Унификация структур данных в области изучения, освоения и использования ресурсов Мирового океана♣

© Е.Д. Вязилов, А.А. Федорцов, А.Е. Кобелев

Всероссийский научно-исследовательский институт гидрометеорологической информации – Мировой центр данных
vjaz@meteo.ru

Аннотация

Представлены методы выделения унифицированных элементов данных при интеграции распределенных гетерогенных информационных ресурсов, применяемые в Единой государственной системе информации об обстановке в Мировом океане. Показаны проблемы идентификации унифицированных элементов. Предложены подходы по созданию многомерных структур с учетом жизненного цикла информации об объектах БД. Приведены примеры многомерных структур данных и реализации БД с элементами многомерности.

1 Введение

В настоящее время создаваемые в разных организациях базы данных (БД) не имеют общей концептуальной основы и по большому счету не совместимы друг с другом. Одновременное использование заложенной в них информации становится трудоемким процессом.

Вопросам стандартизации структур и схем данных в настоящее время уделяется большое внимание, так разработан стандарт ISO 19115, протоколы описания данных – SensorML, TML, MODBUS [11–14] которые упорядочивают представление метаданных и данных на уровне измерительных систем.

Ведутся работы по интеграции данных во многих предметных областях и странах [2, 4, 7–9]. Под эгидой Всемирной метеорологической организации (ftp://ftp.wmo.int/Documents/wis/WIS-TechnicalSpecification-v1-0.doc), Межправительственной океанографической комиссии ЮНЕСКО (http://data.meteo.ru:81/odp/), в рамках Европейского сообщества ведутся большие работы интеграции данных и созданию баз метаданных. Это проекты Sea Search (http://www.sea-search.net), Sea Data Net (http://www.seadatanet.org/), др. [15]). При этом основное внимание в европейских проектах было уделено созданию схем метаданных для таких объектов, как сведения о массивах и БД, проектах, организациях, рейсах научно-исследовательских судов и созданию общего индекса данных.

В Единой государственной системе информации об обстановке в Мировом океане (ЕСИМО, http://data.oceaninfo.ru) реализована технология интеграции [1] разнородных, распределенных информационных ресурсов (ИР, http://data.meteo.ru:81/srbd/). Поддержка этой технологии требует приведения имен локальных атрибутов данных к унифицированным элементам данных. Для этого необходимо развитие типовых моделей данных, стандартизации представления и обмена данными, метаданными.

Унифицированный элемент данных – это поименованный объект данных (атомарная порция), находящийся под управлением информационной системы. Унифицированные элементы данных определяют метаданные (сведениями о данных – страна, организация, платформа, пространственно-временные характеристики) и значения атрибутов морской среды и морской деятельности.

Для отображения атрибутов данных и их представления необходимо назначение соответствий имен атрибутов локальной БД их унифицированным элементам. Обмен и использование атрибутов данных в различных приложениях также осуществляется через унифицированные элементы, которые имеют уникальные имена (идентификаторы). Локальные атрибуты данных представлены в массивах и БД, структурированных (форматированных) или объектных файлах.

Имена унифицированных элементов данных отличаются от атрибутов в локальных системах данных при одном и том же содержании и представлении (единицы измерений, формат и др.). Унифицированные элементы хранятся в словаре (http://data.oceaninfo.ru/udopweb/), используются для унификации локальных атрибутов данных с различным представлением и имеют уникальные имена и соответствующие им атрибуты описания. С помощью унифицированных элементов можно получить информацию, как пользователям, так и программным средствам (идентификатор, краткое имя, пределы изменений, формат хранения поля).

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

2 Именование унифицированных элементов данных

При регистрации ИР требуется установить однозначное соответствие между атрибутами массивов и БД, предоставляемых в качестве ресурсов ЕСИМО, и унифицированными элементами данных. Программный комплекс «Поставщик данных» технологии интеграции E2EDM [10] осуществляет преобразование (маппинг) локальных имен атрибутов данных в унифицированные имена элементов данных. Для этого преобразования каждому локальному имени атрибута требуется найти аналог по содержанию и убедиться, что свойства выбранных пар – единицы измерений и/или системы кодирования значений атрибута, пространственно-временное разрешение измерения или обобщения – совпадают.

