Towards a multilingual, comprehensive and open scientific journal ontology

Éric Archambault
Science-Metrix, 1335 avenue du Mont-Royal E, Montréal, Québec, H2J 1Y6 Canada

Olivier H. Beauchesne
Science-Metrix, 1335 avenue du Mont-Royal E, Montréal, Québec, H2J 1Y6 Canada

Julie Caruso
Science-Metrix, 1335 avenue du Mont-Royal E, Montréal, Québec, H2J 1Y6 Canada

Abstract

This paper describes the development of a new journal ontology to facilitate the production of bibliometric data. A number of approaches have been used to design journal-level taxonomies or ontologies, and the scholarly research and practical application of these systems have revealed their various benefits and limitations. To date, however, no single classification scheme has been widely adopted by the international bibliometric community. In light of these factors, the new classification presented here—featuring a hierarchical, three-level classification tree—was developed based on best-practice taxonomies. Categories were modelled on those of existing journal classifications (ISI, CHI, ERA), and their groupings of journals acted as "seeds" or attractors for journals in the new classification. Individual journals were assigned to single, mutually exclusive categories via a hybrid approach combining algorithmic methods and expert judgment. Notably, the classification was designed to be as inclusive as possible of newer fields of inquiry; general and multidisciplinary journals; and the range of arts and humanities disciplines. The new scientific journal ontology is freely available (it can be found at www.sciencemetrix.com) under a creative commons license and is operational in 18 languages.

Bibliographic Information

This Post-Print is the version of the conference paper accepted for publication.
Published online 2011 in the Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI) (http://www.issi-society.org/publications/issi-conference-proceedings/).
Archambault, Éric; Beauchesne, Olivier H.; Caruso, Julie (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In Noyons, B., Ngulube, P., & Leta, J. (Eds.), *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, pp. 66–77.

Introduction

According to Ruocco and Frieder (1997), the specific goal of classification is to "provide some insight into the organization of the data themselves." The classification scheme is the targeted end result of the agglomerative process, which reflects resemblance among items through either similarity or dissimilarity measures. Glänzel and Schubert (2003) suggest that the "classification of science into a disciplinary structure is at least as old as science itself" and that the classification of scholarly literature into appropriate subject fields is "one of the basic preconditions of valid scientometric analyses." Two levels of aggregation are used to delimit scientific fields using scholarly literature: the classification of journals into fixed sets of fields or specialties and the thematic coding of articles—that is, categorizing research papers into scientific areas (Gómez, Bordons, Fernández, & Méndez, 1996; Leydesdorff, 2002). Developing subject classifications at the journal level is a particularly common approach to field delimitation because aggregated sets of journals are considered to be indices of activity—indicators of the intellectual organization of the sciences and the exchange that takes place between scholars in particular specialities (Leydesdorff, 2004). This paper examines the approaches traditionally used to design journal-level taxonomies (or scientific journal ontologies) and presents a new classification, which is an evolutionary development from existing classifications.

The emergence of journal-based classifications

Any discussion of proto-bibliometric studies must include the pioneering work of Gross and Gross (1927), who were the first to tabulate citations for the purposes of analyzing the scientific literature. Their efforts were followed by a succession of studies aiming to describe the intellectual structure of the scientific literature, most notably Bradford's 1934 paper outlining the link between journal subject sets and searches for papers on specific topics. This was followed by numerous ground-breaking works—such as those by Cason and Lubotsky (1936), Garfield (1963), and de Solla Price (1965)—which proposed methods for managing the growing body of scientific literature. They were not, however, limited to delineating scientific fields; in fact, most sought to determine the structure of the sciences by establishing and analyzing networks, or scientific interrelationships, of researchers. There was a new turn in the seventies with the works of Narin, Carpenter, and Berlt (1972), Carpenter and Narin (1973), and Pinski and Narin (1976), who were concerned with mapping hierarchies or stratification systems among journals to determine the influence measures of journals or journal sets in aggregate.

Narin, Pinski and Gee's (1976) seminal work was the first to suggest that an ex ante subject classification scheme be created for scientific journals, which would allow large quantities of publication information to be summarized, enabling a more straightforward analysis of the development of the sciences. The authors asserted that this classification scheme should be "discriminating enough to show the differences between publications from different disciplines or fields, but not so detailed as to submerge major trends within numerous minor categories."

A wide variety of techniques have been proposed in the literature for analyzing journal-tojournal citation relationships, and journal network analyses have been performed for a variety of purposes since the 1930s. Journal-to-journal networks were defined by Doreian and Fararo (1985) as existing "when the citations made by authors of articles appearing in journals are aggregated by journals." According to the concept of journal networks, journals are nodes, and aggregated citations over articles in these journals are the relations between the nodes. The analyses of journal-to-journal citation relationships have led to the clustering of scientific journals—or partitioning of journals in

scientific disciplines into clusters of related journals— a process based on the concept of structural equivalence, where two journals with the same pattern of receiving citations are equivalent (Doreian, 1988; Doreian & Fararo, 1985; Leydesdorff, 1987).

