Physics Subject Headings (PhySH)†

Arthur Smith

American Physical Society, 1 Research Rd, Ridge, NY 11961,
<apsmith@aps.org>

Arthur Smith is Lead Data Analyst at the American Physical Society. He has worked in information technology roles for the *Physical Review* journals since 1995. He received a PhD in physics from Cornell University in 1991. He has given talks at Taxonomy Boot Camp (associated with the KM World conference), at the PIDapalooza meetings on persistent identifiers, and at a variety of related conferences.

Smith, Arthur. 2020. "Physics Subject Headings (PhySH)." *Knowledge Organization* 47(3): 257-266. 14 references. DOI:10.5771/0943-7444-2020-3-257.

Abstract: PhySH (Physics Subject Headings) was developed by the American Physical Society and first used in 2016 as a faceted hierarchical controlled vocabulary for physics, with some basic terms from related fields. It was developed mainly for the purpose of associating subjects with papers submitted to and published in the *Physical Review* family of journals. The scheme is organized at the top level with a two-dimensional classification, with one dimension (labeled "disciplines") representing professional divisions within physics, and the other dimension (labeled "facets") providing a conceptual partitioning of terms. PhySH was preceded in use by PACS ("Physics and Astronomy Classification Scheme"), which was in turn preceded by more ad hoc approaches, and this history and related vocabularies or categorizations will also be briefly discussed.

Received: 9 September 2019; Revised: 10 October 2019; Accepted: 31 October 2019

Keywords: physics, Physics Subject Headings (PhySH), concepts, classification, Phsyics and Astronomy Classification Scheme (PACS)

† Derived from the article of similar title in the *ISKO Encyclopedia of Knowledge Organization*, version 1.0, published 2019-07-31 Article category: KOS, specific (domain specific).

1.0 Introduction

The study of the natural world has been going on since almost the earliest recorded history, but "physics" as a distinct field, as with most of the other major specializations in science, dates to around the turn of the nineteenth century. The entities studied by physicists range from the simple to the complex and from real-world objects to highly abstract mathematical theories. Physicists try to understand physical systems through detailed observations, reductionist analysis, and mathematical modeling, with a general (though far from uniform) division of labor between those doing experiments with real systems and those working on theories to explain or predict observational phenomena.

As a significant area of scientific research, physics has received some attention in general knowledge organization or subject classification schemes: section 530 of the *Dewey Decimal*, Melvil Decimal, and Universal Decimal Classifications (UDC), subclass QC in the Library of Congress *Classification*, class C in Ranganathan's *Colon Classification* (*CC*) (Satija 2017), and class B in the *Bliss Bibliographic Classification*. These all allow for somewhat detailed subdivisions of the field, but (with the exception of Bliss, discussed later) their main subcategories for physics tend to focus on areas that were of interest 100 years ago, which have generally not been major research fields for many decades. Mechanics, for example, is a relatively minor area of current physics research, but occupies sections 531 to 534 in the UDC. Condensed matter physics, which accounts for close to half of current basic physics research is confined to 538.9. Elementary particle or high energy physics is not even mentioned in the UDC summary table (http://www.udcsummary.info); it is buried deeper in 539.1 which has the label "Nuclear physics. Atomic physics. Molecular physics." These classifications are sufficient for the relatively small number of books published in the field, but are not nearly comprehensive enough to usefully group the millions of scientific papers. Instead, much more detailed systems for knowledge organization have been developed by physicists themselves and particularly by the organizations publishing and providing indexes to physics research in recent decades.

This article describes the development and structure of PhySH ("Physics Subject Headings"), a recently developed controlled vocabulary for physics from the American Physical Society. It also covers some other classification systems from related fields or previously used in physics.

2.0 History of knowledge organization in physics

2.1 Physics classification in the nineteenth and early twentieth centuries

As with much of the science literature, publications in physics have grown exponentially since at least the 1950s. Forman et al (1975) found the number of articles published annually in the field in the late 1800s was on the order of 2000, with Germany the leading nation for physics research at the time. The size of the English-language literature in the field was less than half that total. The new *Physical Review* journal begun at Cornell University in 1893 published only sixty-three papers in the year 1900, several of which were book reviews.

Even with a much smaller literature size than at present, abstracting and indexing was found to be useful: in English, *Science Abstracts* started in 1898 (IET 1998) and included a subject index from the first year. By 1902 *Science Abstracts* had split into a part A (Physics) and B (Electrical Engineering), and after 1941 part A was simply known as *Physics Abstracts*; the publications formed the basis for the computer-based INSPEC (Information Service in Physics, Electrotechnology and Control) service in 1967, still with us today.

