. Our testing indicates that in fact, isolates of *Salmonella* Lezennes classify correctly (not as either or both of the supertypes). However, it must be pointed out that based on set theory and from a supertype perspective, isolates of *Salmonella* Lezennes will be members of multiple supertype classes.

Table 15. Serovars with multiple isolates. *Salmonella* Lezennes and *Salmonella* Blegdam are shown in bold with each of their two supertypes.

Lezennes = I 6,8:z4,z23:1,7	
Bellevue = I **8:z4,z23:1,7**	The shared portion of the antigenic formula in bold. Subtyping occurs because the logical definition of Lezennes includes O_6 but is otherwise identical.
Chailey = I **6,8:z4,z23**:[e,n,z15]	The shared portion of the antigenic formula in bold. Subtyping occurs because the logical definition makes no mention of the H_2 phase antigen (it is not asserted to be absent because it is optional).
Blegdam = I 9,12:g,m,q:–	
Moscow = I 1,**9,12:g,q:–**	The shared portion of the antigenic formula in bold. Subtyping occurs because the logical definition of Blegdam includes H_1:m but is otherwise identical. The optional O:1 antigen is not included in the logical definition.
Enteritidis = I 1,**9,12:g,m:–**	The shared portion of the antigenic formula in bold. Subtyping occurs because the logical definition of Blegdam includes H_1:q but is otherwise identical. The optional O:1 antigen is not included in the logical definition.

9.1.6 Laboratory errors

The SSDO cannot correct for errors in the determination of antigens of a laboratory isolate any more than a laboratory technician with a copy of K-W can correct for such errors. It is possible however to predict the classification outcome based on the nature of the error. Examples are shown in Table 16.

Table 16. Examples of laboratory errors and possible SSDO outcomes.

Error	**Outcomes**
Biochemical identity of the subspecies is incorrect.	1. Isolate may classify as a direct subtype of the incorrect subspecies. 2. Isolate will classify as an incorrect serovar if the same antigenic formula is present in two different subspecies.
False positive serotyping error	1. Isolate may be classified correctly if the error makes it a subtype of the correct serovar) 2. Isolate may be identified as the wrong serovar if the false positive is of a flagellar antigen that the correct serovar does not possess and another serovar does. 3. Isolate may classify as a direct species subtype if the false positive is of a flagellar antigen and no serovar matches the false positive antigen.
False negative serotyping error	1. Isolate may classify as a direct species subtype 2. Isolate may classify as the incorrect serovar if the false antigen formula exists in the scheme.
Antigen substitution occurs (both false positive and false negative)	1. Isolate may classify as a direct species subtype 2. Isolate may classify as the incorrect serovar if the false antigen formula exists in the scheme.

9.1.7 New serovars

If an isolate is serotyped correctly but the resulting formula is not represented by a K-W serovar, the isolate may either classify as a member of the subspecies but NOT as a serovar, or the isolate may classify as a member of an existing serovar. As such, classification of previously undescribed serotypes will be variable but if SSDO does not place an isolate as a member of an existing serovar class, it can be assumed that the isolate is a newly described serotype. If serovar class membership is based on logical subtyping, an additional step (either manually or by algorithm) will be required to determine that a new serotype has been identified.

9.2 Exclusion of K-W serogroups from the ontology

The tables that present names for all[3] serovars are organized by so-called serogroups. These are designated by one or more O antigens present in all serovars identified as members of a given serogroup. In producing the SSDO, we considered including serogroups as concept classes, and in fact, evaluated prototypes that included them. In the final version, we decided not to include serogroups.

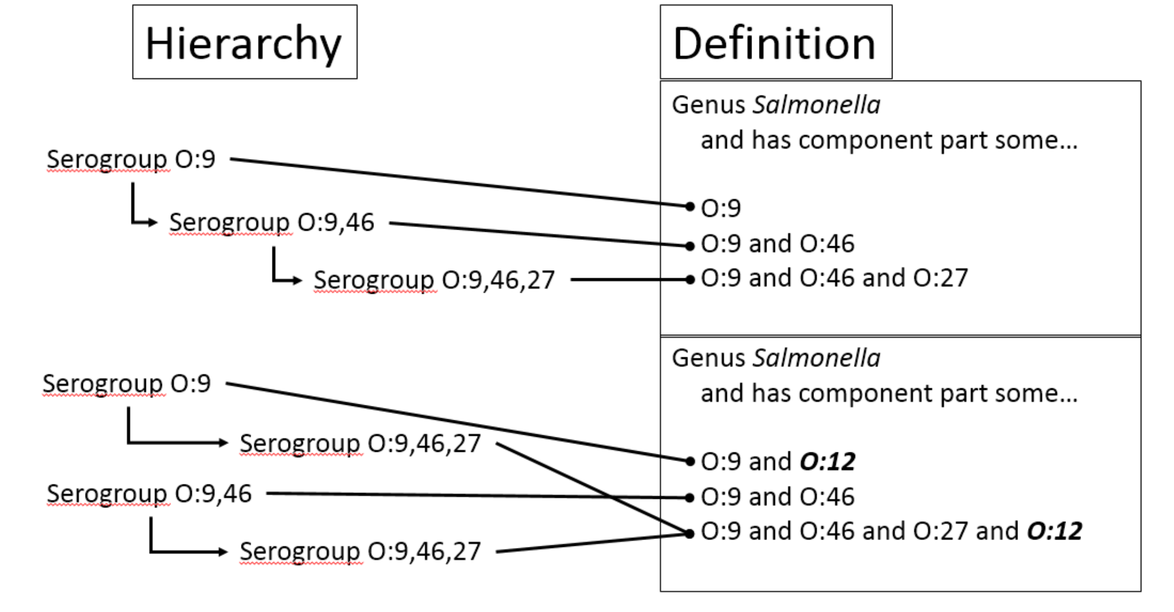
O groups serve a functional purpose during laboratory testing, as they contribute to efficient selection of pooled sera and facilitate manual lookup of serovar names, but they are neither required nor adequate for correct classification (identification) of the serovar of a particular isolate. Serovars are stated subtypes of *Salmonella* subspecies (e.g. *Salmonella enterica* subsp. *enterica*) or species (e.g. *Salmonella bongori*) in any logical classification. This characteristic distinguishes between serovars that share an antigenic formula but are not of the same subspecies. Two such serovars will occupy sequential rows in a K-W O group table, but are not distinct unless one notes the species or subspecies designation.

In addition to the fact that serogroups do little to support classification within an ontology, they also create opportunities for misclassification. Related serogroups exist that present logical difficulties in the scheme. In particular, three separate serogroups require inclusion of the O:9 antigen. By name, these are O:9, O:9,46 and O:9,46,27.

[3] The names of *Salmonella enterica* subsp. *enterica* and their antigenic formulae are presented alphabetically in an additional table.

Interestingly, serovars in groups O:9 and O:9,46,27 include the O:12 antigen, while the O:46 antigen is present in all serovars of groups O:9,46 and O:9,46,27. Unless exceptional steps are taken to create definitions that exclude lists of O antigens, two subtyping arrangements result. If one excludes the O:12 antigen from the definitions, then O:9,46,27 is a subtype O:9,46, which itself is a subtype of O:9. Alternately, the O:12 antigen can be incorporated in the definitions (when it is present in all of the isolates) at which point O:9,46 is no longer a subtype of O:9. In this case O:9,46,27 is a logical subtype of both O:9 and O:9,46. A graphical description of these hierarchies is depicted in Figure 15.

Figure 15. Alternate hierarchies of the O:9 serogroup.



The other logical difficulty is created by the existence of serogroup O:54. The authors of K-W themselves refer to the group O:54 as provisional. Biologically, O_54 is an antigen that is genetically encoded on a plasmid. When the plasmid is present, O_54 is expressed. When the plasmid is absent, O_54 is not expressed, and the antigens that are expressed require the serovar to be placed in a different serogroup. The expression of O_54 may wholly replace the O antigens of a particular serovar, or O_54 may simply be added to the list of O antigens present. The result would be variable classification of serovars bearing the plasmid, with some classifying only as only serogroup O:54 ,while others would have two supertype serogroups. A serovar such as Poeseldorf would classify as both an O:8 and an O:54, while Winnepeg would only classify as an O:54. Table 17 depicts serovars of O:54 and their respective serovar names when O_54 is absent.

Table 17. Members of provisional serogroup O:54, their serovar names, and O groups when O_54 is absent.

