Современные подходы по созданию метаданных[1]

А.Е. Кобелев, Е.Д. Вязилов

ГУ «Всероссийский научно-исследовательский институт гидрометеорологической информации – Мировой центр данных» Росгидромета,
249035, г. Обнинск, ул. Королева, 6
E-mail: vjaz@meteo.ru

Представлено назначение метаданных. Выделены системные, тематические, интерфейсные метаданные и метаданные процессов. Определены недостатки существующих систем метаданных. Рассмотрены места возникновения и использования метаданных. Указаны примеры существующих систем метаданных. Приведены подходы по развитию метаданных (создание централизованно – распределенной схемы организации метаданных, использование систем управления контентом для удаленного ввода метаданных, мониторинг состояние метаданных, агрегация метаданных, широкое применение метаданных пользователями, администраторами баз данных, прикладными системами, др.).

Ключевые слова: классификация, место возникновения, использования метаданных, подходы, методы создания метаданных, агрегация метаданных.

Введение

Существует несколько определений термина метаданные. Метаданные это данные о данных или структурированные данные, которые описывают характеристики объектов-носителей данных, способствующие идентификации, обнаружению, оценке и управлению этими данными. Метаданные это каталоги, поименованные списки, уникально идентифицирующие объекты (приборы, массивы и базы данных, организации поставщики или производители данных и т.п.). В статье предпочтение отдается второму определению.

Основным средством получения сведений о данных, мониторинга их поступления, анализа состояния и развития информационных ресурсов являются метаданные. Исследования по созданию баз метаданных в России и за рубежом ведутся с начала семидесятых годов [1-5]. Независимо от предметной области метаданные рассматриваются на русскоязычном сайте http://metadata.ru, в области океанографии - http://marinemetadata.org/. Можно выделить следующие наиболее известные системы в области исследования окружающей среды:

- Global Change Master Directory (GCMD) - NASA, США, http://gcmd.nasa.gov;
- European Directory of Marine Organisations (EDMO), European Directory of Marine Environmental Data sets (EDMED), European Directory of Marine Environmental Research Projects (EDMERP), Cruise Summary Reports (CSR), European Directory of the initial Ocean-observing Systems (EDIOS) - Панъевропейский проект Sea Data Net, http://www.seadatanet.org/metadata);
- ГЕКАД (ГУ «ВНИИГМИ-МЦД»);
- Централизованная база метаданных Единой государственной системы информации об обстановке в Мировом океане (ЦБМД ЕСИМО, http://data.oceaninfo.ru/meta/).

В ГУ «ВНИИГМИ-МЦД» накоплен большой опыт по сбору, формализации, структуризации, поиску и использованию метаданных [6, 7]. Анализ этого опыта позволяет вы-

[1] *Работа выполнена в рамках проекта РФФИ №* 10-07-00352-а

делить недостатки существующих систем метаданных, провести классификацию метаданных, выделить этапы обработки данных, где появляются метаданные, уровни управления данными, где необходимы метаданные различного уровня обобщения, определить перспективные подходы по развитию метаданных.

Недостатками существующих систем метаданных являются:

– низкая оперативность обновления информации в системах метаданных;
– несогласованный ввод изменений объектов метаданных, в результате метаданные противоречивы, содержат дублирующие блоки (особенно для организаций, экспертов, параметров) и устаревшие записи, объекты метаданных не всегда увязаны между собой;
– недостаточная функциональность и степень автоматизации системы ведения метаданных (не всегда используется удаленный ввод сведений о данных на основе системы управления контентом);
– большинство систем ориентировано на работу с одним объектом метаданных. Наибольшее число систем создано для таких объектов, как сведения о массивах и базах данных, рейсах НИС, ученых и организациях;
– нет единой модели метаданных для всех объектов метаданных;
– службы ведения метаданных разрознены, созданные на международном (MEDI, EDMED, IPY), национальном (GEOSS, ЕСИМО) и ведомственном уровнях (EOSDIS) системы реализованы на разных принципах;
– нет четкого понимания единицы описания метаданных – экземпляра метаданных. За единицу описания сведений о массивах и БД берут рейс или проект, или набор данных. Единица описания – это совокупность параметров, непересекающаяся с другими совокупностями, данные по которым хранятся в БД и массивах. Атрибутами, которые определяют единицу описания метаданных, являются уровень обработанности данных – данные наблюдений, обобщенные, диагностические и прогностические данные; тип системы хранения данных – СУБД, структурированные файлы, объектные файлы данных, программные приложения; организация, пространственное и временное разрешения; методы (приборы) наблюдений;
– неполный набор объектов метаданных – пользователям нужны не только описания массивов, проектов, рейсов, а и сведения о программных средствах, форматах, методах и другие сведения.

Процесс ведения метаданных неэффективен и разрознен, одни и те же объекты поддерживаются различными странами, ведомствами, организациями. Пользователей не устраивает, что метаданные разрознены, недостаточно структурированы (имеется много полей свободного заполнения), противоречивы, содержат дублирующие и устаревшие записи. Объекты метаданных не всегда автоматически связаны между собой. В этом случае даже при хорошо организованной базе метаданных добиться актуальности и достоверности метаданных очень трудно. В существующие системы метаданных необходимо встраивать дополнительные объекты метаданных (описания технологий, сетей наблюдений, методов сбора и обработки и др.).

Назначение и классификация метаданных

Метаданные должны помочь пользователю ответить на следующие вопросы.

