Haas, S., & Hert, C. (2000). Terminology development and organization in multi-community environments: The case of statistical information. . 11th ASIS SIG/CR Classification Research Workshop, 51-74. doi:10.7152/acro.v11i1.12769

Terminology Development and Organization in Multi-Community Environments: The Case of Statistical Information

Stephanie W. Haas* and Carol A. Hert**

*School of Information and Library Science
University of North Carolina
stephani@ils.unc.edu
**School of Information Studies
Syracuse University
Hert_C@bls.gov

1. Introduction

It is becoming commonplace to worry about the vast amount of information that is now available to people on the Web. Searches return thousands of web pages, most of which are not useful, yet there is the underlying belief that the information must be "out there *somewhere*". While it is clear that inadequate document surrogates and search engines are major contributors to this situation, another is that end users often do not know what words to use to describe what they want: if you do not know what to call it, you cannot find it[5]. This can inhibit successful searching even within a single website. If the target concept is described using one term by the document author, and a different term by the end user, the document will not be retrieved unless support tools (e.g., thesauri) exist that can act as intermediaries, mapping between the terms. The problem is exacerbated when the end user and the document author do not think of the topic in the same way; when there is only partial overlapping between their concepts, or when a concept just does not exist in one of their worlds.

In this article, we examine three components of a terminological system that affect the success (or lack thereof) of retrieval in such circumstances. These three components are:

- communities of users, with their varying degrees of understanding of a subject domain, and the language they use to express it,
- the nature of language, specifically the language and terminology used in special domains (Language for Specific Purposes) as contrasted with the language and vocabulary used in everyday situations (Language for General Purposes), and
- the role of warrant in determining what to call concepts, and how best to map among mismatching concepts and vocabularies.

The domain in which we conduct the investigation is that of the United States Bureau of Labor Statistics.

[5] There is also significant evidence that people can not articulate what they do not know and have anomalous states of knowledge (ASKs) (Belkin, Oddy, Brooks, 1982). This paper does not address that particular problem associated with unsuccessful retrieval; however, the authors recognize that terminology knowledge and mapping will not resolve the entire problem of retrieval.

ISSN: 2324-9773

> The Bureau of Labor Statistics (BLS) is the principal fact-finding agency for the Federal Government in the broad field of labor economics and statistics. … [It] collects, processes, analyzes, and disseminates essential statistical data to the American public, the US Congress, other Federal agencies, State and local governments, business, and labor. The BLS also serves as a statistical resource to the Department of Labor. (http://www.bls.gov/blsmissn.htm)

The BLS collects statistics in areas such as jobs and employment, unemployment, wages and benefits, job-related injuries and illnesses, and associated topics.

In company with other federal agencies, BLS has a strong mandate to make its information available to US citizens, and one means by which it does so is via the Web (www.bls.gov). Delivering statistical information to the general public is especially challenging because of the common "terseness" of a table of numbers, the importance of understanding the meaning of the row and column titles, and the frequent lack of understanding of general statistical principles, among other reasons.

The questions that drive our research are related to understanding user language and its relationship to this specific domain in which we see elements of both general and specialized language. In particular, how can the agency improve public access to the information it produces by building mappings among these various language sets?

The organization of the paper is as follows. First, we present the conceptual and theoretical frameworks of our research. There are three areas here. We start by discussing characteristics of different communities of users, the knowledge they have and the information they need to find and use BLS statistical information successfully. Next, we talk about the language(s) that they use, ranging from experts' very technical Language for Specific Purposes (LSP) to very general language (Language for General Purposes, or LGP). The third area is that of warrant; the sources of authority for determining what names or terms will be used for concepts in document representations, indexes, queries, and related information retrieval tools. In the second part of the paper, we describe our preliminary research into BLS knowledge structures and terms, including the nature of these, the communities that use them and the language the communities use to find the information they need. We conclude the paper with observations on the success of various strategies for improving end-user searching of BLS information.

2. Conceptual and Theoretical Frameworks

2.1 Communities of Users

We start by examining the notion of communities of users – users of language and users of information. This is the appropriate place to start, because the communities are both the initial sources of information and the ultimate consumers of it. We do not have a strict definition for communities of users, in part because the concept itself is fluid. Some communities may be quite well-formed and discrete, with explicit membership criteria. Others may be more amorphous, having no explicit criteria, but rather some sense of "having something in common". There is no assumption of physical contiguity or direct communication among members. An individual may belong to many communities simultaneously, some consciously, others because

ISSN: 2324-9773

Haas, S., & Hert, C. (2000). Terminology development and organization in multi-community environments: The case of statistical information. . 11th ASIS SIG/CR Classification Research Workshop, 51-74. doi:10.7152/acro.v11i1.12769

someone else sees similarities among individuals. Communities of users, as we view them, bear some resemblance to the "discourse communities" that Swales (1998) has discussed.

At this point, we can describe several communities of users who play some role in association with BLS information.

> **The BLS experts who create the surveys, organize the resulting data, and produce the statistical products (e.g., tables, time-series, indexes).** This community of users has expertise and experience in the BLS domain. In addition to being in some sense the creators of the BLS information, they are also sophisticated users of it. Their expertise manifests itself in many ways: the tasks they perform, their understanding the concepts and what to call them, and the ability to help others understand BLS information.
>
> **Citizens and companies from whom the BLS gathers data.** When the BLS surveys or interviews members of this community, questions must be phrased in ways that will be understood correctly, otherwise the responses will be meaningless. On one hand, questions such as "how much are you usually paid" or "are you employed full-time" are clear and familiar, and can be expressed in LGP. On the other hand, they may need explanation, also using LGP, to clarify matters such as whether someone should include last week's overtime pay, or that the BLS defines "full-time" as 35 or more hours per week.
>
> **Expert end users from outside the BLS.** Members of this "external" community have some level of statistical knowledge and familiarity with BLS information, and therefore an understanding of BLS concepts and terminology that may be similar to that of the "inside" experts. People in this community include government and private labor economists, journalists on the "labor beat", financial advisors, and academics. Although they can also be considered experts, there may still be some differences between their conceptual structures and language, and those of the BLS. For example, *inflation* is a common economic term, the value of which is based (for the most part) on BLS data. The BLS, however, does not use this term, although it can be roughly equated with the consumer price index, a BLS construct.
>
> **Naïve end users of BLS information.** This community overlaps with the second to some extent, although it is probably larger. We may assume that they have a general knowledge of BLS concepts through their own life experience, and use LGP vocabulary to describe them, for example, when searching for information on the BLS web site. This community is called a community since they share some characteristics (such as knowledge of the web and the similar tasks that require BLS information) but not because they have actively chosen to be a member of this community. (This creates special problems in understanding their language usage which we discuss later.)

