Knowledge Representation, Concepts, and Terminology: Toward a Metadata Registry for the Bureau of Labor Statistics

Final Report to the Bureau of Labor Statistics

Stephanie W. Haas, Ph.D.
School of Information and Library Science
University of North Carolina at Chapel Hill

July, 1999

Executive Summary

The original proposal for this work (Haas, 1998) was to investigate the feasibility of developing an overarching knowledge representation for Bureau of Labor Statistics information that captured its semantics, including concepts, terminology, actions, sources, and other metadata, in a uniformly applicable way.

It was determined during the first few months of this investigation that a promising solution for achieving the overall goal expressed in the proposal already existed in the ISO/IEC 11179 standard for metadata registries (http://metadata.aihw.gov.au/project/rev11179/contents.html), and more specifically in the Census model developed by Dan Gillman of the Bureau of the Census, which instantiates the standard. The focus of the investigation therefore shifted to evaluating the standard and the Census model according to the needs of the BLS, explaining what the process of adopting the model would entail, and exploring the model's conceptual and terminological capabilities. This last work was done in part in collaboration with Dr. Carol Hert. The *pay* concept family was identified as a concept central to the concerns of the BLS, and was therefore used as a case study for pursuing these objectives.

This report contains the following material.

- Criteria for a knowledge representation for BLS information.
- A brief description of the ISO/IEC 11179 standard and the Census model, and a review of how they fulfill the criteria.
- Issues concerned with adopting the Census model.
- A brief introduction to the UML *use case*, along with an example. Use cases can serve as a means of informing people about the Census model, gaining support for its adoption, and determining what information is needed to best utilize it.
- An exploration of the *pay* concept family, and how it is defined within the BLS.
- A discussion of terminological issues in promoting citizen access to and utilization of BLS information through the WWW. Research with Dr. Carol Hert examines the overlap of BLS and end user terms for the *pay* concept family.
- Recommendations resulting from this work along with some suggestions for future work.

Criteria for a knowledge representation for BLS information. Based on the characteristics of BLS information as exemplified by the *pay* concept family, as well as research in metadata structures, conceptual structures and ontologies, and other types of knowledge representations, the knowledge organization adopted by the BLS should satisfy several criteria.

1. The representation should be compatible with other organizations' knowledge structures
2. It should be capable of representing not just core concepts and definitions, such as *pay*, but also the relationships between concepts, including that of modification.
3. It should be able to associate concepts, such as *wage*, with the operational data (often gathered through surveys) that are their instantiations in the real world, such as *dollars/hour*.
4. It should be able to associate information about processes and functions that apply to data elements, such as those used in imputation or aggregation.

5. It should support a wide variety of information processing functions and procedures used both within the BLS, and by the agency to make information available to the public.
6. It should include structures recording terminology and thesaurus relationships among terms.
7. Thesaural relationships should include the familiar ones such as synonym, broader term, or narrower term, that are useful with precisely defined terms and term hierarchies. But they should also include those that are useful in building "crosswalks" between end user terms and agency terms.
8. It should require as little effort as possible initially to populate it with data. If the representation supports incremental population and does not require that all information maintain directly in its structure, the effort may be somewhat reduced.
9. The representation should make updating existing information as simple as possible.

The ISO/IEC 11179 standard and the Census model. Given the criteria drawn both from the literature and from my observations of the BLS data, I recommend the adoption of the ISO/IEC 11179 Specification and Standardization of Data Elements standard. The ANSI standard ANS X3.285, Metamodel for the Management of Sharable Data (which is to be included in a revision of Part 3 of 11179) is also a closely related standard. Dan Gillman of the Bureau of the Census has created a metadata registry for census information which instantiates the standard; I recommend that the BLS adopt this model. "A data registry is a place to keep facts about characteristics of data that are necessary to clearly describe, inventory, analyze, and classify data. A data registry supports data sharing with cross-system and cross-organization descriptions of common data." (ISO/IEC 11179, Section 1.1) A metadata registry is thus a data registry designed to hold the metadata that describes an organization's data. The Bureau of the Census and the BLS deal with very similar kinds of information, indeed they jointly administer and draw from the CPS. In other words, these two agencies are already sharing data; the joint adoption of the registry model will enable them to continue and perhaps expand their cooperation.

Issues concerned with adopting the Census model. There are several advantages in adopting a metadata registry modeled on the ISO/IEC 11179 standard in general, and the Census model in particular.

1. Since one of the goals of adopting a metadata registry is to increase data sharing among agencies (and presumably other entities), there is real advantage to following a standard that is commonly accepted in the knowledge management community.
2. In adopting any standard or model for a technology that is intended to promote the sharing of data (e.g., EDI, network standards, etc.), there is a critical mass of adopters that must be reached to achieve real success. Organizations both within the United States and internationally are adopting this standard.
3. The national and international community of researchers and practitioners in knowledge modeling is very active at this point, so there are readily available resources of information.
4. The Census model matches many of the information needs of the BLS.
5. The Census model has a rich thesaurus/terminology component. . Terminology is an obvious and important access point to an organization's information. The model allows the links between concepts and terms to be associated with a particular context, which is a

good mechanism by which to handle the ambiguity inherent in the BLS information (as illustrated by the *pay* concept family), as well as the ambiguity that is an unavoidable by-product of encouraging use of the BLS information by the public.
6. Every component of the model can be associated with an administering body or authority. This is an important feature in a model of shared data, where different components may be created, changed, or otherwise controlled by different people or programs.
7. Conceptually, the registry model can be viewed as a single unit, however "[t]he structure described by this metamodel may be distributed over several implementations. These implementations may be database, registries, repositories, dictionaries, etc. (ISO/IEC 11179, Section 7.0). This distinction contributes to its flexibility in both implementation and use.
8. The model clearly shows the connections between concepts (e.g., *pay*), formal terms and operational definitions of the concept (e.g., the CPS definition of "usual weekly earnings"), the data items that represent them (e.g., dollars/hour), and the source of the data items (e.g., a specific instance of a survey question).

There are few disadvantages in the adoption of this standard and model.
1. The effort needed to adopt any type of knowledge organization is like that associated with any change of infrastructure, and cannot be completely avoided. It is addressed in more detail in Section 2.3.
2. Although ISO/IEC 11179 is the standard with the most momentum behind it (to my knowledge), it is still undergoing some changes.
3. The metadata repository is like any other component of the information technology infrastructure in that it will require maintenance.

The remaining issues primarily concern the technical and policy-oriented matters likely to arise in the adoption process. Success requires coordinated effort along several dimensions. The BLS must develop a solid understanding of what a metadata registry is and the advantages that will accrue from its adoption in order to get buy-in from pivotal people and departments. Adoption strategies should address the following concerns.
- Education of BLS personnel regarding the uses of a metadata registry.
- Presentation of different views or versions of the Census model for different audiences, emphasizing aspects of special concern to them.
- Development and presentation of use cases (see Section 3) illustrating the role that the registry can play in common processes.
- Discussion of coordination between policy makers (e.g., scope of coverage, access, responsibility) and technical people, both within the BLS and across agencies, especially with Bureau of the Census.
- Identification of technical and policy "champions", who will help maintain the momentum of the adoption process.

The UML *use case* as a means of informing people about the Census model, gaining support for its adoption, and determining what information is needed to best utilize it.
A use case is a narrative, like a story, that describes what an information system should do to perform one of its functions, or to achieve a goal. Initial development depends more on familiarity with the functions and actors, and less with the details of how to make it work or

exactly what the format of the data types should be. Similarly, use cases provide a good basis for "walking through" a procedure, determining what the possible actions are, what resources are needed, and identifying what system components already exist, and what needs to be developed. Because of these characteristics, they are a good tool for informing people what the role of the metadata registry could be, how it relates to other components of the system, and how it could benefit their own work. Use cases also provide a foundation for brainstorming, and developing new uses for the registry. Section 3 includes an example use case.

An exploration of the *pay* concept family, and how it is defined within the BLS. The *pay* concept family was selected as the focal point for the investigation of BLS concepts and terminology for several reasons, including its complexity, and the wide community of users, experts and non-experts, to whom it is of interest.

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. Results of this research include three sets of data.

1. An extensive list of concepts related to the main *pay* concept was compiled from these publications, along with their definitions. Table 1 gives a list of terms which describe different aspects of *pay*.
2. A class hierarchy based on the CPS definition. An overview version is shown in Figure 3 with a full version given in Appendix A. The classes and subclasses in the hierarchy can be used in at least two ways:
   - as a component of the metadata registry proposed in Section 2.
   - as an aid in helping users of BLS data understand how terms are defined by the agency (e.g., in titles or column/row headers of tables), and how meanings may differ slightly from table to table. In essence, it can become the basis of a terminology crosswalk (see Section 5).
3. Information facets associated with the *pay* concept family, which represent the ways in which the concept is used (modified, measured, manipulated, etc.) in the BLS. Preliminary lists are given in Figures 4 and 5.

This concept family illustrates several important notions about concept representations. It is not uncommon for technical terms, even within a single organization, to have multiple, contradictory definitions. Similarly, users outside of the agency are also likely to have different ideas of names and definitions for technical concepts.

Terminological issues in promoting citizen access to and utilization of BLS information through the WWW. The theme that is common to all of these is the recognition that the terminology used by experts in a domain, e.g., by people within the BLS and by statisticians and other experts who regularly use BLS data, is not necessarily the same as that used by non-experts, e.g., the average citizens who use the BLS web page to find information. The importance of this issue for easing access to BLS resources for all citizens must be emphasized – if someone does not know the "correct name" of something, it can be very difficult, if not impossible, to find.

Hert & Haas (forthcoming) reports on a project which had the intent of exploring the relationship between user terminology for a concept (as represented in a search engine's log) and the terminology employed by a specialized agency (i.e., BLS). The specific objectives were:

1. To determine the extent of the overlap between agency (BLS) terminology for the concept of *pay* with user terminology for the same concept as identified in user inputs in a search engine.
2. To determine the extent of the overlap between agency terminology expanded with related terms from two thesauri (WordNet and Websters) for the same concept with user inputs.
3. To compare the extent of the two overlaps.
4. To consider the feasibility of this expansion approach for automatically enhancing agency terminology and/or user queries.

Tables 2 and 3 give the agency terms and their expansions used in this study. Further details and results of this research will be presented in Hert & Haas (forthcoming).

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. Crosswalks can represent mappings between two formal terminologies, or between a formal terminology and general language, as is used in end user queries. This raises an important issue for the contents of the metadata registry. Where and how are crosswalks represented? Participants at the 1999 Open Forum agreed on the need to maintain some level of separation between the "official terminology", and words or terms that may be used to access information from outside.

Recommendations and suggestions for future work.

1. The adoption of the ISO/IEC 11179 Specification and Standardization of Data Elements standard, as instantiated by Dan Gillman's Census metadata registry model. The standard and the model both satisfy the majority of the criteria identified for a knowledge structure for BLS information. Fundamental to the successful adoption of the standard and the model is the development of a solid understanding of what a metadata registry is and the advantages in its use. Other policy and technical issues include getting buy-in from pivotal people and departments, and the coordination of effort between policy makers and technical managers.
2. The development of use cases. As noted in this report, they can serve many purposes in the metadata registry adoption process. In addition, they can be a valuable tool in developing a variety of help facilities to aid in citizen access and use of BLS data, especially from its web site.
3. The determination of where and in what format various kinds of user services should be housed. The use case presented in this report proposed a services module which would contain helps and prompts for searchers, aids to understanding BLS concepts and terms, formatting and presentation tools for retrieved information, and tools to interface between the database and the registry. This is not the only solution, and other possibilities should be explored. This would require some investigation of the current BLS information system structure, and how it would be modified with the adoption of a metadata registry.

4. Additional study of BLS concepts and terms is a vital part of implementing the metadata registry, and also of improving citizen access to and understanding of BLS information. Several activities could be pursued to support this.
    - Continue studying the overlap between BLS terminology and end user terms.
    - Examine different types of concepts and terms.
    - Study the overlap between BLS concepts and terms, and those used in other government agencies.
    - Develop terminology crosswalks for BLS-end user terminologies and BLS-other agency terminologies. Crosswalks can serve as the basis for many kinds of end user help services.

