**Title:** Mapping of two schemes of classification for software classification

**Authors:**

1. Ms. Suvarsha Minj[1]
2. Dr. T. B. Rajashekhar[2]

[1] MLISc., Karnatak University, Dharwad. Working as Project Assistant, Dept of ECE, Indian Institute of Science, Bangalore – 560012
[2] Ph.D., Pune University, Pune. Working as Associate Chairman, National Centre for Science Information, Indian Institute of Science, Bangalore – 560 012

**Abstract:**

SALIS is a repository of open source software along with metadata information. The objective is to empower the Indian academic and developer community to make informed decisions while using open source software. To enable organization and retrieval of the information stored in the repository, a modified CCS (Computing Classification Scheme) classification scheme by the ACM (Association of Computing Machinery) was used. Since a sizeable section of the end users community were familiar with the USPTO classification scheme, a need was felt to classify the software by USPTO scheme also. Instead of classifying by two schemes it was decided to have a mapping or a concordance between the two schemes so that the classification takes place semi-automatically. The approach used to derive a concordance between two diverse classification schemes is described in this paper.

**Keywords:** Classification schemes, mapping, concordance,

**Introduction:**

Despite the advances in information technology, the need to classify information is felt acutely in the electronic age. “Organizing and providing access to the resources on the Internet has been a problem area in spite of the availability of sophisticated search engines and other software tools.” [1] Classifications are used today in web directories (e.g. Yahoo and Google), Internet and intranet information portals, content management and knowledge management applications. There has also been an extensive use of universal library classification schemes to classify information or objects and retrieve them in the electronic environment [2][4]. One example is the UDC adopted in the MIRACLE project classifying Braille music works [3].

The advantages of using a classification scheme are many [5]. Classification schemes provide a means of neatly organizing objects and quickly retrieving them when needed. They are often used to display the content to the user in an organized manner to facilitate browsing. Inexperienced users who are clueless about the content organization can use the scheme as a guide when they approach an information system with their needs. The hierarchical structure acts as a navigation aid.

The structure also enables broadening and narrowing of searches, improving the recall and precision rate of the searches made. It acts as a filter by limiting keyword search results to specific areas. Category-based search overcomes problems associated with keywords like synonyms, homonyms etc. The simplified vocabulary used for searching avoids misguiding the user and wastage of time. Classification can also be used to display the search results effectively by providing a context for search

results. By showing the category hierarchy associated with the result, records can make great sense to a user looking for that single drop in an ocean of information.

Classification schemes also serve another purpose by acting as maps of knowledge content. We are able to comprehend the knowledge stored in a repository not by merely knowing how many objects it has, but also by distributing those numbers across various categories and sub categories. [5] Only then do we fully recognize the strengths and weaknesses of a collection of objects. Thus classifying gives an overview of the quality (distribution) of the content.

The SALIS repository contains software and their metadata, which was classified using the CCS (Computing Classification Scheme) by ACM (Association of Computing Machinery). A need was felt to organize the information using USPTO scheme to facilitate users familiar with this scheme to access the repository contents. To enable this, a mapping between the two schemes was developed.

Mapping between two subject vocabularies have been carried out before in various contexts. Chang and Zeng [6] list out some of these and categorize the different methods used in constructing mappings. The method we have used here is computer-aided mapping, not a completely automated system. We could not rely entirely on automated systems because the two schemes were diverse in their structure. Human intellectual effort was thus needed to finalize the mapping. A description of this methodology is given in this paper. The use of the concordance in the repository browse and search is also shown.

**About Software and Licensing Information System (SALIS)**:

The project is sponsored by the IPR Cell, Ministry of Information and Communication Technology, Government of India. The purpose of the project is to build a database of freeware, shareware, and academic software, their copyright and use and reuse information. The main objective is to promote the use of free or open source software by providing information to software developers about availability of such software. Intellectual property information and an analysis of its implications help in making a judicious selection of the base software and also to provide a databank for Registrar of Copyrights to speed up the copyrighting process.

The outcome of the project is SALIS, a repository of about 300 software in the areas of Communication Networks and Databases. These are mostly free or open source software, with some proprietary software included for providing a comparison. Along with detailed metadata for each software, a copy of the software is also included in the repository, wherever license terms allow redistribution. The repository also includes brief technical review for each software and also Intellectual Property reusability terms.