Early on, clustering was found to take the subject classification of journals "to a level more precise than that of the discipline" (Carpenter & Narin, 1973), and Small & Koenig (1977) noted that while clustering was previously an exercise of theoretical interest, it could also be used as a "practical method for organizing journal sets." Numerous studies have since attempted to determine similarity between and classify journals using top-down and bottom-up cluster analysis procedures (Boyack, Klavans, & Börner, 2005; Chen, 2008; Doreian, 1988; Leydesdorff, 1987; Leydesdorff, 2006; Pudovkin & Garfield, 2002; Small & Koenig; Zhang, Liu, Janssens, Liang, & Glänzel, 2010). Particular grouping techniques or association measures are generally chosen based on the purposes they serve—of which classification is only one possibility—and their acceptability within an application area (Ruocco and Frieder, 1997; Swanson, 1973). Studies by Janssens, Glänzel, and De Moor (2008) and Zhang et al. suggest that hybrid clustering methods, which incorporate textual content and bibliometric information, appear to work best at indicating true document similarities, as they complement each others' strengths and compensate for weaknesses.

The early conceptualizations of journal networks based on citations from one article to another—and by extension from one journal to another (Small and Koenig, 1977)—have been fundamental to the development and popularity of databases such as Thomson Reuters' (formerly ISI) widely used Science Citation Index (SCI), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI), and Journal Citation Reports (JCR). Because the information they contain supports the establishment of relational network structures, they are commonly used to measure the 'influence' and network positions of authors, papers and journals (Carpenter & Narin, 1973; Doreian & Fararo, 1985). The JCR, in particular, has played a central role in the journal classification literature. The JCR's tabulation of citation relationships makes it possible to delineate journals more easily at the disciplinary level, enables researchers to validate journal similarity measures, ensures consistency in journal sets, and continues to play a major role in scientometric evaluations (Carpenter & Narin; Leydesdorff, 2004; Pudovkin & Garfield, 2002; Todorov & Glaenzel, 1990).

There have been many attempts to design new classifications, and most of these attempts, like the one described in the present paper, were evolutionary developments based on previous attempts. For instance, Katz and Hicks (1995) developed a journal classification scheme to examine, among other things, sectoral output and collaborative activity. The authors had originally hoped to use the CHI journal classification scheme, but found that it was not suitable for their purposes; they then developed a hybrid system that was based on classifying the 154 sub-fields of the SCI into the 10 broad Australian Standard Research Classification Scheme fields. This scheme was validated through discussions with various experts and scientists. Katz and Hicks hoped that their scheme might "ultimately lead to the acceptance of a standard journal classification scheme," which in particular "may be useful for developing indirect indicators of the change in interdisciplinary scientific research publications."

Another example is the work of Glänzel and Schubert (2003), who aimed to design a system for classifying scientific journals contained in ISI's Science Citation Indices, to be used for the purposes of scientometric analysis. Documents were classified into categories using a three-step

iterative process. The first, "cognitive," step involved distinguishing sets for the initial scheme based on subjective expert judgment; the second, "pragmatic," step involved classifying the majority of the SCI journal set into the subfields established in the previous step and making adjustments to ensure that multiple assignments were retained within reasonable limits; and the third, "scientometric," step involved classifying articles published in unambiguous core journals into the subfield of the given journals and manually classifying individual articles belonging to the more 'un-assignable' journals. The resulting scheme, which contains 15 fields and 68 subfields, is being used by the Steunpunt Onderwijs & Onderzoek Indicatoren (SOOI) in Leuven, Belgium, to support research evaluations (Rafols & Leydesdorff, 2009).

Many other multidisciplinary and disciplinary science journal databases use classification schemes derived from journal-based subject classifications. In addition to these electronic databases of scholarly literature, subject classification systems have been conceived by an extensive array of entities, including libraries, publishers, national research funding organizations, encyclopaedias, and internet-based information services. To date, no international standard classification scheme exists that supports bibliometric research, and no single classification scheme has been widely adopted by the bibliometric community. Neither do governments use internationally standardized classifications of scientific fields to analyze their research funding (Katz & Hicks, 1995). That is not to say that the field is completely scattered—as noted, the classification used by Thomson Reuters for its Web of Science (WoS) database and its many other bibliographic products is widely used in the field of bibliometrics and scientometrics. Moreover, the US National Science Foundation (NSF) has been using a classification originally designed by Mark Carpenter and Francis Narin at Computer Horizons Inc. (later called CHI Research) since the 1970s, which has had some traction in Canadian bibliometrics as well (e.g., both Science-Metrix and the Observatoire des sciences et des technologies [OST] of the Université du Québec à Montréal use it regularly).