O_54 present	**Formula**	**O_54 absent**	**Formula**
Tonev	I 21,54:b:e,n,x	Minnesota (O:21)	I 21:b:e,n,x:[z_{33}],[z_{49}]
Winnipeg	I 54:e,h:1,5	Ferruch (O:8)	I 8:e,h:1,5
Uccle[‡]	I 3,54:g,s,t:-		N/A
Newholland	I 4,12,54:m,t:-	Banana (O:4)	I 4,[5],12:m,t:[1,5]
Poeseldorf	I 8,$\underline{20}$,54:i:z_6	Kentucky (O:8)	I 8,$\underline{20}$,:i:z_6
Ochsenwerder	I 6,7,54:k:1,5	Thompson (O:7)	I 6,7,14:k:1,5:[R1…]
Canton	I 54:z_{10}:e,n,x	Hadar (O:8)	I 6,8:z_{10}:e,n,x
Barry	I 54:z_{10}:e,n,z_{15}	Mbandaka (O:7)	I 6,7,14:z_{10}:e,n,z_{15},[z_{37}],[z_{45}]

[‡] According to K-W, no serovar has been elucidated that would represent a non-plasmid bearing version of Uccle.

9.3 Exceptions to the K-W Scheme

For some serovars, the K-W scheme notes that "exceptional" strains do not conform to the antigenic formula of the serovar itself. Specific examples include

Salmonella Enteritidis and *Salmonella* Typhi. Additionally, there are literature references to an important serotype (*Salmonella* I 4,[5],12:i-) that is currently unclassified in K-W but is occasionally referred to as a monophasic variant of *Salmonella* Typhimurium.

9.3.1 *Salmonella* Enteritidis

The antigenic formula for *Salmonella* Enteritidis (*Salmonella* I 1,9,12:g,m:–) indicates that this organism is monophasic and the first logical definition incorporated in SSDO included an assertion of this fact. This logical definition resulted in classification of the serovar class *Salmonella* Enteritidis as a subtype of *Salmonella* Berta (*Salmonella* I 1,9,12:[f],g,[t]:–). Subsequent revisions of the logical definition of *Salmonella* Enteritidis incorporated information from a table footnote: "*exceptional strains can have antigen H:1,7 as second phase*." Taken *prima facie*, this would change the published formula from '*Salmonella* I 1,9,12:g,m:– ' to '*Salmonella* I 1,9,12:g,m:[1,7]'. The two logical formulae are shown in Figure 16, the difference being the removal of the axiom that asserted monophasicity (the lack of a second phase) from the original definition. One relatively minor result of the change was that *Salmonella* Enteritidis no longer classifies as a subtype of *Salmonella* Berta because the latter serovar class retains the assertion of absence for being monophasic, while *Salmonella* Enteritidis does not.

An unexpected side effect of the change was noted during detailed evaluation of the placement of individuals as members of classes. In the original configuration, all *Salmonella* Enteritidis individuals created classified correctly as members of the *Salmonella* Enteritidis class. These individuals were also members of the *Salmonella* Berta class based on the subtyping that occurred. When the logical definition was changed an isolate with the formula '*Salmonella* I 1,9,12:g,m:- ' classifies as both a member of *Salmonella* Enteritidis (correct) and *Salmonella* Berta (*Salmonella* I 1,9,12:[f],g,[t]:–), which is no longer associated with subtyping. We considered a solution where the serovar Enteritidis would be defined with a mutually exclusive axiom ('or' construction) that contained the optional antigens (H_1 and H_7) and the assertion of exclusion (and not (hasProperPhysicalPart some Salmonella phase H2)), but this construction is logically incongruous.

Figure 16. Logical constructions for *Salmonella* Enteritidis before and after editing "and not" assertion.

Original definition	**Final definition**
'Salmonella enterica subsp. enterica' and (hCP some O_9) and (hCP some O_12) and (hPPP some Salmonella_phase_H1 and (hCP some H_g)) and (hPPP some Salmonella_phase_H1 and (hCP some H_m)) **and not (hPPP some Salmonella_phase_H2)**	'Salmonella enterica subsp. enterica' and (hCP some O_9) and (hCP some O_12) and (hPPP some Salmonella_phase_H1 and (hCP some H_g)) and (hPPP some Salmonella_phase_H1 and (hCP some H_m))

hCP = hasComponentPart
hPPP = hasProperPhysicalPart
	SH1 = Salmonella_phase_H1
	SH2 = Salmonella_phase H2

9.3.2 *Salmonella* Typhi

The antigenic formula for *Salmonella* Typhi is listed in K-W as '*Salmonella* I 9,12,[Vi]:d:–:[z66]'. Its entry in the K-W scheme reference this footnote:

> "Rare strains can have, as phase 1, H:j instead of H:d (261-nucleotide deletion in gene fliC). Independently, rare strains can have an additional phase H:z66 determined by a plasmid-borne gene."

Although the scheme does not state this directly, there is no indication that these "rare" strains should not be considered members of the serovar class *Salmonella* Typhi.

Because H_j substitutes for (replaces) H_d is consistent with the genetic alteration that causes it to happen, we decided to modify the logical definition to reflect this fact (Figure 17). Without any particular appreciation for K-W's use of the word "rare" we suggest that the antigenic formula of *Salmonella* Typhi could actually be '*Salmonella* I 9,12,[Vi]:{d}{j}:–:[z66]'.

Figure 17. Logical constructions for *Salmonella* Typhi before and after adding "or" assertion.

Original definition	Final definition
'Salmonella enterica subsp. enterica' and (hCP some O_9) and (hCP some O_12) and (hPPP some SH1 and (hCP some H_d)) and not (hPPP some SH2)	'Salmonella enterica subsp. enterica' and (hCP some O_9) and (hCP some O_12) and (hPPP some SH1 and (hCP some H_d) **or (hCP some H_j))** and not (hPPP some SH2)

hCP = hasComponentPart
hPPP = hasProperPhysicalPart
	SH1 = Salmonella_phase_H1
	SH2 = Salmonella_phase H2

9.3.3 Monophasic *Salmonella* Typhimurium

CDC's 2012 *National Enteric Disease Surveillance: Salmonella Annual Report* listed the top 20 *Salmonella* serovars in order of reporting frequency by public health laboratories in the United States. The second most often reported serovar was *Salmonella* Typhimurium (n=5704). The fifth most frequently reported serovar is the only one represented by its antigenic formula '*Salmonella* I 4,[5],12:i:-'. The appearance of this serovar on the list is interesting for two reasons. First, this serovar is not represented in the K-W scheme. Second, its antigenic formula is not only quite similar to Typhimurium (*Salmonella* I 1,4,[5],12:i:1,2), but it has been referred to as the "Monophasic variant of Typhimurium." (CDC, 2011b) In SSDO, '*Salmonella* I 4,[5],12:i:-' classifies as a subtype of *Salmonella enterica* subsp. *enterica* rather than as a member of a named serovar class. From our perspective, this is as it should be. As it has not been officially recognized as either a variant of *Salmonella* Typhimurium or given status as a serovar in its own right, the classification behavior is correct.

9.3.4 *Salmonella* Montevideo

For the serovar *Salmonella* Montevideo, K-W inexplicably provides three distinct antigenic formulae in the scheme. The first appears in serogroup O:7 (*Salmonella* I 6,7,14:g,m,[p],s:[1,2,7]), the second in serogroup O:54 (*Salmonella* I {6,7,14}{54}:g,m,s:-), and the third in the alphabetical listing of *Salmonella* I serovars (*Salmonella* I 6,7,14,[54]:g,m,[p],s:[1,2,7]). For the listing in the two serogroups, K-W provides a footnote that indicates that O_54 is substituted for O_6,7,14 when plasmid-controlled factor O_54 is present. In order to create a logical definition, we first attempted to reconcile the three formulae. If the footnotes are correct, the formula indicating simple optionality of O_54 (the third listing that includes [54]) is simply incorrect and was ignored. We then reconciled the two remaining formulae to (*Salmonella* I {6,7,14}{54}:g,m,[p],s:[1,2,7]). The logical definition asserts that the H_2 phase antigen is now optional.

As a consequence of our reconciliation of *Salmonella* Montevideo, multiple classification of a single *Salmonella* Montevideo isolate occurs between serovars Montevideo and Menston (*Salmonella* I 6,7:g,s,[t]:[1,6]). Unfortunately, by omitting the assertion for exclusion (monophasicity), a specific isolate with the formula '*Salmonella* I 6,7,14:g,m,p,s:1,2,7' will classify as both a member of *Salmonella* Montevideo (correct) and *Salmonella* Menston, which is inappropriate. The reason for dual classification (instead of a subtype relationship) is that *Salmonella* Montevideo has the axiom that asserts O_6,7 or O_54. As serovar Menston is defined by O_6,7, description logic dictates that all subtypes of serovar Menston must have, at a minimum, O_6,7. When O_54 is present in an isolate of *Salmonella* Montevideo, O_6,7 is not. Therefore, serovar Montevideo cannot be a subtype of Menston, even as their antigenic formulae are closely related. For reasons similar to this, other instances of serovar Montevideo classify in multiple locations as noted in Table 13b.