The *Physical Review* also had its own end-of-volume indexes before 1900, but these were initially just alphabetized lists of author names and words pulled from article titles. In 1923, the author index and an "Analytic Subject Index" were separated; the latter was a flat alphabetized list of less than 100 classifying terms across all of physics, with a list of matching articles provided under each term.

2.2 Physical Review hierarchical index

By 1964 the *Physical Review* had grown to about the size of the entire worldwide physics literature of the year 1900 with 1873 published articles. The end-of-volume alphabetic subject lists seemed no longer sufficient, and a new hierarchical classification with decimal notation was introduced by the editors. The ordering and hierarchy imposed by the notation had a logic that physicists would appreciate, with higher numbers corresponding generally to higher energy phenomena. At the bottom was 0 (for general physics), and at the top 60 (for particles and fields). Individual terms had whole number or one-digit-after-the-decimal codes, for example "classical mechanics" was 2 (under "0 - general"), "magnetohydrodynamics" was 33.4 (under "33—plasma physics" which was part of "30—physics of fluids"), 54.9 was "nuclear fission" (under "54—nuclear reactions and scattering" and "50—physics of nuclei"), etc. There was also a special classification code above the top subject-related code—70—for errata (articles published as corrections to previously published articles). The decimal format resembled and may have been inspired by the Mathematics Subject Classification codes which had used a two-digit + one or two-digit + two decimal notation starting in 1940 (https://mathscinet.ams.org/mathscinet/help/field_help.html#mscp).

This subject notation system soon had a profound effect directly on the structure of the journal. The 3,377 articles published in 1970 were split among four new journals: *Physical Review A* which covered the topics under headings 0, 10, 20 and 30 (general, atomic, molecular, optical and fluid physics and related topics), *Physical Review B* had the topics under 40 (solid state physics), *Physical Review C* had 50 (nuclear physics), and *Physical Review D* had 60 (particles and fields). These divisions remain today, although in 1995 *Physical Review A* was further split, adding a new journal *Physical Review E* to cover some of the general, interdisciplinary, and fluid physics topics. *Physical Review Letters*, a weekly publication for short high-impact papers that had been started in 1958, also in 1970 started sorting articles according to this decimal notation with general physics papers at the front and particle physics papers at the back of each issue.

2.3 Physics and Astronomy Classification Scheme (PACS)

Since the problem of classification in physics was not unique to the *Physical Review* journals published by the American Physical Society, in the early 1970s several organizations got together to develop a more uniform system, which became the Physics and Astronomy Classification Scheme (PACS). The process was described in an editorial by Krumhansl and Trigg in April 1975 in *Physical Review Letters*:

> Several years ago, negotiations began between the American Institute of Physics (AIP), the major publishers of physics literature in English, and the Institution of Electrical Engineers (IEE), publishers of the principal English-language abstract journal, *Physics Abstracts*. Eventually other discussants were brought in, and the whole business was brought under the aegis of the Abstracting Board of the International Council of Scientific Unions (ICSU-AB) ... this scheme has been accepted by ICSU-AB, the AIP, and the IEE, as well as some non-English abstracting journals ... Reflecting this change, the grouping of papers in *Physical Review Letters* will also be changed to incorporate the PACS headings.

PACS outwardly resembled the Physical Review Hierarchical Index with hierarchy organized using a decimal notation that at the top level went from 00 to 99, with the ten top headings (00, 10, 20, etc.) representing the major subject areas. However, it was internally significantly more detailed, having two digits after the first ".", and then another "." character followed by a final "+" (indicating no further hierarchy) or "-" (indicating there were additional terms at a lower level), or an upper-case letter for those lower-level terms. All of this was followed by a final lower-case letter used as a checksum (to limit errors in manual retyping of the codes). In the final few editions of PACS, an additional layer in the hierarchy was added by abandoning that final checksum letter, using a '-' character to indicate there were sublevels, and then the lower-level codes were added with a lower-case version of the first letter, and a final lower-case letter that was simply incremented to indicate different sub-classifications.

PACS also largely reversed the orientation of the earlier index, ordering roughly by the distance scale involved rather than by energy, so that "10" now represented the highest energy "elementary particles and fields," "20" was now "nuclear physics," while the burgeoning field of condensed matter physics took both the "60" and "70" sections. As suggested by Krumhansl and Trigg in their July 1975 editorial, this necessitated a substantial reordering of the *Physical Review Letters* table of contents. Those top-level groupings for physics defined by PACS have had a much broader impact as they have been widely used in other classification and related analysis work on physics, even very recently. For example, Desale and Kumbhar (2017) base their "first order array divisions" (Table 3.4) largely on those top-level PACS codes. Radicchi and Castellano (2011) in their analysis of citation patterns in physics similarly subdivide the analysis according to that top-level PACS code.