Still, the general sentiment towards designing an internationally accepted standard can be summed up by Glänzel and Schubert (2003): "After many centuries of constructive but yet inconclusive search for a perfect classification scheme, the only sensible approach to the question appears to be the pragmatic one: what is the optimal scheme for a given practical purpose?" Despite the many practical and theoretical impediments that would have to be overcome, some researchers are nevertheless aware of the importance of pursuing the use of common subject classifications, at the very least for the sake of enabling international comparisons. Gómez et al. (1996) wrote that such standards should be "flexible enough to both enable international comparisons and meet the needs of local studies." Although we harbour no illusions about the difficulty of establishing an internationally accepted standard, the classification developed as part of this project contains three very important features: 1) it is available, at the time of writing, in 18 languages; 2) it is openly available, can be used freely in research and education, and can be downloaded for free and modified and improved by anyone with an interest in doing so; and 3) comprises the vast majority of the output of scientific journals that exist in both Scopus and the WoS, which means that it can be used in the majority of bibliometric studies.

Limits of journal-based classifications

Journal classification methods have shown many significant uses and benefits through the years, but one of their largest strengths is that they are relatively easy and low cost to implement and use. Nevertheless, existing schemes, as well as the methods used in their development and maintenance,

also have noted pitfalls and shortcomings. Even the most widely used schemes are believed to provide only "crude approximations" and "superficial perceptions" of categories of research, relatedness between journals or the diversity of publications (Pudovkin & Garfield, 2002; Tijssen, 2010). Although this is not meant to be a comprehensive catalogue of limits associated with journal-based categories, the following factors are generally thought to affect the validity of existing journal subject classifications. Many of these were identified in Gómez et al.'s 1996 paper "Coping with the problem of subject classification diversity".

- Disciplines are subject to a very fast rate of change and, as a result, so are journals. Bibliometricians have therefore long been faced with the problem of delineating journal sets consistently and over time (Carpenter & Narin, 1973; Narin et al., 1972; Narin et al., 1976; Leydesdorff, 2004). For instance, in most ex ante schemes, new journals are placed into existing frameworks, or they are disregarded altogether (Leydesdorff, 2006). To mitigate these problems, authors have stressed the importance of updating subject classifications of journals in order to "overcome the birth of new journals and to identify the emergence of new disciplines" (Gómez et al., 1996).
- Delimiting a field at the journal level will not generate results as accurately as doing so at the article level, primarily because journals contain articles that may have one primary subject but that ultimately involve a broad range of themes. This means that subject delimitations based on journal classifications are likely to contain articles that have a weak relation with the target subject and will also miss other pertinent articles (Gómez et al., 1996).
- Reference-based classification systems are highly subject to inconsistencies and even misrepresentation due to significant variations in publication activity and citation habits among subfields (Glänzel & Schubert, 2003).
- The increasing interdisciplinarity of the research continues to challenge efforts to set boundaries between disciplines. Researchers have frequently observed that articles and journals do not always fit snugly into a single category; for instance, Glänzel and Schubert (2003) concluded that authors' activity is often not limited to a single subfield but "usually covers a range of subfields with varying weights," and Boyack et al. (2005) found that while over half of the ISI categories corresponded closely with the clusters based on inter-journal citation relations, the remainder could not be assigned unambiguously. Because of this, researchers commonly assign journals into multiple subfields or into sets of multidisciplinary journals, which allows them to deal with a single journals' potential range of topics but results in the problem of multiple counts. Sometimes, bi-disciplinary journals can be split into two parallel fields, but more often, journals belong to two non-parallel fields, making it necessary to determine the primary and secondary fields and/or attribute weights to those fields based on the strength of the journal's affiliation to them (Glänzel & Schubert; Narin et al., 1976). Of course, not only is this procedure highly complex, but the resulting mixed sets make it more challenging for researchers to obtain valid conclusions, make comparisons between studies, or allow for reproducibility (Gómez et al., 1996). Additionally, when journals are commonly classified as multidisciplinary, schemes can no longer yield a complete picture of the research output of a given field, particularly when papers with the highest impact in a field are published in such journals (Bornmann, Mutz, Neuhaus, & Daniel, 2008).
- Existing classification schemes also display large variations in their degrees of specificity and aggregation. Bornmann et al. (2008) noted that the level of aggregation offered by

reference standards is a very important criterion in their selection. Journal classification schemes generally exhibit a higher level of aggregation that is suitable for macro-level analyses, but databases do not offer a level of discipline aggregation that is suitable for the purposes of most bibliometric analyses (especially those of highly specialized fields of research), and the largest and most diverse networks are the most likely to exhibit an over-aggregation of results (Klavans & Boyack, 2008). As a result, researchers carrying out such investigations may decide to make their own subject delimitation decisions to achieve a suitable level of aggregation for analyses, leading to a greater accuracy of results but reduced comparability with other studies (Gómez et al., 1996; Narin et al., 1976).