9.3.5 Flagellar (H) antigens of the e,n,x/e,n,z_{15} complex

K-W creates an interesting conundrum in their handling of the so-called e,n,x/e,n,z_{15} complex. K-W specifically states that:

> "Most e,n,x phases of S. enterica subsp. enterica (subspecies I) strains contain both factors x and z_{16}, and their formula is in fact e,n,x,z_{16}. Exceptionally, e,n,x phases can occur without z_{16} and with z_{17}. In contrast, factor x never occurs in strains of subspecies salamae (II) and diarizonae (IIIb) even if the WKLM scheme indicates e,n,x or e,n,x,z_{15}. The so-called factor x in their formula is in fact z_{16}.
>
> All e,n,z_{15} phases have, in addition, factor z_{17}. Factor z_{17} never occurs with e,n,x,z_{15} (which is in fact, e,n,z_{15},z_{16})." (Grimont & Weill, 2007)

The K-W approach to the e,n,x complex essentially deprecates H_z_{16} in the antigenic formulae of serovars. The published antigenic formulae either substitute H_x for H_z_{16} or simply omit H_z_{16}. We are left with two modeling options to accommodate this fact. Either our definitions could specify the H_z_{16} antigen where it was present or we could assert an equivalence between H_x and H_z_{16}. The first approach would classify almost all serovars correctly but would create definitions that do not align with the stated antigenic formula. More importantly, for serovars of *Salmonella enterica* subsp. *enterica*, it is not possible to know which are actually e,n,x,z_{16} and which are the rare e,n,x,z_{17}. The second approach, while not biologically true, maintains the integrity of the naming scheme and does in fact, classify correctly.

9.4 Classification of serovars that include identical individual isolates

While the SSDO properly assigns all distinct individuals correctly, there are isolates for which the serovar cannot be resolved based on their antigenic formulae alone. Depending on the serovars it may not be possible to resolve the serovar for any isolate, most of the isolates or a few of them. When these instances have been recognized, K-W provides additional characteristics by which they can be distinguished.

9.4.1 *Salmonella* Miami ≡ *Salmonella* Sendai

Salmonella Miami and *Salmonella* Sendai share identical antigenic formulae (*Salmonella* I $\underline{1}$,9,12:a:1,5). In SSDO, they have the same logical definition and classify as equivalent. As this is the case no single serovar class can be assigned on the basis of an isolate's antigenic formula. K-W indicates that these can be differentiated biochemically:

> "Serovars Miami and Sendai are both kept in this scheme because they might be different. Biochemical characters formerly used for their differentiation (xylose, arabinose, rhamnose, H_2S) can only be used to define biovars. The differentiation is now based on an essential character: Sendai, which is adapted to man, is auxotrophic, i.e. does not grow on a minimal medium with glucose or on

Simmons's citrate agar. On the contrary, Miami, which is ubiquist, is prototrophic, i.e. grows on such minimal media."

9.4.2 *Salmonella* Cholerasuis ≡ *Salmonella* Paratyphi C ≡ *Salmonella* Typhisuis

Similar to the situation with Miami and Sendai, three biovars of *Salmonella* Cholerasuis, *Salmonella* Typhisuis and *Salmonella* Paratyphi C share the antigenic formula '*Salmonella* I 6,7:c:1,5'. It should be noted that the serotype formula of *Salmonella* Paratyphi C may include the optional Vi antigen. K-W takes a similar approach to these biovars in that tables are provided to describe the biochemical, serum agglutination (using antisera unrelated to the *Salmonella* O and H antigen scheme) and host adaptation. As is the case for *Salmonella* Miami and *Salmonella* Sendai, isolates of these serovars cannot be classified on the basis of O and H antigens alone.

9.4.3 *Salmonella* Paratyphi B

Although K-W has deprecated the use of named variants of serovars, the serovar *Salmonella* Paratyphi B is associated with this footnote: *"L(+) tartrate (=d-tartrate) positive variant is often referred to as var. Java."*

Interestingly, this variant appears among the top twenty serovars for incidence in the United States. (CDC, 2014) Based on antigenic formula alone, SSDO will classify both L(+) tartrate negative and positive variants identically.

We recognize that a logical model could be developed that would include attributes and values unrelated to the serovars. In fact, we expect that SNOMED CT will develop such a model so as to organize their entire hierarchy of bacterial species. We consider this model and any related solution to be beyond the scope of our current work.

Primarily this is because attributes that convey biochemistry-based definitions will, by necessity, exist above this scheme. Distinguishing between these "biovars of serovars" will be depend on definitions the serovars will inherit from antecedent content in a larger hierarchy. In SSDO we chose not to create definitions that would allow auto-classification of *Salmonella* species and subspecies for these reasons. As such, definitions to distinguish *Salmonella* Miami and *Salmonella* Sendai would be similar.

9.4.5 Equivalent classes not referenced in K-W

Previously we discussed serovars of the same or similar antigenic formulae (e.g., Miami and Sendai) that were specifically identified by the authors of the K-W scheme. Classification of SSDO leads to identification of other equivalent serovar classes. These cases are slightly different than those specifically pointed out in K-W. For example, while it is true that the serovar is equivocal for some isolates (e.g., a '*Salmonella* I 6,14:d,e,n,x' could be either Charity or Lindern), other isolates can be identified using the tables. SSDO will not be able to determine the correct serovar even if the optional antigen is present. It will however, return information that says the isolate is a member of two serovar classes. Additional means would be required to complete the identification. For the same reasons that equivalence exists between *Salmonella* Charity and *Salmonella* Lindern (Figure 18), two serovars of *Salmonella enterica* subsp. *salamae* (II 16:g,[m],[s],t:[1,5]:[z42] ≡ II 16:g,[m],[s],t:[e,n,x]) and three of *Salmonella enterica* subsp. *diarizonae* (IIIb 60:i:[e,n,x,z15]:[z50] ≡ IIIb 60:i:[z35]:[z50] ≡ IIIb 60:i:[2]:[z50]) were also identified as logically equivalent.

Figure 18. Logical constructions for equivalent serovar classes (Charity ≡ Lindern).

Charity = I [1],6,14,[25]:d:e,n,x	'Salmonella enterica subsp. enterica' and (hCP some O_6) and (hCP some O_14) and (hPPP some (SH1 and (hCP some H_d))) and (hPPP some (SH2 and (hCP some H_e))) and (hPPP some (SH2 and (hCP some H_n))) and (hPPP some (SH2 and (hCP some H_x)))
Lindern = I 6,14,[24]:d:e,n,x	'Salmonella enterica subsp. enterica' and (hCP some O_6) and (hCP some O_14) and (hPPP some (SH1 and (hCP some H_d))) and (hPPP some (SH2 and (hCP some H_e))) and (hPPP some (SH2 and (hCP some H_n))) and (hPPP some (SH2 and (hCP some H_x)))

hCP = hasComponentPart
hPPP = hasProperPhysicalPart
	SH1 = Salmonella_phase_H1
	SH2 = Salmonella_phase H2

9.5 Isolate vs. serovar information

One primary outcome of the entire approach to *Salmonella* serotype naming is that individual organisms with distinct and fixed antigenic formula are named and grouped in such a way that they often lose this distinctiveness. This is the basic effect of "optional" antigens. Ironically, the K-W document points out that the

> "Presence or absence of accessory O factors (underlined or in square brackets) does not interfere with serovar identification. These factors are only interesting as epidemiological markers within a given serovar."

Identifying an isolate as a member of a serovar class with optional antigens results in loss of information. Whether this loss of information actually affects epidemiologic investigations would seem to be related to the odds the two outbreaks are caused by isolates from the same serovar with different antigenic formulae. The chances of such an occurrence should increase as the incidence of a particular serovar increases and as the number of concurrent outbreaks increases. Ironically, national surveillance programs that are most likely to see concurrent outbreaks caused by organisms from the same serovar have lost access to the "epidemiological markers within a given serovar." However, local investigations can still benefit from certain knowledge of the antigenic formula of a specific isolate from a specific patient.

Specifically in this investigation, we were limited in our ability to perform a robust assessment because actual PHL isolate instance information is not available. For a serovar such as *Salmonella* Senftenberg with its 8 optional antigens, we could either test all possible combinations or test the two extremes as we did (no optional antigens present or all optional antigens present). The testing we did confirmed that the SSDO logic functioned as expected. Unfortunately, we resorted to testing isolate antigenic formulae that may or may not actually exist.

10. Conclusions

We began this project by attempting to answer a question regarding the functionality of our online SNOMED CT browser. Although the direct answer to the question was a technical matter related to the text indexing and retrieval settings in our databases, testing the solution made us look more closely at the K-W scheme, the fidelity of SNOMED CT in representing the scheme and the functionality of various digital artifacts provided by WHOCC-Salm, IHTSDO, and CDC. It seemed to us that without the introduction of a new approach, reliable identification will remain a multi-step process that depends on identifying the serovar name using the K-W document (pdf) and finding the SCTID using a tool like our browser or using Microsoft Excel and the table distributed by PHIN-VADS.