Each of these user communities carries its own conceptual structures and understanding of BLS information, as well as its own tasks. Along with shared concepts and shared tasks, a community of users also shares language. In a more general community, the language itself is general, and is, in fact, referred to as "Language for General Purposes" (LGP). LGP is widely used and understood, and contains vocabulary for talking about widely shared concepts and actions. A community of experts shares a more specialized language, referred to as "Language for Special Purposes" (LSP), which contains terms for communicating about their shared expertise. Experts can communicate among themselves using their special language. For successful communication with non-experts, i.e., those outside their community, adjustments in language must occur. Cabré (1999) refers to this as adjusting the "concentration" of LSP to suit the participants. Any effort to make BLS information more accessible, easier to find, and easier to understand by all its users must include a recognition of these differences, as a first step to providing bridges between them.

2.2 Language for Special Purposes

In this section, we describe characteristics of a model LSP and its accompanying terminology. We then contrast that with that of less-well-differentiated domains, such as that of BLS information. Finally we discuss why these less-than-ideal (and probably more common) characteristics can create hurdles for finding and using BLS information.

LSPs have existed for some time. As long as there have been experts conversing among themselves, there have been specialized languages. Formal descriptions of LSP characteristics were developed along with the rise of terminology as a discipline (see, e.g., Sager, 1990, or Cabré, 1999). LSP is usually associated with specific domains, such as medicine, mechanical engineering, or the law. In contrast, LGP is general, everyday language, used by everybody to discuss general, everyday things. Its community of users could be considered to be all-inclusive, with the necessary shared knowledge being that gained from ordinary life experiences. Membership in the user community of an LSP, on the other hand, carries with it some assumption of specialized knowledge.

LSP encompasses more than just the words and concepts used in the domain, and includes register, style, and genre. The focus of attention in LSP and related research, however, is frequently on the words and phrases used to identify concepts in the domain. Terminology in an LSP may have a strong element of planning involved; for example, in chemistry there is a formal scheme for constructing the names of new chemicals. Other LSPs may adopt existing words; the terms *charm* and *color*, used in particle physics, is an example of this type of adoption.

A planned, well-formed terminology has little, if any, ambiguity. Each term refers to one concept, and concepts are named by only one term. As new concepts are developed, there is a process that is accepted by the community for developing new terms as needed. Concept definitions are also clear and unambiguous. Relationships among concepts and their terms (e.g., broader, narrower, part-of, or more domain-specific relationships) are also clear and well-documented. Finally, the concepts and terms are shared and understood by the community of users associated with that domain. "[T]erminology is a set of useful, practical communication units which are assessed according to criteria of economy, precision, and suitability." (Cabré, 1999, p. 11) These are characteristics of the ideal terminology, of course, but many domains approach this level, especially those whose concepts are quite different from everyday experience (Haas, 1997). An outsider (someone who is not a member of the community of users) can recognize the technical terms of such domains because they are different from the everyday vocabulary that occurs in LGP. (This is what *jargon* commonly refers to.)

ISSN: 2324-9773

In other domains, (including, we claim, that of the BLS) the concepts of interest are closely related to everyday experience, and the terms have thus been adopted from everyday vocabulary. Their technical uses may have specific definitions and different (usually narrower) coverage than the LSP uses, but the outside user (i.e., one who is not a member of one of the expert communities of users) may have difficulty recognizing that these changes have occurred. In the hypothetical extreme case of this, all LGP vocabulary would be "double-coded" and have a technical meaning in the LSP as well. In reality, we see a somewhat more muddled situation. Some of the domain concepts are quite similar to those of everyday experience. The LSP uses existing LGP words for these concepts, although the meanings may have been somewhat changed. Other LGP concepts and words that might seem to be closely related to LSP ones are not used or recognized in the LSP. For example, the market basket on which the BLS calculates the consumer price index (CPI) includes costs for *medical care*. A non-expert end user might search for costs related to *health care*, however, and not find the table containing the desired information. Yet these phrases would be considered synonyms in LGP. For the outside user who is attempting to find information, it is difficult to predict which everyday words will be helpful in the search, which will retrieve something, but not necessarily the desired information, and which will be unsuccessful. In effect, the end user may be using technical terminology, but not realize it.

Riggs (1993) comments on the phenomenon whereby existing words are borrowed and given technical meanings, especially in the social sciences. This can occur as an attempt to draw some type of analogy with the LGP concept. It also occurs, however, when the general concept and the technical concept are closely related. If the technical concept is commonly more narrowly defined than the general one, for example, the coverage of the word as it becomes a term in the LSP is similarly narrowed, thereby creating an ambiguous term. Riggs also points out that "a polyseme produces ambiguity if, within a given discourse community (research field, discipline or specialty) it has more than one meaning" (p. 197). This occurs in the BLS expert community. The agency is divided into several programs, each of which has responsibility for collecting and disseminating a variety of statistical data. There is overlap among these areas, however, which leads to the possibility for the nuances of the concepts and terms to be different among the program areas. For example, information about earnings and compensation are collected by two programs, one that surveys employers about their employee costs, and one that surveys individuals about their wages and salaries. These two perspectives on the information result in slightly different definitions for the general concept of "how much someone earns". As we have seen, however, ambiguity can also be produced across discourse communities where their concepts and terms overlap.

In describing the nature of LSP and terminology in the BLS world of information, we have shown that many of its concepts and terms are drawn from general language and experience, either directly, or with slight modifications. There is another dimension to this language, however, that Person (1998) described as "subfield" terms (see also Haas, 1997). In this situation, concepts and terms that are central to one discipline are also crucial to work in another. In most cases, the first provides tools and procedures for the second. Pearson's example is computational linguistics, and the use of statistical terms and concepts in its discourse. The meanings and usage of the statistical terms are consistent with their use in the statistical domain.