Knowledge Representation, Concepts, and Terminology:
Toward a Metadata Registry for the Bureau of Labor Statistics

Final Report to the Bureau of Labor Statistics
(Purchase Order #OPS-184298)

Stephanie W. Haas, Ph.D.
School of Information and Library Science
University of North Carolina at Chapel Hill

July, 1999

1. Introduction

The original proposal for this work (Haas, 1998) was to investigate the feasibility of developing an overarching knowledge representation for Bureau of Labor Statistics information that captured its semantics, including concepts, terminology, actions, sources, and other metadata, in a uniformly applicable way.

Specific objectives laid out in the proposal were to:

> 1. Gather and analyze information associated with published BLS information, including concepts, terms, definitions, information sources, metadata, associated actions, and other information pertinent to the development of a knowledge organization.
>
> 2. Determine an appropriate knowledge representation for storing and displaying the information to be incorporated into the knowledge organization.
>
> 3. Prototype a knowledge organization for at least a portion of the BLS information.

We determined that the best approach would be to identify a concept central to the concerns of the BLS, and use it as a case study for pursuing these objectives. The *pay* concept family was selected to be this focal point, and is discussed in Section 4.

It was determined during the first few months of this investigation that a promising solution for achieving the overall goal expressed in the proposal already existed in the ISO/IEC 11179 standard for metadata registries (http://metadata.aihw.gov.au/project/rev11179/contents.html), and more specifically in the Census model developed by Dan Gillman of the Bureau of the Census, which instantiates the standard. The focus of the investigation therefore shifted to evaluating the standard and the Census model according to the needs of the BLS, explaining what the process of adopting the model would entail, and exploring the model's conceptual and terminological capabilities. This last work was done in part in collaboration with Dr. Carol Hert.

This report contains the following material.

- Criteria for a knowledge representation for BLS information.
- A brief description of the ISO/IEC 11179 standard and the Census model, and a review of how they fulfill the criteria.
- Issues concerned with adopting the Census model.
- A brief introduction to the UML *use case*, along with an example. Use cases can serve as a means of informing people about the Census model, gaining support for its adoption, and determining what information is needed to best utilize it.
- An exploration of the *pay* concept family, and how it is defined within the BLS.
- A discussion of terminological issues in promoting citizen access to and utilization of BLS information through the WWW. Research with Dr. Carol Hert examines the overlap of BLS and end user terms for the *pay* concept family.
- Recommendations resulting from this work along with some suggestions for future work.

2. The Knowledge Model

The overall goal of the original proposal was to determine an appropriate knowledge representation for storing and displaying the information to be incorporated into the knowledge organization. Once the best representation is identified, the real work of adopting it, populating it with the agency's information, and incorporating it into the agency's information processes can be planned. The appropriate level at which this kind of information should be modeled is the *conceptual* level.

> A conceptual data model describes how relevant information is structured in the natural. In other words, it is how the human mind is accustomed to thinking of the information. It is one layer more abstract than a logical data model that describes a particular computer-based system. The conceptual data model provides an excellent place to start modeling data within the sphere of interest. It is also the most viable level at which to integrate different data models because object representational differences are excluded. (ISO/IEC 11179, Foreward).

2.1 Criteria for the Model

The metadata and knowledge management research community has drawn extensively from research in ontology. The scope and purpose of an ontology is generally extended when it moves from being a static representation of a set of concepts and their relationships, to being the living repository of an organization's operational information. In addition, an important impetus for the "metadata movement" today is the need to provide a basis for sharing information both between departments within an organization (what is often called an "enterprise model"), and with other organizations.

Gruber (1995) discusses the requirements for such ontologies. First, he states the importance of evaluating the quality of an ontology based on its intended purpose. "Formal ontologies are viewed as designed artifacts formulated for specific purposes and evaluated against objective design criteria." (p. 907) One could debate the probability of developing truly "objective design criteria", but it is clear that the designers should have a good idea of (1) the kinds of

information that will be stored in it, (2) who the intended users are, and (3) what they will be doing with it. He suggests that design criteria should include clarity, coherence, extendibility, minimal encoding bias, and minimal ontological commitment. Especially with the goal of information sharing, these criteria all aim to promote flexibility of use. Gruber also mentions the importance of consistency within the ontology. In many situations, even within a single organization, there may not be consistency in the use of concepts and terms. The BLS *pay* concept family (see Section 4) is an example of this. In this sort of situation, the goal may instead be consistency within each context of use, and clear identification of when each definition applies. Guareno (1997) adds the importance of domain knowledge in ontological modeling; this may be especially true when the model is intended to be shared.

Based on the characteristics of BLS information as exemplified by the *pay* concept family, as well as research in metadata structures, conceptual structures and ontologies, and other types of knowledge representations, I determined that the knowledge organization adopted by the BLS should satisfy several criteria.

1. The representation should be compatible with other organizations' knowledge structures. This is vital for supporting data sharing among organizations. This criterion suggests that a "home-grown" structure is not ideal – using an existing structure, especially one conforming to an international standard (such as ISO/IEC 11179) is preferable. The Web page for the current ISO/IEC 11179 draft includes a good discussion about the role of data models in data sharing. Organizations frequently express concern that a standardized structure will not really fit their information needs – they might have to "warp" the information to fit the structure. There are two responses to this very real concern. First, the standards usually allow for varying amounts of "depth" in following the standard. As long as the broad outlines conform, the details may be omitted. Second, the ISO/IEC 11179 standard specifically allows for tailoring the model to local information needs.

   > It is not expected that this schema will completely satisfy all users. For example, scientific data requires metadata attributes not addressed in this standard. Each industry or each application may find a need to extend this schema. Such extensions shall be considered conformant if they do not violate any of the rules inherent in the structure and content as specified by the metamodel in this standard. Entities, relationships, and attributes may be added to this conceptual data model, but the core schema shall not be altered. (ISO/IEC 1179, Section 6.2)

   A related note is that as more and more organizations discuss their information structures and requirements, it seems that there are more issues in common than not. The discussions at the Open Forum on Metadata Registries in February 1999 highlighted this.

2. It should be capable of representing not just core concepts and definitions, such as *pay*, but also the relationships between concepts, including that of modification. For example, in searching for information about *pay*, a person's query could include modifications concerning profession, location, or time, e.g. *annual salary of teaching in New York in 1995*. Concepts found to modify *pay* are listed in Section 4.1.3.

3. It should be able to associate concepts, such as *wage*, with the operational data (often gathered through surveys) that are their instantiations in the real world, such as *dollars/hour*.

4. It should be able to associate information about processes and functions that apply to data elements, such as those used in imputation or aggregation.

5. It should support a wide variety of information processing functions and procedures used both within the BLS, and by the agency to make information available to the public. Examples of functions within the agency might include recording changes in definitions, formulae, or survey questions, creating a glossary, or determining which program was responsible for maintaining a particular data set. Examples of functions which make information available include publishing data along with their definitions, limitations, or other associated information, developing tools to aid end user access to agency data and supporting information, and providing various help and education facilities, so that end users can learn more about the information the agency has.

6. It should include structures recording terminology and thesaurus relationships among terms. Terms that are ambiguous within the BLS, or between technical and nontechnical uses should be clearly identified as such, and the contexts in which the different definitions (and all the other information connected with the usage) apply should be described. This information also clearly separates the "authority" terms and definitions sanctioned by the agency from the end user terms and definitions.

7. Thesaural relationships should include the familiar ones such as synonym, broader term, or narrower term, that are useful with precisely defined terms and term hierarchies. But they should also include those that are useful in building "crosswalks" between end user terms and agency terms. These relationships are likely to be somewhat "fuzzier", as the end user definitions themselves are often defined in a more general way. Such relationships might include *overlaps*, *near synonym*, or *often confused with*. The structure that these relationships produce is frequently not the familiar one-to-many structure that results from the broader term – narrower term relationship, but rather a many-to-many mapping that creates a complex network. However, the crosswalk structure lets end users start with whatever terms they know, and then provides guidance to steer them to the appropriate agency terms, and thus, the information they are seeking. This important point is discussed further in Section 5 and in Hert & Haas (forthcoming).

8. It should require as little effort as possible initially to populate it with data. There is no escaping the fact that populating a knowledge model with all of an organization's metadata is a significant undertaking, however there are a couple of characteristics that could "ease the pain". First, the representation should support incremental population, allowing its builders to insert information by topic or type and use it at various points of stability during its construction. For example, a thesaurus may be useful after some portion of its terms have been entered, as long as the portion is relatively independent and does not have many "dangling references". The idea of partitioning its development so that different teams can build it in parallel goes along with this notion. Second, the

representation should not require that all information be maintained directly in its structure. There should be minimal duplication of existing information. Rather, the contents of the representation could be pointers to the existing resources. This should speed the development of the knowledge structure, and also has useful properties for maintenance (see #9).

9. The representation should make updating existing information as simple as possible. In some cases, updates may be automatic (or nearly so), such as adding metadata for new editions of surveys. Other kinds of updates, such as redefining a term or concept, will require human effort. Note the importance of representing information once within the agency, whether in a resource outside of the structure with links to it, or just within the structure itself. (See item #8.)

2.2 Recommendation to adopt ISO/IEC 11179 and the Census Model

Given the criteria drawn both from the literature and from my observations of the BLS data, I recommend the adoption of the ISO/IEC 11179 Specification and Standardization of Data Elements standard. The ANSI standard ANS X3.285, Metamodel for the Management of Sharable Data (which is to be included in a revision of Part 3 of 11179) is also a closely related standard. Dan Gillman of the Bureau of the Census has created a metadata registry for census information which instantiates the standard; I recommend that the BLS adopt this model. According to the standard, a data registry is defined in this way.

> A data registry is a place to keep facts about characteristics of data that are necessary to clearly describe, inventory, analyze, and classify data. A data registry supports data sharing with cross-system and cross-organization descriptions of common data. Units of shareable data have precise identifiers, meanings, structures, and values.They are consistently deployed among users and systems and are centrally administered within an organization. Data registries may be organized into federations for interchange among many enterprises. A data registry assists users of shared data to have a common understanding of a unit of data's meaning, representation, and identification. Just as a data registry may assist users in understanding like units of data, registries also assist in the understanding of differences of similar but different units of data. (ISO/IEC 11179, Section 1.1)

A metadata registry is thus a data registry designed to hold the metadata that describes an organization's data.

An Open Forum on Metadata Registries was held in Washington D.C. in February, 1999. Another Forum will be held in Santa Fe in January 2000. These interesting meetings provide a way for organizations of all kinds to meet and discuss not only the development of the standards, but also the practical issues that drive the decisions. The list of topics that will be discussed at the next Open Forum can be seen at http://www.sdct.itl.nist.gov/~ftp/l8/sc32wg2/2000/events/openforum/index.htm.

2.2.1 The Census metadata registry model

The Census metadata registry model was developed by Dan Gillman of the Bureau of the Census. The Bureau of the Census and the BLS deal with very similar kinds of information,

indeed they jointly administer and draw from the CPS. In other words, these two agencies are already sharing data; the joint adoption of the registry model will enable them to continue and perhaps expand their cooperation. Jim Carpenter of the BLS has studied the model and agrees that it is appropriate for BLS data. Specifications of the model are available from Dan Gillman or Jim Carpenter.

2.2.2 Advantages

This section discusses the advantages of adopting a metadata registry modeled on the ISO/IEC 11179 standard in general, and the Census model in particular.

1. Since one of the goals of adopting a metadata registry is to increase data sharing among agencies (and presumably other entities), there is real advantage to following a standard that is commonly accepted in the knowledge management community. Among other entities, the Dublin Core is adopting the ISO/IEC 11179 representation for metadata elements (http://www.dstc.edu.au/RDU/DC-Agent/review2413.html). The Dublin Core is widely recognized as the standard structure for descriptive metadata. GILS (Government Information Locator Service) is also using 11179-compliant data definitions, and has a chart showing the equivalencies between its definitions and 11179 definitions (http://www.gils.net/element2.html). Statistics Canada, which has been working with its metadata registry for some time, also based it on this standard.