Our first task then, was to determine whether and to what extent SNOMED CT faithfully replicated K-W serovar content. In general we have shown that errors in SNOMED CT are relatively small in number. Our analysis also revealed that SNOMED CT does contain a limited number of gaps in its coverage. Of greater concern are challenges associated with the organization of the content. We believe that the primary difficulties public health laboratories encounter are created through the policies and processes of the three organizations in the information chain.

Any attempt to automate serovar identification needs to be based on the presence of reliable information in available digital artifacts. CDC has determined that SCTIDs are to be submitted as the serovar identifiers. Unfortunately, K-W names the serovars in two different ways (binomial and formulaic representations) and the formulaic representations may include reference to optional antigens. Laboratories identify isolates with fixed antigenic formulae (an isolate has no optional antigens) and these are not represented directly in K-W, SNOMED CT, or PHIN-VADS. Finally, these three important information artifacts are distributed in ways that contribute little to solve the look-up problem for laboratories.

Viewed from the perspective of laboratory personnel attempting to determine the proper SCTID of an isolate, each link in the chain presents a different set of problems. The K-W pdf cannot be searched for proper representations of complete antigenic formulae because these representations are in tabular form and do not include the essential punctuation. SNOMED CT cannot be searched for the antigenic formulae of isolates because they are not present in all cases. PHIN-VADS cannot be searched for any alternate representations (synonyms) for the serovar names, whether formulae or binomials, because they are not present. In addition, all three share the limitations imposed by inefficacy of string-based searching.

To address shortcomings present in existing systems and to inform a possible restructuring of SNOMED CT, we represented the content in K-W as an ontology. Although ontologies are normally thought of in terms of the "classification of things," the most attractive feature of ontologies for organizing *Salmonella* serovars is that the "things" are described (and organized) by their characteristics. As such, K-W is a characteristics based scheme. In order to find a serovar using the K-W document (pdf), a person essentially locates antigenic characteristics of the isolate, one after another, until the formula of the isolate aligns with that of the serovar. At that intersection, the isolate has a name, and that name can be used to locate an SCTID in another resource.

As it is currently structured and populated, SSDO reliably places isolates in their proper serovar classes with very few exceptions. There are two sets of conditions under which the ontology does not or cannot readily classify an isolate. The first is when two serovars actually share identical antigenic formulae. These are well documented in K-W, and alternate means are provided for a laboratory to make the right serovar assignment depending on biochemical characteristics of the isolate. In order to correct this problem, it would be necessary to expand the scope of SSDO to incorporate biochemical characteristics. An ontology that classifies bacteria to the species level will necessarily be based on biochemical characteristics, and its model should be adopted by SSDO when practicable.

Secondly, the ontology may place an isolate as a member of two unrelated serovars. Typically, the isolate is a proper member of one of the two serovars, and this can be determined by inspection. Rarely, an isolate may not be a proper member of either serovar. Regardless, neither SSDO nor a person scanning the K-W tables for a match will find one. In this case, the formula of the isolate either represents a laboratory error, or the isolate is a member of a serovar that has not been previously described. For the rare instances of multiple classification, relatively simple best-fit validation can determine the correct serovar.

Given that SSDO does not include serogroups and serovar subtyping occurs, one might question whether SSDO represents the logic of K-W. We recognize that SSDO does not completely parallel the K-W scheme as rendered in the pdf document distributed by WHOCC-Salm. However, both K-W and SSDO represent the characteristics of the serovars equally well and differ only in presentation. Serogroups are a useful organizing principle for human beings locating serovars in enormous documents, but are not necessary for proper classification of an isolate as a member of any given serovar. Optionality conveyed through punctuation for a human reader is replaced by logical definitions and the predictable behavior of description logics.

At the present time, SSDO cannot be fully implemented in SNOMED CT. Specifically, the description logic employed by SNOMED CT cannot accommodate the exclusive "or" constructions and does not allow for negation, employed in SSDO to prevent subtyping of biphasic serovars under monophasic serovars. Although an SSDO based on our original Prototype II could be implemented without either of these logic constructions, doing so creates new problems. In the end, we are left with an ontology that does classify *Salmonella* isolates correctly and can evolve into a stand-alone application.

Ultimately, we feel that this work will assist PHLs in reporting *Salmonella* outbreak data to the CDC by allowing the proper serotype designations and corresponding SCTID to be identified quickly and accurately. We feel this model may apply to other microorganisms and eventually play a role in fully automating laboratory reporting.

11. References

Beißwanger, E., Schulz, S., Stenzhorn, H., & Hahn, U. (2008). BioTop: An Upper Domain Ontology for the Life Sciences - A Description of its Current Structure, Contents, and Interfaces to OBO Ontologies. *Applied Ontology*, 3 (4):205-212. Retrieved from http://www.imbi.uni-freib

Brenner, F., & McWhorter-Murlin, A. (1998). *Identification and serotyping of Salmonella.* Atlanta, GA: Centers for Disease Control and Prevention.

CDC. (2011a). *National Salmonella Surveillance Annual Data Summary,2009.* Atlanta, Georgia: US Department of Health and Human Services.

CDC. (2011b). *National Salmonella Surveillance Overview.* Atlanta, Georgia: US Department of Health and Human Services.

CDC. (2011c). *PHIN Vocabulary Access and Distribution System (VADS) Content Version: 2011.06.17*.

CDC. (2011d). *PHIN Vocabulary*. Retrieved from http://www.cdc.gov/phin/activities/vocabulary.html

CDC. (2013). *Revised Recommendations for the Interpretation of Salmonella Serovar Effective date: 7/25/2013.* Atlanta, Georgia: US Department of Health and Human Services.

CDC. (2014). *National Enteric Disease Surveillance: Salmonella Annual Report, 2012.* Atlanta, Georgia: US Department of Health and Human Services.

Euzéby, J. (1997). List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. *Int. J. Syst. Bacteriol.*, 47, 590-592. Retrieved from http://www.bacterio.net

Fields, P. (2006). Salmonella Serotyping. *10th Annual PulseNet Update Meeting.* Miami, FL: Association of Public Health Laboratories. Retrieved from http://www.aphl.org/conferences/proceedings/Documents/2006_10th_Annual_PulseNet_Update_Meeting/22_Fields_Salm_Sero.pdf

Fitzgerald, C., Collins, M., van Duyne, S., Mikoleit, M., Brown, T., & Fields, P. (2007). Molecular Determination of Common Salmonella Serogroups. *J. Clin. Microbiol*, 45(10):3323.

Gašević, D., Djurić, D., & Devedžić, V. (2009). Model Driven Engineering and Ontology Development (2nd ed.). Springer.

Gene Ontology Consortium. (2001). Creating the Gene Ontology Resource: Design and Implementation. *Genome Res., 11*, 1425-1433.

Gene Ontology Consortium. (2014). Retrieved from Gene Ontology Documentation: www.geneontology.org/GO.doc.shtml

Grimont, P., & Weill, F.-X. (2007). *Antigenic formulae of the Salmonella serovars, 9th Edition.* Paris, France: WHO Collaborating Centre for Reference and Research on Salmonella. Retrieved from http://www.pasteur.fr/ip/portal/action/WebdriveActionEvent/oid/01s-00003

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowl. Acq.*, 5:199–220.

Guibourdenche, M., Roggentin, P., Mikoleit, M., Fields, P. I., Bockemühl, J., Grimont, P., & Weill, F.-X. (2010). Supplement 2003-2007 (No. 47) to the White-Kauffmann-Le Minor scheme. *Res Microbiology*, 161, 26-29.

Hogan, W. (2008). Aligning the Top Level of SNOMED-CT with Basic Formal Ontology. In R. Cornet, & K. Spackman (Ed.), *Representing and sharing knowledge using SNOMED. Proceedings of the 3rd International Conference on Knowledge Representation in Medicine.* Phoenix, AZ: Nature Precedings. Retrieved from http://dx.doi.org/10.1038/npre.2008.2373.1

Horridge, M. (2011, March 24). *Protégé OWL Tutorial.* Retrieved from http://owl.cs.manchester.ac.uk/publications/talks-and-tutorials/protg-owl-tutorial/

IHTDSO. (2015, January). *SNOMED CT International Release Files*. Retrieved from http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html

IHTSDO. (2012, January). *SNOMED CT Editorial Guide.*

Krötzsch, M., Simančík, F., & Horrocks, I. (2012). A description logic primer. *Computing Research Repository (CoRR)*, abs/1201.4089.

Lapage, S., Sneath, P., Lessel, E., Skerman, V., Seeliger, H., Clark, W., & editors. (1992). *International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision.* Washington (DC): ASM Press.

Lolis, E., & Bucala, R. (2003, August). Therapeutic approaches to innate immunity: severe sepsis and septic shock. *Nature Reviews Drug Discovery, 2*, 635-645. Used with permission from Nature Publishing Group; letter attached.