- The number and diversity of subject classifications and databases have led to difficulties in, for example, performing inter-country comparative analyses, combining two or more classification schemes from different information sources and establishing comparability among studies (Gómez et al., 1996).

Manual vs. automated, and mutually exclusive vs. overlapping journal taxonomies

The earliest subject classification schemes largely involved subjective analysis and heuristic methods—journals were classified through human assignment as well as visual examination of cross-citation patterns among journals (Chen, 2008; Glänzel & Schubert, 2003; Pudovkin & Garfield, 2002). It was assumed that subjective approaches that incorporated human intelligence and expertise could lead to more useful, flexible schemes. However, these subjective schemes have been heavily criticized for being "inadequate" in many fields and "subject to the vagaries of time" (Pudovkin & Garfield), with an output that relies on assumptions of hierarchies among subject subdivisions and therefore varies from expert to expert (Bensman & Leydesdorff, 2009; Chen; Glänzel & Schubert; Leydesdorff & Rafols, 2009). Leydesdorff (2006) surmised that "one cannot develop a conclusive classification on the basis of analytical arguments." Many of these concerns, coupled with the exponential growth of document bases and the corresponding "computationally prohibitive" execution time needed to operate on such large document sets, has led to an increased focus on designing automated systems for subject classification (Ruocco & Frieder, 1997).

As early as 1973, Carpenter and Narin touted the practical advantages of an automated, "objective" journal classification system, where the "ability to classify large sets of publications may aid in analysis of scientific capability." Soon after, Small and Koenig (1977) stated that the need for such an objective scheme was motivated by aesthetic and practical considerations, namely the "challenge of doing algorithmically what has been a very nontrivial task intellectually—the classification of journals," which the author described as an "almost pure problem in numerical taxonomy, that of partitioning a population on the basis of shared characteristics."

However, an automated classification scheme—one that is considered fast, complete, and reliable enough for widespread use—has yet to become available (Chen, 2008). As there are few generally accepted measures of quality (for example, of large sets of journals based on quantitative citation data), and relevance may shift greatly according to specific purposes, many authors still rely on independently derived manual classifications of journals to determine whether the clusters they generate make sense, agree with alternatively derived classifications, and are useful to their purposes. Additional problems are encountered with respect to the optimal levels of aggregation and the stability of clusters over time. In their analysis of the ISI (i.e., Thomson Reuters) classification in nanoscience and nanotechnology at various levels of aggregation, Rafols and Leydesdorff (2009) discovered that machine algorithms cannot be expected to create categories

that are balanced in terms of size. Nonetheless, expecting a "balanced" set of categories might be a faulty premise, as the work of Sylvan Katz suggests that power-law distributions may be ubiquitous in journal ontologies.

Another important feature that distinguishes different types of taxonomies is whether they propose a mutually exclusive set of categories or whether journals are allowed to be simultaneously classified in several categories. The classification favoured by Thomson- Reuters is of the overlapping style, whereas that originally developed by CHI Research and still regularly updated for the NSF is of the mutually exclusive type. There are at least two reasons why mutually exclusive categories would be preferred: 1) metrics obtained by counting output classified in a mutually exclusive classification are simpler—totals are the sum of sub-totals at the category level (whereas this total is greater than the sum when journals and their papers are counted in more than one category); and 2) this solution is more parsimonious in the sense that each journal should (ideally) be classified in the category that best represents its intellectual contribution. That is not to say that one mechanism currently exists to determine the best field for every journal, but at least the end goal is clear and unambiguous.

The most commonly encountered objection to the use of a mutually exclusive ontology is that knowledge production is a complex web of interaction, and a mutually exclusive classification misses out on this aspect. Although this paper's authors certainly agree that knowledge production reflects a complex web of interaction between disciplines, we would argue that all non-mutually classifications developed to date have failed to demonstrate how they systematically and a priori address the complexity of this web. Indeed, most of the classifications available are not well documented generally, if at all, and do not include criteria to show why one or more categories were chosen for a particular journal. Most classifications choose one main category for the majority of journals, but some journals are allocated to two, three, four or even more categories. What guides the selection of one category in one case and more than one category in another? In fact, it is often quite impossible to determine a neat cutting point when using more than one category. This point is illustrated in the following figure representing the extent to which the journal *Image and Vision Computing* is attracted to and attracted by different subfields (Figure 1). A parsimonious rule based on the most representative category would classify this journal as belonging to the *Artificial Intelligence & Image Processing* subfield (in the ontology presented in the figure and in the present paper). In contrast, the WoS classification suggests each of these subfields: 1) *Optics*; 2) *Engineering, Electrical & Electronic*; 3) *Computer Science, Artificial Intelligence*; 4) *Computer Science, Software Engineering*; 5) *Computer Science, Theory & Methods*. This undocumented assignation, in addition to being opaque, certainly makes the computation of metrics substantially, and unwarrantedly, more complex.