2. In adopting any standard or model for a technology that is intended to promote the sharing of data (e.g., EDI, network standards, etc.), there is a critical mass of adopters that must be reached to achieve real success. U. S. Government agencies have a growing awareness of the importance of sharing data, and the need for coordination among them in developing the means for doing so. The Bureau of the Census and the BLS have a head start in this area, since they share the CPS. By adopting the Census metadata registry model, they could serve as a model themselves for other agencies. Other countries have the same need for sharing data, and have developed registries based on ISO/IEC 11179. Canada and Australia, for example, can serve as an inspiration for U.S. agencies, demonstrating the advantages of such registries, and also serving as a source of ideas for how to carry out the adoption process.

3. The national and international community of researchers and practitioners in knowledge modeling is very active at this point. Many organizations have implemented metadata registries, and incorporated them into their day-to-day operations. Others have recognized the need for sharing data within themselves, requiring a coherent model – what is sometimes referred to as an enterprise model. When you add to this the need for sharing data across organizations, and thus either a more globally shared model or a means of "translating" between models, it is clear that organizations that do not have such models will be left out.

4. The Census model matches many of the information needs of the BLS. Although it may require some further tweaking or development of details, it is already in existence. There is no purpose in reinventing either the "standards" wheel or the "model" wheel.

Furthermore, there is existing expertise on the model in the BLS and the Bureau of the Census.

5. The Census model has a rich thesaurus/terminology component. This is recognized as a crucial area in the ISO/IEC 11179 standard, and has received (and will continue to receive) a great deal of attention in the Open Forum meetings. Terminology is an obvious and important access point to an organization's information. The model allows the links between concepts and terms to be associated with a particular context, which is a good mechanism by which to handle the ambiguity inherent in the BLS information (as illustrated by the *pay* concept family), as well as the ambiguity that is an unavoidable by-product of encouraging use of the BLS information by the public. As the standard states, "It is recognized that a component may have many names that will vary depending on discipline, locality, technology, etc." (ISO/IEC 11179, Section 7.2.1)

6. Every component of the model can be associated with an administering body or authority. This is an important feature in a model of shared data, where different components may be created, changed, or otherwise controlled by different people or programs. This means that the metadata registry can become the place to find out who is responsible for any particular aspect of the data.

7. Conceptually, the registry model can be viewed as a single unit, however "[t]he structure described by this metamodel may be distributed over several implementations. These implementations may be database, registries, repositories, dictionaries, etc." (ISO/IEC 11179, Section 7.0). This distinction contributes to its flexibility in both implementation and use.

8. The model clearly shows the connections between concepts (e.g., *pay*), formal terms and operational definitions of the concept (e.g., the CPS definition of "usual weekly earnings"), the data items that represent them (e.g., dollars/hour), and the source of the data items (e.g., a specific instance of a survey question). These connections serve many uses in creating and using the data, and it is useful to have them explicitly laid out in a shared registry. In this way, they are available to all who need them, from agency experts who depend on codes and skip patterns when interpreting data to the creators of online help for users looking for information via the Web site. This satisfies the criteria of concerning the representation of concepts, their related facets, operational data, processes and functions, and terminology.

9. The Census model satisfies Gruber's (1995) design criteria quite well. Clarity must be seen in light of the complexity of the data being represented – oversimplification causes as many problems as over-complication does. The specifications and notations used in the ISO/IEC 11179 standard are reasonably straightforward. I have given examples of its coherence above. The standard recognizes the need for local modifications or extensions to the model. The notion of a metadata registry itself is based on the need for minimal encoding bias and minimal ontological commitment. The registry provides the "definitions" of encoding and representation that is used in the organization. The standard states what those "definitions" should look like. The content is determined by the needs

of the organization. This helps in part with the criteria about populating and updating the registry. The Census model provides the organizational structure for the registry – the possibility that the actual contents may exist in a variety of forms and places should not be a problem.

2.2.3 Disadvantages

Once the commitment to adopting some form of knowledge structure has been made, there are few disadvantages of adopting the ISO/IEC11179 standard in general, and the Census model in particular. The effort needed to adopt any type of knowledge organization is like that associated with any change of infrastructure, and cannot be completely avoided. It is addressed in more detail in Section 2.3.

1. Although ISO/IEC 11179 is the standard with the most momentum behind it (to my knowledge), it is still undergoing some changes. For example, one open issue is that of cultural adaptability – what is the best way of incorporating multilingual/multicultural information into the model? As the standard is adopted by more organizations, and a better understanding of it spreads, it should be expected to evolve, exploring more peripheral areas. Therefore, it will be necessary to monitor the changes, and even contribute to the discussion surrounding them, both during the adoption period, and after. BLS already has interested people involved in the standards effort.

2. The metadata repository is like any other component of the information technology infrastructure in that it will require maintenance. In order to maintain its effectiveness, it must accurately describe the information produced by the agency. As definitions change, new versions of surveys are written, new data tables are created, or the responsibility for the data itself changes hands, the changes must be reflected in the registry. The registry itself must have designated "owners" who are responsible for its upkeep. In a successful adoption process, a point comes where the registry is the only place where changes are recorded – there is no duplication of data – and maintenance becomes somewhat easier. During the transition, however, there is likely to be duplication, requiring on vigilance to keep inconsistencies from creeping in.

3. As noted in section 2.1, flexibility is a vital characteristic of a metadata registry. This applies to the storage of metadata, but also applies to how the metadata can be used. In the database world, for example, businesses have been using databases for years to ease the storage and retrieval of records. Recently, however, the use of data mining and knowledge discovery techniques have created new uses for existing collections of records – deeper analysis for finding trends, recognizing patterns, and making predictions. Similarly, there are many obvious uses for a metadata registry today, but it is probable that as a registry becomes an integral part of an organization's information infrastructure, new uses will be discovered. A lack of clarity in the registry's structure or flexibility of access could impede such uses.

2.3 Remaining issues

The remaining issues primarily concern the technical and policy-oriented matters likely to arise in the adoption process. Success requires coordinated effort along several dimensions. The BLS must develop a solid understanding of what a metadata registry is and the advantages that will accrue from its adoption in order to get buy-in from pivotal people and departments.

In February 1999, Cathryn Dippo, Jim Carpenter, Dan Gillman (Census), Ron Graves (Statistics Canada), and I met to discuss possible adoption strategies. Crucial aspects of whatever approach is taken include these points.

- Education of BLS personnel regarding the uses of a metadata registry.
- Presentation of different views or versions of the Census model for different audiences, emphasizing aspects of special concern to them. Ron Graves noted the importance of this in the Statistics Canada adoption process.
- Development and presentation of use cases (see Section 3) illustrating the role that the registry can play in common processes. This report recommends use cases because they are flexible and easy to understand, however there are other means of "telling the story" that could be used, if preferred.
- Discussion of coordination between policy makers (e.g., scope of coverage, access, responsibility) and technical people, both within the BLS and across agencies, especially with Bureau of the Census.
- Identification of technical and policy "champions", who will help maintain the momentum of the adoption process.

Once the idea of the Census metadata registry has been accepted, the next step is to plan the implementation and deployment of the registry. As mentioned above, much of the required information already exists and can be pointed to from the registry. Items that may need further development include:

- The conceptual structures, terminology, and thesaurus. The example of the *pay* concept, with the accompanying agency and end user terminology (Hert & Haas, forthcoming) illustrates what this might encompass.
- Crosswalks mapping internal agency terminology with itself (in cases of ambiguity), with other agency's terms, and with external end user terms.
- Use cases. These are helpful to illustrate the possible roles of the registry and are discussed in Section 4.
- Presentation of and/or access to metadata information to various user populations. Internal BLS users may be more familiar with the contents of the registry, and also have more need for access to all its various parts (note that access is not the same as being the responsible authority for the data). External end users may benefit most from presentations based on the registry contents, such as a thesaurus, or term/concept mappings, for example, as help in searching for information.

3. Use Cases

A use case is a narrative, like a story, that describes what an information system should do to perform one of its functions, or to achieve a goal. More formally, "a *use case* is a description of a set of sequences of actions, including variants, that a system performs to yield an observable result of value to an actor"(Booch, Rumbaugh, & Jacobson, 1999, p. 222.). The collection of use cases for a system describes its repertoire of functions – what should happen in order to get things done. Use cases are a component of UML, the Unified Modeling Language, which is rapidly being adopted in the information industry as a way of describing information system specifications. (See http://www.rational.com/uml/index.jtmpl?borschtid=08223005192316484400 for further information on the UML. Booch et al., 1999 and Alhir, 1998 are also good sources.)

Use cases can describe sequences of actions at any level of generality. A set of use cases at different levels of generality can describe a function at a high level, and then break it down into more detail, e.g., different "versions" of that function. For example, one could have a general use case that describes the sequence of actions for accepting a user question and providing an answer. More specific use cases could then describe different kinds of questions, e.g., requests for data sets, requests for publications, requests for explanation or clarification of a term or concept, etc.

A use case describes *what* should happen, not *how* it should happen. A use case describes these aspects of the function.

- Who the actors are. An actor may be a person or a component of the system, such as a database or a metadata registry.
- What the actors have or know as they enter the process. For example, a user looking for a specific piece of information may know the appropriate agency term, or may know a common name for it.
- The expected sequence of events that should occur under normal conditions.
- Alternative sequences of events, such as what happens if an error occurs. For example, what should happen if the user enters a term that has no match in the database?
- What the results of the process should be. In the search example, the result could include a specific piece of information, or a pointer to the location where that information could be found (e.g., a url). Results could also include an opportunity for the user to modify the query and try again, or to obtain online help.

The basic form of the use case, that of a story or sequence of events, is a very approachable kind of model. Initial development depends more on familiarity with the functions and actors, and less with the details of how to make it work or exactly what the format of the data types should be. Similarly, use cases provide a good basis for "walking through" a procedure, determining what the possible actions are, what resources are needed, and identifying what system components already exist, and what needs to be developed. Alternative use cases can be developed for the same function, to promote discussion of their comparative advantages. Because of these characteristics, they are a good tool for informing people what the role of the metadata registry could be, how it relates to other components of the system, and how it could

benefit their own work. Use cases also provide a foundation for brainstorming, and developing new uses for the registry.

3.1 Sample use case: *Handle Web Queries*

This use case describes the actors and steps necessary for accepting queries from the BLS Web page and returning an answer. Two versions are presented, and the advantages of each are described. Note that this is not intended to be a final proposal, rather, it serves as an example of how a use case can be used to promote discussion of necessary structures and procedures, and illustrates a possible role of the metadata registry.

Actors:

1. Web site user.

   Since this use case describes handling queries, we'll omit the casual browser who has no particular information needs. If it seems necessary, more specific use cases could be developed around different types of users and user needs, for example, the frequent "power user" who knows what he or she wants and how to find it vs. the occasional user who may need more feedback or guidance.

   Characteristics:

   - Some level of searching skill (see above).
   - Some level of domain familiarity (see above).
   - Information need or question.
   - Terms or words that express it (to a greater or lesser extent).
   - Ability to recognize the answer (to a greater or lesser extent).
   - Ability to make use of the answer (to a greater or lesser extent), e.g., to interpret a table in an appropriate way. (Affecting this ability is somewhat out of the scope of this particular use case, but the same user will be an actor in another use case, which describes providing information, reference materials, etc. to users with questions about the information they've found.)

2. Database.

   This actor name is in the singular, and is used as though there is a single, monolithic database that contains all the data the agency produces. However, that is just for convenience; for the purposes of the use case, the precise implementation does not matter. Contents of this conceptual database include surveys, data files, news releases, publications, and the other "products" of the agency.

   Characteristics:

   - Assume that it is already in existence, and that no direct modification is necessary for implementing the metadata registry. Rather, the registry (or the services module described below) will provide mappings between data locations in the database, and descriptive material (metadata) that relates to it.
   - Assume predictable growth. With each new survey instance, publication, etc., there is more data in the database. This is in contrast to the registry, where the main type of growth (after its full development) will be adding pointers or mappings to new instances of known types, e.g., to the newest version of a survey question.

