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

Code Number: 008-122-R
Division Number: IV
Professional Group: Classification and Indexing
Joint Meeting with: -
Meeting Number: 122
Simultaneous Interpretation: -

Обеспечение взаимодействия между предметными словарями и схемами организации знаний: Методологический анализ

Lois Mai Chan,
School of Library and Information Science, University of Kentucky
USA

Marcia Lei Zeng
School of Library and Information Science, Kent State University
Kent, USA

Аннотация:

Неоднородная среда информационного поиска в WWW выявила недостаток взаимодействия между различными системами. Во время тематического поиска пользователи сталкиваются не только с различными словарями и схемами, но и с различными языками. В результате за последние несколько лет появилось много проектов, направленных на улучшение взаимодействия между предметными словарями и схемами организации знаний и работающих с разными словарями и языками. В этом докладе сделана попытка проанализировать методы, используемые в этих проектах. В начале дан краткий обзор, а затем подробно рассмотрены пути и методы, используемые в последних исследованиях.

1. ВВЕДЕНИЕ

В открытой среде Интернета и Web источники информации разнородны, заиндексированы с помощью различных словарей и организованы в соответствии с различными схемами. Как

достигать наилучших поисковых результатов при поиске в разных областях – особая проблема в информационной сфере. В информационном поиске пользователи обычно не осведомлены (да и нет в этом необходимости) о закулисных механизмах соответствия терминов их запроса словарям, применяемым различными системами. Идеальным подходом было бы обеспечить универсальный целостный поиск, а не поиск в отдельных базах данных или фондах. Чтобы сделать возможным такой подход, важно обеспечить взаимодействующие внутри одного поискового аппарата различные системы организации знаний, такие как контролируемые словари и классификационные схемы.

2. ОБЗОР ПРОЕКТОВ И ПРИМЕРОВ ВЗАИМОДЕЙСТВУЮЩИХ СЛОВАРЕЙ

Прежде, чем мы рассмотрим методы, давайте кратко изложим, что за последнее время предпринято в достижении взаимодействия между различными предметными словарями (включая контролируемые и неконтролируемые словари) и системами организации знаний. Сюда включены исследования в области установления взаимодействия между словарями на одном или разных языках, между различными классификационными схемами и взаимодействия между контролируемыми словарями и классификационными схемами. Эти усилия привели к отображению и интеграции существующих систем организации знаний или к созданию новых систем в совместном использовании информации в сетевом окружении. Проекты меняются в зависимости от целей отображения и методов, используемых в достижении этих целей. Ниже приведены проекты, объединенные по сходству задачи:

2.1. Среди контролируемых словарей на одном языке

1. Между *Library of Congress subject headings* (LCSH) и *Medical subject headings* (MeSH) – Northwestern University (Olson, 2001).
2. Среди различных контролируемых словарей – H.W. Wilson Company (Kuhr, 2001).
3. Среди различных тезаурусов на немецком языке, используемых для индексирования литературы по математике и физике, а также по общественным наукам – CARMEN (Content Analysis, Retrieval, Metadata: Effective Networking) (CARMEN WP12, 2000).

2.2. Среди многочисленных предметных словарей на разных языках и классификационных систем

1. Среди тезаурусов, классификационных систем, систем кодирования и списков контролируемых терминов в биомедицинской области – UMLS (Unified Medical Language System) Metathesaurus (National Library of Medicine, 2001).
2. Среди служб, применяющих различные словари индексирования, используемые такими организациями, как архивы, учебные заведения высшего образования и повышения квалификации, библиотеки, музеи, the National Grid for Learning, the Resource Discovery Network и т.д. – HILT (High-Level Thesaurus Project). (HILT, 2000; Nicholson, Wake and Currier, 2001a).
3. Среди словарей, используемых системами при вводе данных (например, указатели к *BIOSIS Concept Codes*, *INSPEC Thesaurus*, *U.S. Patent and Trademark Office Patent Classification и т.д.),* для отображения их в словарях, используемых при формировании поисковых запросов. – University of California Berkeley DARPA Unfamiliar Metadata Project (Buckland et al., 1999).
4. Среди локальных классификационных схем и общей схемы (DDC (*Dewey Decimal Classification*) – Renardus project (Koch, Neuroth, and Day, 2001).
5. Среди четырех контролируемых словарей и схем: *Polish Thematic Classification* (PTC), дескрипторы на базе *Thesaurus of Common Topics* (TCT), *Universal Decimal Classification*

(UDC), и *Subject-Heading Language* (SHL) Национальной библиотеки в Варшаве – Polish Project (Scibor and Tomasik-Beck, 1994).

6. Среди контролируемых словарей, используемых каталогами четырех национальных библиотек на трех языках: английском, французском и немецком – MACS (Multilingual Access to Subjects) (Freyre and Naudi, 2001).
7. Среди словарей для многоязычной базы данных о французском наследии – Мериме (См. статистику, представленную в Doerr, 2001.).