This repository can be searched or browsed on the Internet (URL: http://salis.ece.iisc.ernet.in). SALIS is also available on CD-ROM.

**SALIS Classification Scheme:** For effective organization and retrieval of the software, it was necessary to classify the software according to a scheme of classification. The SALIS classification system is an adaptation of the classification and keyword system called the 'Computing Classification System' developed by Association for Computing Machinery (ACM). Last modified in 1998 the ACM scheme is one of the comprehensive schemes for classification of entities in the field

of Computing. ACM classification has been used predominantly for categorizing papers submitted to ACM journals, particularly to group reviews in ACM Computing Reviews journal. As there was no software-specific classification available, and since software, especially open source software, could be viewed in a very general sense as a kind of publication, we decided to adapt the ACM classification. [8]

We also studied INSPEC and Compendex database classification schemes to identify relevant categories in these schemes that need to be included in the SALIS scheme. Small sections of INSPEC scheme were also added as subclasses of SALIS scheme. By and large, however, the SALIS classification scheme is quite similar to the ACM scheme. Since the scope of repository was defined to cover the domains of Communication Networks and Databases, appropriate portions of ACM covering these subject domains were considered (C and H). The subclasses in these main divisions were included without much change. The SALIS classification scheme has two main classes, which are further divided into subclasses (see Appendix-A showing a portion of the SALIS scheme). Further, the classification scheme was not kept rigid but flexible to accommodate new classes due to the rapid developments in the subject domains. This way, the classification scheme would remain more reflective to the changes in the particular domains of study.

**Advantages of SALIS Scheme:**

1. **Is hierarchical:** The ACM scheme is a well-developed scheme based on computing theory. ACM classification offered us most of the advantages of hierarchical schemes mentioned in the beginning for organizing software in SALIS. It has neatly organized categories like computer-communication

networks, Information Systems etc., and further sub-categories under these headings. The categories and sub-categories are arranged in a hierarchical manner.

2. **Is easier to comprehend:** Because of the hierarchical arrangement the scheme is easier for a new user to approach and understand.

**Disadvantages of SALIS Scheme:**

1. **Is not meant for software packages:** The scheme is based on theory and is originally meant for classification of documents. (For example, if a classification scheme for food were to classify food theoretically as proteins, vitamins, carbohydrates, minerals etc., then under which category would a veg pizza fall?) Similarly, it is difficult to classify application software using SALIS scheme strictly under one subject category, as many software contain features and perform functions described under many categories. Another problem is that most software tend fall under the same broader category and thus it is difficult to highlight the subtle differences in the features of these products.
2. **Is not updated regularly:** Compared to USPC scheme, the ACM scheme is updated less frequently. However while designing the SALIS scheme a provision was made to update the scheme if the need was felt.

**USPTO Classification Scheme (USPCS):**

The United States Patent and Trade Office (USPTO) provides access on the Internet to all patents filed in the United States. To classify these documents, it has a

scheme of classification. The Manual of Classification arranges all areas of technology into a classification scheme.

The U.S. Patent Classification System (USPCS) provides for the storage and retrieval of patent documents, which a Patent Examiner needs to review when examining patent applications. A fundamental principle of the USPCS is that each class, or part thereof, is created by first analysing the claimed disclosures of the U.S. patents and then creating various divisions and subdivisions on the basis of that analysis. All similar subject matter is gathered together in large groupings to create classes. These classes are then subdivided into smaller searchable units called subclasses.

Unlike traditional classification schemes, the USPCS does not represent the body of knowledge in a systematically hierarchical fashion. A variety of rationales have been developed over the years to subdivide the Patent and Trademark Office's (PTO) classified files into searchable units. Collections of art based on each of the following rationales can be found in the system as it exists today.

**Industry or Use:** This approach divides art on the basis of the industry employing the art or the use to which a device is put.

**Proximate Function:** To avoid the tendency to fragment art based on its industry or use, the USPTO uses the fundamental, direct, or necessary function as one of the primary bases of classification. Proximate function means that similar processes or structures that achieve similar results by applying similar natural laws to similar substances are considered to have the same fundamental utility and are grouped together.