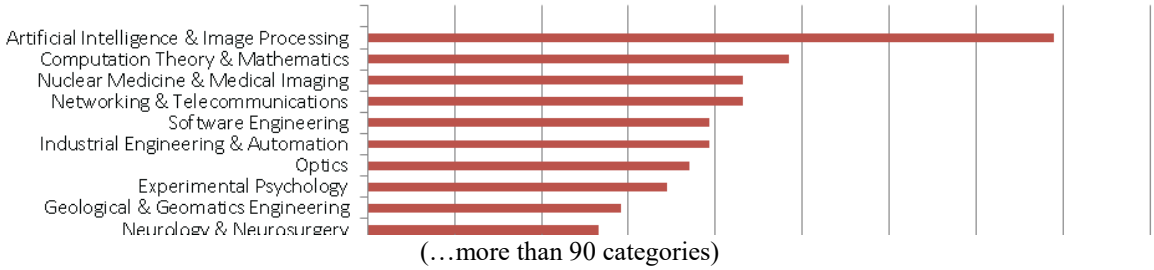

Figure 1. Attractiveness of the *Image and Vision Computing* journal, by subfield

Design goals for a new journal ontology

Leydesdorff (1987) warned that the large variety of computational and statistical methods used by researchers to classify the scholarly literature, if used unreflectively, may provide "results that are largely dependent on the choice of options offered by the computer programme." One can add that many other choices made along the way shape the form, function, and usability of a journal-based ontology. The prior experiences of ontologies' designers certainly play an important role. This is particularly the case in ontologies such as that proposed here, which used a hybrid method to obtain an evolutionary development based on best-practice taxonomies. The hybrid aspect of the method lies in its use of a mixture of algorithmic methods in addition to expert judgment, the latter being used more frequently in the classification of journals with few and/or not highly specific linkages (references and citations) to the categories used.

That the central aim of this classification is to facilitate the production of bibliometric data imposes some constraints. In particular, our prior experience in scientometrics suggested the development of a hierarchical, three-level classification tree. This would afford flexibility in terms of aggregation and disaggregation. One of our design goals was therefore to classify journals into fields, which would subsequently be grouped into subfields, and these into domains. The domains would distinguish between the social sciences, the health sciences, and so forth. The scientific *fields* would comprise categories such as chemistry, physics, and biology whereas *subfield*s, or scientific specialities, would comprise items such as anatomy, evolutionary biology, analytical chemistry, and so forth.

An important design goal of this ontology was to make place for both modern fields, such as ICT and enabling and strategic S&T (such as bioinformatics, nanotechnology and energy), and for general and multidisciplinary journals. After several iterations, we feel that we have achieved an appropriate balance between both including new research areas and acknowledging the still predominantly structuring effect of the more traditional fields. Another design goal included having a comprehensive representation of the arts and humanities, which are often underrepresented in journal and discipline classifications.

Assigning journals to subfields

It was decided from the outset to use existing journal classifications as a source of inspiration for categories and as "seeds" that would determine where journals would initially appear before being classified algorithmically. Please note that although other classifications were used as seeds, these were not used at all following the first pass of classification, and a total of seven iterations were made before obtaining the resulting ontology. Thus, an important step in the building of the

taxonomy involved carefully examining existing journal classifications in order to 1) produce a taxonomical tree and 2) gather lists of journals and their previously determined links with specific knowledge subdomains. After careful consideration, the following journal classifications were used as seeds:

- The US NSF: A tried and tested mutually exclusive classification used in the Science & Engineering Indicators since the 1970s. It was originally designed by CHI Research.
- The WoS classification: A widely used, constantly updated and authoritative classification, but with important overlaps between categories and possessing only one hierarchical level (although, internally, Thomson Reuters aggregates these classes).
- The Australian Research Council (ARC) Evaluation of Research Excellence (ERA) classification: A modern and newly designed scheme, hot off the press. Journals were assigned to fields by interested researchers. It has significant overlaps and, also, being so new, it has not yet been extensively tested (http://www.arc.gov.au//_journal_list.htm).

From the ERA classification, the most original idea we borrowed was that of creating the field *Built Environment and Design*. Otherwise, most of the field headings were largely borrowed from the NSF classification, which we felt was tried and tested as well as logical, if somewhat outdated in some respects. In addition to these classifications, which were examined in great depth, the following classifications were also examined to anchor our work in best practices:

- The revised Field of Science and Technology (FOS) Classification in the Frascati Manual, produced by the Working Party of National Experts on Science and Technology Indicators of the Organisation for Economic Co-operation and Development (OECD).
- The European Research Council (ERC) classification.