Критерием выделения унифицированных элементов является возможность интеграции разнородной информации в рамках одной модели описания ИР [5] с минимизацией числа используемых локальных имен атрибутов.

В ЕСИМО для преобразования структурированных данных, хранящихся в БД и структурированных файлах данных, используются разделы словаря «Метаданные» – **M** и «Данные» – **P**.

Для объектных файлов (карты, графики, документы) используется унифицированный элемент с префиксом **R**, показывающий отношение к рубрике. Унифицированные элементы данных для объектных файлов используются в том случае, если есть возможность задания элемента на весь объект (например, jpeg-файл с унифицированным элементом «Давление воздуха», представленное в виде изолиний на карте).

Атрибуты метаданных специфицируют характеристики производства (получения, описания) данных. В метаданных определены унифицированные элементы, связанные с описанием ИР. Это идентификаторы структурной единицы данных, платформы наблюдений, инструмента измерений, метода обработки; проекта, географического объекта.

В разделе «Метаданные» выделены подразделы:

- общей идентификации данных;
- принадлежности стране, организации, автору, проекту и др.;
- спецификации производства (получения) данных – платформа, метод, прибор и др.;
- географические характеристики – районы, высота над уровнем моря, глубина, толщина слоя, расстояние, местоположение платформы (широта, долгота), направление движения платформы;
- временные характеристики (год, месяц, день, время начала и окончания события).

В разделе «Данные» выделены подразделы:

- морская среда;
- морская деятельность;
- социально-экономическая информация (СЭИ).

Атрибуты данных, специфицирующие отдельное свойство природного процесса (явления):

- могут быть измеренными значениями, или вычисленными статистическими, или новыми величинами, полученными на основе теоретических или эмпирических закономерностей – каждый атрибут (измеренный или вычисленный), включается в состав унифицированных элементов;
- указывают интенсивность явления – этот показатель является общим унифицированными элементом для многих унифицированных элементов;
- показывают аномалию (отклонение от многолетнего среднего значения), экстремальные значения, тенденцию, характеристику тенденции, количество зарегистрированных явлений (ситуаций) – включается в состав унифицированных элементов.

Значения атрибутов могут визуально наблюдаться, определяться путем химического анализа в пробе, измеряться, регистрироваться в виде порядкового номера. Эта информация указывается в описании метода определения атрибута и представляется в отдельном объекте метаданных [3, 4] – «Методы».

Для указания этих свойств атрибутов используется следующая структура имени унифицированного элемента:

ANNNN_SS_ I (M, Q),

где **A** – раздел словаря (P, M или R),

NNNN – числовой код, определяющий атрибут наблюдений (измерений, вычислений);

SS – числовой код, отражающий название статистической характеристики значения атрибута (измеренное значение, минимум, максимум, среднее, повторяемость, вероятность, аномалия, продолжительность явления, число случаев, сумма, тенденция);

I – интенсивность явления, выраженная в значении атрибута;

M – метод получения атрибута;

Q – признак качества значения атрибута.

Примеры унифицированных элементов представлены ниже:

P0295_00 – температура воды, измеренная;

P0295_01_Q – признак качества температуры воды, минимальной.

Для упорядочения и улучшения поиска унифицированных элементов в словаре для поля «Название» применяется следующий принцип записи. В названии атрибута первым записывается основное слово, характеризующее элемент, а затем уточняющее, например, давление воды, измеренное; температура воды максимальная.

3. Проблемы определения унифицированных элементов

3.1 Множество атрибутов, имеющих одинаковое назначение

В распределенных ИР бывают ошибки нормализации данных. В таблицах одного объекта данных или метаданных может быть несколько значений атрибутов для даты, страны, организации, персоны, проекта и других атрибутов. Или, например, одни и

те же атрибуты (погода между сроками, наблюдения в один срок на разных площадках), имеющие несколько значений, хранятся в виде отдельных полей таблиц с фиксированными позициями. Это объясняется тем, что разработчики БД в локальных источниках данных не всегда выходят на необходимый уровень нормализации данных. Проще ввести два или несколько новых атрибутов, характеризующих разные процессы жизненного цикла (ЖЦ) объекта, чем создавать более универсальные структуры данных. Примерами нескольких значений атрибутов в ИР являются:

- страна автора проекта, производитель, владелец судна;
- даты проектирования, постройки, передачи в другую организацию, списания судна;
- даты рассмотрения, принятия, утверждения, подписания, ратификации документа.