ISSN: 2324-9773

They refer to sets of concepts that are distinct from computational linguistics terms, but they are necessary to research and discussion within computational linguistics. This situation also occurs in the case of the BLS terminology. Statistical concepts and terms are an important part of the vocabulary used in conjunction with terminology from labor and economics. Statistical terms and concepts are generally distinct from those of general language. For example, *sampling error* is not a common phrase in everyday conversation. Even phrases that may be vaguely familiar to the general public are recognizable as jargon; with that recognition comes the understanding that a technical definition exists, even if it is unknown to the non-expert. For example, many people may know that the phrase *seasonal adjustment* is something that is done to employment rates because of events such as hiring extra sales clerks for Christmas shopping. But people also realize that the experts must have a more precise definition. In the BLS world, this and many other statistical concepts have both a verbal description and a technical definition in the shape of the formula that calculates them.

Cabré (1999) suggests that constructing a terminology starts by defining a concept, and then naming it, whereas lexicography starts with a word and associates it with a concept. If we accept this strong claim, then this supports our proposition that the LSP language used in the BLS domain does not consist of a "pure" terminology, but rather a combination of BLS and statistical terminology, and modified and unmodified general vocabulary. This presents several problems in helping end users (of any community) find the information they need. There is no existing well-constructed terminology. If such a terminology were constructed, many of the terms and definitions would merely duplicate general vocabulary, while others would require slight modification. Still others would be specific to the BLS domain. In providing search aids to end users (e.g., suggesting preferred terms), it might not be clear (to the end user or an information retrieval system) which query terms need expansion or substitution and which do not.

2.3 Sources of Warrant for Document Descriptions and Search Terms

The preceding sections have explored two of the areas that contribute to our theoretical framework. In this section we discuss the third area, that of warrant. The notion of warrant has been used by classifiers, indexers, and terminologists as a means of providing evidence from which to derive categories, controlled vocabularies or terminologies. The underlying idea is that if a category, word, or definition has been accepted by a discipline or community of users as a de facto standard, that is justification for its inclusion in the "official" set of categories, words, or terms. The interesting question then becomes what the critical community (or communities) of users is, and how to recognize "acceptance". At one extreme, a non-expert inventing a word that no one else in the domain subsequently adopts is not ā good authority for incorporating the word into the terminology. At the other extreme, a word that appears in specialized dictionaries or textbooks, is used by all experts in the field, and has stood the test of time should be accepted. It is the areas in between that can be difficult.

Several scholars have addressed the problem of appropriate sources of warrant. Beghtol (1995) discusses several: literary, scientific/philosophical, educational, and cultural. Literary warrant is based on the usage of concepts and terms in published literature of the domain. The assumption is that published authors' use of the term (with some level of frequency and distribution across authors) provides the evidence that the term should be included in the controlled vocabulary or terminology. Cabré (1999) identifies a similar idea in discussing the types of documentation

ISSN: 2324-9773

needed to construct a terminology. She defines *extraction documentation* as "the corpus from which terms will be selected" (p. 134). Among the crucial characteristics of extraction documentation are that it must be complete, up-to-date, and written by highly regarded authors in the field.

Scientific/philosophical warrant is based on consensus of acceptance within the field as a whole. The assumption here is that use of the concept and term is stable over time and is in accord with the general conceptual framework of the field as a whole. This type of evidence is especially clear in fields with well-established principles of organization of core concepts and names. Chemistry, for example, has the periodic table of elements, and guidelines for naming the basic elements as well as new compounds created from them. This is closely related to educational warrant, which reflects how information and knowledge in a domain is passed on to new students of the field. The organizational structure used in education should accurately reflect the experts' view of the subject, and also help learners understand what the important concepts and process are, and how they are related. Cultural warrant reflects a more general societal acceptance of subjects and ideas, as well as what to call them.

Lancaster (1986) discusses literary warrant (which he also calls bibliographic warrant), but he also describes another important source of warrant.

> There is another requirement that is frequently overlooked, however. This can be referred to as *user warrant*: A term is justified for inclusion in an index only if it is of interest to users of the information service. User warrant is especially important in establishing the appropriate level of specificity in the vocabulary. (p. 26)

User warrant states the importance of looking at the community of users of the target information, and determining how they identify it. Improving access to the information thus means taking the users' words for the concepts and somehow establishing links between them, even if (or especially when) the users' words are not necessarily the same as the experts' terms.

In our research context of attempting to improve public access to BLS information, it is clear that the role of user warrant is crucial. Because of the multiple communities of users, however, user warrant does not provide as much guidance in vocabulary identification as might be desired. Similarly, the differing communities of experts, even within the agency itself, mean that literary and scientific/philosophical sources can disagree. The result is that rather than envisioning a single column of expert terms neatly corresponding to a column of general end user terms in a one-to-one (or even one-to-many) relationship, we must expect a complex web of many-to-many relationships. It might be more fruitful to picture a cluster of closely related BLS terms corresponding to a similar cluster of user terms.

2.4 Summarizing the Theoretical Frameworks

The sections above have provided overviews of the three theoretical areas which are relevant to understanding the nature of terminological problems in this domain. At the core, BLS information is produced by experts and made available to experts and non-experts alike. Each community of users has information needs or tasks that should incorporate BLS information (including the production of the information itself), some level of conceptual knowledge or

ISSN: 2324-9773

familiarity, and some words or terms with which to name the concepts. Each community is also a legitimate source of warrant for words or terms to use in supporting the retrieval and use of BLS information. Our research investigates the problems that arise when these various communities of users attempt to access and use BLS information, how different conceptual and linguistic structures contribute to the problems, and how they can contribute to the solution – providing support for end users.

3. Overview of Research

In this part of the paper, we report on a project concerned with learning more about the words or terms communities of users employ to identify BLS information (especially in various kinds of searching tasks), and with seeking ways of improving user access. This project provides a baseline for understanding the extent to which the domain of interest has the properties described above, the relationship among the vocabularies employed by different user communities, and the role of sources of warrant in enhancing terminologies. Along with providing a baseline portrayal of the domain, our work has enabled us to investigate the feasibility of particular methodologies and approaches for understanding a very complex terminological system. It has also provided us with avenues for future work.