3. Metadata registry.

For the purposes of this use case, assume that this is the Census model (see Section 2). Characteristics:

- Since this is a use case describing user queries coming through the Web, it is possible that there are parts of the registry that are deemed "off limits" to the users.
- The parts likely to be most pertinent to this use case are the terminology/thesaurus, crosswalks, definitions of agency terms, descriptions of data and publications available to the public, and other resources for providing help. The name of the person or position who is the administrator for a particular data element or collection, for example, is less likely to be of interest to an outside end user.

4. The services module.

This module contains information about processes, procedures, mappings between the metadata registry and the database (possibly), and the other services associated with handling queries. Note that this may or may not become an actual system component – its contents may be distributed in any number of "containers". For the purpose of this use case, however, it is considered to be a single actor. (This is similar to the treatment of the database actor.)

This is where the two versions of this use case diverge. In Version 1, all actions between the user and the other components occurs through the services module (Figure 1). In Version 2, the user interacts directly with each actor (Figure 2).

In Version 1, the services module mediates all the actions involved in user querying. The database and registry are relatively static storage areas, and all procedural knowledge is in the services module.
Characteristics of Version 1.

- Contains procedures for accepting queries from end users, matching with appropriate resources (database or metadata registry), displaying results, giving help, and other user services.
- Contains mappings (e.g., metadata-to-filename) between registry and database.
- Can merge query results drawn from the database and the registry.
- Contains some level of diagnostic procedures for providing help. For example, if a query returns no results, it could offer to let the user browse through the thesaurus or show the appropriate part of the crosswalk (see Section 5) between end user terms and agency terms.
- Because it sits between the user and the database and registry, provides access control to them. (This could be seen as either an advantage or a disadvantage.)

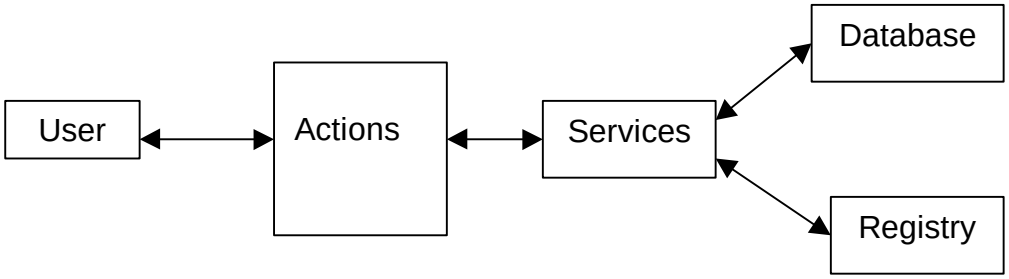


Figure 1: Version 1 of the *Handle Web Queries* use case.

In Version 2, the database and the registry are involved directly in the query actions, without depending on the services module. This might tend to fragment the procedural knowledge, especially between the registry and the services (if we assume that there will be no or only minimal changes to the database). For example, the registry might become more active in providing help to end users. I think I have a preference for Version 1.

Characteristics of Version 2.

- Contains procedures for accepting queries from end users, displaying results.

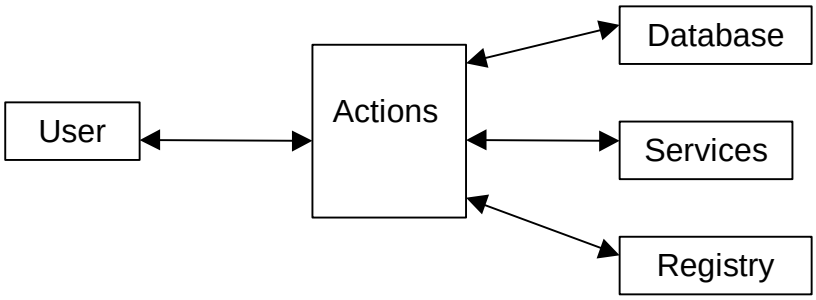


Figure 2: Version 2 of the *Handle Web Queries* use case.

Actions:

These actions are all part of handling user queries, but they do not always occur as a single sequence of steps. Some, like displaying the overview/map of the available data, may be requested as a stand-alone action by a user, with no subsequent query. The same action could also occur after a first attempt at a query. Others, like entering a query, viewing results, and then refining the query, are more likely to occur as a sequence. Note that there may be loops in the actions, reflecting loops in the user's querying process. I have assumed Version 1 of the services module.

- Display overview/map of data and/or collections. The overview is based on metadata registry information, but is mediated by the services module, in terms of the type of

display, the amount of detail, etc. This relates to Gary Marchionini's work, and also to Carol Hert's work on the A-Z index and the metadata that searchers use.

- Display definition and/or related terms of an agency term. This is based on metadata registry information, specifically the concept structure and thesaurus. The form of the display may be chosen via the services module.
- Display crosswalk of terms. This could be the crosswalk between end user (common) words and agency terms, which gives suggestions for other terms to use in a query, or it could be a crosswalk between different agencies' terms. The crosswalks are part of the registry, the display is controlled through the services module.
- Display more general information about the agency, its mission, the kinds of information that are available, etc. This information is probably in the database; as before, display is controlled through the services module.
- Accept query from end user. The services module can provide different kinds of queries, e.g., A-Z, keyword, restrictions on date, region, etc. This information comes from the registry. For example, the metadata might show that a query on wages can be qualified by region, sex, race, etc., but not by individual skill level. Other kinds of prompts, look-up tools, etc. can also be based on the registry.
- Execute query. The services module examines the mapping between the registry and the database to produce the result set.
- Present result set. The data itself comes from the database. The services module can have a default display, depending on the size of the result set and the type of result. In addition, it could present options to the user for how results should be displayed. The metadata registry plays a role here in determining what kinds of display are possible.
- Explain/expand result. The end user can select a term, a column or row heading, etc., and get some explanation of what it means. This is based in part on the definition/term display actions listed earlier.
- Refine query. The end user can modify the query based on the result set (either the size of the set or its content). The services module provides tools and/or suggestions or prompts for how to reformulate the query. For example, if the initial query included the word *pay*, the services module could draw upon the user-agency crosswalk to show the user the possible meanings, and allow him or her to select the appropriate meaning. If a result set is very large, the services module could suggest means of narrowing the query, for example by adding terms limiting region, sex, race, etc. Some information could be static, (e.g., what it means to combine terms with AND), and stored directly in the services module. Other information, such as possible ways of limiting a particular query, would be drawn from the registry. The actual submission of the refined query could resemble the initial submission.

3.2 Additional use cases.

In section 2.3, I discussed the role that use cases could play in informing agency personnel about the benefits of a metadata registry, as well as in the actual development of the registry and (in this example) the services module. The example use case presented here focuses primarily on the parts of the registry concerned with terminology, definitions, and the demographic categories for which data is available. Other use cases could focus on different parts of the registry. For example, the work that Carol Hert and John Bosley are doing

regarding the metadata that BLS experts use in interpreting the data could easily be made into a use case. I recommend that as part of the education effort, and as a means of encouraging brainstorming for other ways the metadata registry could be used, that more use cases be developed and circulated.

The use case presented here proposed the services module as an actor, which would house information for conducting interactions between end users, the metadata registry, and the database. If one assumes that (1) the database structure is quite stable, and it grows in a predictable manner, and (2) once the metadata registry is established it will also be fairly stable, there are great advantages to storing the potentially more volatile interaction services in something (or some things) resembling the services module. As we develop different ways of handling queries, displaying information, and providing help, the services could change regularly, and in some cases, dramatically. Good practice in system design would suggest that these changes should be kept separate from the more stable core of information. (Recall that having a single actor called the services module does not dictate that its components be physically housed in the same place.) Additional use cases should be developed that focus on the roles that services play in the BLS information system, and present alternatives to a services module.

4. The *pay* Concept Family

The *pay* concept family was selected as the focal point for the investigation of BLS concepts and terminology.

- It is a fairly complex concept in general, and one that is of interest to a wide community of users.
- 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 acknowledge within the BLS that different programs use slightly different definitions of the concept.
- Data associated with the *pay* concept is 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.
- Finally, this is a concept that is of interest to a wide range of users of the BLS and Fedstats web sites; 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.

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.

4.1 *Pay* data and metadata

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.

4.1.1. Publications

Publications examined included regular news releases, monthly and quarterly print publications, the *Handbook of Methods*, 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.

4.1.2 Names and definitions

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 terms which describe different aspects of *pay*.

- 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*).

Examples of these definitions from a variety of sources are given here.

The CPS collects data on "hourly wage rate and usual weekly earnings" from "currently employed wage-and-salary workers". (CPS history, http://www.bls.census.gov/cps/bhistory.htm) Definitions can be found in several places in the CPS documentation.

- "Usual weekly earnings. Data represent earnings before taxes and other deductions, and include any overtime pay, commissions, or tips usually received." (CPS concepts, http://www.bls.census.gov/cps/bconcept.htm)
- The CPS interviewer manual (http://www.bls.census.gov/cps/intmanb5.htm) defines *gross pay* as follows. "The total dollar amount usually received by the wage earner before deductions for federal/state income taxes, social security, union dues, et cetera…Include piece rate income as earnings…Also count college assistantships and fellowships and on the job training as earnings. Do not include pay in kind, such as food or lodging for work, or expense accounts as earnings."
- The same source defines *hourly earnings* as "the hourly rate as stated by the employer, expressed precisely in dollars and cents. It does not include tips, commissions, or any other non-hourly wages."

- Finally, the CPS Glossary (http://www.bls.census.gov/cps/bglosary.htm) contains this definition. "Income Sources – Wages and Salary. Money wages or salary is defined as total money earnings received for work performed as an employee during the income year. It includes wages, salary, Armed Forces pay, commissions, tips, piece-rate payments, and cash bonuses earned, before deductions are made for taxes, bonds, pensions, union dues, etc."

Note that the CPS concept of *income* differs from that of *pay* in that it includes unearned income such as rent, interest and so on.

The establishment survey collects information on "gross payrolls and the corresponding paid hours" from employers.

- "Aggregate payrolls include pay before deductions for Social Security, unemployment insurance, group insurance, withholding tax, salary reduction plans, bonds, and union dues. The payroll figures also include pay for overtime, shift premiums, holidays, vacations, and sick leave paid directly by the employer to employees for the pay period reported. They exclude bonuses, commissions, and other lump-sum payments (unless earned and paid regularly each pay period or month), or other pay not earned in the pay period concerned (e.g., retroactive pay). Tips and the value of free rent, fuel, meals, or other payment in kind are not included." (*Handbook of Methods*, Ch. 2)
- The Guide to Respondents for the NCS (http://www.bls.gov/special.requests/ocwc/oclt/ncsocs/ncs/ncbr0002.pdf) gives these instructions.

  "Record the current wage rate:

  Inclusions:
  - Straight-time hourly pay or salary
  - Cost-of-living allowances
  - Income deferred via 401(k)-type plans
  - Incentive plans including commissions, production bonuses, and piece rates

  Exclusions
  - Bonuses not tied directly to production
  - Shift differentials
  - Premium pay for overtime, holidays, weekends
  - Payments by third parties, e.g., tips"

The Occupational Compensation Survey overview includes this definition of the *pay data* it collects.

- "Pay data exclude premium pay for overtime and for work on weekends, holidays, and late shifts. Pay increases – but not bonuses – under cost-of-living allowance clauses and incentive payments, however, are included in the pay data." (http://www.bls.gov/ocsglanc.htm)

The Employee Cost Trends Program gives the following definition of wages and salaries.