Some examples of the PACS codes included "52.30.+r Plasma flow; magnetohydrodynamics" (under "52. The physics of plasmas and electric discharges," and "50. Fluids, Plasmas and Electric discharges"); "24.80.+y Fission" (under "24. Nuclear reactions and scattering, general," and "20. Nuclear physics"), and "72.10.Jp Thermoelectric effects" (under "72.10.-d Electronic conduction in metals and alloys," "72. Electronic transport in condensed matter," and "70. Condensed matter: electronic structure, electric, magnetic, and optical properties").

PACS changed substantially over time. There were at least twenty-three distinct versions from 1975 to 2010, growing from about 1,600 to almost 5,300 categories (some were also deleted on the way). The most dramatic year-to-year change was from 1976 to 1977, when there was a slight reordering and about 30% more terms were added at once.

As a classification scheme, PACS was designed to allow a single code to characterize the full scope of a work, so the "leaf nodes" or narrowest classes of the scheme tended to combine multiple aspects or facets of the subject at hand, leading to some combinatorial complexity. As a result, the labels were often not unique, with meaning determined not just by the label attached to the code in the scheme but to the hierarchy in which it sat. For example, the 1975 scheme had a second entry with exactly the same "thermoelectric effects" label with code "72.20.Pa." which placed it under "72.20.-i Conductivity phenomena in semiconductors and insulators"—i.e., the same phenomenon but in a different sort of system from the metals and alloys of the other code. PACS was not consistent in whether it was the system or the phenomenon or some other aspect that provided the finest layer of detail within a class; in contrast to the thermoelectric example we can look at the case of fullerenes, first introduced into the scheme with the 1993 edition, and which appeared in ten different classifications by the 2010 edition—see Table 1.

Code	Label	Parent
61.48.-c	Structure of fullerenes and related hollow and planar molecular structures	61. Structure of solids and liquids...
68.35.bp	Fullerenes	68.35.B- Structure of clean surfaces
68.55.ap	Fullerenes	68.55.A- Nucleation and growth
71.20.Tx	Fullerenes and related materials; intercalation compounds	71.20.-b Electron density of states and band structure of crystalline solids
72.80.Rj	Fullerenes and related materials	72.80.-r Conductivity of specific materials
73.61.Wp	Fullerenes and related materials	73.61.-r Electrical properties of specific thin films
78.30.Na	Fullerenes and related materials	78.30.-j Infrared and Raman spectra
78.40.Ri	Fullerenes and related materials	78.40.-q Absorption and reflection spectra: visible and ultraviolet
78.66.Tr	Fullerenes and related materials	78.66.-w Optical properties of specific thin films
81.05.ub	Fullerenes and related materials	81.05.U- Carbon/carbon-based materials (part of 81.05.-t Specific materials: fabrication, treatment, testing, and analysis)

Table 1. Fullerenes in the 2010 PACS edition.

This sort of complexity made PACS difficult for ordinary physicists to use; few practicing researchers could remember even a few of the codes relevant to their research. Full-text search and other electronic capabilities by the early 2000s had made some of the other purposes of the classification no longer relevant, so the owners of the classification at the American Institute of Physics (AIP) decided to make 2010 the last edition of PACS.

2.4 Other related vocabularies and classification schemes

The *BlissBibliographic Classification*, one of the general knowledge classification schemes mentioned in the introduction (and widely used by UK libraries), has been undergoing a major revision since 1969 (Mills and Broughton 1977) with the new version denoted *BC2*. Both the original (*BC1*) and new classifications have a prominent physics class with the physics schedule for *BC*2 published in 1999 (Bliss Classification Association 1999). The Bliss classification interestingly uses the same terminology of "disciplines" and "facets" that was (apparently independently) adopted by PhySH, though in *BC2* the disciplines are only the highest-level classes (such as "physics" itself), and not narrower subject areas. *BC2* facets are organized with a technique of "inversion" and "retroactive compounding" so that more general groupings appear first, and there are many detailed ordering rules. The notation is strictly ordered but does not reflect the hierarchy, which is noted separately. *BC2* has nine main facets in its "standard citation order," of which the main ones used in physics are "operations & agents of operations" (methodology and techniques), "processes & properties," "parts," and "types;" for example under particle physics ("BM") there are elementary particle types listed such as leptons ("BNM") and quarks ("BNR"). However, many topics of current interest in physics research are not to be found in the *BC2* physics class: examples are "quantum information," "nonlinear systems," "optical trapping," "graphene," and "nanoscopics" or "mesoscopics" generally. "Fullerenes" are not listed under physics but can be found in the chemistry ("C") section with the rather lengthy notation "CGF LMG JQU."