Для решения этой задачи атрибуты, имеющие несколько значений, должны записываться вместе с дополнительным атрибутом «ЖЦ». Чтобы определить стадию ЖЦ таких атрибутов, как страна, организация, персона, проект необходимо описать стадии ЖЦ объекта и выделить необходимые метаданные.

Ниже представлены стадии ЖЦ, которые относятся к атрибутам:

- страна – проектирование, производство, владение, администрирование;
- организация – производство, владение, разработка, эксплуатация;
- персона – владение, разработка;
- проект – выполнение, участие.

Структура хранения таких данных в многомерной модели выглядит следующим образом:

Объект: идентификатор *(прибор, судно, организация, др.)*
Объект: ЖЦ *(проектирование, разработка, эксплуатация, др.)*
Дата: значение
Атрибут: название (*страна, организация, персона, проект, др.)*
Атрибут: значение (*идентификатор страны, организации, персоны, проекта)*

Такая структура применима для хранения значений атрибутов в различных каталогах объектов (судов, товаров и т.п.).

Для ЖЦ документа используются такие стадии как создание, модификация (редактирование), обсуждение (рассмотрение), согласование, принятие, ратификация, утверждение с одной и другой стороны, подписание, вступление в силу документа, передача в печать, издание, доступность, передача на хранение, прекращение срока действия, доставка, уничтожение. Хранение множества дат для документа должно производиться в отдельной таблице:

Объект: идентификатор (*документа)*
Объект: ЖЦ
Дата: значение

Если требуется уточнить условия применения, указать признак качества даты, метод записи времени, то для этого можно ввести дополнительные атрибуты метаданных.

3.2 Большое количество однородных атрибутов в локальных БД

Иногда одному атрибуту дают разные названия и хранят в одной или разных таблицах, например, для оценки производства различных товаров в СЭИ представлено более ста названий атрибутов. Для работы с такими данными необходимо выделить атрибут «Класс». Классом может быть товар, промышленный объект, объект вылова, отходы, данные, экономические показатели и др. Чтобы правильно описать объект, необходимо выделить стадии ЖЦ производства и определить состояние объекта. Примеры стадий ЖЦ для различных объектов приведены ниже.

Товар (продукция): проектирование, создание (производство), ввод в действие – пуск, сдача, заказ, поставка, продажа, хранение, монтаж, постановка на учет, обслуживание, вылов, сдача, продажа, списание, уничтожение (утилизация).

Состояние транспортного объекта: отправление, переход, заход, прибытие, стоянка у причала, на рейде, ремонт, дрейф, производство наблюдений, погрузка, разгрузка, списание.

Выбросы (**отходы**): создание, хранение, обезвреживание, утилизация.

Проект: инициация; планирование; выполнение; контроль и мониторинг; завершение.

Данные: агрегация – расчет статистических характеристик, прогноз, классификация, результаты сравнения – аномалии.

Еще одним вариантом использования множественных имен атрибута, который встречается в БД, является тип географического объекта. Например, географический объект можно представить в виде географический названий материков, стран, федеральных округов, субъектов РФ, населенных пунктов, морей и океанов, др. Для этого атрибута необходимо хранить в таблице БД два атрибута – роль географического объекта (страна, район, др.) и значение географической области (Индийский океан).

4 Модель информационного пространства

Основной идеей типизации структур данных является использование минимальной атомарной единицы хранения. Сейчас в большинстве ИР атомарной единицей является группа атрибутов, измеренных или вычисленных для одной точки в определенный момент времени или за какой-то период. Если за атомарную единицу хранения принять отдельное значение атрибута, то многие атрибуты метаданных могут быть вынесены в отдельную таблицу «Каталог объекта», а изменяющиеся значения в пространстве и во времени – в таблицу «Факты».