- "Wages and salaries are defined as the hourly straight-time wage rate. For workers not paid on an hourly basis, straight-time earnings are divided by the hours worked. Straight-time wage and salary rates are total earnings before payroll deductions and include production bonuses, incentive pay, commissions, and cost-of-living allowances. Not

included in straight-time earnings are nonproduction bonuses, such as lump-sum payments provided in place of wage increases, shift differentials, and premium pay for overtime and weekend work; these payments are included in the benefits component.
Benefits include: paid leave – vacations, holidays, sick leave, and other leave; supplemental pay – premium pay for work in addition to the regular work schedule (such as overtime, weekends, and holidays), shift differentials, nonproduction bonuses, and lump sum payments provided in place of wage increases; insurance – life, health, short-term disability, and long-term disability; retirement and savings – defined benefit and defined contribution plans; legally required benefits—Social Security, Federal and State unemployment insurance, and Workers' Compensation; and other benefits—severance pay and supplemental unemployment benefits."
(http://www.bls.gov/news.release/ecec.tn.htm)

The technical notes for the Occupational Employment Statistics Survey (OES) gives this definition of *wages*.

- "Wages for the OES survey are straight-time, gross pay, exclusive of premium pay. Included are base rate, cost-of-living allowances, guaranteed pay, hazardous-duty pay, incentive pay including commissions and production bonuses, and on-call pay. Excluded are back pay, jury duty pay, overtime pay, severance pay, shift differentials, nonproduction bonuses, and tuition reimbursements."
(http://www.bls.gov/oes/oes_tec.htm)

The *Handbook of Methods* defines *total wages* in the context of unemployment insurance in the Covered Employment & Wages Program in this way.

- "Total wages, for purposes of the quarterly UI reports submitted by employers in private industry in most States, include gross wages and salaries, bonuses, stock options, tips and other gratuities, and the value of meals and lodging, where supplied. In some of the States, employer contributions to certain deferred compensation plans, such as 401(k) plans, are included in total wages. Total wages, however, do not include employer contributions to Old-age, Survivors'' and Disability Insurance (OASDI); health insurance; unemployment insurance; workers' compensation; and private pension and welfare funds.
…
For Federal workers, wages represent the gross amount of all payrolls for all pay periods paid within the quarter. This gross amount includes cash allowances and the cash equivalent of any type of remuneration. It includes all lump-sum payments for terminal leave, withholding taxes, and retirement deductions." (Ch. 5)

Finally, the Current Employment Statistics program provides this information about *payroll* in the CES technical notes.

- "The payroll is reported before deductions of any kind, e.g., for old-age and unemployment insurance, group insurance, withholding tax, bonds, or union dues; also included is pay for overtime, holidays, vacation, and sick leave paid directly by the firm. Bonuses (unless earned and paid regularly each pay period); other pay not earned in the pay period reported (e.g., retroactive pay); tips; and the value of free rent, fuel, meals, or other payment in kind are excluded. Employee benefits (such as health and other types of

insurance, contributions to retirement, etc., paid by the employer) are also excluded." (http://stats.bls.gov/cestn1.htm#CONCEPTS)

Clearly, there are many different ways of categorizing the components of the *pay* concept. For example, one categorization of *pay* concepts could be based on whether they are included in or excluded from definitions of income used in three surveys, the CPS, and the NCS. Table 1 lists these inclusions/exclusions as explicitly given in the BLS publications.

Table 1

Term	NCS	CPS
Base rate	I	I
Hourly rate	I	
Union rate, scale	I	
Apprentice rates	I	
Journey level rates	I	
Helpers' rate	I	
Probationary rate	I	
Beginner rate	I	
Entrance rate, hiring rate	I	
Pay-for-knowledge, skill-based pay, knowledge-based pay, multi-skill compensation	I	
Educational pay differential	I	
Flat rate	I	
Guaranteed rate	I	
Piece rate	I	I
Stint rate	I	
Superannuated rate	I	
Temporary rates, experimental rate, trial rate	I	
Tonnage rate	I	
Blue circle rate	I	
Flagged rate, red circle rate, out of line	I	
Salary	I	I
Straight-time earnings	I	
Production bonus	I	
Commission, commission payment	I	I
At-risk pay	I	
Incentive earnings	I	
Call-in pay, reporting pay	I	
Cost of living adjustment	I	
Hazard pay	I	
Longevity pay		
Portal to portal pay		
Deadhead pay	I	
Nonproduction bonus	E	
Attendance bonus		
Back pay		
Christmas bonus	E	
Profit-sharing, cast profit-sharing	E	

Dismissal pay, severance pay		
Premium pay	E	
Supplemental pay	E	
Double time	E	
Overtime	E	I
Shift differential, shift premium	E	
Holiday pay	E	
Holiday premium pay	E	
Penalty rate	E	
Moving allowance, relocation allowance		
Per diem allowance		
Subsistence allowance		
Referral bonus	E	
Tool allowance	E	
Vacation pay		
Draw account	E	
Tips	E	I
Uniform allowance	E	
Jury duty pay		
Holiday bonus		
Tuition reimbursements		
Push money	E	
Profit-sharing distributions	E	
Stock bonus	E	
Free room & board		

Table 1: Terms associated with the *pay* concept family, and whether they are included in the definition of *income* for the CPS and NCS. *I* indicates explicit inclusion, *E* indicates explicit exclusion, and no entry indicates no explicit mention was found.

Based on the CPS definition, I designed a class hierarchy. An overview version is shown in Figure 3. A full version is given in Appendix A. The classes and subclasses in the hierarchy can be used in at least two ways.

- This hierarchy could be a component of the metadata registry proposed in Section 2. Note that the registry can represent multiple "views" of the same (or closely related) families of concepts, which accurately reflects the situation within the BLS itself, and among agencies. This is accomplished by associating a concept with the context in which it is used. (See Section 7.2.1 of the ISO/IEC 11179 standard.)
- The hierarchy can also aid in helping users of BLS data understand how terms are defined by the agency (e.g., in titles or column/row headers of tables), and how meanings may differ slightly from table to table. In essence, it can become the basis of a terminology crosswalk (see Section 5).

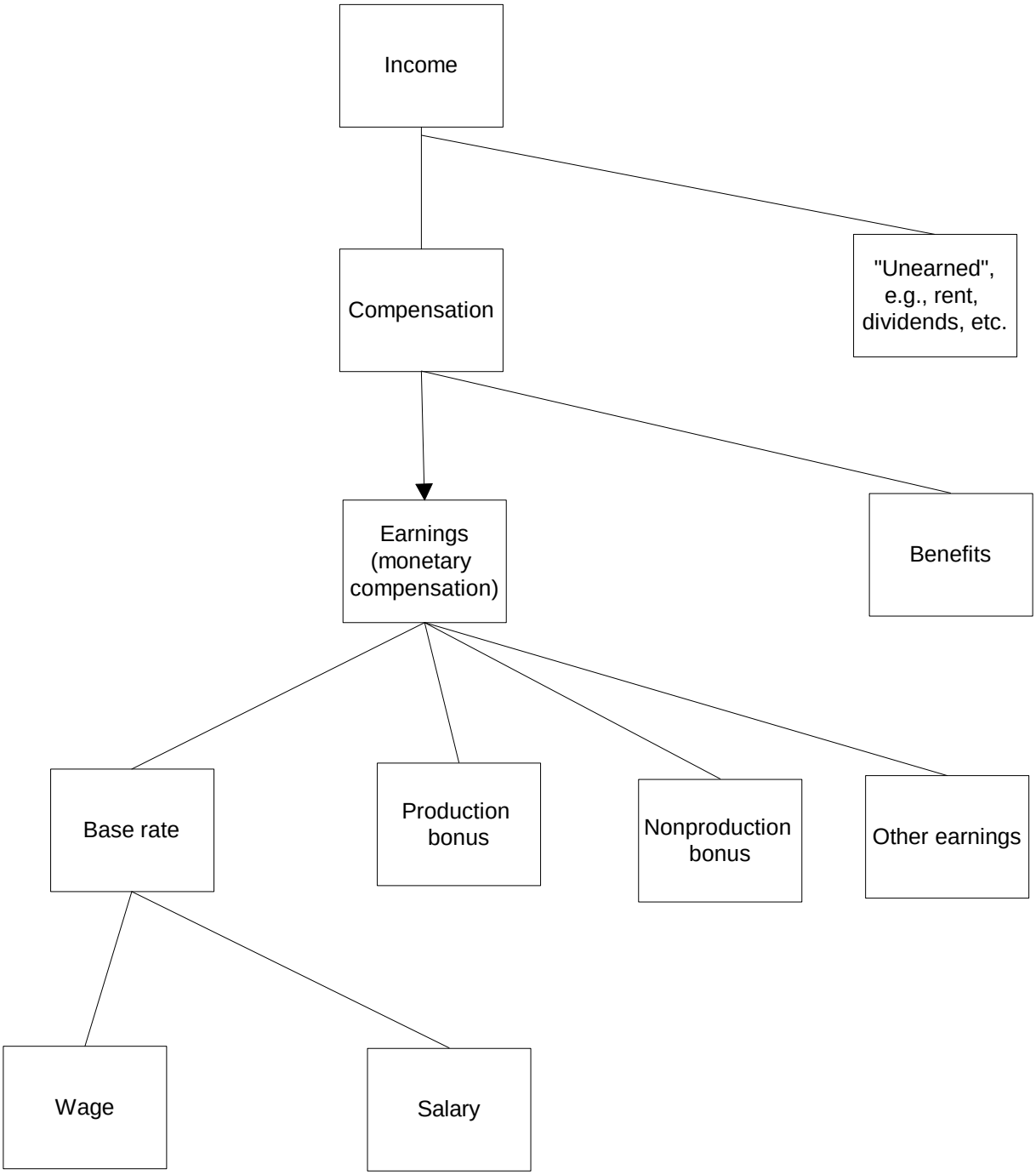


Figure 3: Class hierarchy for the *pay* concept family, based on CPS definitions and usage.

4.1.3. Information facets associated with the *pay* concept family

The information facets associated with the *pay* concept family represent the ways in which the concept is used (modified, measured, manipulated, etc.) in the BLS. Many facets were discovered in BLS publications and in conversation with agency personnel, as well as with others who use BLS information. Many of these facets are incorporated into the metadata registry itself, such as the *organization and contact person in charge of an administered component*. Others can be thought of as concepts in their own right, such as *geographic location*, or *index*. The list given in Figure 4 should not be considered complete at this point; among other things, the work that Carol Hert and Gary Marchionini are doing regarding the presentation of information to outside users should be incorporated. In this list, the facets are roughly divided into two groups: those associated with gathering and publishing the data (administration) and those associated with the end use of the data. Obviously, there is some overlap between these two groups.

Data administration
- Data source - survey, questions, skip patterns.
- Sources of error
- Sampling techniques
- Publications
- Domain of values
- Change history (of definitions, procedures, etc.)
- Stewardship/administration

Data end use
- Location, e.g., from which data was gathered, to which it applies, regions
- Time – point, e.g., at which data is released publicly, at which it was gathered, and
  span e.g., during which data was gathered, which time series covers
- Functions to be applied to data, e.g., aggregation, imputation, seasonal adjustment
- Demographic categories
- Terminology, thesaural entries

Figure 4: Facets associated with the *pay* concept family drawn from BLS publications.

In addition, a preliminary list of descriptive facets associated with *pay* terms by users of the Fedstats Web page (http://www.fedstats.gov/) was gathered from the 11/98 collection of queries submitted to the search engine. Figure 5 shows a categorized list of these facets. Note the overlap with the facets gathered from BLS publications, especially those categorized as end-use facets. Again, Carol Hert's work with outside users as well as BLS experts should also be examined in this light. <*Cite her work with John Bosley*.>

Statistical concepts.
index, mean, average, median, projection, growth, trend, percentage, per capita

Time concepts
current, 1992 (specific year), hourly, annual,
NOTE: some of these terms are identifiers of "which edition" of the statistic is wanted (e.g., specific year). The others identify the units in which the data should be reported (annual, hourly).