Similar to PACS, the purpose of *BC2* as a classification where items can be placed in a logical linear order results in the same subdivisions appearing among a variety of higher-level topics. With *BC2* the "filing order" does appear to be more consistent and useful than the one chosen by PACS, but it leads to a similar combinatorial complexity of the scheme. The full *BC2* physics schedule (Bliss Classification Association 1999) comes to eighty-nine pages with 5,231 classifications.

The Institute of Electrical Engineers (IEE, now the "Institute of Engineering and Technology," IET) was one of the organizations that helped develop PACS, but soon diverged with what is now their proprietary Inspec classification (IET 1998), which combines a modified version of the PACS codes with additional keywords, providing a controlled vocabulary that their indexers could use to help retrieval of relevant literature. AIP replaced PACS for their own purposes with an internally developed thesaurus with the primary goal of automated classification. This meant a rules-based system to look at abstracts or full text of articles and pick out terms or concepts that were likely to apply.

In closely related fields to physics, while there is conceptual overlap, the focus is often very different. In astronomy (which was nominally covered by PACS in the "90" section) there are a number of different controlled vocabularies along different conceptual dimensions, exemplified by the faceted search system of the Astrophysics Data System (https://ui.adsabs.harvard.edu); one aspect of that is the identifiers for specific celestial objects (a planet, star, galaxy), but specific objects or locations are not usually important to physics outside of astrophysics and geophysics.

The Mathematics Subject Classification (MSC) codes have been previously mentioned; these have always included many concepts from physics and some physics journals historically used these codes rather than PACS (which they resemble). Examples of physics terms from MSC are 81T55 ("casimir effect"), 76W05 ("magnetohydrodynamics and electrohydrodynamics"), or 82D55 ("superconductors").

Chemists have established a number of different taxonomies of molecules and compounds and other chemical systems, including the International Union of Pure and Applied Chemistry International Chemical Identifier (InCHI) and the Chemical Entities of Biological Interest" (ChEBI). In the life sciences there has been a proliferation of persistent identifiers for concepts, for example with the Identifiers.org database (https://registry.identifiers.org/registry) which allows a particular protein, gene, or other item of interest to be identified with a "prefix: identifier" notation, enabling simpler indexing and searching. The National Library of Medicine's "Medical Subject Headings" (https://www.nlm.nih.gov/mesh/) is a widely used curated vocabulary of biomedical terms.

2.5 Development of PhySH

PhySH had its origins in a November 2011 workshop in Boston (https://sites.google.com/site/physicsclassification 2011/) where the participants discussed what should follow PACS for classification and knowledge organization in physics, as described by Smith (2019). Two basic models stood out: centrally organized and comprehensive classification systems such as what the National Library of Medicine did in the life sciences, or more independent and interoperable "vocabularies," often relying on the Simple

Knowledge Organization System (SKOS) (Miles and Bechhofer 2009) design where each concept has a unique Uniform Resource Identifier (URI). There was also much discussion of hierarchy (Smith 2019); the organization of parent-child and other relations depends in large part on the purpose for which the vocabulary is being created, so good clear conceptual terms should be a first priority, with hierarchy a secondary component.

The purpose that the *Physical Review* journals had in mind was somewhat different than the concerns of other parties who had used PACS, according to Smith (2019). Assigning submitted articles to editors with the right expertise, grouping related articles together when published, and assisting editors in finding referees were viewed as more important than indexing automation or improving search capabilities for end-users. So, the new scheme needed to be relatively easy for authors to assign to their papers from the start, and for editors to check and correct if needed. In part, that meant the new scheme needed to be openly available to anyone (not proprietary); it was also hoped that the scheme would be easier to use and simpler (ideally smaller) than PACS.

Work on the new vocabulary began in earnest at the American Physical Society in late 2013. The idea of grouping the terms with a faceted structure (such as "physical systems" or "techniques") and filtering based on major research areas of physics (such as "nuclear physics" or "condensed matter") allowed a reasonable partitioning of the vocabulary so that different groups, including editors and outside consultants, could work on pieces relatively independently. The work was substantially complete by the first half of 2015, at which point serious internal testing began, and after some more feedback the APS started using the new system in late 2015 (with initially only editors tagging papers). There was then an "unveiling" in January 2016 (Conover 2016) and the *Physical Review* journals started requiring authors to supply PhySH terms in the first half of that year. Shortly after that point, PhySH had completely replaced PACS in handling of submitted manuscripts.