The intent of this project was to explore the relationship between user words for a concept (as represented in a search engine's log) and the terminology employed by the agency (as represented in its published documents). In this way, we were attempting to find the degree of overlap between the languages used by two different communities; the agency experts who authored the documents, presumably using the domain's LSP, and the external users of the agency's information[6]. The specific objectives were:

1. To determine the extent of the overlap between BLS terminology and an enhanced agency terminology (enhanced using WordNet (Fellbaum, 1998, (http://www.cogsci.princeton.edu/~wn/); and *Webster's New World Thesaurus*, 1985) for the concept of "pay" with user words for the same concept as identified in user inputs to a search engine.
2. To consider the feasibility of this approach for automatically enhancing agency terminology and/or user queries.

An additional benefit of this work was to investigate new metrics for assessing the utility of terminology enrichment.

The *pay* concept family was selected as the focal point for the investigation of BLS concepts and terminology. The *pay* concept is useful in our investigations because:

- It is a fairly complex concept in general, and one that is of interest to a wide community of users; high school students thinking about careers, employers needing to set pay scales, employees wanting to see where they stand in relation to national or regional averages, labor specialists, etc.

[6] Note that we cannot determine the level of expertise of these users from the search logs, and therefore cannot state whether they were experts using LSP, or non-experts expressing their needs in LGP.

ISSN: 2324-9773

Haas, S., & Hert, C. (2000). Terminology development and organization in multi-community environments: The case of statistical information. . 11th ASIS SIG/CR Classification Research Workshop, 51-74. doi:10.7152/acro.v11i1.12769

- Within the BLS, it falls within the jurisdiction of a couple of programs, notably Employment and Earnings and Compensation and Working Conditions.
- It is commonly acknowledged within the BLS that different programs use slightly different definitions of the concept.
- Data associated with the *pay* concept are gathered via different surveys, including the Current Population Survey (CPS), the National Compensation Survey (NCS), and the Current Employment Statistics Program (CES).
- The CPS is jointly administered by the BLS and the Bureau of the Census, so the *pay* concept family also illustrates successful data and metadata sharing between two agencies. Similarly, the CES is conducted by the State employment security agencies in cooperation with the BLS, requiring data and metadata sharing between federal and state agencies.
- The *pay* concept family is also used by numerous other federal agencies, such as the Internal Revenue Service and the Equal Employment Opportunity Commission, which have their own definitions and needs in its representation. As such, it can serve as the focal point of a future investigation of inter-agency uses and definitions of the concept.

Because of the variety of understandings those interested in the *pay* concept family bring to the pertinent BLS statistics and publications, it was therefore chosen as a useful starting point for this investigation into knowledge representation and terminology. Figure 1 shows a portion of the structure of the *pay* concept.

4. Methodology

In order to investigate the problem, we

- developed a conceptual map of terms employed by BLS for the pay concept
- expanded the resultant set of terms with terms from two thesauri
- gathered one month of logs from a search engine and parsed these logs into sessions
- manually searched each agency term (and term on expanded list) in the session logs
- calculated overlaps between the agency terminology set for the concept and the user session terminology.

We provide details below.

4.1 Conceptual Map of BLS Pay Terms

BLS publications, both print and on the Web site, were searched for terms used to name the *pay* concept and its closely related concepts, definitions given for those concepts, and types of information (information facets) associated with them. Publications examined included regular news releases, monthly and quarterly print publications, the *Handbook of Methods* (U.S. BLS, 1997), the CPS questions and supporting documents, and guidelines covering the NCS survey. Special attention was paid to paragraphs which defined the concepts and those which differentiated them from related concepts.

An extensive list of concepts related to the main *pay* concept was compiled from these publications, along with their definitions. See Table 1 for a list of agency terms that describe different aspects of *pay*. These include:

ISSN: 2324-9773

- Terms that describe some form of baseline pay, such as *wage* or *salary* and their variants (e.g., *hourly wage, apprentice wage, annual salary*).
- Terms that describe additional means of calculating pay. These may serve as the baseline amount itself (e.g., *commission* (in some cases), *piece rate),* or they may be added to it (e.g., *overtime, double time, commission* (in some cases)).
- Terms that describe monetary compensation in addition to baseline pay (e.g., *push money, tip*).

4.2 Expanding with Thesauri

The set of terms relating to the *pay* concept ("agency terms") were expanded with terms gathered from the Web-accessible version of WordNet 1.6 (http://www.cogsci.princeton.edu/~wn/) and the *Webster's New World Thesaurus* (1985).

- The appropriate sense of multi-sense agency terms was chosen by hand; if more than one sense seemed applicable in the *pay* domain, all applicable senses were used. Similarly, if a multi-word term did not appear as a whole, its constituent words were looked up and appropriate senses (if any) were used.
- Synonyms were obtained from the synset of the word or term. Only one generation of synset was used; that is, only the synset of the agency term, not the synsets of words in the first synset.
- Broader terms were obtained from the hypernym of the appropriate sense(s) of the agency term. Only one generation of hypernym was used; that is, only the parent term, not more distant ancestors.
- Narrower terms were obtained from the hyponyms of the appropriate sense(s) of the agency term.
- Coordinate terms were not used, since they were almost always too far from the desired meaning.

Table 1 shows the agency terms expanded with terms from the two thesauri.

4.3 Identifying User Terminology for the Same Concept

The second part of the methodology entailed identifying the user terminology for the *pay* concept. This involved several steps, starting with manipulating search logs to identify sessions, extracting user terms employed by a user within a search session, and then determining which of these sessions were related to the *pay* concept.

The research team received the November 1998 logs from the FedStats (www.fedstats.gov) search engine. These logs include the IP address associated with a given query, a time stamp, the search query, and databases searched (the FedStats engine enables a user to specify which agency websites to search), and information about the results received (number of pages found)[7]. It was necessary to perform some "cleaning" of the logs. In particular, we removed log entries that represented a user displaying additional pages of results rather than inputting queries (these entries were differentiated by a code). Once these were removed, the analyst normalized the

[7] Full details of the parsing methodology are available in Hert (1999) and the limitations of log analysis are considered as well.

ISSN: 2324-9773

query strings by removing extra spaces and by making all entries lower case. No additional normalization or stemming was done in this analysis.

For additional analyses, it was necessary to parse the queries into sessions. A *session* was defined as a series of inputs from an IP address with each input occurring less than 30 minutes from the last input.