Geographical concepts
U.S., alabama, florida, dc, california, Los Angeles, boston, las vegas local area, cities, county, state, foreign countries, area

Demographic concepts.
Gender: gender, women, {men, male}
Age: age, elderly
Life stage: retirement, disability
Marital status: single, divorce
Race: race, {black population, african american}
National origin: vietnamese american
Education: {education, education level, educational and attainment}, college graduate
Occupation: {occupation, job,},
General type of industry: by industry, business<??>, non-profit, executive, congressional, federal, hotel, construction, farm, medical,
Specific job titles: physician, pediatrician, dentists, correction officer, attourney, {president, presidential}, librarian, draftsperson, computer operators, accountants

Figure 5: Categorized list of facets associated with the *pay* concept family drawn from user queries. Example terms are shown for each category, with synonyms or near synonyms in braces.

4.2. Summary

In this presentation of the *pay* concept, I have used two main sources: the BLS publications and experts, and input from users outside the agency. Characteristics of this concept family illustrate several important notions about such concept representations.

- It is not uncommon for technical terms, even within a single organization, to have multiple, contradictory definitions. This may occur because of historical reasons, different sources of the term and its associated data, or differences between "casual" and "precise" usage. Therefore, although it may be desirable to have one "authoritative" definition, it is probably unrealistic to expend much effort toward that end. Rather, the conceptual representation in the metadata registry must capture the multiple meanings and uses, identify the contexts in which various definitions hold, and draw the attention of the users to the potential ambiguity.
- Similarly, users outside of the agency are also likely to have different ideas of names and definitions for technical concepts. They may be working with definitions that are technically precise and correct in another context, or with "commonsense" or "naïve"

definitions. Either way, the registry should be able to support multiple views of concepts. In this situation, routines (e.g., stored in the services module) based on the representation could help users recognize the ambiguity or imprecision, and guide them to an understanding that is a better match for the authoritative definition(s).

- The BLS is fortunate to have an extensive collection of publications in which definitions already exist. The effort to populate the registry will be expended more in finding and organizing links to existing definitions than in creating definitions for existing terms.
- As we suspected at the outset, the *pay* concept family is an interesting one. It has many components and associated terms, it is somewhat ambiguous, and there are clear differences between a naïve, common definition and the technical definitions used in determining what data is included in the related statistics produced by the BLS. It might be interesting to contrast it with a less commonly-used, more specialized concept, such as *seasonal adjustment* or *consumer price index*. If these are indeed more precise terms, one would expect to find fewer definitions within the agency, and fewer terms associated with it both inside and outside of the agency. If outside users are familiar with the concept at all, they may know only a single name for it. Their understanding of its meaning may or may not be deep (or even correct), but there may be fewer meanings floating around.
- The conceptual structure proposed for the *pay* concept in Figure 3 fits easily into the ISO/IEC 11179 structure, and therefore into the Census model. The concept structure, along with the definitions and contexts in which they are valid are a useful part of the standard.

5. Terminology and Crosswalks

This part of the report focuses on terminological issues as they relate to public access to BLS information, and the role the metadata registry can play in supporting access. There are two related efforts involved. The first concerns work done with Dr. Carol Hert, which is presented in detail in her report (Hert, 1999) and in our forthcoming article. The second involves the Open Forum on Metadata Registries, whose participants have a strong interest in terminologies and mapping among them.

The theme that is common to both of these efforts is the recognition that the terminology used by experts in a domain, e.g., by people within the BLS and by statisticians and other experts who regularly use BLS data, is not necessarily the same as that used by non-experts, e.g., the average citizens who use the BLS web page to find information. The importance of this issue for easing access to BLS resources for all citizens must be emphasized – if someone does not know the "correct name" of something, it can be very difficult, if not impossible, to find.

As illustrated in the talk I gave at the Open Forum in February, 1999 (Haas, 1999; see Appendix B or ftp://sdct-sunsrv1.ncsl.nist.gov/x3l8/sc32wg2/openforum/A5.htm for the slides from it), there are several possible relationships that can exist between the end user terminology and the agency terminology:

- Same word(s), same meaning
- Same word(s), different meaning

- Same word(s) some relationship between meanings (e.g., broader term, narrower term, part-of, or a domain-specific relationship)
- Different word(s) same meaning
- Different word(s), different meaning
- Different word(s), some relationship between meanings (e.g., broader term, narrower term, part-of, or a domain-specific relationship)
- Different word(s), some overlap in meaning

This theoretical set of relationships is somewhat complicated in situations where the "official" terms themselves have some ambiguity, as happens with the *pay* concept family in the BLS. The existence of these different terminologies (e.g., end user and agency) is well established. The challenge is to recognize what relationships, or mappings, exist among terms, how and where to organize and store this information, and how best to deploy it to help end users find the information they want.

5.1 The terminology research

Hert & Haas (forthcoming) reports on a project which had the intent of exploring the relationship between user terminology for a concept (as represented in a search engine's log) and the terminology employed by a specialized agency (i.e., BLS). The specific objectives were:

5. To determine the extent of the overlap between agency (BLS) terminology for the concept of *pay* with user terminology for the same concept as identified in user inputs in a search engine.
6. To determine the extent of the overlap between agency terminology expanded with related terms from two thesauri (WordNet and Websters) for the same concept with user inputs.
7. To compare the extent of the two overlaps
8. To consider the feasibility of this expansion approach for automatically enhancing agency terminology and/or user queries.

The list of *pay* terms was drawn from agency publications; a list is shown in Table 1. The user terms were taken from the FedStats search engine log containing inputs for the month of November, 1998. Hert (1999) and Hert & Haas (forthcoming) give details regarding the processing of the search logs to extract the user searches.

In information retrieval, *query expansion* is often performed to increase the probability of finding documents relevant to the query. This process is motivated by the same conditions we find with BLS terminology. The words originally used in the query by the end user, and those used in the documents or document surrogates may not match. In domains where there is a technical terminology with specific definitions, the overlap of authority terms with words used in a query by a non-expert user may be quite small. If the terms in the query are expanded, usually by "OR-ing" synonyms together, then there is a greater likelihood of finding a match.

WordNet (Fellbaum, 1998; http://www.cogsci.princeton.edu/~wn/) is a structured thesaurus developed by George Miller and his colleagues that has been used in a number of information retrieval projects. Different parts of speech (nouns, verbs, adjectives, adverbs) have different

information structures in the thesaurus. I will describe just the noun structures, since terms associated with the *pay* concept were nouns. Each noun is divided into *senses*, one for each major meaning of the word. For example, *lemon* has four senses, representing the fruit, the color, the tree, and the flavor. Each word sense is associated with a *synset*, a set of words that mean the same thing (in at least one of their senses). The synset for *pay* contains *wage, pay, earnings, remuneration,* and *salary*. Another word associated with the *pay* concept in BLS terminology is *compensation*. In WordNet, there are three senses given for compensation, whose synset (for the appropriate sense) contains only *recompense*.
Nouns are organized into hierarchies of broader and narrower terms (*hypernyms* and *hyponyms*, respectively). Because the WordNet hierarchy is fairly shallow, there can be dramatic changes in meaning when moving up or down the hierarchy. For example, the hypernym of *wage* is *regular payment*, and wage's hypernym list includes *minimum wage* and *sick pay*. *Coordinate terms* are sibling terms; children of the same parent. E.g., the coordinate terms of *wage* (other kinds of *regular* payment) include *stipend*, but also *installment plan*.

Voorhees (1994) describes using WordNet for expanding text queries for the TREC conferences. She found two interesting results. First, query expansion using WordNet was more effective when the initial query was very brief. Expansion gave little benefit when the query already contained many terms. This implies that query expansion may be a useful strategy for queries coming into the BLS via the Web page, since they tend to be very brief. The second thing that Voorhees found, which has been confirmed by many researchers since, is that expanding a query using a general purpose thesaurus must be done in very small steps, or else the new terms can easily start leading the query in very inappropriate directions. Suppose word W1 is included in the initial query. An word in its synset, S1, may be a useful addition to the query. A synonym of that synonym, S2, may mean something rather different from the original W1, and taking further steps (e.g., a synonym of the synonym of the synonym, S3) will end up in a completely different part of the "meaning space". Similarly, adding a narrower term associated with W1 may be a useful strategy, but adding a broader term may bring in meanings that are not helpful at all.

Aslandogan et al. (1997) also used WordNet to aid retrieval in an image database, expanding both the queries and the metadata associated with the images. They dealt with the term overgeneration problem by ensuring that expanded terms were in the same general category (e.g., *artifact, animal*, etc.) as the original term. They found that using synonyms of all senses, or synonyms, broader terms, and narrower terms (hypernyms and holonyms) improved retrieval performance slightly over using no expansion.

In the case of the BLS *pay* terms, the words themselves are generally common words, but they are often used in specific ways (see Section 4). 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. Terms relating to the *pay* concept ("agency terms") were gathered from BLS publications and data presentations (see Table 1).

Expansion terms were gathered from the Web-accessible version of Wordnet 1.6 (http://www.cogsci.princeton.edu/~wn/).

- 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 hypernym 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.

The online version of Websters thesaurus is also a general language thesaurus, and was used to provide additional expansions. It is not as highly structured as WordNet. Synonyms were selected from those agency terms that had entries. Terms that duplicated those already found in WordNet were eliminated.

Table 2 shows the agency terms and their possible expansions found in WordNet. Table 3 shows the agency terms and their possible expansion found in Websters.

Many of the agency terms, especially the multi-word ones, did not have entries in either thesaurus. This is not surprising, since we were looking for technical terms in general-purpose thesauri. Another approach to query expansion uses specialized domain thesauri as a source for related terms. For example, Srinivason (1996), discusses the expansion of MEDLINE queries using the MeSH thesaurus. Once the thesaurus section of the BLS metadata registry is developed, this could also be investigated as a source of additional terms.

There are three ways that query expansion can be done.

1. Completely automatically. The retrieval system generates additional terms based on terms in the initial query and includes them in the query with no intervention by the user.
2. Semi-automatically, also known as “interactive expansion”. The retrieval system generates additional terms based on terms in the initial query and presents them to the user. The user then chooses which of the expansion terms should be included in the query.
3. Completely by hand. The user must think up alternative terms to include in the query.

Automatic expansion is a very attractive idea, but can result in the system retrieving documents that are far afield from the what the user was looking for. Interactive expansion, or “suggestion” gives the user more control over what the retrieval system is doing.

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. What we were really attempting to do was build a crosswalk between agency terminology and user terms (see Section 5.2), which could be stored permanantly in the BLS metadata registry or the services module (or its equivalent). In addition, some of the terms that were found expressed concepts that were quite different from the BLS *pay* concept family. For example, the term "blood money" is given as a hyponym of *payment* in WordNet. As a result, we decided to use the interactive expansion model, choosing terms from those suggested by the two thesauri.

Details and results of this research will be presented in Hert & Haas (forthcoming).

Table 2: Agency terms and their hypernyms (broader terms), synsets (synonyms), and hyponyms (narrower 1.6. Sense number are shown after agency terms that had more than one entry. Terms that seem especially p italics; terms that seem less than helpful are lined through. A hypen means there was no entry.