Модель информационного пространства (ИП), объединяющего разнообразную информацию, представлена на рис. 1.

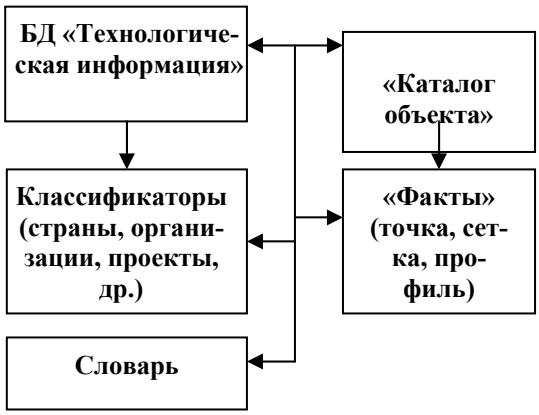

Рис. 1. Модель информационного пространства

Состав основных компонент такой модели включает:

- каталог объекта;
- таблица фактов;
- словарь унифицированных элементов;
- классификаторы;
- технологическая информация, определяющая состояние и связи объектов.

Каталоги включают справочные сведения о различных объектах по классам (метаданные, товары, отходы, документы, др.). Так как каталоги для конкретных объектов очень сильно отличаются по составу атрибутов, то для хранения каталогов эффективнее применять многомерную модель хранения данных. Таблица «Каталог объектов» имеет следующую структуру данных:

Идентификатор класса объекта (*каталога****)***
Идентификатор экземпляра объекта
Имя атрибута *(унифицированный элемент каталога*)
Значение атрибута (*каталога****).***

Факты отражают пространственно-временные координаты объектов, находящихся на различных стадиях ЖЦ. Данные могут быть представлены в виде точки, сетки, профиля. Эти типы данных имеет свою структуру данных [6]. Более подробно об этих структурах сказано в разделе «Примеры реализации». Таблица фактов должна включать такие атрибуты как класс объекта, ЖЦ, дата и время регистрации значения атрибута, имя атрибута, его значение. Кроме того, в зависимости от ситуации могут использоваться дополнительные атрибуты – страна, организация, персона, проект (если значения этих атрибутов изменяются в различных событиях).

Сведения об унифицированных элементах включают имя, временной и пространственный масштабы, полное и краткое наименование атрибута на русском и английском языках, точность наблюдений или расчетов, единицы измерения, диапазон изменчивости (мин, мах), вертикальное разрешение (мин и макс высота/глубина), используемый классификатор для значений атрибутов (международный, национальный), метод определения атрибута, международный код атрибута, описание.

Технологическая информация включает дату ввода, все даты редактирования, кто вводил и редактировал, показатели использования данных, другую информацию и предназначена для мониторинга состояния таблиц «Каталог объектов» и «Факты». Эта информация может также храниться в двух таблицах «Каталог» и «Факты».

Многомерные структуры данных для таблиц «Каталог объектов» и «Факты» представлены ниже.

Таблица «Каталог объектов»

Объект: Класс
Объект: тип
Объект: процесс
Страна: ID
Организация: ID
Персона: ID
Дата: значение
Атрибут: ID
Атрибут: значение

Таблица «Факты» для класса «Отходы»

Объект: Класс
Отходы: Тип
Отходы: Процесс
Страна: ID
Организация: ID
Персона: ID
Проект: ID
Геообъект: Тип
Геообъект: Значение
Дата: Значение
Атрибут: ID
Атрибут: Значение
Отходы: Класс опасности

Для каждого объекта составляется своя пара таблиц «Каталог объекта» и «Факты». То есть физически это будут разные таблицы, а по структуре они не будут отличаться.

5 Структура словарной базы данных

Создание и ведение ИП требует управления семантическим обеспечением (классификаторами, словарями). Для этого необходимо иметь в словарной базе (рис. 2):

- список объектов, включенных в ИП;
- список технологий, которые используют словарную базу (технология интеграции E2EДМ, централизованная база метаданных, web-портал, другие компоненты);
- список задач, для которых необходимо использование словарной базы (ввод, визуализация, конвертирование данных);
- список программных средств, в которых используются классификаторы;
- список объектов метаданных, для которых необходима словарная база;
- список таблиц, где используются классификаторы;
- список необходимых классификаторов для каждой технологии;

- сведения о классификаторах (кодовых таблицах);
- кодовые таблицы;
- сведения об атрибутах (унифицированных элементах).