McQuiston, J., Herrera-Leon, S., Wertheim, B., Doyle, J., Fields, P., Tauxe, R., & Logsdon, J. (2008). Molecular phylogeny of the salmonellae: relationships among Salmonella species and subspecies determined from four housekeeping genes and evidence of lateral gene transfer events. *J. Bacteriol*, 190:7060-7067.

McQuiston, J., Waters, R., Dinsmore, B., Mikoleit, M., & Fields, P. (2011). Molecular Determination of H Antigens of Salmonella by Use of a Microsphere-Based Liquid Array. *J. Clin. Microbiol*, 49(2):565.

Microsoft Corp. (2013). Microsoft Excel. Redmond, WA.

Mikoleit, M., & Fields, P. (2006). *Conventional Serotyping of Salmonella. National Salmonella Reference Laboratory Standard Operating Procedure.* Atlanta, GA: Centers for Disease Control and Prevention.

Mirhaji, P. (2009, June). Public Health Surveillance Meets Translational Informatics: A Desiderata. *Journal of Laboratory Automation, 14*(3), 157-170.

National Center for Biomedical Ontology. (2015). *Systematized Nomenclature of Medicine - Clinical Terms*. Retrieved from BioPortal: http://bioportal.bioontology.org/ontologies/SNOMEDCT

Open Biological and Biomedical Ontologies Foundry. (2011). Retrieved from OBO Foundry Principles: http://www.obofoundry.org

Oracle Corp. (2013). MySQL Server 5.5. Redwood Shores, CA, USA.

Public Health Information Network, CDC. (2013, 10). *Meaningful Use Fact Sheet: Reportable Lab Results.* Retrieved from http://www.cdc.gov/phin/library/PHIN_Fact_Sheets/FS_MU_RLR.pdf

Regenstrief Institute, Inc. . (2015). *LOINC Background*. Retrieved from https://loinc.org/background

Santamaria, S., Fallon, M., Green, J., Schulz, S., & Wilcke, J. (2012). Developing the Animals in Context Ontology. *Proc 2012 ICBO.* Graz, Austria.

Scallan, E., Hoekstra, R., Angulo, F., Tauxe, R., Widdowson, M., Roy, S., & et.al. (2011). Foodborne illness acquired in the United States---major pathogens. *Emerg Infect Dis*, 17(1): 7-15.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., . . . Rosse, C. (2005, April). Relations in biomedical ontologies. *Genome Biology, 6*, R46.

Smith, B., Kumar, A., & Bittner, T. (2005). *Basic Formal Ontology for Bioinformatics.* Saarland University, Institute for Formal Ontology and Medical Information Science. Saarbrücken, Germany: IFOMIS Reports.

Sneath, P. (2003). *A short history of the Bacteriological Code.* Retrieved from http://icsp.org/misc/history.html

Stanford Center for Biomedical Informatics Research. (2014). *Protege*. Retrieved from http://protege.stanford.edu

Tindall, B., Grimont, P., Garrity, G., & Euzéby, J. (2005). Nomenclature and taxonomy of the genus Salmonella. *Int. J. Syst. Evol. Microbiol*, 55:521–524.

Tsarkov, D., & Horrocks, I. (2006, April 29). FaCT++ description logic reasoner: System description. In U. Furbach, & N. Shankar (Ed.), *Proc. 3rd Int. Joint Conf. on Automated Reasoning (IJCAR'06) LNAI. 4130*, pp. 292-297. Berlin Heidelberg: Springer-Verlag.

University of Manchester. (2012, July). FaCT++ reasoner. (1.6). Manchester, England, UK. Retrieved from http://owl.cs.manchester.ac.uk/tools/fact/

VTSL. (2012). Retrieved from http://vtsl.vetmed.vt.edu/browser

Weblite Solutions Corp. (2013). PDF to Spreadsheet Pro, Version 1.3.4. Burnaby, BC, Canada. Retrieved from http://solutions.weblite.ca/pdf-to-spreadsheet/

Wilcke, J. (2014). Chair of the Organism and Infectious Disease Model Project Group, IHTSDO.

Appendices

Appendix A. *Salmonella* O groups and associated antigens

(CDC. National *Salmonella* Surveillance Overview. Atlanta, Georgia: US Department of Health and Human Services, CDC, 2011.)

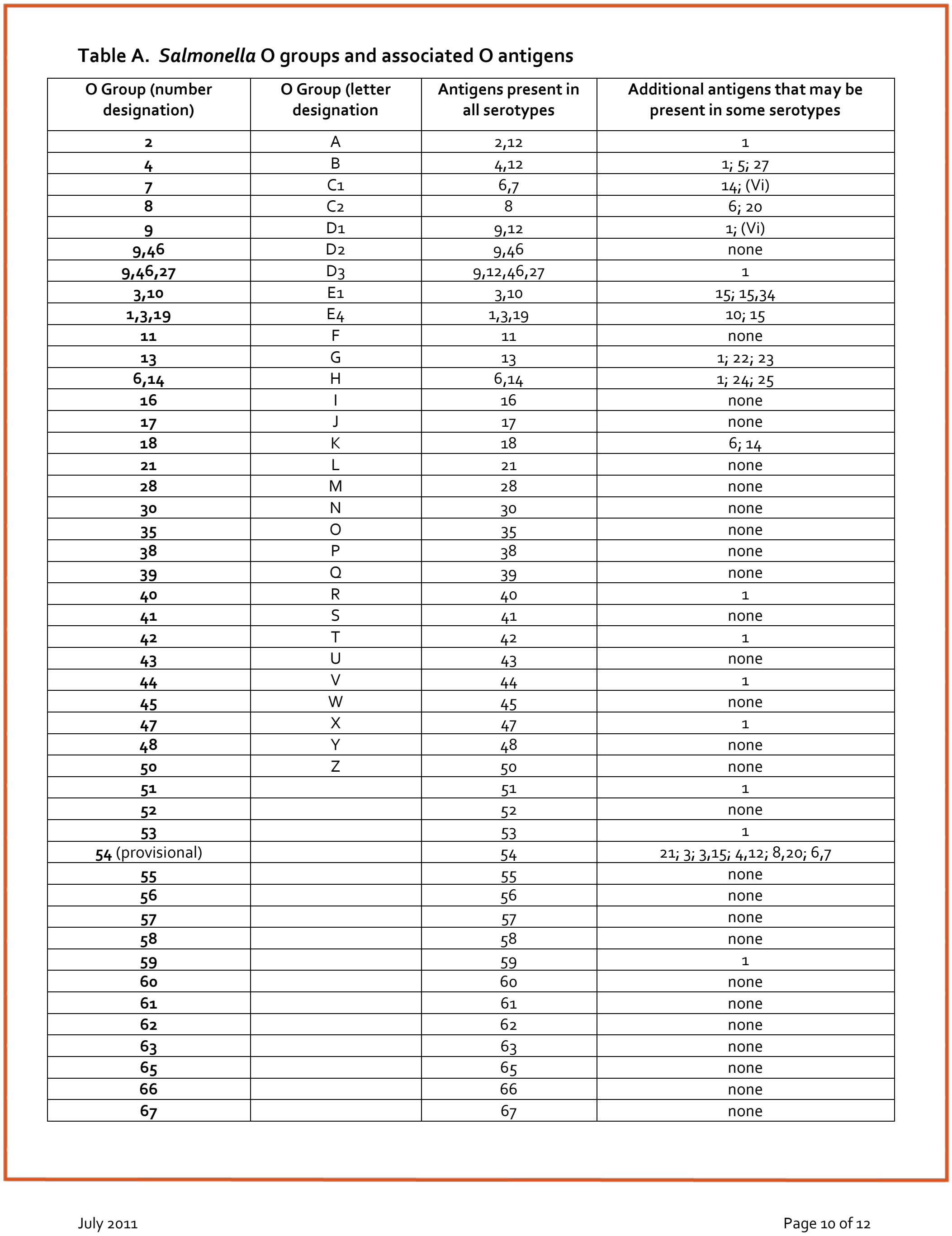


Table A. *Salmonella* O groups and associated O antigens

O Group (number designation)	O Group (letter designation	Antigens present in all serotypes	Additional antigens that may be present in some serotypes
2	A	2,12	1
4	B	4,12	1; 5; 27
7	C1	6,7	14; (Vi)
8	C2	8	6; 20
9	D1	9,12	1; (Vi)
9,46	D2	9,46	none
9,46,27	D3	9,12,46,27	1
3,10	E1	3,10	15; 15,34
1,3,19	E4	1,3,19	10; 15
11	F	11	none
13	G	13	1; 22; 23
6,14	H	6,14	1; 24; 25
16	I	16	none
17	J	17	none
18	K	18	6; 14
21	L	21	none
28	M	28	none
30	N	30	none
35	O	35	none
38	P	38	none
39	Q	39	none
40	R	40	1
41	S	41	none
42	T	42	1
43	U	43	none
44	V	44	1
45	W	45	none
47	X	47	1
48	Y	48	none
50	Z	50	none
51		51	1
52		52	none
53		53	1
54 (provisional)		54	21; 3; 3,15; 4,12; 8,20; 6,7
55		55	none
56		56	none
57		57	none
58		58	none
59		59	1
60		60	none
61		61	none
62		62	none
63		63	none
65		65	none
66		66	none
67		67	none

July 2011 Page 10 of 12

Appendix B. H (flagellar) antigens of *Salmonella*

(CDC. National *Salmonella* Surveillance Overview. Atlanta, Georgia: US Department of Health and Human Services, CDC, 2011.)