3.0 PhySH structure and usage

3.1 PhySH Concepts

PhySH uses an adapted version of the SKOS model (Miles and Bechhofer 2009) for controlled vocabularies and thesauri, so that every one of the 3,079 assignable terms in PhySH is a "concept" with a unique preferred label and a stable identifying URI. Having a URI identifier allows the label to be modified without any worry about losing assignments or relationships. Concepts may also have alternate labels to allow easier lookup when there are several different words that express the same thing or when a concept is intended to include several slightly different specializations. The URIs for the concepts in PhySH are actually Digital Object Identifiers of the form:

https://doi.org/10.29172/<id>

where <id> is an otherwise meaningless string consisting of the letters "a"-"f," digits, and the "-" character. As a specific example, "fullerenes" has the URI:

https://doi.org/10.29172/b755f66bb30d4ec1a7a12e31e5f675cb

The hierarchy within the realm of concepts is indicated using the standard "skos:broader" and "skos:narrower" relations, with "skos:related" also used for terms that are related but not in a parent-child manner. Note that the SKOS model allows arbitrary depth of hierarchy and multiple parents for any concept; rather than being a simple tree, the hierarchy is described as a directed acyclic graph, with a partial ordering from the broadest to the narrowest concepts. A concept with multiple parents is viewed as having a single meaning—it does not matter through what path the concept was located, it should mean the same thing. In general, PhySH concept labels are intended to be clear, unambiguous, and independent of their parent or sibling terms.

There are a number of custom RDF predicates that PhySH uses to indicate the special relations between concepts and the "disciplines" and "facets" they are assigned to. There is also a special predicate used to identify deprecated concepts. These are concepts that may be duplicative or otherwise considered no longer relevant and are expected to be deleted in a future update. Deprecated concepts are not included in the statistics provided here.

3.2 PhySH disciplines

The seventeen PhySH disciplines (see Table 2) are also identified by the same kind of URIs with preferred labels, but within the SKOS framework each discipline is treated as a "concept scheme" rather than a regular "concept." Most of the disciplines have a relatively manageable number of concepts of at most a few hundred, but several are quite large. Condensed matter physics has historically been the largest coherent subfield, and as noted earlier was covered by two top-level PACS codes from the start of the PACS scheme in 1975. Some of the disciplines correspond to the other top-level PACS codes ("10" in PACS is "particles & fields" in PhySH for example), but about half are new to PhySH: "accelerators & Beams" and "physics education research" correspond to relatively new specialty journals in the *Physical Review* family, while "quantum information" and some of the others represent distinctive areas of recent growth in research activity.

PhySH Discipline	Number of concepts
Accelerators & Beams	103
Atomic, Molecular & Optical	387
Biological Physics	672
Condensed Matter & Materials Physics	1139
Fluid Dynamics	145
General Physics	133
Gravitation, Cosmology & Astrophysics	104
Interdisciplinary Physics	61
Networks	185
Nonlinear Dynamics	52
Nuclear Physics	134
Particles & Fields	239
Physics Education Research	19
Plasma Physics	338
Polymers & Soft Matter	862
Quantum Information	43
Statistical Physics	549

Table 2. The PhySH disciplines, as of version 1.1.1—these counts only include concepts directly listed in the discipline and their narrower terms.

PhySH facet label	Number of disciplines	Number of concepts
Physical Systems	13	846
Professional Topics	2	12
Properties	4	53
Research Areas	17	1622
Techniques	13	752
– Computational Techniques	9	52
– Experimental Techniques	11	381
– Theoretical & Computational Techniques	2	226
– Theoretical Techniques	10	222

Table 3. The PhySH facets.

The sum of the concept counts in Table 2 (5,165) is considerably more than the total number of concepts in this version of PhySH (3,079). This is because the disciplines do not in themselves constitute an exclusive partitioning of the conceptsthere is considerable overlap. The physical systems studied in one discipline may be studied in others, and techniques are even more widely shared.

3.3 PhySH facets

The facets in PhySH (Table 3) are similarly identified by URIs and preferred labels. However, as common cross-cutting classifications across all the disciplines (SKOS concept schemes), they do not strictly fit within the SKOS framework at all. A SKOS-compatible version of PhySH is provided in which discipline-facet pairs serve as the "top concepts" for each discipline. For example, "nuclear physics research areas" and "nuclear physics techniques" are top concepts within the "nuclear physics" concept scheme, while "fluid dynamics research areas" and "fluid dynamics techniques" are similar top concepts within "fluid dynamics." These pairs are not strictly part of PhySH itself, rather PhySH uses custom RDF classes and predicates (separate from the SKOS definitions) to define the facets and their relationships to disciplines and concepts.

The "techniques" facet has been divided into subfacets for the major classes of techniques; in the earliest versions of PhySH there was a three-fold split into "computational," "experimental," and "theoretical." The fourth subfacet ("theoretical & computational") was a partial merger created after finding significant overlap between the "computational" and "theoretical" subfacets. Aside from these subfacets, at the top level the PhySH facets do almost completely partition the concepts, with the total (using the 752 for "techniques" as a whole, rather than the numbers from its subfacets) amounting to 3,285, relative to the 3,079 dis-

tinct concepts. In other words, there is less than 10% overlap of concepts between the top-level facets in PhySH.