2.3. Между контролируемым словарем и универсальной классификационной системой

1. Между LCSH и LCC (*Library of Congress Classification*) –*Classification Plus* (продукт CD-ROM) и *Classification Web* (интерфейс на базе Web, находящийся в процессе разработки), Library of Congress.
2. Между LCSH и DDC (Vizine-Goetz, 1996).
3. Между UDC and GFSH (*General Finish Subject headings*) (Himanka and Vesa, 1992).

2.4. Между классификационными системами

1. Между MSC (the American Mathematical Society (AMS) *Mathematics Subject Classification*) и Разделом 510 в DDC – State University of New York in Albany, New York. (Iyer and Giguere, 1995).
2. Между SAB (*Klassifikationssystem för svenska bibliotek)* и DDC – Swedish Royal Library (IFLA, 2001:34).

2.5. Новая система для разных языков

В проекте HEREIN (Европейская информационная сеть по политике в области культурного наследия) сформировался язык-посредник, тезаурус, состоящий из терминов, взятых из сообщений по политике в области культурного наследия в Европе. Он был создан без прямого обращения к терминам или структуре любого существующего тезауруса. – The HEREIN Project (http://www.european-heritage.net/en/index.html, щелкнуть на кнопке Thesaurus).

3. МЕТОДЫ, ИСПОЛЬЗУЕМЫЕ ДЛЯ ДОСТИЖЕНИЯ И УЛУЧШЕНИЯ ВЗАИМОДЕЙСТВИЯ

Понятие словарной совместимости не является новым. Задолго до наступления века электроники библиотечные и информационные специалисты изучали и применяли различные методы для уменьшения противоречий между различными словарями, использующимися в одной системе. Эти методы почти полностью опирались на интеллектуальные возможности. С появлением современных компьютерных методов для достижения и улучшения взаимодействия компьютерная технология начала с успехом использоваться в сетевом окружении. Далее перечислены ставшие общепризнанными как традиционные, так и новые методы.

1. Словообразование/Моделирование – Создается специализированный или упрощенный словарь наряду с существующим более полным словарем в качестве исходного или модели.
2. Перевод/Редакция – Создается контролируемый словарь, состоящий из терминов, переведенных с одного языка на другой с изменениями или без изменений.
3. Отображение (интеллектуальное) – Создается система отображения, состоящая в основном из установления эквивалентности между терминами в различных

контролируемых словарях или между вербальными терминами и классификационными индексами. Такое отображение обычно требует огромных затрат умственного труда.

4. Отображение (автоматическое) – Создается система отображения, которая полностью или частично опирается на компьютерную технологию.
5. Соединение – Создается список терминов, связанных с другими терминами, которые не являются понятийными эквивалентами, но тесно связаны лингвистически. Считается, что такие связи улучшают результаты поиска.
6. Переключение – Создается язык или схема переключения, служащие посредником для перехода к эквивалентным терминам в различных словарях.

4. МЕТОДЫ, ИСПОЛЬЗУЕМЫЕ В ХРАНЕНИИ И УПРАВЛЕНИИ СВЯЗЯМИ

После создания системы отображения требуется найти средство для хранения и поддержки связей, чтобы управлять большим количеством терминов индексирования и их сложными отношениями. С этой целью было изучено и использовано несколько вариантов:

1. Авторитетные записи – Специальные поля в форматах авторитетных записей могут использоваться для хранения связей.
2. Конкордансы – Разработка конкордансов требует выделения одного главного словаря/схемы и одного или более связанных словарей/схем.
3. Семантическая сеть – Семантическая сеть состоит из организованной структуры, служащей в качестве основы или базовой сети. Каждая единица в сети представляет собой понятие, вокруг которого группируются идентифицированные эквивалентные термины из различных словарей.

5. ОБСУЖДЕНИЕ

При анализе методов, используемых во многих рассмотренных выше проектах, возникает ряд общих проблем.

5.1. Общие проблемы в системе отображения

5.1.1. Отображение многоязычных словарей

В основе многоязычного предметного словаря лежит отображение или установление эквивалентности. Соотношение «один к одному» между терминами в различных словарях и различных языках является идеальным, но маловероятным. Различные языковые выражения для одного и того же понятия, разные степени конкретности, многозначные термины – это лишь немногие трудности, с которыми сталкиваются при попытке отображения словарей или создания многоязычных или многоотраслевых словарей. Сложные требования и процессы установления соответствий между терминами, которые часто неточны, оказывают воздействие на следующие аспекты словарного отображения (Koch, Neuroth, и Day, 2001): структура просмотра, вывод на экран, нетематические классы, компромисс между последовательностью, точностью и применимостью. Различные уровни отображения/соединения могут сосуществовать в одном проекте, а именно те, которые определены проектом MACS: терминологический уровень (предметная рубрика), семантический уровень (авторитетная запись) и синтаксический уровень (приложение) (Freyre and Naudi, 2001).