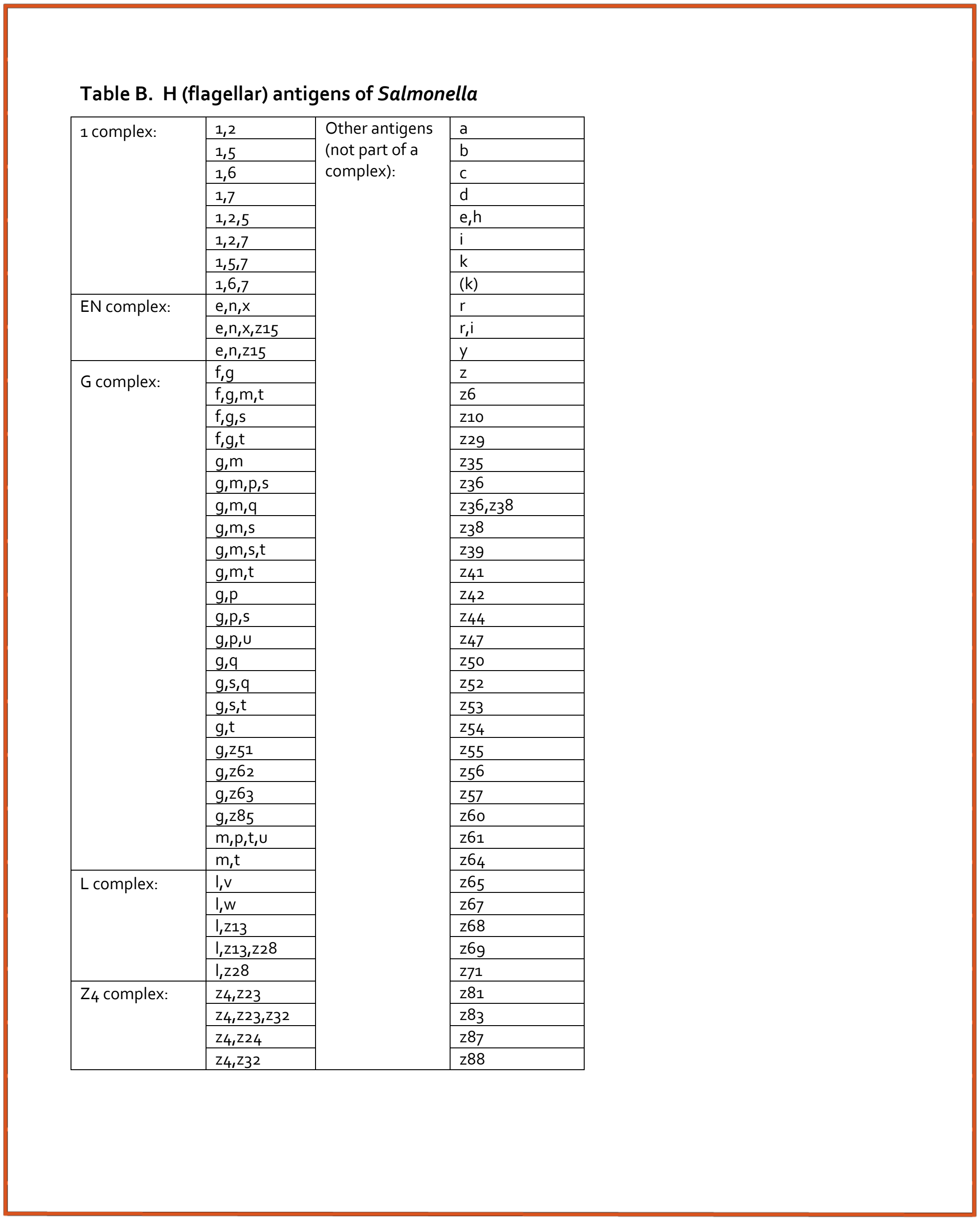


Table B. H (flagellar) antigens of *Salmonella*

1 complex:	1,2	Other antigens (not part of a complex):	a
	1,5		b
	1,6		c
	1,7		d
	1,2,5		e,h
	1,2,7		i
	1,5,7		k
	1,6,7		(k)
EN complex:	e,n,x		r
	e,n,x,z15		r,i
	e,n,z15		y
G complex:	f,g		z
	f,g,m,t		z6
	f,g,s		z10
	f,g,t		z29
	g,m		z35
	g,m,p,s		z36
	g,m,q		z36,z38
	g,m,s		z38
	g,m,s,t		z39
	g,m,t		z41
	g,p		z42
	g,p,s		z44
	g,p,u		z47
	g,q		z50
	g,s,q		z52
	g,s,t		z53
	g,t		z54
	g,z51		z55
	g,z62		z56
	g,z63		z57
	g,z85		z60
	m,p,t,u		z61
	m,t		z64
L complex:	l,v		z65
	l,w		z67
	l,z13		z68
	l,z13,z28		z69
	l,z28		z71
Z4 complex:	z4,z23		z81
	z4,z23,z32		z83
	z4,z24		z87
	z4,z32		z88

Appendix C. Creating SSDO

A primary goal of this work was to create a functional ontology that included all K-W serovars. Because there are nearly 2600 *Salmonella* serotypes, there was a strong desire on our part to automate the ontology creation process. An additional desirable outcome was to align our ontology with the content in SNOMED CT. One product of our previous work assessing the *Salmonella* content in SNOMED CT was the table created from the K-W pdf file. We also knew that a very large portion of the K-W serovars were associated with SNOMED CT codes and were also properly represented by SNOMED's "preferred terms" for the serovar classes. We saw an opportunity to leverage a conversion utility distributed with SNOMED CT RF2 release files. SNOMED distributes a Perl script that converts RF2 tables into an OWL file that can subsequently be imported directly by Protégé.

The Perl script expects to see these SNOMED CT tables: Concepts, Descriptions, Identifier, Relationships, StatedRelationships and TextDefinitions. Only Concepts, Descriptions and StatedRelationships must have content. In each required table there are fields that must be populated and fields that may or may not be. Here we demonstrate the steps required for creation of these tables using *Salmonella* Aachen as an example.

As shown in Figure X.1, the base table for these processes was a simple two column table. The first column was a text representation of the K-W serovar name (e.g., "Salmonella Aachen"), and the second column was a text representation of the K-W antigenic formula for each serovar (e.g., "17:z35:1,6"). Non-textual figures, such as parentheses, brackets, and braces were included in the formulae. The use of underscoring could not be reproduced.

Figure X.1 Base table for SSDO creation.

Name	**Formula**
Salmonella Aachen	17:z35:1,6
…	…

1. Concepts Table

Creation of the concepts table took several steps. The first processing step was an inner join query that included our K-W table and the SNOMED CT descriptions table. The query returned a table with serovar names in one column and SNOMED concept identifiers in the other. When the serovar was correctly represented in SNOMED, the query returned an SCTID. When SNOMED did not contain the serovar, the SCTID field was blank. For each serovar that did match (did not have an SCTID), a mock SCTID was created. The SCTID column became our test concept table. To it were added effectiveTime, active, moduleId and definition status. Rows were added to this table for the antecedents of the serotypes up to and including “Genus *Salmonella*” (the root concept of the resulting ontology). Rows were also added for the antigen values, the four flagella types and the attributes. These were not present in SNOMED CT and mock SCTIDs were created for them. Although the Perl script did not require values in any field except id, the tables were populated for the sake of completeness. One effective time was added to all these rows, the value for active was set to one, and definition status of all concepts was the same. A depiction is shown in Figure X.2.