The subfacet organization is probably unnecessarily complex. A better organization (perhaps to be considered for a future PhySH release) would replace the "techniques" facet and its subfacets with two top-level facets, one for "experimental," and one for the merged "theoretical & computational." These two subfacets have no overlap in the current version and seem to be sufficient to satisfy the requirements of classification.

The facets in PhySH are in principle similar to the "PMEST" facets of Ranganathan (Satija 2017) but in practice somewhat distinct and perhaps not as generalizable. Desale and Kumbhar (2017) in their chapter five explicitly use Ranganathan's approach and assign many physics concepts to the "P," "M," and "E" facets; these assignments partially align with the PhySH facets, so that "E" seems to be closely related to PhySH "techniques," "M" to "properties," and "P" involves both "research areas" and "physical systems." This is similar to the facets of the Bliss classification, at least with the BC2 "operations & agents of operations" corresponding to PhySH techniques. BC2's "processes & properties" overlap more with PhySH "research areas" (and also "properties"), and "parts" and "types" correspond more to the "physical systems" facet in PhySH. So, these two distinct library-oriented approaches to a faceted classification of physics show considerable overlap, but also some divergence from what was done with PhySH. Note that all of these approaches avoid use of anything resembling Ranganathan's "S" and "T" facets in physics as physical concepts are supposed to be applicable universally in space and time.

3.4 Using PhySH

PhySH with its labeled concepts should be easier for regular physicists to understand and remember than the alphanumeric codes used by PACS. The disciplines, facets, and hierarchy as a whole are also designed to make browsing fruitful, but textual searching is easy enough and more frequently used in practice. However, it should be noted that more PhySH terms are typically needed to characterize a scientific article than was true for PACS.

PACS was indeed a classification (Hjørland 2017). That is, in principle there was one best code for each indexed document. That was why different facets were combined, such as with the fullerene examples in Table 1. For example, 73.61.Wp on electrical properties of thin-film materials containing fullerenes and related materials combines the concepts of "electrical properties," "fullerenes" and "thin-film materials." In PhySH, each of those distinct concepts has their own separate entry: "electrical properties," "fullerenes," and "thin films." All three (plus anything else applicable) would need to be assigned to the associated document using PhySH, to provide the same level of characterization of the research. But note that both "electrical properties" and "thin films" have refinements—narrower terms that could be more applicable, such as "bilayer thin films." The PhySH approach allows for more precision on what effect is being studied, what techniques are being used, or what systems being considered—but it is at the cost of requiring multiple terms to characterize the research.

Note also that it is accepted and expected for a concept at (almost) any level of the hierarchy to be used in assignments, whereas with classifications one usually only assigns codes from the lowest possible level. There are no concepts labeled as "other xxx" in PhySH; in PACS there were many such as "61.43.Er Other amorphous solids," catchall categories whose meaning changed over time (narrowing as more subclasses were added to the higher "disordered solids" class). With PhySH one just uses the parent term (in PhySH it is "amorphous materials"), which encompasses all the various types.

One notable difference between PACS and PhySH is in their ability to define a sorting order for sibling terms (concepts with the same "broader" or parent concept). With an alphanumeric notation as used by PACS, the notation determines the order, and so if some particular order made sense, it was possible to enforce it through the notation. For PhySH the underlying identifiers are intended to be meaningless strings, so the only natural order is that of alphabetization of the preferred labels for concepts. For the most part, the order is not important, but in a few exceptions judicious selection of labels helps to establish a preferred order. One example is within the "physical Systems" facet, "0-dimensional systems," "1-dimensional systems" etc. are naturally ordered, but would not be so if the verbal instead of the numeric form were used as the starting word.

With SKOS there is no enforced limit to hierarchy depth or to the number of siblings at one level in the hierarchy. PhySH has a few concepts with paths as deep as nine levels (starting from the disciplines as level one, the discipline-facet pairs as level two, and so on). The vast majority of concepts are found at levels four and five (two or three below the discipline-facet usually set for filtering). Some concepts have twenty or more direct "children" but usually the count is much less.

PhySH includes a number of terms like "fullerenes" that refer to specific materials or other specific physical systems, but only those with a high degree of interest to physicists. It was anticipated that more detailed identifiers for specific physical systems—for example, a comprehensive list of elements, isotopes, crystallographic structures, astronomical objects, etc. would be better handled by outside vocabularies built with those domains in mind. So far PhySH has not been combined with these other vocabularies in practice, but the ability is there in principle and should be relatively

straightforward with concepts identified by their unique URI's as is the practice with SKOS.