**Effect or Product:** This rationale collects art into industrial or trade groupings based on the result produced by the art. This result may be tangible (e.g., the product of a manufacturing process) or intangible (e.g., the communication of sound at a distance).

**Structure:** Simple subject matter which may have no apparent functional characteristics is classified together based upon the structural configuration or physical makeup of the object. [9]

**Advantages of USPC Scheme:**

1. **It is current:** Is based on discloser of innovation to the USPTO. Hence the latest developments get reflected in the scheme immediately.
2. **It is popular and thus familiar***:* USPTO is a well-known for it's' database of patents and trademarks. Legal, information and other professionals working in the areas of intellectual property are known to regularly access the site for information. Hence a set of users are already familiar with this scheme. Enabling access to SALIS thorough this interface would take advantage of this familiarity to provide better access to the database.

**Disadvantages of USPC Scheme:**

1. **It is not hierarchical:** USPCS is not a hierarchical scheme. Similar subjects are found scattered through out the scheme and not put under a single subject heading or not even as a series of headings grouped together. This is because of the nature in which the scheme was formed and is still updated.
2. **It is not easy to comprehend:** USPCS was meant originally for patent examiners and not for end users unfamiliar with the scheme. Hence it is difficult for a new user to follow the USPCS.

**Need for mapping:**

Compared to the ACM classification, which is the basis of the SALIS scheme, USPCS and IPC (International Patent Classification scheme of WIPO) tend to be more popular with users coming from industrial, trade, commerce and intellectual property domains. A need was felt to support browsing and searching the SALIS database using USPC and IPC schemes also. As a concordance between USPC and IPC already existed, we decided to develop a concordance between SALIS scheme and USPC, and use this concordance to support USPC-based searching and browsing in SALIS database.

**Challenges in preparing the concordance:**

**1. Difference in approach:** Since the ACM scheme was built for classifying documents, it concentrates on computing theory. The USPCS is however for classifying patent applications. Hence the approach is more towards applications.

**2. Difference in structure:** ACM is organized in a hierarchical fashion where as USPCS does not pay much attention to grouping related subjects together or under one heading.

**3. Difference in detail:** While ACM does not divulge into details, USPCS has classes representing very narrow subject areas. Following example illustrates this difference.

In ACM we have only one class

C.6 Local and Wide-Area Networks, with the following descriptors

C.6 Local and Wide-Area Networks

- Access schemes
- Buses
- Ethernet (e.g., CSMA/CD)
- High-speed (e.g., FDDI, fiber channel, ATM)
- Internet (e.g., TCP/IP)
- Token rings
- Local area networks (MIT)

- Wide area networks (MIT)
- Metropolitan area networks (MIT)

Where as the USPCS has the following subclasses in the main class 370 Multiplex Communications:

| | |
|---|---|
| 901 | **WIDE AREA NETWORK** |
| 902 | . Packet switching |
| 903 | .. OSI Compliant Network |
| 904 | ... Integrated Services Digital Network (ISDN) |
| 905 | ... Asynchronous Transfer Mode (ATM) |
| 906 | ... Fiber Data Distribution Interface (FDDI) |
| 907 | ... Synchronous Optical network (SONET) |
| 908 | **LOCAL AREA NETWORK** |
| 909 | . Token ring |
| 910 | . Carrier sense multiple access (e.g., Ethernet, 10Base-T) |
| 911 | . Bridge (e.g., brouter, bus extender, etc.) |

**4. Vastness :** While we could confine ourselves to two main classes in ACM dealing with our subject, we had to take all the classes of USPCS into consideration for the concordance.