Agency Term	Hypernym	Synset	
Wage (subclass of base rate)			
wage	regular payment	*pay, earnings, remuneration, salary*	*strike pa half-pay minimu pay pac home p*
hourly rate	-	-	-
rate 1 <added>	charge	charge per unit	pay rate rate, rat ~~repaym~~ ~~rate, ta~~ ~~of intere~~ ~~freight,~~ ~~rate of~~ ~~depreci~~ ~~exchan~~ ~~excursi~~ ~~lineage~~
union rate	-	-	-
scale	~~standard, criterion, measure~~, ~~touchstone~~	scale of measurement, graduated table, ordered series	index, ~~l~~ scale, w ~~scale, w~~
apprentice rates	-	-	-
journey level rate	-	-	-
helpers' rate	-	-	-
probationery rate	-	-	-
beginner rate	-	-	-
entrance rate	-	-	-
hiring rate	-	-	-

Table 2, continued.

pay <added> <same synset as wage>	regular payment	*wage, pay, earnings, remuneration, salary*	*strike pa half-pay minimu pay pac home p*
pay-for-knowledge	-	-	-
skill-based pay	-	-	-
knowledge-based pay	-	-	-
multiskill compensation	-	-	-
educational pay differential	-	-	-
flat rate-	-	-	-
guaranteed rate	-	-	-
piece rate	-	-	-
stint work	-	-	-
superannuated rate	-	-	-
temporary rates	-	-	-
experimental rate	-	-	-
trial rate	-	-	-
tonnage rate	-	-	-
blue circle rate	-	-	-
flagged rate	-	-	-
red circle rate	-	-	-
out of line rate	-	-	-
Salary (subclass of base rate)			
salary <same synset as wage>	regular payment	*wage, pay, earnings, remuneration, salary*	*strike pa half-pay minimu pay pac home p*
straight-time earnings	-	-	-
Base Rate (subclass of earnings)			
Production bonus (subclass of earnings)			

Table 2, continued

production bonus	-	-	-
bonus 2 <added>	payment	*bonus, incentive*	~~windfall~~ ~~godsen~~
commission 2	fee	-	-
commission payment	-	-	-
payment <added>	cost	-	*royalty;* ~~overpay~~ ~~subscri~~ ~~blood m~~ ~~refund;~~ ~~support~~ ~~money;~~ ~~benefit;~~ ~~remittar~~ ~~repaym~~ ~~paymen~~ ~~premiur~~ ~~installm~~ ~~deposit;~~ ~~paymen~~ ~~paymen~~
at-risk pay	-	-	-
incentive earnings	-	-	-
incentive 2 <same synset as bonus>	payment	*bonus, incentive*	~~windfall~~ ~~godsen~~
call-in pay	-	-	-
reporting pay	-	-	-
cost of living adjustment	-	-	-
adjustment 2<added>	recompense	*allowance*, adjustment	*cost-of-* depreci deducti *adjustm*
cost-of-living allowance <added>	*allowance*, adjustment	-	-
hardship allowance	-	-	-
allowance 1 <added>	share, portion, percentage	-	~~privy pu~~
allowance 2 <added>	reimbursement	-	travel al

Table 2, continued

allowance 3 <added>	recompense	allowance, adjustment	*cost-of-* depreci deducti *adjustm*
hazard pay	-	-	-
high time pay	-	-	-
longevity pay	-	-	-
portal to portal pay-	-	-	-
payments for income deferred due to participation in a salary reduction plan			
deadhead pay	-	-	-
Nonproduction bonus (subclass of earnings)			
nonproduction bonus	-	-	-
attendance bonus	-	-	-
back pay	-	-	-
bereavement pay	-	-	-
profit-sharing	share, portion, part, percentage	-	-
cash profit-sharing	-	-	-
year-end bonus	-	-	-
dismissal pay	-	-	-
severance pay	-	-	-
premium pay	-	-	-
supplemental pay	-	-	-
double time	*wage, pay, earnings, remuneration, salary*	-	-
overtime	-	-	-
shift differential	-	-	-
shift premium	-	-	-
holiday premium pay	-	-	-
penalty rate	-	-	-
Moving allowance	-	-	-
relocation allowance	-	-	-
paid absence allowance	-	-	-

Table 2, continued

Other earnings (subclass of earnings)			
per diem allowance	-	-	-
subsistence allowance	-	-	-
referral bonus	-	-	-
tool allowance	-	-	-
vacation pay	-	-	-
pay in lieu of vacation	-	-	-
draw account	-	-	-
tips 2	fringe benefit, perquisite, perk	*gratuity, tip*, ~~baksheesh~~, ~~bakshish, bakshis, backsheesh~~	~~Christm~~
uniform allowance	-	-	-
holiday bonus	-	-	-
tuition reimbursements	-	-	-
reimbursement <added>	*compensation*	-	allowan
bilingual pay differential	-	-	-
safety bonus	-	-	-
contract-signing bonus	-	-	-
make-up pay	-	-	-
push money	-	-	-
retroactive pay	-	-	-
royalty 1	payment	-	-
profit-sharing distributions	-	-	-
free room & board	-	-	-
Earnings (subclass of compensation, monetary meaning)			
earnings (monetary only) 1.	income	*net income, net*, net profit, lucre, profit, profits, *earnings*	earning profit; ki profit, g margin; dividenc

Table 2, continued

earnings 2. <same synset as wage>	regular payment	*wage, pay, earnings, remuneration, salary*	*strike pa half-pay minimu pay pac home p*
Benefits (subclass of compensation)			
Note: as of 5/18/99, we aren't working with benefit terms/concepts)			
Compensation (subclass of income)			
compensation	recompense	-	overcon *compen* emolum ~~damage~~ ~~indemni~~ ~~redress~~ ~~reparati~~
Unearned income (subclass of income)			
Note: includes rent, pensions, etc.			
unearned income 1	income	unearned income, unearned revenue	-
Income			
income	financial gain	-	*disposa* ~~money,~~ revenue net, net profits, proceed payoff; revenue unearne revenue governn *capita i*

Table 3: Agency terms and synonyms found in Websters. Potentially useful terms not already found in WordNet are in italics.

Agency Term	Synonym
Wage (subclass of base rate)	
wage	-
hourly rate	-
rate <added, WN>	Scale, fixed amount, valuation, allowance
union rate	-
scale <only other senses>	-
apprentice rates	-
journey level rate	-
helpers' rate	-
probationery rate	-
beginner rate	-
entrance rate	-
hiring rate	-
pay <added, WN>	Profit, proceeds, interest, return, *recompense*, indemnity, reparation, rake-off, reward, perquisite, consideration, defrayment, compensation, salary, payment, *hire*, *remuneration*, commission, fee, *stipend*, earnings, settlement, consideration, reimbursement, *recompensation*, reckoning, satisfaction, *honorarium*, emolument, time, time and a half, double time, overtime
pay-for-knowledge	-
skill-based pay	-
knowledge-based pay	-
multiskill compensation	-
educational pay differential	-
flat rate-	-
guaranteed rate	-
piece rate	-
stint work	-
superannuated rate	-
temporary rates	-
experimental rate	-
trial rate	-
tonnage rate	-
blue circle rate	-
flagged rate	-
red circle rate	-
out of line rate	-

Table 3, continued

Salary (subclass of base rate)	
salary	Wage, wages, reompense, *payroll* (?)
straight-time earnings	-
Base Rate (subclass of earnings)	
Production bonus (subclass of earnings)	
production bonus	-
bonus <added, WN>	Gratuity, reward, special or additional compensation
commission	Remuneration, royalty, salary, *fee*, stipend, indemnity, rake-off
commission payment	-
payment <added, WN>	Recompense, reimbursement, restitution, subsidy, return, redress, refund, remittance, reparation, disbursement, cash, salary, wage, fee, sum, pay-off, repayment, indemnification, requital, defrayment
at-risk pay	-
incentive earnings	-
incentive	-
call-in pay	-
reporting pay	-
cost of living adjustment	-
adjustment <added, WN>	-
cost-of-living allowance <added WN>	-
hardship allowance	-
allowance 1 <added, WN> <only 1 sense>	Salary, wage, commission, fee, recompense, hire, quarterage, pittance, recompense, stipend, pernsion, settled rate, bounty
allowance 2 <added, WN>	-
allowance 3 <added, WN>	-
hazard pay	-
high time pay	-
longevity pay	-
portal to portal pay-	-
payments for income deferred due to participation in a salary reduction plan	
deadhead pay	-
Nonproduction bonus (subclass of earnings)	
nonproduction bonus	-
attendance bonus	-
back pay	-
bereavement pay	-
profit-sharing	-
cash profit-sharing	-

Table 3, continued

year-end bonus	-
dismissal pay	-
severance pay	Pittance, stipend, allotment
premium pay	-
supplemental pay	-
double time	-
overtime	-
shift differential	-
shift premium	-
holiday premium pay	-
penalty rate	-
Moving allowance	-
relocation allowance	-
paid absence allowance	-
Other earnings (subclass of earnings)	
per diem allowance	-
subsistence allowance	-
referral bonus	-
tool allowance	-
vacation pay	-
pay in lieu of vacation	-
draw account	-
tip	Reward, gift, compensation, fee, lagniappe, pourboire, handout
uniform allowance	-
holiday bonus	-
tuition reimbursements	-
reimbursement <added, WN>	Compensation, restitution, recompense
bilingual pay differential	-
safety bonus	-
contract-signing bonus	-
make-up pay	-
push money	-
retroactive pay	-
royalty	-
profit-sharing distributions	-
free room & board	-
Earnings (subclass of compensation, monetary meaning)	
earnings (monetary only) .	Net proceeds, balance, receipts
earnings	-
Benefits (subclass of compensation)	
Note: as of 5/18/99, we aren't working with benefit terms/concepts)	

Table 3, continued

Compensation (subclass of income)	
compensation	Remuneration, recompense, indemnity, remittal, buyback, commission, gratuity, reimbursement, allowance, salary, stipend, wages, hire, earnings, settlement, honorarium, defrayment, fee, quittance, reckoning, bonus, premium, profit
Unearned income (subclass of income)	
Note: includes rent, pensions, etc.	
unearned income	-
Income	
income	Earnings, salary, wages, livelihood, returns, profit, dividends, assets, proceeds, benefits, receipts, gains, commission, drawings, royalty, honorarium, income after taxes, net income, gross income, taxable income, bottom line, cash, pickings, take

5.2 Terminology Crosswalks

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. Crosswalks can represent mappings between two formal terminologies, or between a formal terminology and general language, as is used in end user queries. 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.

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*). In a crosswalk, the relationships are between different terminologies. Betsy Humphreys (1999), in her talk for the Open Forum, gave a very interesting presentation on the difficulties that can be encountered in constructing the UMLS Metathesaurus, a crosswalk of medical terminologies, even when the terminologies involved cover essentially the same subject domain.

This distinction between an authoritative thesaurus and a crosswalk introduces an important issue for the contents of the metadata registry. Where and how are crosswalks represented? Participants at the 1999 Open Forum agreed on the need to maintain some level of separation between the "official terminology", and words or terms that may be used to access information from outside. In the use case presented in Section 3, there are two possibilities for locating the BLS-end user crosswalk; in the registry itself, or in the services module. If it were in the registry, it could either be a separate conceptual structure, labeled with its context of use (i.e., user queries), or it could be incorporated into the thesaurus, but clearly marked as "user synonyms" in some way. If it were in the services module, there wouldn't be any difficulty in keeping it separate. Another advantage to putting it in the services module would be the ability to tightly couple it with the help facilities. Further, there is some evidence that the set of words that non-experts user for technical concepts may change relatively quickly, and therefore need frequent updating. If there is an assumption that the registry is relatively stable, and the services module is more volatile, that would be another reason for including the crosswalk there. This topic is expected to be of continuing interest in the coming year, and one that I intend to address in my talk at the 2000 Open Forum.

6.0 Conclusions

6.1 Recommendations and future work

This section summarizes the recommendations resulting from this work along with some suggestions for future work to support their implementation.

1. The adoption of the ISO/IEC 11179 Specification and Standardization of Data Elements standard, as instantiated by Dan Gillman's Census metadata registry model. The standard and

the model both satisfy the majority of the criteria identified for a knowledge structure for BLS information. Fundamental to the successful adoption of the standard and the model is the development of a solid understanding of what a metadata registry is and the advantages in its use. Other policy and technical issues include getting buy-in from pivotal people and departments, and the coordination of effort between policy makers and technical managers. Publicity and educational efforts could include:

- Observing examples of successful adoption in peer organizations, e.g., Statistics Canada.
- Developing and circulating use cases, as a means of (1) illustrating the purposes of a metadata registry, (2) identifying information that must be generated for populating the registry, and (3) generating ideas for fitting the registry into the BLS information systems as seamlessly as possible.
- Encouraging continued participation in national and international knowledge management fora.
- Planning for the incremental development and deployment of the registry.

2. The development of use cases. As noted above, they can serve many purposes in the metadata registry adoption process. In addition, they can be a valuable tool in developing a variety of help facilities to aid in citizen access and use of BLS data, especially from its web site. Candidate use cases include:

- Handling web queries from a different types of users
- Handling web queries for different kinds of data or resources (e.g., a request for an index or table, a request for a specific publication, a request for "information on a topic", or a request for an explanation of previously retrieved information.
- Those focusing on different parts of the registry, e.g., the use of metadata by BLS experts.