Figure X.2 Concepts Table.

id	effectiveTime	active	moduleId	definitionStatusId
Antecedents of the serotype and the serotype				
27268008	20020131	1	900000000000445007	900000000000074008
110378009	20020131	1	900000000000445007	900000000000074008
398508004	20030731	1	900000000000445007	900000000000074008
114638004	20040131	1	900000000000445007	900000000000074008
Antigen values				
80259500c	20150601	1	332351000009108	900000000000074008
90277400c	20150601	1	332351000009108	900000000000074008
90265600c	20150601	1	332351000009108	900000000000074008
90265900c	20150601	1	332351000009108	900000000000074008
Flagella type values				
90265000c	20150601	1	332351000009108	900000000000074008
90265100c	20150601	1	332351000009108	900000000000074008
Attributes				
90282300c	20150601	1	332351000009108	900000000000074008
90282400c	20150601	1	332351000009108	900000000000074008

moduleId

- 900000000000445007 | International Health Terminology Standards Development Organisation maintained module
- 332351000009108 | Veterinary Terminology Services Laboratory maintained module

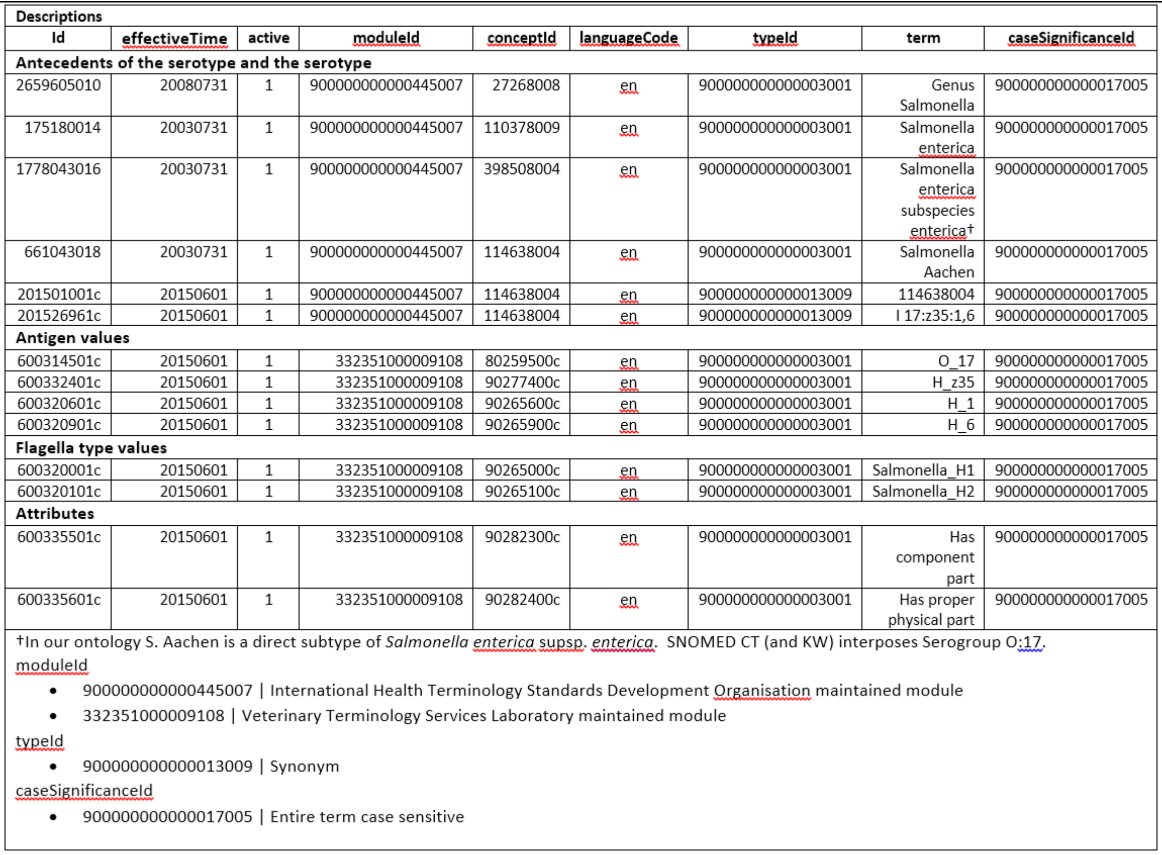
definitionStatusId

- 900000000000074008 | Necessary but not sufficient concept definition status
- 900000000000073002 | Sufficiently defined concept definition status

2. Descriptions Table

Creation processes for the descriptions table was identical to those used for the concepts table except that the term, the descriptionId and the conceptId were included in the output file. An example is depicted in Figure X.3.

Figure X.3 Descriptions Table.



Descriptions								
Id	effectiveTime	active	moduleId	conceptId	languageCode	typeId	term	caseSignificanceId
Antecedents of the serotype and the serotype								
2659605010	20080731	1	900000000000445007	27268008	en	900000000000003001	Genus Salmonella	900000000000017005
175180014	20030731	1	900000000000445007	110378009	en	900000000000003001	Salmonella enterica	900000000000017005
1778043016	20030731	1	900000000000445007	398508004	en	900000000000003001	Salmonella enterica subspecies enterica†	900000000000017005
661043018	20030731	1	900000000000445007	114638004	en	900000000000003001	Salmonella Aachen	900000000000017005
201501001c	20150601	1	900000000000445007	114638004	en	900000000000013009	114638004	900000000000017005
201526961c	20150601	1	900000000000445007	114638004	en	900000000000013009	I 17:z35:1,6	900000000000017005
Antigen values								
600314501c	20150601	1	332351000009108	80259500c	en	900000000000003001	O_17	900000000000017005
600332401c	20150601	1	332351000009108	90277400c	en	900000000000003001	H_z35	900000000000017005
600320601c	20150601	1	332351000009108	90265600c	en	900000000000003001	H_1	900000000000017005
600320901c	20150601	1	332351000009108	90265900c	en	900000000000003001	H_6	900000000000017005
Flagella type values								
600320001c	20150601	1	332351000009108	90265000c	en	900000000000003001	Salmonella_H1	900000000000017005
600320101c	20150601	1	332351000009108	90265100c	en	900000000000003001	Salmonella_H2	900000000000017005
Attributes								
600335501c	20150601	1	332351000009108	90282300c	en	900000000000003001	Has component part	900000000000017005
600335601c	20150601	1	332351000009108	90282400c	en	900000000000003001	Has proper physical part	900000000000017005

†In our ontology S. Aachen is a direct subtype of *Salmonella enterica* supsp. *enterica*. SNOMED CT (and KW) interposes Serogroup O:17.

moduleId

- 900000000000445007 | International Health Terminology Standards Development Organisation maintained module
- 332351000009108 | Veterinary Terminology Services Laboratory maintained module

typeId

- 900000000000013009 | Synonym

caseSignificanceId

- 900000000000017005 | Entire term case sensitive

3. Stated Relationships Table

The stated relationships table was made through a process of additive steps. The following Figures X.4 through X.13 represent the steps taken through the course of development while providing an explanation of each step.

Figure X.4 Table created by adding an identifier field (ID) to the existing table.

ID	**Name**	**Formula**
0001	Salmonella Aachen	17:z35:1,6
…	…	…

Figure X.5 Antigenic formulae were separated into columns for each antigen group (O, H_1, H_2, H_other) using ‘text to columns’ function of Excel. A colon was the delimiter for this step. Columns were relabeled to reflect the content they contained.

ID	**Name**	**O**	**H_1**	**H_2**	**H_other**
0001	Salmonella Aachen	17	z35	1,6	
…	…	…	…	…	…

Figure X.6 One table was created for each of the four antigen groups (H_2 shown). These tables included the identifier and name from the original table plus the antigen class column from the previous table.

ID	**Name**	**H_2**
0001	Salmonella Aachen	1,6
…	…	…

Figure X.7 Antigen lists were separated into columns for the antigen groups (O, H_1, H_2, H_other) using the “text to column” function of Excel. A comma was the delimiter for this step. Columns were relabeled to reflect the content they contained.

ID	**Name**	**H_2a**	**H_2b**	**…**
0001	Salmonella Aachen	1	6	
…	…	…	…	…

Figure X.8 Use Matrix Convert Macro[†] to convert rows to columns.

ID	Name	H_2
0001	Salmonella Aachen	1
0001	Salmonella Aachen	6
…	…	…

[†] Oboyski, Peter. (November 2013) Matrix Converter. http://nature.berkeley.edu/~oboyski67/download/MatrixConvert.txt

This Excel Add-In converts (transcribes, transforms) Microsoft Excel spreadsheet data in a matrix format to data in columns for use in relational databases or data analysis. (This is the opposite of what a pivot table does.)

Figure X.9 Create column for type and populate with "Has component part". Create column for antigen phase and populate with table antigen.

ID	Name	type	Antigen Phase	Antigen
0001	Salmonella Aachen	Has component part	H_2	1
0001	Salmonella Aachen	Has component part	H_2	6
…	…	…	…	…

Figure X.10 Merge four tables, add RoleGroup column, index the rows for each serovar to populate the RoleGroup column using function (=countif(A$1:A1,A1)) The range assertion 'A$1:A1' allows range to change with each cell. The value in the "A" cell for the row is the criteria.