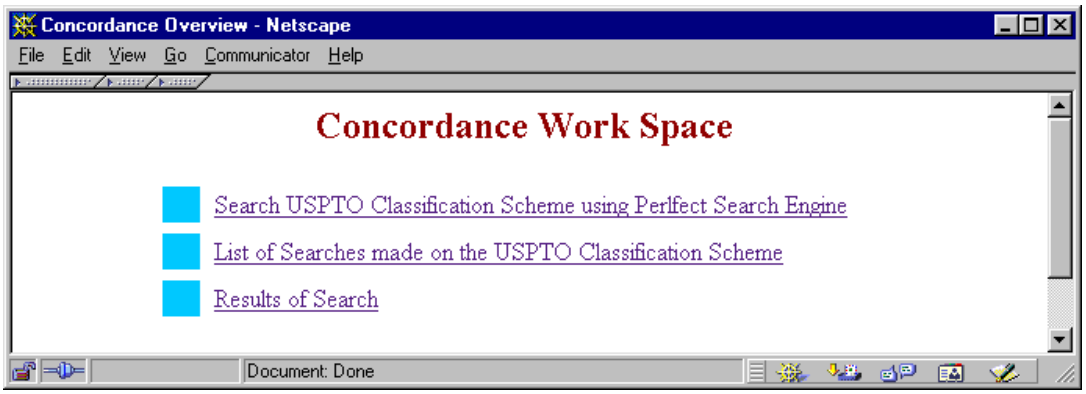
**Concordance Preparation Methodology:**

Developing the concordance between SALIS scheme and USPTO involved the following steps:

1. Downloading of the entire USPTO Classification Scheme in html format
2. Identification of keywords in each of the SALIS classes
3. Installation of a indexing and search engine software
4. Searching of the USPTO scheme using keywords of the SALIS classes
5. Listing of possible matching classes in USPTO for the SALIS classes
6. Processing of these lists to find the most probable classes that match

7. Preparation of concordance

Figure 10: Concordance Work Space




**1. Downloading of the entire USPTO Classification Scheme in html format**

The entire USPTO classification Scheme of more than 900 is available in html format at the USPTO website (URL: http://www.uspto.gov/go/classification/selectnumwithtitle.htm). Each class including its subclasses is represented in a single HTML file. These files were downloaded using an automated downloading tool.

**2. Identification of keywords in each of the SALIS classes**

When the concordance was to be constructed, more than 200 software in the database were already classified by the SALIS scheme. Other than descriptors (standard keywords) given under each class of the SALIS scheme, the staff also assigned free keywords, which appropriately describe the software. Since these keywords are found

to be appropriate description of a particular class, the SALIS database was searched for keywords under each class. For example, what are all the keywords the staff have used to describe all the software listed under C.2 - Network Architecture and Design?

**3. Installation of an indexing and search engine software**

Manual browsing of the USPSC for matching classes was not a feasible option for two reasons : (a) The method of development of the USPTO scheme does not ensure that a hierarchical structure of the scheme is maintained as in most other schemes. This scatters similar classes through out the scheme making it difficult to trace all of them out manually.

(b) USPTO is very vast with more than 900 classes and numerous subclasses.

Hence an indexing and search engine was installed on the server, which exclusively searches the USPTO scheme. Perlfect is a free search engine written in PERL and distributed under the GNU GPL license. This was installed and configured to search the files downloaded earlier.