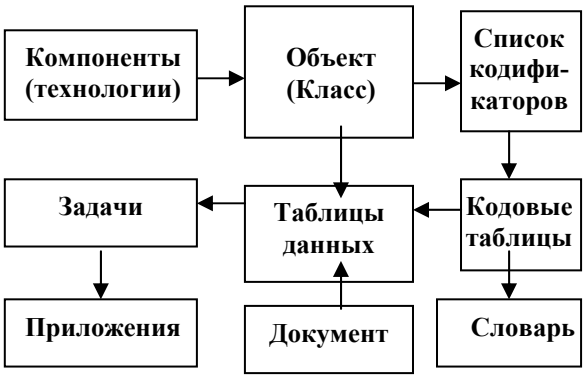

Рис. 2. Структура словарной БД

Число используемых классификаторов в ЕСИМО составляет около 300 единиц. Если для каждого классификатора создавать свою таблицу, то число таблиц резко увеличится, поэтому используется структура хранения классификаторов, включающая идентификатор классификатора, код объекта и значение кода. Для повышения эффективности и удобства использования классификаторов для каждой компоненты создаются рабочие словари. Состав классификаторов и кодов в них назначается в зависимости от потребностей технологий.

6 Примеры реализации

В рамках разработки интегрированной БД ЕСИМО созданы четыре типа хранения структур данных (данные, измеренные на профиле в случайных точках, данные в виде временных рядов, сеточные данные и каталоги объектных файлов). Эти структуры активно используют различные классификаторы.

Данные о временных рядах оформляются в виде двух таблиц – сведения о временном ряде и матрицы значений. Сведения о временных рядах включают идентификатор ряда, название станции, широту, долготу, дату начала и окончания наблюдений, горизонт, масштаб временного и пространственного осреднения, кто создал, когда, дату последней редакции, пополнения данных, название исходного массива. Матрица временного ряда значений включает идентификатор ряда, год, месяц, день, время, значение.

Сеточные данные хранятся в виде трех таблиц: сведения о сетке, сведения о поле, значения атрибута в узлах сетки. Сведения о сетке включают уникальный код, тип сетки, ссылку на источник данных, идентификатор модели, размерность сетки и шаг сетки в град по X, Y, широту и долготу точки, временной масштаб осреднения (сутки, месяц, год, другой), пространственный масштаб осреднения (квадрат, трапеция, точка), шаг при отрисовке изолиний.

Сведения о поле включают идентификатор поля, ссылку на сетку, значение года, месяца, дня, часа, минут наблюдений, имя атрибута, тип уровня, значение уровня. Данные в узлах сетки включают идентификатор точки, широту, долготу, значение атрибута.

Данные измерений по профилю состоят из двух таблиц: сведения о профиле и данные по профилю. Сведения о профиле включают идентификатор профиля и пространственно- временные координаты профиля. Данные по профилю включают идентификатор профиля, уровень, имя атрибута и значение.

Сведения об объектных файлах хранятся в виде каталога, имеющего ссылки на файлы с документами, картами и рисунками.

7 Заключение

Унификация процессов управления данными в гетерогенной среде, выработка наборов правил, применение типового инструментария и тем самым снижение расходов на администрирование данных возможна только через рассмотрение стадий ЖЦ объектов, сведения о которых отражаются в БД.

Необходимо создавать каталоги для организации хранения и поиска объектных файлов (документов, графических файлов, др.), а также для любой информации, представленной в виде сведений о каких-либо объектах.

В сущностях, для которых создаются БД, необходимо выделять общие поля типа идентификатор, дата, время создания, редактирования, хранения, название георайона, другие, в т.ч. поля для организации связи между объектами. При работе с атрибутами необходимо создавать обобщенные форматы полей (целое число, число с плавающей запятой, строка).