After the preliminary parsing and manipulation of the search logs, the analyst developed summary counts of actions recorded in the logs, the number of queries (number of actions minus number of next page commands), the frequency with which each normalized query string appeared in the log, and the number of unique queries (total queries minus all duplicate queries). An interactive website was developed to present summary statistics as well as provide the ability for a user to search the logs for a particular word (and see all queries that included that word) and to search the sessions for a particular word or string.

At this point a few definitions are in order. A *query* is the word or phrase (normalized as explained below), along with any Boolean operators, that appears in the log. There can be multiple instances of a query appearing in the log. An *instance* of a query is one entry in the log file. A *term* is the phrase or word used by the agency for a concept as well as a part of a query that is separated by a Boolean operator from another part of a query. Terms may consist of single or multiple words. Thus the user query "Catholic priests and salaries" consists of two terms, "Catholic priests" and "salaries."

In order to compare agency terminology with user vocabulary, we had to find user sessions and queries that represented searches for the concept of interest (pay and wages). Given the total number of queries and our inability to understand user intent from log entries, the researchers defined user queries related to the concept as those queries that were part of a user session in which an agency term for the concept was employed. If a user used no agency terms, the session was not identified. At this point, we have no measure of how many such unidentified sessions might exist—to identify them would involve extensive use of thesauri and clustering algorithms to bring together potentially related words and phrases, which would enable us to infer user intent even if agency words were not used. Such activities were beyond the scope of this exploratory analysis.

An example may clarify the process of identifying relevant user queries. A term on the agency list was "salary." Searching the database by session, 94 sessions are found that included the term "salary". The queries that make up these sessions are all considered relevant to the concept of "salary."

The analysts searched the session database (via the website listed above) for all terms on both the agency list and the expanded agency list of terms. The identifier of each session (source, session #) that utilized agency terminology was entered into an Excel spreadsheet and the number of queries, total terms, and agency terms was recorded for each session, as were the set of terms employed by the user during the session. Since sessions might use multiple agency terms (and thus would be identified more than once by the above process), duplicate sessions were removed from the database prior to further analysis.

ISSN: 2324-9773

In order to simplify coding, analysts were instructed not to interpret user queries in an effort to "understand what the user was doing." A limitation of the use of logs is that we can never know exactly what the user was thinking, why he or she input the various terms, etc. so rather than assume that there are some cases in which we can tell, the set of coding rules purposely attempted to limit analyst interpretation of the queries. Data from the spreadsheets were then used to develop frequency distributions and scatterplots.

4.4. Findings

The first analyses that were done compared user terminology with the unextended list of agency terms. Table 2 provides the frequency distribution of the proportion of agency terms to number of terms used in the session. Most sessions' (89% of sessions) terminology consists of 50% or less agency terminology.

The same analyses were done comparing the user terminology with the extended list of agency terms. In this case, the proportion that was calculated was the proportion of terms in the extended list to the total terms. Table 2 provides the frequency distribution. In the case of the extended term set, 73% of the sessions use 50% or less of the extended agency terminology. 19.8% of the extended term set sessions use all extended terms compared to only 8.7 of the sessions used only agency terms.

5. Discussion

The study described provided us with several insights into the relationships among user and agency terminologies, the feasibility of using our techniques on a wider basis, and enabled us to suggest further avenues of research and strategies for improving user access. We learned that:

- agency terminology does not map well to user vocabulary (as evidenced in a search log),
- general purpose (LGP) tools for enhancement did not sufficiently extend the LSP,
- further sources of warrant need to be exploited, and
- the need for the agency to create mappings from user vocabulary to agency terminology is critical.

ISSN: 2324-9773

Haas, S., & Hert, C. (2000). Terminology development and organization in multi-community environments: The case of statistical information. . 11th ASIS SIG/CR Classification Research Workshop, 51-74. doi:10.7152/acro.v11i1.12769

5.1 User and Agency Terminology Comparisons

The results indicate that users are employing many terms for the pay and wage concept and that many of these are not used by the agency itself. While this study has not assessed the user terminology in a qualitative fashion to determine the nature of the additional terms (how many are misspellings, incorrect terms, etc.), it does appear that there is a mismatch between the two languages. (A finding not detailed here is that many terms from the agency list do not appear in any user queries at all.) When the agency set is extended there appears to be increased matching but it is still low. Generally, what this would suggest is not that either party is performing ineffectively, but that instead, mappings between the two sets of terminology might be made. A strategy employed in many information retrieval systems is to include a controlled vocabulary of terms—a set of terms that the agency uses to describe its documents -- and to make this available to users. Users thus have the option of searching free text (as they currently do) or using the vocabulary as it is used by the agency. A less-visible strategy is to translate behind the scenes without informing the user but this requires that a user query be interpreted. Either of these approaches would have to be based on an information structure that recorded the mappings between the agency and user vocabularies. We shall say more about the requirements for this crosswalk structure later.

5.2 Enhancing an LSP with LGP tools

Our work suggests that utilizing vocabulary from general purpose thesauri only marginally improves the match with user terms and does not enhance this LSP terminology in meaningful ways.

- Many of the agency terms, especially the multi-word ones, did not have entries in either thesaurus.
- Many of the terms consist of multiple general words, and one way they gain their technical meaning is by combining with other words. Finding synonyms of the component words is not helpful in matching the whole term itself.
- Even when the basic concept was similar between the BLS usage and the thesaurus entry, synonyms and narrower terms could diverge in meaning from the agency term in surprising ways. For example, *blood money* is given as a hyponym of *payment* in WordNet.

In the case of the BLS *pay* terms, the words themselves are mostly common words, but they are often used in specific ways. Our idea was that by using non-specialized, general language thesauri, we could create overlap between the user queries and the agency documents where none existed before. The problem we encountered was that the LGP resources did not contain BLS terms as entry points. This finding is not surprising if we recall the characteristics of the BLS language that we discussed in Section 2. Its combination of highly technical terms and LGP words that have been modified into technical terms means that it cannot be treated as a traditional LSP or as general language, but rather must be viewed as a complex mix of language types.