**4. Searching of the USPTO scheme using keywords of the SALIS classes**

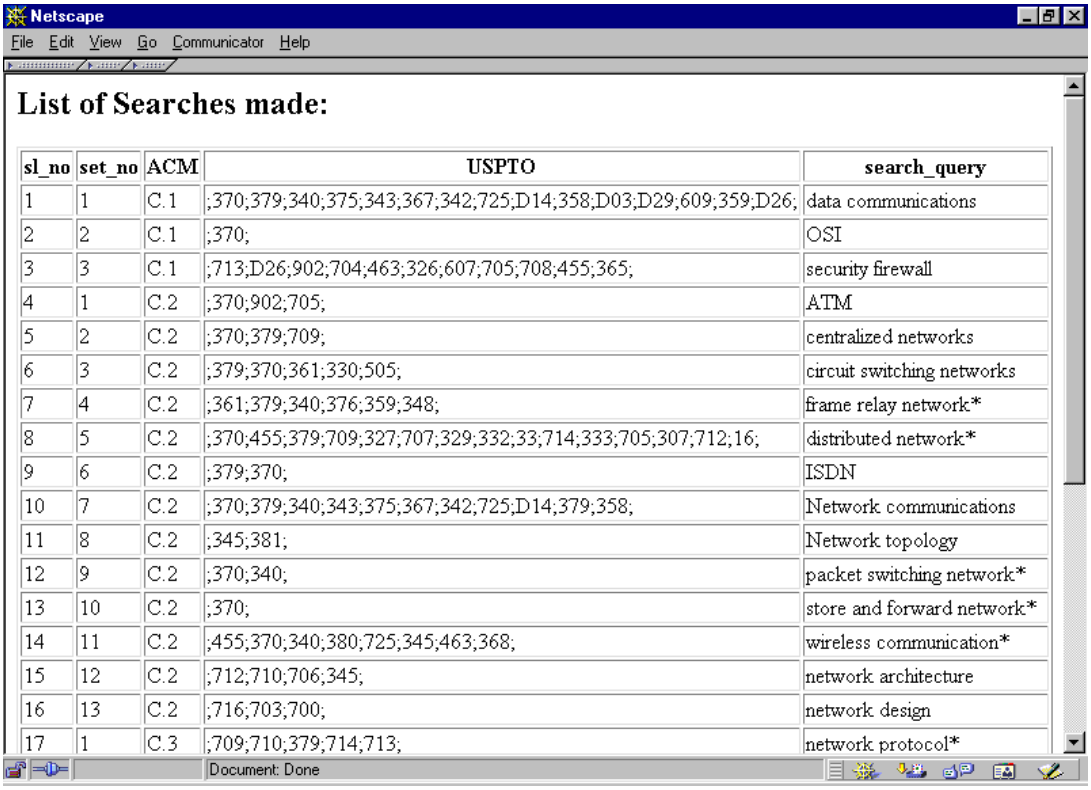
The list of keywords obtained in Step 2 of this process were used to search the USPCS html pages to find where matching words are found. Appropriate Boolean operators were used and the search queries were noted down.

**5. Listing of possible matching classes in USPSC for the SALIS classes**

The first thirty results of the search done on USPCS were scanned manually to find truly matching classes. This manual checking was required, as it was observed that all the results were not found to be appropriate. For example, searching for ‘Network Protocols’ would also list out the USPSC equivalent for ‘Building units and

construction elements', which is completely irrelevant to us. The results thus found appropriate were stored in a database using a web interface.

Figure 11: List of searches made on USPTO (Matches found and Search query)




| sl_no | set_no | ACM | USPTO | search_query |
|---|---|---|---|---|
| 1 | 1 | C.1 | ;370;379;340;375;343;367;342;725;D14;358;D03;D29;609;359;D26; | data communications |
| 2 | 2 | C.1 | ;370; | OSI |
| 3 | 3 | C.1 | ;713;D26;902;704;463;326;607;705;708;455;365; | security firewall |
| 4 | 1 | C.2 | ;370;902;705; | ATM |
| 5 | 2 | C.2 | ;370;379;709; | centralized networks |
| 6 | 3 | C.2 | ;379;370;361;330;505; | circuit switching networks |
| 7 | 4 | C.2 | ;361;379;340;376;359;348; | frame relay network* |
| 8 | 5 | C.2 | ;370;455;379;709;327;707;329;332;33;714;333;705;307;712;16; | distributed network* |
| 9 | 6 | C.2 | ;379;370; | ISDN |
| 10 | 7 | C.2 | ;370;379;340;343;375;367;342;725;D14;379;358; | Network communications |
| 11 | 8 | C.2 | ;345;381; | Network topology |
| 12 | 9 | C.2 | ;370;340; | packet switching network* |
| 13 | 10 | C.2 | ;370; | store and forward network* |
| 14 | 11 | C.2 | ;455;370;340;380;725;345;463;368; | wireless communication* |
| 15 | 12 | C.2 | ;712;710;706;345; | network architecture |
| 16 | 13 | C.2 | ;716;703;700; | network design |
| 17 | 1 | C.3 | ;709;710;379;714;713; | network protocol* |

For example, 13 keyword searches were made for class C.2- Network Architecture and Design. The matching USPCS class numbers were noted down for each search made, as shown in Fig 11.