Although we felt that the ERC classification was too disaggregated for our purposes and would be extremely difficult to operationalize, we borrowed the idea to have one branch of the classification grouping all studies of the past (in the manner of *SH6 The study of the human past*) and made that field quite encompassing (including anthropology, archaeology, classics, history, and palaeontology; although we recognise that palaeontology is much closer to the natural sciences, we included it for the internal coherence of that field). The OECD classification served as an inspiration for the naming of several of the domains and fields, but could not be followed for both practical and ontological reasons. For instance, the OECD divides biotechnology into many different fields, and we knew from experience that a journal classification could not warrant such a level of disaggregation.

Both the taxonomical tree and the list of journals and their association with research subfields served as the starting point in the building of the new classification. In particular, the journals served as seeds in the new classification—that is, the groupings of journals in previous taxonomies—which would serve as "attractors" for journals in a subsequent stage of analysis. This is illustrated in Table 1, which shows how subfields in the seed classifications (and hence, their associated journals) were attributed to a new subfield called *Dairy & Animal Science*.

Table 1. Example of consolidation of journals from three classifications

Source	Subfield	New_Field	New_Subfield
WoS	Agriculture, Dairy & Animal Science	Agriculture, Fisheries & Forestry	Dairy & Animal Science
ERA	Animal Production	Agriculture, Fisheries & Forestry	Dairy & Animal Science
NSF	Dairy & Animal Sci	Agriculture, Fisheries & Forestry	Dairy & Animal Science

Using the NSF, Thomson Reuters (WoS), and ERA classifications, journals were assigned to "work-in-progress" subfields, which were in turn grouped into "work-in-progress" fields and domains. Following a review of this seed table, the *Languages & Culture, General & Miscellaneous* subfield was judged as being too broad and was thus unassigned. A subfield dealing with energy was not present in the initial subfield list, but was deemed necessary to conform to contemporary ontological expectations. This subfield was compiled using a list of journals dealing with energy matters, which drew from a list of journals that was obtained from previous contract work (see Archambault & Coté, 2009).

This resulted in a table comprising all of the classification subfields, and all associated journals were then put in a single table (more than 40,000 non-unique entities, due to the overlapping of subfields between classifications and the non-mutually exclusive nature of WoS and ARC classifications). This table was used as seed data to be processed by a purpose-built classifying engine. The initial idea was to use a classification process that would have assigned subfields to journals following the similarity of the journal to that subfield either in the references, citations, or text (in addresses, titles, or abstracts). A good seed was thus paramount to a satisfactory final classification. From this merged list, only journals with at least one article in Scopus or in the WoS were kept.

Classification engine

During an exploratory, pilot phase, two methods were used to classify journals into the most appropriate subfield. The first classification method used citations and references of each journal's articles. Assuming a power-law relationship between citations and number of papers, we log transformed values and performed regression analyses of the citations and references to the number of articles in a subfield for each journal. The expected values of citations or references inferred from the regression model for each subfield was then compared using a ratio. The subfield having the highest ratio was presumed to be the subfield that best described the journal similarity. However, this approach proved to be inconclusive—journals were often assigned to categories substantially different from what one would have expected based on their titles. The second classification method used the addresses of articles in each journal. Addresses containing words such as "school", "department", "institute", and "centre" or their abbreviations were pooled and grouped. From this list, only the top 3,000 distinctive and relevant words (such as "Mechanical", "Engineering", etc.) were kept. All addresses contained in a journal were grouped to form a vector. The same method was applied to the addresses contained in each subfield. Each journal was then compared to the vectors describing the subfields. The subfield with the highest cosine similarity to the journal was then presumed to be the subfield best describing the journal. Once again, this approach did not yield sufficiently precise results upon examination of the journal titles.

These approaches were therefore set aside in favour of a more straightforward process. Pilot tests—in which the log of the product of references made by a journal to each subfield to the number of citations received by that journal from each subfield, divided by the log of the number of articles

in each subfield—showed promising results. Each journal was then assigned to the subfield with the highest score. This method was substantially faster, more stable, and less sensitive to errors or inaccuracies present in the input data. Although the initial idea was to have several iterations of that program parse through all the journals and assign them to the best category in an unsupervised manner, examination of the results of the third iteration showed that general categories were progressively being emptied—the program was forcing cohesion in a somewhat artificial manner. In fact, after a few runs it appeared that some of the subfields were beginning to contain several subfields. Hence, it was decided that the results of the first iteration of the program be visually examined, and when needed, manually corrected. The whole list of journals was thus parsed by hand to catch errors and to force some of the journals into newer subfields, such as nanotechnology and biotechnology.