Another approach to query expansion uses specialized domain thesauri as a source for related terms. For example, Srinivasan (1996) discusses the expansion of MEDLINE queries using the MeSH thesaurus. This approach maintains LSP characteristics (notably meanings) of the initial

ISSN: 2324-9773

Haas, S., & Hert, C. (2000). Terminology development and organization in multi-community environments: The case of statistical information. .
11th ASIS SIG/CR Classification Research Workshop, 51-74. doi:10.7152/acro.v11i1.12769

terms, but would probably not be as successful at providing mappings between the technical terms and the everyday words that non-experts might use to find relevant information, since the LGP words and meanings are not included in the specialized thesaurus.

5.3 Building Crosswalks

Since we started with the agency terms and found their synonyms, our strategy could be viewed more as expanding the document representation than query expansion. This is something that could be done once, as each document was added, rather than having to be done on the fly as each query is entered. Another way of viewing this is as an attempt to build a crosswalk between agency terminology and user terms, which could be stored permanently in the BLS metadata registry, services module, or some equivalent.

A terminology crosswalk shows the correspondences between two (or more) terminologies that are different, but whose coverage of concepts overlaps to some extent. It is somewhat analogous to a bilingual dictionary, providing equivalent terms in one "language" for those in another. The GILS table listing terms from ISO/IEC 11179, GILS, Dublin Core, and MARC (http://www.gils.net/element2.html) is an example of a type of crosswalk, as is the UMLS Metathesaurus (http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html). Crosswalks can represent mappings between two formal terminologies, or between a formal terminology and general language, which is what we believe is called for here.

A crosswalk is similar to the thesaurus of a terminology that shows synonyms, broader terms, etc., but with one crucial difference. In the thesaurus of a single terminology, the included terms belong to the terminology itself, unless it is explicitly indicated that they should not be used (e.g., *use instead,* or *deprecated*). In a crosswalk, the relationships are between different terminologies. What we propose here is a crosswalk that maps the agency terms with the end user vocabulary. Further, to handle the many-to-many mapping problem referred to in 2.3, we suggest that the mappings be made among clusters of related concepts and terms, rather than on a term-by-term basis.

5.4 Directions for Future Research

This works points in the direction of further investigations. Naturally, we need to assess the relationships among user vocabularies and agency terminologies for additional concepts. Our techniques might be employed to determine the level of overlap among terminologies thereby pointing to times when terminological enhancement or mapping adds value.

We also plan to continue to investigate a variety of sources of user warrant. User warrant is an important source in improving access to BLS information. There are two difficulties involved in drawing upon it. First, there are many different communities of users (some of which were described earlier) who all require access to it. Their names for domain concepts overlap to some extent (although their understanding of the concepts may vary), but it is important to allow for their differences. Second, it can be very difficult to gather data from some of these communities that tells us what the important names and concepts are. The general user community, for example, is extremely diffuse and it can be hard to structure experiments that capture their needs and questions in a natural way. In our research, some of the approaches we have tried include interviewing reference intermediaries, studying web search logs, observing end users using the web pages, and scanning general media sources.

ISSN: 2324-9773

We have also considered the other possible sources of warrant for finding words and terms that should be included in any tool designed to improve retrieval of BLS information. Literary sources provide important information on the concepts and terms used within the agency. These include internal documents such as the *Handbook of Methods* (U.S. BLS, 1997) or instructions to employees who administer the surveys, regular publications that include extensive tables and related articles written by agency experts, the tables and accompanying notes that are available on the Web page, and the press releases that come out on a periodic basis (e.g., the monthly unemployment figures). One difficulty with using agency documents is that since they are usually produced by a single program within the agency, variations in usage and meaning of an ambiguous term must be inferred by comparing documents produced by different programs. Explicit discussion of differences are rare.

The use of scientific/philosophical and educational warrant are more problematic. First, different programs within the agency use slightly different terms and definitions. This means, for example, that there is no single definition of *wage*. The second difficulty arises from the nature of the concepts they deal with. The strictly statistical concepts have a well-established framework that is used by statistical experts both inside and outside the agency. Actual formulas for specific statistical concepts (which may be considered a form of definition) may vary from domain to domain, but the core of the concept is the same. Other important concepts, drawing from general language and experience, have less well-organized structures. Further, the structure of a concept such as *pay* (see Figure 1) that is used within the agency differs slightly from that drawn from other agencies (such as the IRS), as well as from that recognized by the general community of users.

Finally, work on fully conceptualizing and implementing a crosswalk in this domain remains to be done. Preliminary investigations suggest that a crosswalk based on multiple mapping strategies may be helpful because of the kinds of user words, agency concepts, and question structures that we have identified.

Some concepts fall into obvious clusters of related meanings. This is the case, for example, with many demographic concepts such as sex, age, race, or education. If a user wants to see information broken down by sex, there are only two possible categories, and probably a finite number of ways of expressing them. Age is a little more challenging from the user perspective. It can be expressed as a specific number of years (*20 year olds*), a range of years (*20-25 years of age*), as a label that is more or less well-defined (*baby-boomers, Gen-X*), or as rather vague description (*older workers*). The agency perspective, however, provides limitations on possible mappings because the data is provided in fixed age-groupings; what starts as a vague description must be operationalized as one or more of these. A user who is interested in earnings of Gen-Xers, for example, may need to settle for information about people aged 25 to 34 which presumably includes the required data. Providing users with lists of the available divisions in the age concept cluster could help them determine the best mapping for their words when direct mapping based on numeric ages is not possible. In a sense, these clusters correspond to the closed class parts of speech in English (e.g., determiners or conjunctions). It is feasible to identify most, if not all, of the common ways of expressing these concepts.

ISSN: 2324-9773

Many user queries include some sort of job title or industry name, such as *social worker* or *hotel industry*. The BLS utilizes extensive classification systems for these, and there are many situations where the best strategy would be to guide the users to these existing classifications, and let them identify the best match to their query. This strategy would also eliminate the futile attempt to list all possible job titles. To continue the analogy to English, these clusters correspond to open class parts of speech, such as nouns and verbs. New jobs and job names are constantly being created in the world, but the BLS classification structures change more slowly.

Finally, there are many user queries that are so common, the mapping between user words and BLS terms has become "institutionalized". The example of *inflation – consumer price index* is perhaps the most common. Unfortunately, even these mappings are not always straightforward. Often a disclaimer or explanation is also needed, because the LGP word does not correspond precisely to the LSP term. For example, *inflation* may also be used so as to incorporate other factors, such as producer prices.