4.0 Governance, updates, and impact

PhySH is publicly available from a primary website at physh.org and also in a variety of downloadable formats through github (https://github.com/physh-org/PhySH/). There have been a number of changes to PhySH since its first use in 2015-2016, including the first public release in 2018 (version 1.0); the latest version as of this writing is 1.1.1. It remains owned and governed by the American Physical Society, but it is available for any other person or organization to use under the Creative Commons CC-0 1.0 license (CC0). Suggestions for improvement are welcome and may be submitted through the github site as github "issues." The APS also receives suggestions from its authors and editors through online forms linked to the manuscript handling system. Hundreds of suggestions have been acted upon, with many new terms added and over one hundred concepts in the latest version marked as "deprecated." Guidelines for suggestions are provided including when a new concept may be added, types of changes allowed for existing concepts, rules for concept labels, and the possibility of structural changes.

An internal group within APS reviews these change requests, generally on a discipline-by-discipline basis. In 2018, the condensed matter & materials and fluid dynamics disciplines were reviewed, and some additional minor changes were made in a few of the other disciplines.

So far, the only significant publicly known user of PhySH has been the American Physical Society which developed it, and has been using it since 2016 to index all articles published in the *Physical Review* journals, about 70,000 so far. APS has previously provided the PACS indexing data of published articles for use by researchers (for example Radicchi and Castellano (2011) used this in their analysis), and it is expected similar research may be conducted with the PhySH terms in future.

5.0 Evaluation and considerations for the future

The type of ordering imposed by PACS and the more general classification schemes such as UDC, *BC2*, or *CC* was necessary when our knowledge was primarily to be found in paper documents stored on library shelves. The general idea was to file items on similar topics together, and to file more general items before more specific ones if possible. The complexities of assigning one linear position to a document that covered multiple topics is exemplified in the "filing order" and consequent "citation order" specifications for the Bliss classification. Many of the detailed classes in PACS or *BC2* are the result of combining two, three, or even more elementary concepts from several different facets, or from different arrays in *BC2* terminology (logical divisions) within a single facet. This enables grouping of documents on several different dimensions, but it is preferential to only one such possible grouping which the classification designers had some reason to believe would be most helpful.

In the paper-based world, the problem of arbitrariness in grouping was handled through indexes which allow alternate paths to find relevant documents. With electronic documents there is no longer any need to maintain a linear sequence and any grouping can be created dynamically. In the *BC2* introduction (Mills and Broughton 1977), there is a relevant discussion in section 4.8:

> Virtually all problems of information indexing are problems caused by compound classes. In conventional indexing, the linking (intersection, coordination) of elementary terms to form compound classes is done at the time of indexing. In coordinate indexing, this is done only after receipt of a request. When the question is received then the search is made for the particular combination of elementary terms making up the search prescription. So it may be said that in conventional indexing coordination to form compound classes is done before (pre-) receipt of any particular request whereas in coordinate indexing it is done only after (post-) receipt of a request.

The label "concept" for the elementary ingredients of a SKOS vocabulary suggests that these should always be considered as elementary terms, to be combined through such a post-coordination process. PhySH is indeed largely designed this way so that the individual concepts are mostly elementary terms and are intended to be combined dynamically in searches, but even so there are many apparently compound terms in the vocabulary. Some of these, for example "atomic & molecular processes in external fields," are justified as providing a logical grouping for a large number of more specific elementary concepts such as "photoemission" or the "stark effect." "Photon & charged-lepton interactions with hadrons" on the other hand has no narrower terms below it, and perhaps could have been better represented by post-coordinating the two or three elementary concepts involved. Nevertheless, over fifty articles have been published in the *Physical Review* journals and indexed with this PhySH concept, so it may be justified as a useful precoordinated term.

This brings us to the issue of "literary warrant," discussed by Barité (2018). A term that specifies over fifty relevant documents deserves to be included in the vocabulary. The guidelines for contributions (American Physical Society 2018) specify criteria relating to a literary warrant for new terms:

> Please consider if the concept you're about to suggest is really needed. If a closely related concept already exists, would adding the new proposal as an alias suffice? Adding a new alias to an existing concept is preferred over adding a new concept .... A new concept may be needed where a significant body of work (dozens of papers per year, say) is associated with it, and not distinguished by any existing concept in PhySH.

Whether these criteria were applied in the creation of PhySH in the first place may be questioned. Over 3,000 of the 3,079 concepts have been used so far on published articles, leaving about 2.5% of PhySH concepts that have not been used at all. Almost all of these unused concepts (some examples are "leaves," "kinesiology," and "atactic polymers") come from the "biological physics" or "polymers and soft matter" sections of PhySH. These terms may eventually be justified by articles published in other journals in the field so that the reason for no articles is simply that the *Physical Review* journals publish relatively few articles in these areas. But at least for now the literary warrant of some of the terms is questionable. Note that another roughly 300 (about 10%) have been used less than five times, while about half the remainder have been used fifty or more times.