Six more passes were then made—firstly by algorithm, followed by manual validation—and these were helped by a number of tools that were used in parallel, such as the Subfield Affinity Index (SAI) and cosine similarity with an extended set of addresses. The basic equation used in the algorithm was also refined in subsequent passes:

$$\frac{1+\log(Ref.Subfield \rightarrow Jrnl.)}{1+\log(Ref\ Jrnl.)} \times \frac{1+\log(Cit.Subfield \rightarrow Jrnl.)}{1+\log(n\ Cit.Subfield)}$$

In the end, three more passes proved necessary in order to classify the approximately 34,000 journals and conference proceedings that appear at least once in Scopus and/or WoS. A subset of these—specifically, those journals with at least 30 papers, 30 references, and 30 citations—is available on Science-Metrix' website, in addition to a series of visualization tools (e.g., Figure 2).

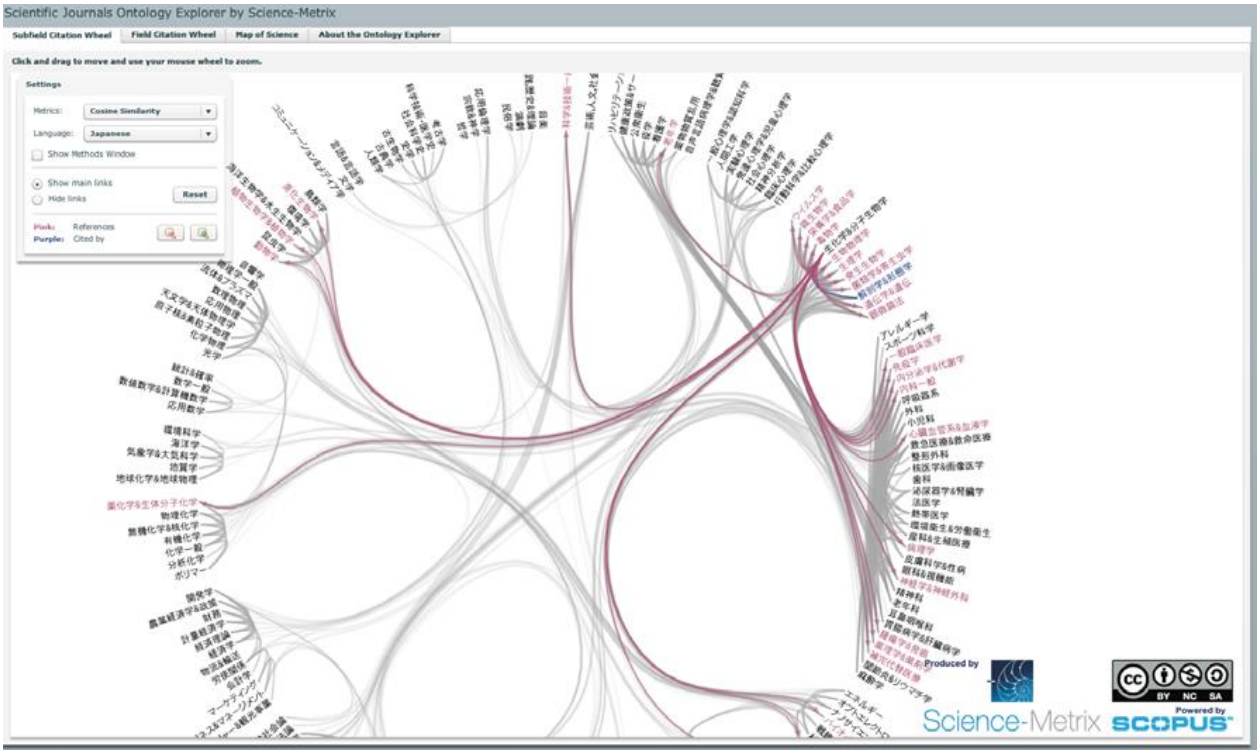

Figure 2. Screenshot of the ontology explorer showing subfield names in Japanese, available at www.science-metrix.com

Conclusion

A great deal of experimentation was performed to obtain a method of algorithmically assigning journals to what appears to be the most representative category. We experimented with a scale-adjusted analysis of attraction, with the use of cosine similarity based on author addresses. Our goal was to find a relatively simple method to classify most of the journals in a single category using an algorithm that would produce repeatable results. However, having looked at the use of factorial analyses by Leydersdorff to classify journals, we were not operating under the illusion that this approach would be faultless and that human intervention would be unnecessary in classifying each and every journal. From the outset, we decided that it would also be necessary to use expert judgment to finalize the work. In the end, it took substantially more work than initially expected, with alternating iterations using an algorithmic approach followed by manual fine-tuning.