In building crosswalks, we must keep in mind some important practical considerations. Coverage will always be an issue. We are starting by gathering terms from the users' perspective(s), in an attempt to aid as many searches as possible with a minimum amount of effort (the familiar "80-20 principle"). Second, maintenance of the crosswalk will always be necessary. As the world changes, what users look for and the words they use will change. Currency of the crosswalk must be balanced with the time and effort the BLS can reasonably put toward maintenance. For example, the information age has created new jobs and ways of working. Some, such as *webmaster* or *telecommuting* may be here to stay. Others, such as *cyberjobs* may be more trendy. The former may be worth incorporating in the crosswalk, the latter may be gone before the opportunity for addition occurs.

6. Conclusion

We opened this paper by presenting three theoretical frameworks that we believe help illuminate the problems that can be encountered in the BLS and similar domains that have multiple communities of users, and the resulting problems of language and warrant. In attempting to improve access to BLS information, we feel a bit like Goldilocks:

- Tools based on LSP are too specific and lack the necessary connection to LGP.
- Tools based on LGP are too general and do not include LSP concepts or terms to the required level of specificity.

Our current approach to obtaining one that is "just right" is to create a crosswalk mapping between the BLS LSP and LGP. This paper has presented activities relevant to creating a crosswalk, given our theoretical framework. Theories and practices for such work are in their infancy; we have explored some options here and have suggested some of the problems and additional avenues to pursue. We are currently pursuing some of these and expect to provide information on our investigations in future papers. Given the increasing likelihood that data previously accessible only to experts, will become more available to a wide range of users, strategies for enabling user understandings and language to be mapped to the agency perspective will be critical.

Haas, S., & Hert, C. (2000). Terminology development and organization in multi-community environments: The case of statistical information. . 11th ASIS SIG/CR Classification Research Workshop, 51-74. doi:10.7152/acro.v11i1.12769

7. Acknowledgements

The research reported in this paper was funded by the United States Bureau of Labor Statistics. Cathy Dippo, Fred Conrad, John Bosley, and Michael Levi of BLS provided both logistical and research support. At Syracuse University, we wish to thank Michelle Turecek and Freya Anderson for coding assistance. John Fieber at Indiana University performed the log analysis. At the University of North Carolina at Chapel Hill, we wish to thank James Damron for assistance in term collection, and Gary Marchionini for many useful discussions. Finally, we thank Dagobert Soergel for his re-working of tables which significantly clarified our presentation here.

References

Beghtol, C. (1995). Domain analysis, literary warrant, and consensus: The case of fiction studies. *Journal of the American Society for Information Science*, 46, 1, 30-44.

Belkin, N. J, Oddy, R. N., and Brooks, H. M. (1982). ASK for information retrieval: Parts 1 & 2. *Journal of Documentation* 38(2): 61-71; 38(3):145-164.

Cabré, M. (1999). *Terminology: Theory, Methods and Applications.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Fellbaum, C. (Ed.) (1998). *Wordnet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Haas, S. W. (1997). Disciplinary variation in automatic sublanguage term identification. *Journal of the American Society for Information Science,* 48, 1, 67-79.

Hert, C. A. (1999). *Federal Statistical Website Users And Their Tasks: Investigations Of Avenues To Facilitate Access: Final Report to the United States Bureau of Labor Statistics.* Available at: http://istweb.syr.edu/~hert/BLSphase3.PDF

Lancaster, J. W. (1986). *Vocabulary Control for Information Retrieval, 2nd ed.* Arlington, Va. : Information Resources Press.

Person, J. (1998). *Terms in Context.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Riggs, F. (1993). Social science terminology: Basic problems and proposed solutions. In H. Sonneveld & K. Loening, (eds.) *Terminology: Applications in Interdisciplinary Communication.* Amsterdam/Philadelphia: John Benjamins Publishing Company, 195-222.

Sager, J. (1990). *A Practical Course in Terminology Processing.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

ISSN: 2324-9773

Haas, S., & Hert, C. (2000). Terminology development and organization in multi-community environments: The case of statistical information. . 11th ASIS SIG/CR Classification Research Workshop, 51-74. doi:10.7152/acro.v11i1.12769

Swales, J. (1998). *Other Floors, Other Voices: A Textography of a Small University Building.* Naweh, NJ: Lawrence Erlbaum Associates.