ID	Name	type	Antigen phase	Antigen	Role Group
0001	Salmonella Aachen	Has component part	O	O_17	1
0001	Salmonella Aachen	Has component part	H_1	H_Z35	2
0001	Salmonella Aachen	Has component part	H_2	H_1	3
0001	Salmonella Aachen	Has component part	H_2	H_6	4
…	…	…	…	…	…

Figure X.11 Create a copy of the table, then delete the antigen phase column.

ID	Name	type	Antigen	Role Group
0001	Salmonella Aachen	Has component part	O_17	1
0001	Salmonella Aachen	Has component part	H_Z35	2
0001	Salmonella Aachen	Has component part	H_1	3
0001	Salmonella Aachen	Has component part	H_6	4
…	…	…	…	…

Figure X.12 In the duplicate table replace “Has component part” with “Has proper physical part” and delete the antigen column. Also remove any row that corresponds to an O antigen. The O antigens do not need a row that depicts the attribute “Has proper physical part”.

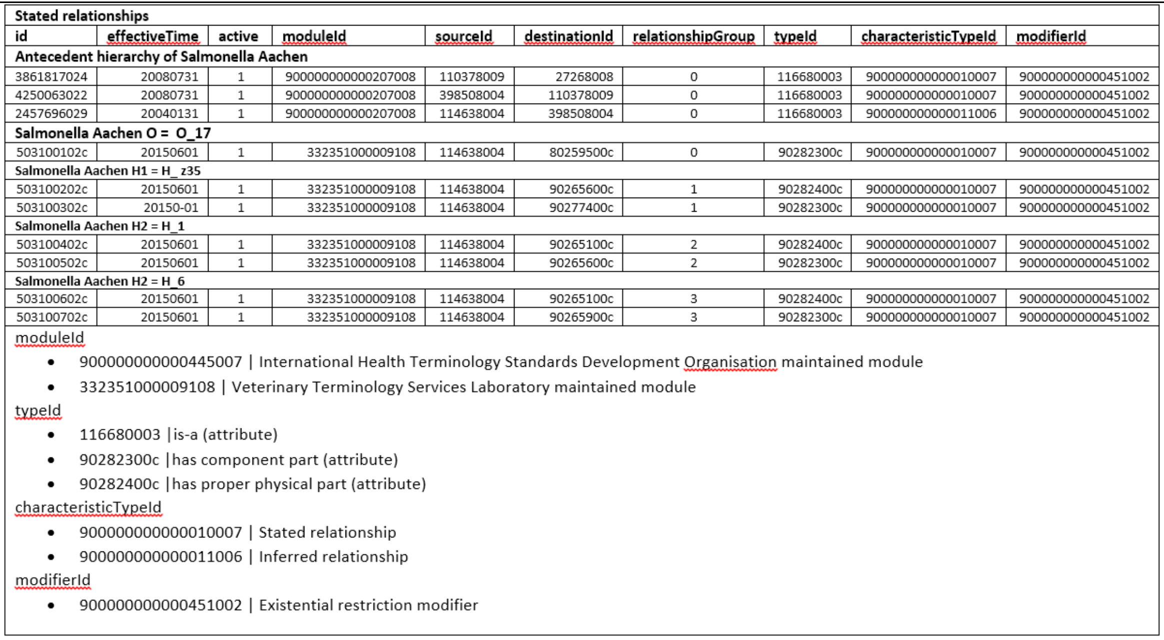
ID	**Name**	**type**	**Antigen phase**	**Role Group**
0001	Salmonella Aachen	Has proper physical part	O	1
0001	Salmonella Aachen	Has proper physical part	H_1	2
0001	Salmonella Aachen	Has proper physical part	H_2	3
0001	Salmonella Aachen	Has proper physical part	H_2	4
…	…	…	…	…

Figure X.13 Change column names to SNOMED relationship column names, create unique identifiers for each row, and substitute SCTID for terms where they exist (e.g. Salmonella Aachen = 114638004). Create them where they do not (some serovars and all attributes and values are not present in SNOMED). HCP = 90282300c, HPPP = 90282400c, Salmonella_O_Antigen = 90265600c, Salmonella_H_1 Antigen = 90265100c, O_17 = 80259500c, H_z35=90277400c, H_1=90265600c, H_6=90265900c. Note there is only one row for Role Group 1 (the O antigen row).

ID	**SourceId**	**typeId**	**destinationId**	**Role Group**
0001	114638004	90282300c	80259500c	1
0002	114638004	90282300c	90277400c	2
0003	114638004	90282300c	90265600c	3
0004	114638004	90282300c	90265900c	4
0005	114638004	90282400c	90265600c	2
0006	114638004	90282400c	90265100c	3
0007	114638004	90282400c	90265100c	4
…	…	…	…	…

Finally add columns to complete relationships table and populate with identifiers (values in these columns are not used in the ontology) to complete the Stated Relationships table as shown in Figure X.14.

Figure X.14 Stated Relationships Table.



Stated relationships

id	effectiveTime	active	moduleId	sourceId	destinationId	relationshipGroup	typeId	characteristicTypeId	modifierId
Antecedent hierarchy of Salmonella Aachen									
3861817024	20080731	1	900000000000207008	110378009	27268008	0	116680003	900000000000010007	900000000000451002
4250063022	20080731	1	900000000000207008	398508004	110378009	0	116680003	900000000000010007	900000000000451002
2457696029	20040131	1	900000000000207008	114638004	398508004	0	116680003	900000000000011006	900000000000451002
Salmonella Aachen O = O_17									
503100102c	20150601	1	332351000009108	114638004	80259500c	0	90282300c	900000000000010007	900000000000451002
Salmonella Aachen H1 = H_z35									
503100202c	20150601	1	332351000009108	114638004	90265600c	1	90282400c	900000000000010007	900000000000451002
503100302c	20150-01	1	332351000009108	114638004	90277400c	1	90282300c	900000000000010007	900000000000451002
Salmonella Aachen H2 = H_1									
503100402c	20150601	1	332351000009108	114638004	90265100c	2	90282400c	900000000000010007	900000000000451002
503100502c	20150601	1	332351000009108	114638004	90265600c	2	90282300c	900000000000010007	900000000000451002
Salmonella Aachen H2 = H_6									
503100602c	20150601	1	332351000009108	114638004	90265100c	3	90282400c	900000000000010007	900000000000451002
503100702c	20150601	1	332351000009108	114638004	90265900c	3	90282300c	900000000000010007	900000000000451002

moduleId

- 900000000000445007 | International Health Terminology Standards Development Organisation maintained module
- 332351000009108 | Veterinary Terminology Services Laboratory maintained module

typeId

- 116680003 |is-a (attribute)
- 90282300c |has component part (attribute)
- 90282400c |has proper physical part (attribute)

characteristicTypeId

- 900000000000010007 | Stated relationship
- 900000000000011006 | Inferred relationship

modifierId

- 900000000000451002 | Existential restriction modifier

4. Perl Script

We were able to automate population of the ontology by leveraging a modified Perl script written by Kent Spackman of IHTSDO. We modified the script so as align with having only *Salmonella* information displayed in the final output. This Perl script was included in the July 2014 SNOMED CT International Release. Its purpose is to take SNOMED CT specific tables and convert them into an OWL file, which can be read and modified by Protégé. The three tables needed for the conversion are Concept, Description and Relationship. These are text-based files containing all pertinent *Salmonella* information (in the form of SCTIDs) that were developed for this project using a combination of Microsoft Excel and Access. For those data elements that did not already have a SNOMED assigned corresponding SCTID (e.g. individual antigens, serotypes not found within SNOMED, etc.), a collision-proof dummy SCTID was created and used.

Appendix D. Copyright Permission Letter

Page 51 (Figure 8)

Lolis, E., & Bucala, R. (2003, August). Therapeutic approaches to innate immunity: severe sepsis and septic shock. *Nature Reviews Drug Discovery, 2*, 635-645.

Adapted by permission from Macmillan Publishers Ltd: *Nature Reviews Drug Discovery* Lolis & Bucala, Copyright 2003; letter attached.

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Jun 24, 2015

This is a License Agreement between Jeffry C Alexander ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3655440525548
License date	Jun 24, 2015
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Drug Discovery
Licensed content title	Therapeutic approaches to innate immunity: severe sepsis and septic shock
Licensed content author	Elias Lolis and Richard Bucala
Licensed content date	Aug 1, 2003
Volume number	2
Issue number	8
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no

Figures	Figure 1 | Structural features of the cell wall that distinguishes the Gram-positive from the Gram-negative bacteria, which are two principal classes of pathogenic bacteria.
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Considerations for Automating Salmonella Serovar Identification within the Electronic Public Health Reporting Environment
Expected completion date	Aug 2015
Estimated size (number of pages)	140
Total	0.00 USD
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the British Journal of Cancer, the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK:
   [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the British Journal of Cancer, the following credit line applies.
   Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information. Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:
   Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

   Note: For translation from the British Journal of Cancer, the following credit line applies.
   Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK:[JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.