**6. Processing of these lists to find the most probable classes that match**

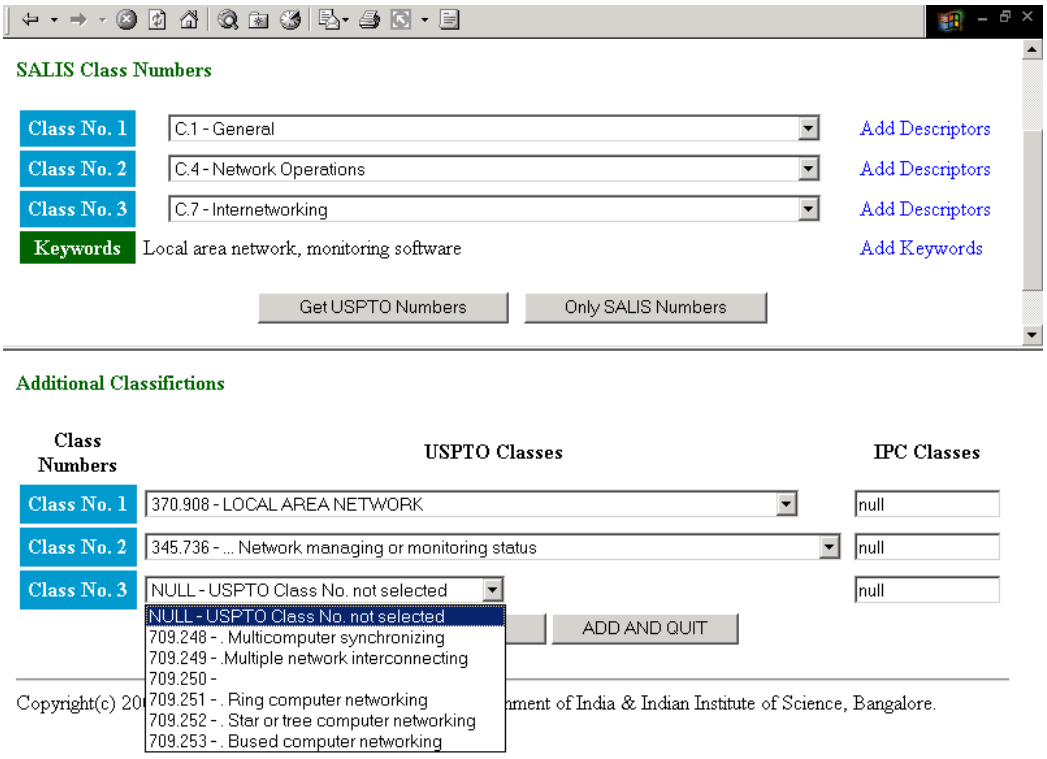
The list of matching classes was processed using programs to find out which classes in USPCS occur most often for the different keywords used to do a search. This method was used to weed out the irrelevant classes and include only those that were most appropriate.

**7. Preparation of concordance**

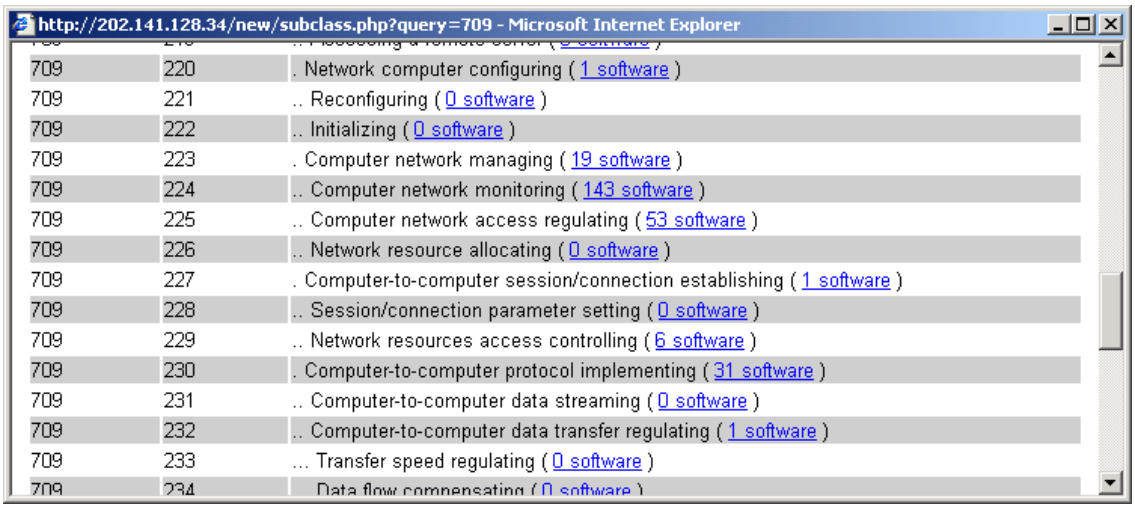
After obtaining the list of USPCS classes from step 6, the final step was to manually look through the classes and find if these actually matched the SALIS headings. Also the subclasses in the main class corresponding to the subjects represented in SALIS were noted down. This output was taken as the concordance between SALIS and the USPCS. A portion of the concordance for the SALIS category xxxx and USPCS is shown in Figure xxx.