Предложенные структуры данных могут быть использованы в других предметных областях, при этом не надо добавлять новые сущности и поля.

Литература

[1] Белов С.В, Бритков В.Б. Интеграция информационных ресурсов в задачах исследования морской среды // Информационные технологии и вычислительные системы. 2008, Вып. 1. С. 73–82.

[2] Вавилов К., Щербина С. Web – Интеграция // Издательство «Открытые системы». Электронный журнал «Открытые системы». 2001. Вып. 1. 12 с. http://www.osp.ru/os/2001/01/043.htm

[3] Вязилов Е.Д. Консолидация метаданных в области наук об окружающей среде // Журнал «Вычислительные технологии» Т. 10, Спецвыпуск. СВ-Томск, 2005. С. 30–38.

[4] Вязилов Е.Д. Михайлов Н.Н., Карпенко Г.А., Кобелев А.Е. Широкий комплекс метаданных, как основа мониторинга и управления данными // 6-я Российская научно-техническая конференция «Современное состояние и проблемы

навигации и океанографии («НО-2007»), 23–25 мая 2007, С-Петербург, ГНИНГИ. С. 493–496.
[5] Вязилов Е.Д. О стандартизации структур данных в области морской среды // Электронный журнал «Новости ЕСИМО». 2007. Вып. 30. ftp://meteo.ru/resource/magazine/ news30.mht
[6] Вязилов Е.Д., Михайлов Н.Н. Интеграция гетерогенных информационных ресурсов в области морской деятельности // Журнал «Вычислительные технологии» Т. 10, Спецвыпуск. СВ-Томск, 2005. С. 21–29.
[7] Ершова Г., Сафонов В. Интеграция в книжном мире. 21.10.2003. Журнал «Директор ИС», № 10. 2003 // Издательство «Открытые системы». http://www.osp.ru/cio/2003/10/080.htm
[8] Журавлева О.В., Лисовский К.Ю., Томусяк Г.С., Томусяк Э.С. Функциональные методы обработки слабоструктурированных данных и их применение для построения электронных библиотек // Сборник трудов Третьей Всероссийской конференции по электронным библиотекам. КНЦ РАН, Петрозаводск, сентябрь 2001. http://www.pair.com/lisovsky/misc/biblio/index.html
[9] Игнатович Н Брокер интеграции приложений // Журнал «Открытые системы», 2003. № 09. Издательство «Открытые системы». http://www.osp.ru/os/2003/09/008.htm
[10] Belov S., Mikhalov N., Vyazilov E. A model of the distributed marine information resources – approaches and decisions // ICES-IOC Study Group on the Development of Marine Data Exchange Systems using XML, Gothenburg, Sweden 26–27 May 2003. C. 79–100. http://www.ices.dk/reports/OCC/2003/SGXML03.pdf
[11] Botts Mike. OpenGIS® Sensor Model Language (SensorML), Implementation Specification, Version 1.0 // Open Geospatial Consortium Inc., 2006. – 118 c.
[12] Havens Steve. OpenGIS® Transducer Markup Language, Implementation Specification, Version 1.0.0 // Open Geospatial Consortium Inc., 2006. 258 pp. http://xml.coverpages.org/OGC-06-010r6-TransducerMarkupLanguage-TML.pdf
[13] ISO 19115. – 2003. 146 pp.
[14] MODBUS application protocol specification. Version 1.1b // Modbus-IDA, December 2006, http://www.modbus.org/docs/Modbus_Application_Protocol_V1_1b.pdf.
[15] Ocean Biogeographic Information System // The Census of Marine Life.2003. http://www.iobis.org/.

Unification of data structure for field research, exploration and resources using of World Ocean

E. Vyazilov, A. Fedortsov, A. Kobelev

The coding methods of system elements are presented in report for integration of distributed heterogenic information resources in Unified state system of information for World Ocean. The selecting problems of system element are showed. The approaches for development multi dimensional structures with using of life cycles of information about data base objects are proposed. The examples of multi dimensional data structures and creating of such data base are cited.

♣ Работа выполнена в рамках проекта РФФИ № 07-01-00662-а.