5.1.2. Интегрирование взглядов разных культур.

Предположим, что в конкордансе все языки равны, тогда возникает вопрос, могут ли взгляды определенной культуры, выраженные через контролируемый словарь или классификацию,

соответствующим образом быть перенесены на другую культуру в процессе отображения. Hudon (1997) отмечает следующие проблемы, связанные с многоязычными системами:

1) проблема расширения языка с целью соответствия иной понятийной системе до такой степени, когда он становится едва узнаваемым самими носителями языка;

2) проблема перенесения целой понятийной системы из одной культуры в другую независимо от того, подходит она или нет;

3) проблема дословного перевода терминов с входного языка на бессмысленные выражения на выходном языке, и т.д.

Автор характеризует данные проблемы как проблемы управления, лингвистические/семантические проблемы и технологические проблемы.

5.1.3. Отображение систем с различными структурами

В терминах макроструктур контролируемых словарей и классификационных систем имеются существенные различия. Есть гарантия, что в тезаурусах, составленных в соответствии с ISO 2788 и другими национальными стандартами, структура и «грамматика» словаря остаются последовательными или совместимыми. Построение словарей предметных рубрик и классификационных схем, с другой стороны, определяется существующими образцами и примерами. Скорее всего, если имеется десять различных универсальных систем, то будет и десять различных принципов организации. В результате, системы организации знаний отличаются одна от другой по структуре, семантическим и лексическим характеристикам и особенностям нотации и записи (Iyer and Giguere, 1995). Например, они могут охватывать разные предметные области, в разных пределах и с разным охватом; они могут иметь семантические отличия, вызванные различиями в концептуальном структурировании; могут различаться их уровни конкретности и использование терминологии; могут быть различны синтаксические особенности, такие как порядок слов в терминах и выбор резервного использования рубрики.

Эти несоответствия с самого начала представляют проблемы для любой попытки отображения. Например, создание конкорданса или перевода между тезаурусом и классификацией или среди различных систем иногда становится невозможным или требует больших затрат. Особенно это становится реальным, когда выходная система имеет более высокий уровень конкретности, чем входная или другие системы.

5.2. Методологические варианты

Для проектов, имеющих целью создание взаимодействия между выбранными системами организации знаний с целью удовлетворения новых требований пользователя в сетевой среде, необходимо принять одно важное решение – это выбрать соответствующий метод. Первым сложным вопросом, требующим ответа, является: присоединить, отобразить или создать новую систему? Возможные варианты похожи на те, которые предложил Riesthuis (2001) как различные подходы к созданию многоязычных тезаурусов:

1. перевод
2. слияние
3. создание с нуля

Внутри каждого из этих подходов существует множество возможностей, как предполагают исследователи HILT в двухмерной сети (Nicholson, Wake and Currier, 2001b). Они предлагают три основных варианта:

- Использование или создание единой схемы (LCSH, UNESCO, на основе DDC, на основе UDC, полностью новой);
- Отображение существующих схем (LCSH, UNESCO, на основе DDC, на основе UDC);
- Отображение существующих схем в сокращенном виде для создания единой полной схемы.

К перечисленным выше вариантам можно добавить следующие решения:

- дополнительная структура тезауруса;
- новые предметные специальные микротезаурусы;
- отображение среди существующих отраслевых специальных микротезаурусов;
- многоязычность;
- общий контроль;
- автоматизированные методы;
- методы искусственного интеллекта;
- подготовка пользователей;
- универсальные средства в помощь пользователям;
- пользовательские карты;
- последовательность в применении терминов, обеспечиваемая подготовкой и мониторингом;
- обученные библиотекари в помощь пользователю для оптимизации поиска

Выбор основного подхода плюс любые комбинации решений могут принести разные конечные результаты и потребовать разных затрат времени и ресурсов. Каждый метод в сочетании с другими процессами имеет свои за и против. Необходимо провести всестороннее исследование и выявить проблемы, которые могут возникнуть при применении определенного метода.

6. ЗАКЛЮЧЕНИЕ

1. Чему мы научились у проектов?
2. Какие проблемы еще не решены?
3. Что требуется, с точки зрения интеллектуального и технического подходов, для дальнейшего продвижения в данной области?