**Use of the concordance:**

**1. Content addition:** In the SALIS content management system, the concordance is useful to the data entry staff. While classifying the software, they need to be familiar with the SALIS classification system, which is easier to follow due to its hierarchical nature. Once classified by the SALIS system the content management interface allows them to view the mapped USPSC classes for the selected SALIS classes. They can choose the appropriate USPSC class, thus classifying software by both the schemes.




**2. Content browsing:** Classifying by both has also made possible browsing of the content by both classification schemes.




**Conclusion:**

A combination of automated and manual procedures were used in compiling the concordance. Despite these efforts, it was found that the concordance does not

match exactly as it is ideally expected to do. The two systems differed not only in the structure and organization but also the depth with which they treated the subjects. The purpose for which they were built is also different. Considering these factors an exact concordance cannot be expected. One way of overcoming this problem is to allow the data inputting staff to modify the concordance as and when they find errors. For this a concordance management system was also put in place. Using this, the staff will be able to add, modify or delete records in the concordance itself.

The exercise of building a concordance between the two schemes was a useful effort. The SALIS team members now spend lesser time and effort in classifying software. They need to familiarize themselves with only one classification scheme. A user familiar with any of these schemes is able to search or browse the SALIS database. This will be a useful point of access to the end user.

**References:**

1. Devadason, F et al., "Search interface design using faceted indexing for Web resources", ASIST 2001: Proceedings of the 64th ASIST annual meeting, V 38, 2001: 224 –238.

2. Saeed, H and Chaudry, AS, "Potential of bibliographic tools to organize knowledge on the Internet: The use of Dewey Decimal Classification scheme for organizing Web-based information resources", knowledge organization, V. 28 no. 1 (International society of knowledge organization 2001): 17 – 26.

3. Adcock, L, "Building a virtual music library: Towards a convergence of classification within Internet-based catalogues" knowledge organization, V. 28 no. 2 (International society of knowledge organization 2001): 66 - 74

4. Zins, C and Guttmann, D, “Structuring Web bibliographic resources: An exemplary subject classification scheme”, knowledge organization, V. 27 no. 3 (International society of knowledge organization 2000): 143-159

5. Traugott Koch and others, “The role of classification schemes in Internet resource description and discovery”. A DESIRE project deliverable. 19 Feb 1997. http://www.ukoln.ac.uk/metadata/desire/classification/

5. Andersson, Bjarne and Hegna, Knut, “Crossing the Border:Subject search across library catalogues - attempting to match subject descriptions by a quantitative method.” VINE, V. no. 114, 1999, p.56-66.

Preliminary version: http://www.rub.ruc.dk/~bas/virt/crossingX.html

6. Lois Main Chan and Marcia Lei Zeng, “Ensuring interoperability among Subject vocabularies and Knowledge organization Schemes: A Methodologial Analysis”,

68th IFLA Council and General Conference, Aug. 18-24, 2002

http://www.ifla.org/IV/ifla68/papers/008-122e.pdf

7. http://www.bl.uk/services/information/patents/class.html

8. http://www.nypl.org/research/sibl/pattrade/TradePat/Patents/psearchtool.htm

9. http://www.uspto.gov/web/offices/pac/dapp/sir/co/ovrvw.htm#.index