The concentration of unused terms in two disciplines also suggests examining other potential inconsistencies between the disciplines in PhySH. The issue of diverging subfacets has already been mentioned (some disciplines have separate "computational techniques" and "theoretical techniques" groupings, while others use a combined subfacet). There is also quite inconsistent treatment of the "properties" facet, which is so far not used by many of the disciplines. For example, "symmetries" in the "properties" facet of the "particles & fields" discipline is closely related to "symmetries in condensed matter," but that has been placed under the "techniques" facet in "condensed matter & materials physics." The "research areas" facet in that discipline contains several top-level terms that actually use the word "properties:" "electrical properties," "structural properties," etc. It is not clear whether this inconsistent placement is deliberate or simply a consequence of different people making different decisions about use of the facets. Perhaps some of these issues will be improved upon in future releases of PhySH.

Smith (2019) discusses the potential impact of "artificial intelligence" on the need for classification and indexing systems like PhySH and counters that rather, the manual work of classification is an essential ingredient to successful automation in scientific research. PhySH was specifically designed to be easy for those with field expertise (the authors and editors) rather than classification experts or automated tools to assign useful indexing concepts to articles. With the scientific literature continuing to grow beyond the capacity of any individual to perceive all of it, good indexing and classification will continue to be important long into the future.

6.0 Conclusions

Indexing and classification has a long history in the field of physics. For scientific articles it was largely ad hoc until the advent of PACS in the 1970s, as a classification scheme with decimal/alphanumeric notation. PACS established a widely used standard for both classification and ordering of the subfields of physics for several decades. However, the rigidity and complexity of PACS was not a good fit for the online world of the early twenty-first century, and in the last few years a new approach following the philosophy of the SKOS knowledge organization system produced PhySH, a publicly available (CC-0) controlled vocabulary and thesaurus for physics managed by the American Physical Society.

PhySH disciplines are like the traditional subdivisions of physics, expanded to include new fields that have developed in recent decades. PhySH facets provide a cross-cutting classification by conceptual type, similar to the facets of Ranganathan's *Colon Classification*. PhySH concepts are identified by unique URIs while their unambiguous labels aid human work in indexing and searching. The vocabulary is still developing and has been made openly available for use and feedback. There are some obvious improvements still needed, particularly in the organization of the facets, but it can be used now for indexing and classification in physics. As the system is largely compatible with the SKOS model for knowledge organization, it will be interesting to see in future how PhySH and other SKOS vocabularies can work together in organizing research in the physical sciences.

References

American Physical Society. 2018. "PhySH contribution guidelines." https://physh.org/contribute

Barité, Mario. 2018. "Literary warrant." *Knowledge Organization* 45: 517-36.

Bliss Classification Association. 1999. *Bliss Bibliographic Classification. 2nd ed. Class B Physics. Full Schedule.* Available at: http://www.blissclassification.org.uk/ClassB/B_sched.pdf

Conover, Emily. 2016. "New Physics Classification Scheme Unveiled." *APS News* 25, no. 2: 3.

Desale, Sanjay and Rajendra Kumbhar. 2017. *Methodology to Develop Depth Classification Scheme for Physics.* Erfurt: Lambert Academic.

Forman, Paul, John L. Heilbron and Spencer Weart. 1975. "Physics circa 1900: Personnel, Funding, and Productivity of the Academic Establishments." *Historical Studies in the Physical Science* 5: 1-185.

Hjørland, Birger. 2017. "Classification." *Knowledge Organization* 44: 97-128.

IET (The Institute of Engineering and Technology). 1998. "History of Science Abstracts and Inspec." https://www.theiet.org/publishing/library-archives/the-iet-archives/iet-history/history-of-science-abstracts-and-inspec/

Krumhansl, James A. and George L. Trigg. 1975. "Indexing and Classification." *Physical Review Letters* 34: 1065-6.

Miles, Alistair and Sean Bechhofer, eds. 2009." SKOS Simple Knowledge Organization System Reference." https://www.w3.org/TR/2009/REC-skos-reference-20090818/

Mills, Jack and Vanda Broughton. 1977. "Introduction." Transcription from *Bliss Bibliographic Classification*. 2nd ed. "Introduction and Auxiliary Schedules." London: Butterworths, [1-107]. http://www.blissclassification.org.uk/Class1/introduction.pdf

Radicchi, Filippo and Claudio Castellano. 2011. "Rescaling Citations of Publications in Physics." *Physical Review E* 83: 046116.

Satija, Mohinder P. 2017. "*Colon Classification* (*CC*)." *Knowledge Organization* 44: 291-307.

Smith, Arthur. 2019. "From PACS to PhySH." *Nature Reviews Physics* 1: 8-11.