Исходя из приведенных в данном докладе примеров, можно суммировать следующие тенденции, формирующие основное направление:

1. Необходимость во взаимодействии систем организации знаний является неизбежной проблемой, а также процессом в современной сетевой среде.
2. В достижении взаимодействия систем организации знаний используются разные методы. Система переключения может быть необходима, а может быть и нет. Идеальным вариантом может быть построение конкорданса между словарями, или он не будет использоваться. А также вполне вероятно, что взаимодействие может быть более успешным через предметные авторитетные записи различных онлайновых систем.
3. Отображение словарей все еще требует огромных интеллектуальных усилий, поэтому для содействия в управлении большими файлами предметных данных и в управлении связями используются компьютерные технологии. Компьютерные системы отображения высокого уровня являются экспериментальными и тестирующимися. Отображение, осуществляемое человеком, и автоматическое отображение будут сосуществовать и в будущем.
4. Начаты многочисленные проекты для отображения многих языков и структур. Выявлены разнообразные методы, и проведены эксперименты с ними. Можно с уверенностью сказать, что будет намного больше многоязычных продуктов и служб. Большинство из них охватит множество структурированных систем, таких

как тезаурус, классификация, предметные рубрики, термины индексирования, предназначенные для записей баз данных.

Бесспорна необходимость в согласовании различных предметных словарей в сетевой среде. Обнадеживают результаты от недавних усилий по достижению взаимодействия между словарями разных видов и в разных языках. Остается вопрос: полностью ли используются технологические возможности в наших попытках улучшить предметный доступ к бесчисленным источникам, имеющимся сейчас в сетевой среде?

БЛАГОДАРНОСТЬ

Выражаем глубокую признательность ведущим исследователям обсуждаемых в данном докладе проектов взаимодействия: Traugott Koch (Sweden and Denmark), Patricia Kuhr (USA), Martin Kunz (Germany), Max Naudi (France), Dennis Nicholson (UK), and Tony Olson (USA), которые великодушно уделили время и ответили на наши вопросы, предоставили данные по своим проектам или просмотрели доклад.

ЛИТЕРАТУРА

Buckland, M., et al. (1999). Mapping entry vocabulary to unfamiliar metadata vocabularies, *D-Lib Magazine*, *5(1).* http://www.dlib.org/dlib/january99/buckland/01buckland.html, (last accessed Feb. 5, 2002).

CARMEN. WP12: Cross concordances of classifications and thesauri. http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.en *(last accessed Feb. 5, 2002)*

Doerr, Martin. (2001) Semantic problems of thesaurus mapping. *Journal of Digital information*, 1 (8). http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/#Nr.52

Freyre, Elisabeth and Max Naudi. (2001) MACS: Subject access across languages and networks. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

HILT. (2000) *HILT: High-Level Thesaurus Project Proposal*. http://hilt.cdlr.strath.ac.uk/AboutHILT/proposal.html. (Last accessed Feb.5, 2002)

Himanka, Janne and Kautto Vesa. (1992) Translation of the Finish Abridged Edition of UDC into General Finish Subject Headings. *International Classification* 19(3):131-134.

Hudon, Michele. (1997) Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge concepts. *Knowledge Organization* 24(2): 84-91.

IFLA Section on Classification and Indexing. (2001) *Newsletter* Nr.24, December 2001.

Iyer, Hermalata and Mark Giguere. (1995). Towards designing an expert system to map mathematics classificatory structures. *Knowledge Organization* 22(3/4):141-147.

Koch, Traugott, Heike Neuroth, and Michael Day. (2001) Renardus: cross-browsing european subject gateways via a common classification system (DDC). In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on*

Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001. Dublin, OH: OCLC. http://www.lub.lu.se/~traugott/drafts/preifla-final.html (last accessed Feb.5, 2002)

Kuhr, Patricia S. (2001) Putting the world back together: mapping multiple vocabularies into a single thesaurus. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

National Library of Medicine. (2001) *Fact Sheet*: *UMLS ® Metathesaurus ®* Last updated 2001. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html. (last accessed Feb. 5, 2002)

Nicholson, Dennis and Susannah Wake. (2001a) HILT: Subject retrieval in a distributed environment. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

Nicholson, Dennis, Susannah Wake, and Sarah Currier. (2001b) High-Level Thesaurus Project: investigating the problem of subject cross-searching and browsing between communities. In *Global Digital Library Development in the New Millemnnium: fertile ground for distributed cross-disciplinary collaboration*, edited by Ching-Chih Chen. Beijing: Tsinghua University Press, 2001.

Olson, Tony. (2001) Integrating LCSH and MeSH in information systems. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

Riesthuis, Gerhard J.A. (2001) Information languages and multilingual subject access. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC.

Scibor, Eugeniusz and Joanna Tomasik-Beck. (1994) On the establishment of concordances between indexing languages of universal or interdisciplinary scope (Polish Experiences). *Knowledge Organization* 21(4): 203-212.

Vizine-Goetz, Diane. (1996) Classification Research at OCLC. *Annual Review of OCLC Research*, pp. 27-33.