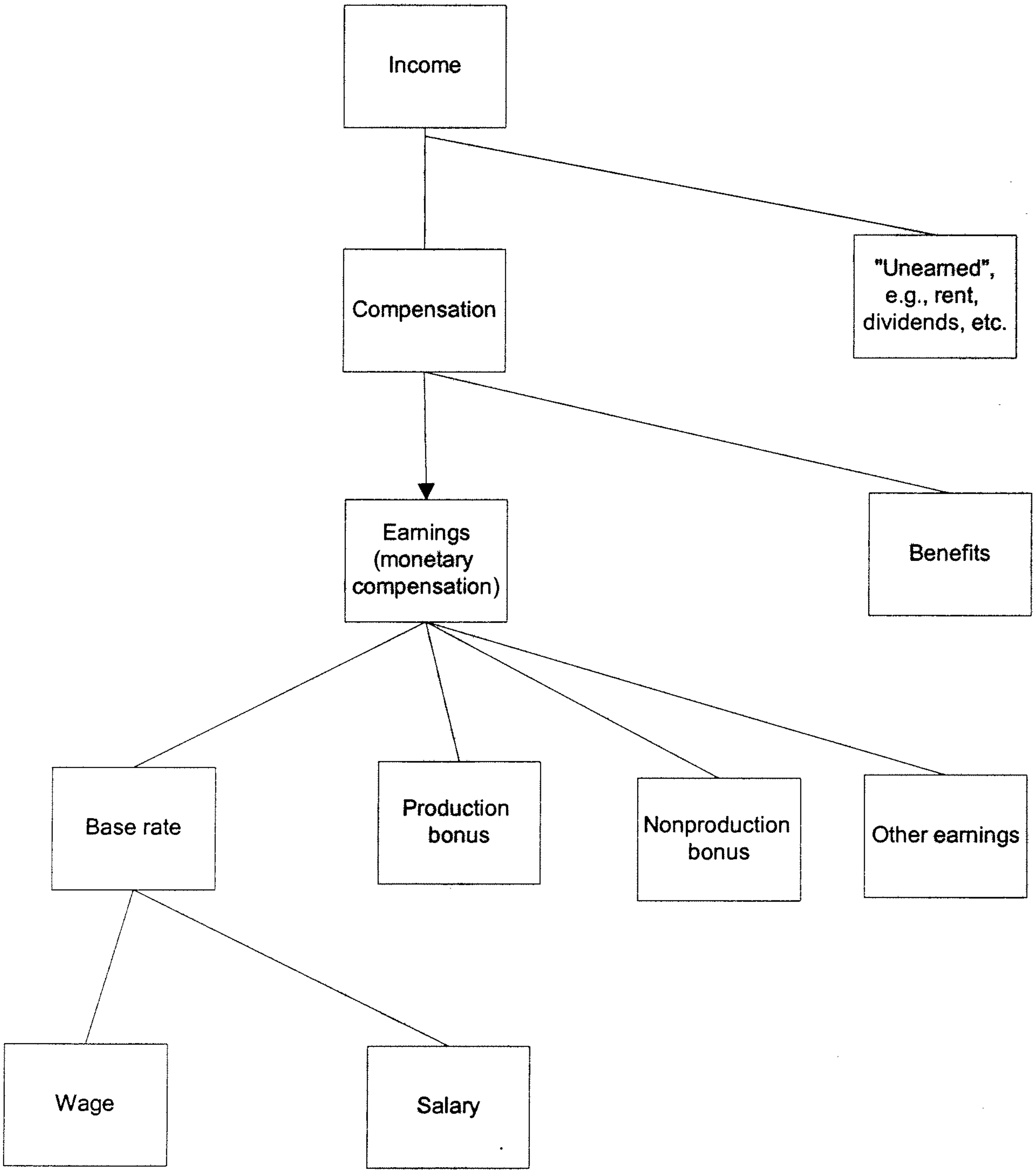
United States. Bureau of Labor Statistics. (1997). *BLS handbook of methods*. Washington, DC :U. S. Dept. of Labor, Bureau of Labor Statistics : For sale by the US G. P.O., Supt. of Docs.

Websters New World Thesaurus. (1985). Simon & Schuster, Inc. On *Toolworks Reference Library* [CD-ROM]. (1991). The Software Toolworks, Inc.

ISSN: 2324-9773

Haas, S., & Hert, C. (2000). Terminology development and organization in multi-community environments: The case of statistical information. . 11th ASIS SIG/CR Classification Research Workshop, 51-74. doi:10.7152/acro.v11i1.12769

Figure 1. Portion of the structure of the *pay* concept family.



ISSN: 2324-9773

Table 1. Expanded list of BLS *pay* terms. Added terms are in italics.

Accumulation
Adjustment
Allowance
Apprentice rates
At-risk pay
Attendance bonus
Back pay
Base rate
Beginner rate
Bereavement pay
Bilingual pay differential
Blue circle rate

Bonus
Call-in pay

Charge

Charge per unit

Cleanup
Commission
Commission payment
Compensation
Contract-signing bonus

Cost
Cost of living adjustment

Cost-of-living allowance
Deadhead pay
Deduction
Depreciation allowance
Discount
Dismissal pay
Disposable income
Dividend
Double time
Draw account

Earning per share
Earnings
Educational pay differential

Emolument
Experimental rate
Fee
Financial gain
Flagged rate
Flat rate
Free room and board
Fringe benefit
Government income
Government revenue
Gratuity
Gross
Gross profit
Gross profit margin
Gross sales
Guaranteed rate

Half-pay
Hardship allowance
Hazard pay
High time pay
Hiring rate
Holiday bonus
Holiday premium pay
Honoraria
Honorarium
Hourly rate
Incentive
Incentive earnings
Income
Index
Issue
Journey level rate

Killing
Knowledge-based pay
Living wage
Longevity pay

Lucre
Make-up pay
Margin
Markup
Merit pay
Minimum wage
Moving allowance
Multiskill compensation
Net
Net income
Net profit
Net sales
Nonproduction bonus
Out of line rate

Overcompensation
Overtime
Paid absence allowance
Pay
Pay in lieu of vacation

Pay packet
Pay rate
Pay-for-knowledge
Payment
Payment rate

ISSN: 2324-9773

Payments for income deferred due to participation in a salary reduction plan

Payoff
Payroll
Penalty rate

Per capita income
Per diem allowance
Percentage
Perk
Perquisite

Personal income
Piece rate
Portal to portal pay

Portion
Premium pay
Probationary rate

Proceeds
Production bonus
Profit
Profits
Profit-sharing
Profit-sharing distributions
Push money
Rate of pay
Rate of payment
Receipts

Recompense
Red circle rate
Referral bonus
Regular payment
Reimbursement
Relocation allowance

Remuneration
Reporting pay
Retroactive pay
Return
Revenue
Royalty
Safety bonus
Salary
Straight-time earnings
Scale

Seasonal adjustment
Severance pay

Share
Shift differential
Shift premium

Sick pay
Skill-based pay
Stint work
Stipend
Strike pay
Subsistence allowance
Superannuated rate
Supplemental pay

Take care, takings
Take-home pay
Temporary rates

Tip
Tips
Tonnage rate
Tool allowance
Trial rate
Tuition reimbursements
Unearned income

Unearned revenue
Uniform allowance
Union rate
Vacation pay
Wage
Wage scale
Wage schedule
Windfall profit
Workmen's compensation
Year-end bonus

Yield

ISSN: 2324-9773

Table 2. Frequency Distribution of Agency Term Proportion

Proportion of agency terms to total user terms	Original agency terminology		Extended agency terminology	
	No. of searches	Percent	No. of searches	Percent
.00 - .09	10	1	13	1
.10 -. 19	101	12	101	8
.20 - .29	167	19	211	17
.30 - .39	202	23	234	18
.40 - .49	11	1	14	1
.50 - .59	288	33	360	28
.60 - .69	18	2	62	5
.70 - .79	0	0	19	1
.80- .89	1	0	6	0
.90 - .99	0	0	0	0
1.00	76	9	253	20
All	874	100	1273	99

ISSN: 2324-9773

Blank

ISSN: 2324-9773

ISSN: 2324-9773