Although we are confident, given the care put into this classification, that it is highly accurate overall, it remains that the number of journals involved is extremely large considering the limited means available for this project. There is also little doubt that classification errors remain, especially for the more fringe journals, as well as for those that have very broad scope and therefore for which neither a mathematical nor a manual classification may produce definitive results. The ontology has now been released under a creative commons licence (www.science-metrix.com), and anyone is welcome to use it in their research, education, and librarianship endeavours. It is our desire that researchers and practitioners will provide feedback that will help the classification to progress into an even more precise and useful tool and that each journal that is not yet classified in its optimal category will find its way there.

Acknowledgments
Science-Metrix acknowledges the partial financial contribution of the European Commission in the completion of this research project.

References

Archambault, E. & Côté, G. (2009). *Bibliometric analysis of energy research at the world level and benchmarking of CanmetENERGY.* Report produced by Science-Metrix for Natural Resources Canada.

Bensman, S. J. & Leydesdorff, L. (2009). Definition and identification of journals as bibliographic and subject entities: Librarianship versus ISI Journal Citation Reports methods and their effect on citation measures. *Journal of the American Society for Information Science and Technology, 60*(6), 1097-1117.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. (2008). Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8*, 93-102.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*, 351-374.

Bradford, C. (1934). Sources of information on specific subjects. Engineering (London), 137, 5-86.

Carpenter, M. P., & Narin, F. (1973). Clustering of scientific journals. *Journal of the American Society for Information Science, 24,* 425-436.

Cason, H. & Lubotsky, M. (1936). The influence and dependence of psychological journals on each other. *Psychological Bulletin, 33*, 95-103.

Chen, C. M. (2008). Classification of scientific networks using aggregated journal-journal citation relations in the Journal Citation Reports. Journal of the American Society for Information Science and Technology, 59(14): 2296-2304.

de Solla Price, D. (1965). Networks of scientific papers. *Science, 149*, 510-515.

Doreian, P. (1988). Testing structural equivalence hypotheses in a network of geographical journals. *Journal of the American Society for Information Science, 39*(2), 79-85.

Doreian, P., & Fararo, T. J. (1985). Structural equivalence in a journal network. *Journal of the American Society for Information Science, 36,* 28-37.

Garfield, E. (1963). Citation indexes in sociological and historical research. *American Documentation, 14,* 289-291.

Glänzel,W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*(3), 357-367.

Gómez, I., Bordons, M., Fernández, M. T., & Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics, 35*(2), 223-235.

Gross, P. L. K. & Gross, E. M. (1927). College libraries and chemical education. *Science, 66*, 385-389.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics, 75*(3), 607-631.

Katz, J. S., & Hicks, D. (1995). *The classification of interdisciplinary journals: A new approach (Version 2.0).* In M.E.D. Koenig & A. Bookstein (Eds.), Proceedings of the Fifth Biennial Conference of the International Society for Scientometrics and Informatics (pp. 245-254), Medford: Learned Information.

Leydesdorff, L. & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348-362.

Leydesdorff, L. (1987). Various methods for the mapping of science. *Scientometrics, 11*, 291-320.

Leydesdorff, L. (2002). Dynamic and evolutionary updates of classificatory schemes in scientific journal structures. *Journal of the American Society for Information Science and Technology, 53*(12), 987-994.

Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in the Journal Citation Reports. *Journal of Documentation, 60*(4), 371-427.

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology, 57*(5), 601-613.

Narin, F., Carpenter, M., & Berlt, N. (1972, October). Interrelationships of scientific journals. *Journal of the American Society for Information Science, 23*(5), 323-331.

Narin, F., Pinski, G., & Gee, H. H. (1976, January/February). Structure of the biomedical literature. *Journal of the American Society for Information Science, 27*(1), 25-45.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics. *Information Processing & Management, 12*(5), 297-312.

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology, 53(*13), 1113-1119.

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology, 60*(9), 1823-1835.

Ruocco, A. S., & Frieder, O. (1997, October). Clustering and classification of large document bases in a parallel environment. *Journal of the American Society for Information Science, 48*(10), 932-943.

Small, H. G., & Koenig, M. E. D. (1977). Journal clustering using a bibliographic coupling method. *Information Processing & Management, 13*(5), 277-288.

Swanson, R. W., (1973, January-February). On clustering techniques in information science. *Journal of the American Society for Information Science, 24*(1), 72-73.

Tijssen, R. J. W. (2010, September). Discarding the 'basic science/applied science' dichotomy: A knowledge utilization triangle classification system of research journals. *Journal of the American Society for Information Science and Technology, 61(*9), 1842-1852.

Todorov, R., & Glaenzel, W. (1990). Computer bibliometrics for journal classification. *Information Processing and Management: An International Journal Archive, 26(*5), 673-680.

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics, 4*, 185-193.