3. The determination of where and in what format various kinds of user services should be housed. The use case presented in this report proposed a services module which would contain helps and prompts for searchers, aids to understanding BLS concepts and terms, formatting and presentation tools for retrieved information, and tools to interface between the database and the registry. This is not the only solution, and other possibilities should be explored. This would require some investigation of the current BLS information system structure, and how it would be modified with the adoption of a metadata registry.

4. Concept structure and terminological issues. The *pay* concept family clearly illustrated the complexity that attempts to harmonize a single concept can uncover. Yet developing a good understanding of BLS concepts and terms is a vital part of implementing the metadata registry, and also of improving citizen access to and understanding of BLS information. Several activities could be pursued to support this.

- Continue studying the overlap between BLS terminology and end user terms. This will lead to a better understanding of the problems that citizens have in finding the information they need, and therefore to a better foundation for providing various kinds of search tools and help facilities.
- Examine different types of concepts and terms. The *pay* concept family is familiar to most people, even if their understanding of it may differ somewhat from that of BLS experts. Further, the terms associated with the concept are also common. More "technical" concepts, such as *seasonal adjustment* or *consumer price index* may have

different characteristics of use and understandings. Studying associated concepts, especially those that can appear in a modification relationship, is an important part of this examination, since these findings can contribute to the development of improved searching and help facilities.

- Study the overlap between BLS concepts and terms, and those used in other government agencies. This is an additional step toward data sharing among agencies. The *pay* concept family is a good candidate for this expanded effort, in that it comes into the domain of many different agencies.
- Develop terminology crosswalks for BLS-end user terminologies and BLS-other agency terminologies. Crosswalks can serve as the basis for many kinds of end user help services.

6.2 Dissemination of results.

The following presentations and articles have been delivered or are planned.

Haas, Stephanie W. (1999) "What Everyone Calls It": Mapping Users' Words and Terminologies. Invited talk, Open Forum on Metadata Registries, February 16-19, 1999, Washington, D.C.
Slides from this talk are in Appendix B or from ftp://sdct-sunsrv1.ncsl.nist.gov/x3l8/sc32wg2/openforum/A5.htm.

Hert, Carol & Haas, Stephanie W. (forthcoming). Terminology paper, title TBA.

Haas, Stephanie W. (2000) Title TBA. Invited talk, Open Forum on Metadata Registries, January 17-21, 2000, Santa Fe, New Mexico.

Haas, Stephanie W. (2000). The Role of Knowledge Representation in Managing Statistical Information. Panel presentation "The Human Side of Data Dissemination, organized by Frederick Conrad and Cathryn Dippo, ICES 2, June, 2000.
Abstract: The importance of knowledge representation in understanding and managing complex information resources is gaining increasing recognition. A clear, unambiguous representation of concepts, rules, sources, ranges, of values, etc., if fundamental to the organization and presentation of information in an intelligible and useful way. Providing citizen access to statistical information is a particular challenge, in that users have a wide range of expertise and interest to support them in their quest for answers. The knowledge representation can guide system designers in determining what kind of information to provide and how best to do it, and can also serve as a resource for the users themselves, showing them how the "statistical world" is put together.

7. Selected Bibliography

Alhir, S. (1998). *UML in a Nutshell*. Sebastopol, CA: O'Reilly & Associates, Inc.

Aslandogan, Y., Their, C., Yu, C., Zou, J. & Rishe, N. (1997). Using semantic contents and WordNet in image retrieval. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 286-295.

Booch, G., Rumbaugh, J., & Jacobson, I. (1999). *The Unified Modeling Language User Guide*. Reading, MA: Addison Wesley.

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

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43, 907-928.

Guareno N. (1997). Understanding, building, and using ontologies. *International Journal of Human Computer Studies*, 48, 293-310.

Haas, S. W. (1998). *Investigation into the Requirements and Structure of a Knowledge Organization for BLS Published Information*. Proposal to the Bureau of Labor Statistics, August 1998.

Haas, S. W. (1999) *"What Everyone Calls It": Mapping Users' Words and Terminologies. Invited talk, Open Forum on Metadata Registries*, February 16-19, 1999, Washington, D.C.
Slides from this talk are in Appendix B or from ftp://sdct-sunsrv1.ncsl.nist.gov/x3l8/sc32wg2/openforum/A5.htm.

Hert, C. A. (1999) *Federal Statistical Website users and Their Tasks: Investigations of Avenues to Facilitate Access.* Final Report to the Bureau of Labor Ststistics, July, 1999.

Hert, C. A.. & Haas, S. W. (forthcoming). Terminology paper, title TBA.

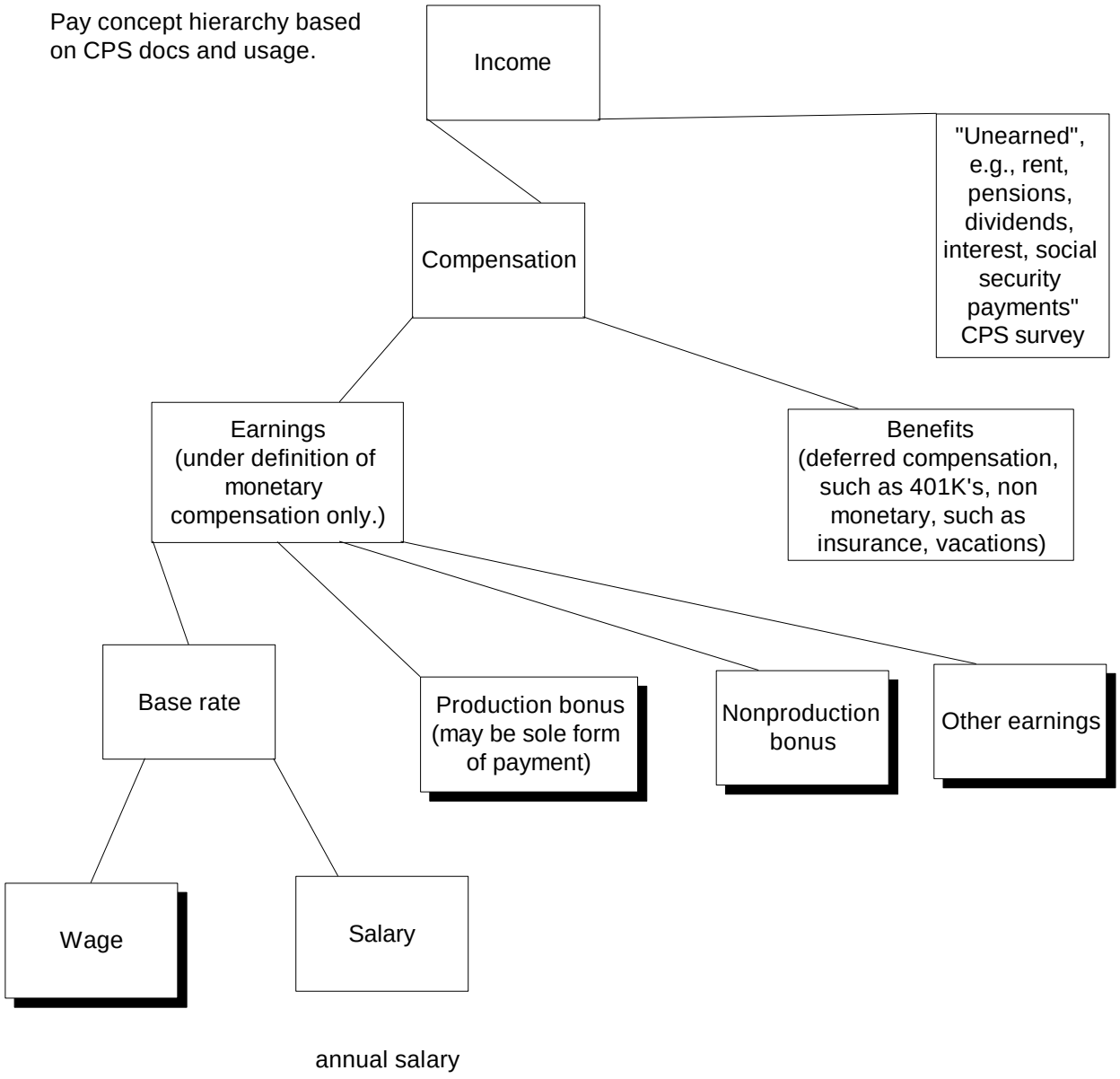
Humphrys, B. (1999). *Perspectives to Gain from the Unified Medical Language System (UMLS)*. Presentation given at the Open Forum on Metadata Registries, February 16-19, 1999, Washington, D.C. ftp://sdct-sunsrv1.ncsl.nist.gov/x3l8/sc32wg2/openforum/A5.htm.

ISO/IEC 11179 Specification and Standardization of Data Elements. Draft Revision (WG2 BNE 012) [Dated: 1998-06-12]. http://metadata.aihw.gov.au/project/rev11179/contents.html.

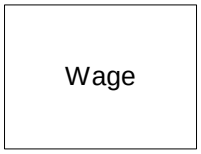
Srinivasan, P. (1996) Query expansion and MEDLINE. *Information Processing & Management*, 32, 4, 431-443.

Voorhees, E. (1994). Query expansion using lexical-semantic relations. *Proceedings of the seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 61-69.

Appendix A. Expanded Concept Hierarchy for *pay*.
Subsequent pages expand the shadowed boxes.



pay-wage.
terms associated with the wage concept.



hourly rate
straight time earnings
union rate, scale
apprentice rates
journey level rate
helpers' rate
probationery rate
beginner rate
entrance rate, hiring rate
pay-for-knowledge, skill-based pay, knowledge-based pay, multiskill compensation
educational pay differential
flat rate
guaranteed rate

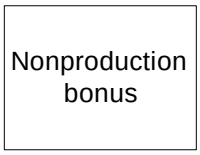
superannuated rate
temporary rates, experimental rate, trial rate
piece rate
stint work
tonnage rate
blue circle rate
flagged rate, red circle rate, out of line rate

pay-productionbonus.
terms associated with production bonuses.
These are generally included as part of base
salary for NCS establishment survey,
excluded (with some exceptions) from CPS. In some
cases, these types of payment may be base form,
e.g., a commission-only pay rate.

Production bonus

commission, commission payment
at-risk pay
incentive earnings
call-in pay, reporting pay
cost of living adjustment
hardship allowance
hazard pay
high time pay
logevity pay
portal to portal pay
payments for income deferred due to participation
in a salary reduction plan
deadhead pay

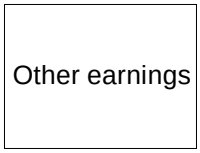
Pay-nonproductionbonus.
terms associated with non-production bonuses. These are generally not included as part of base salary for NCS establishment survey,
some are included in CPS.



attendance bonus
back pay
bereavement pay
Christmas bonus
profit-sharing, cash profit-sharing
year-end bonus
dismissal pay, severance pay (and synonyms)
premium pay
supplemental pay
double time
overtime
shift differential, shift premium
holiday pay
holiday premium pay
penalty rate
moving allowance, relocation allowance
paid absence allowance

per diem allowance
subsistence allowance
referral bonus
tool allowance
vacation pay
pay in lieu of vacation
year-end bonus
draw account
tips
uniform allowance
jury duty pay
holiday bonus
tuition reimbursements

pay-otherearnings.
other terms associated with
some type of earning. Some of these
are listed as being excluded from NCS earnings,
but not explicitly classified as non-production
bonuses. There are some possible
near-synonyms of these terms in the
other classes, however.



bilingual pay differential
safety bonus
contract-signing bonus
makeup pay
out of town work payments
push money
retroactive pay
royalty
profit-sharing distributions
stock bonus
free room & board

Appendix B. Slides from Open Forum Presentation.

These are also available from ftp://sdct-sunsrv1.ncsl.nist.gov/x3l8/sc32wg2/openforum/A5.htm