Где получить необходимые данные? Ответ: организация, другой источник данных.

Где находятся источники данных? Ответ: город, страна, субъект РФ, название; географический район; наблюдательные платформы; приборы; проекты; рейсы.

Кто и как представляет данные? Ответ: имена и адреса организаций, специалистов; проект, наблюдательная платформа; разработчик формата, ПС.

Что ищет пользователь? Ответ: параметры, методы, интервалы - частота взятия проб, методы наблюдений, контроля качества информации, алгоритм.

Как найти необходимые данные? Ответ: описание массивов и БД, проектов, рейсов НИС, размещение инструмента на наблюдательной платформе, др.).

Чем и где измерены значения параметров? Ответ: описание прибора; сведения о калибровке инструмента; тип и название платформы; владелец платформы.

Когда были измерены (получены) данные? Ответ: ведения о массивах, станциях.

Каким образом можно быстро разобраться в составе и структуре файлов и БД? Форматы хранения, сбора и обмена данными.

Что измеряют эти параметры? Ответ: единый словарь параметров.

Как рассчитывается тот или иной показатель и т.д.? Ответ: сведения о методах.

Каков минимально необходимый объем данных по пространственным и временным масштабам? Ответ: изученность района, период наблюдений, др.

Какие наиболее эффективные методы статистической обработки данных под конкретную задачу? Ответ: сведения о программных средствах.

Какие методы и формы представления информационной продукции? Ответ: сведения о web – ресурсах, интерфейсах, программных средствах, методах.

Метаданные появляются на различных этапах обработки данных:

Производство наблюдений. Здесь нужны сведения о наблюдательных сетях и методах определения различных параметров среды, способах и местах поверки приборов, описания платформ, сведения об измерительных средствах.

Сбор данных. На этом этапе необходимы сведения о технологиях сбора, форматах передачи данных, описания передаваемых и поступивших данных, проектах, параметрах телекоммуникационной системы.

Каталогизация данных. Создаются описания массивов данных, организаций - владельцев данных, пользователей, форматов сбора, наблюдательных проектов, параметров, методов сбора, первичной обработки, контроля данных, появляются сведения о единицах учета данных – рейсах научно-исследовательских судов, полетах спутников и др.

Накопление данных. Появляются сведения о технологиях, массивах и баз данных, методах контроля, обмена данными, технологиях, форматах, проектах и программах.

Хранение и защита данных. Нужны сведения о технологиях хранения, защиты;

Использование данных. Пользователю передаются сведения о методах использования данных, пространственно – временных координатах наблюдений, типовых запросах.

Анализ и ассимиляция данных. На этом этапе необходимы сведения о платформах, инструментах, качестве данных, методах наблюдений, методы первичной обработки данных.

Прогнозы состояния среды. Здесь тоже необходимы сведения о качестве данных, методах прогнозирования, обобщения, определения качества данных.

Климатическая обработка. В этом случае нужны сведения о методах обработки, контроля и анализа данных, программных средствах и технических средствах.

Моделирование. Необходимы сведения о моделях, методах, форматах выходных данных.

Распространение данных. При этом собираются сведения об экранных формах представления информации, форматах передачи данных, публикуемых документах, в т.ч. в Web-среде.

Поддержка решений. На этом этапе необходимы сведения об услугах и регламенте их предоставления, свойствах атрибутов.

Метаданные используются на различных уровнях управления данными:

- **локальном** – наблюдательная платформа (отдельная организация), здесь необходима детальная информация в виде сведений о рейсах НИС и их состоянии (в обработке, на каком носителе и т.п.), о состоянии изученности того или иного географического района по различным параметрам;
- **региональном** – проекты, экспедиции (организации), необходимы сведения о каждом наборе данных, единице сбора, учета и обмена данными (рейс, месячный поток данных от прибрежной станции);
- **национальном** – сведения о мореведческих организациях, массивах данных независимо от носителя, программных средствах обработки, форматах сбора и обмена на уровне страны, наблюдательных платформах, наблюдательных сетях и др.;
- **международном** - сведения о международных соглашениях, массивах данных, предназначенных или переданных в международный обмен, включая сведения о рейсах и станциях, форматах обмена данными, программных средствах их обработки.

На всех уровнях управления данными имеются как справочные сведения одного класса (сведения о массивах данных, источниках данных, форматах и т.п.), так специфические объекты для каждого уровня.

Метаданные можно разделить на:

– **системные метаданные** - используются для функций извлечения, преобразования, загрузки, управления, документирования, ограничения доступа к БД;
– **описательные метаданные (тематические)** - представляют смысловое содержание данных (период наблюдений, объем данных в логических и физических единицах; представляются в виде общих сведений о БД, сведениях об источниках данных и сведения о единицах сбора и хранения данных - рейсах, станциях, др.;
– **интерфейсные метаданные (метаданные сервисов)** - используются для описания экранов и создания отчетов;
– **метаданные процессов** - отражают информацию о характеристиках системы обработки данных (статистика загрузки БД, информация о календарном планировании и обработке, трафик, скорость доступа).

Перспективные подходы по развитию метаданных

Из перспективных подходов по развитию метаданных можно предложить несколько новых решений:

– создание широкого спектра взаимосвязанных объектов метаданных;
– централизованно – распределенная схема организации метаданных;
– использование систем управления контентом для удаленного ввода метаданных;
– мониторинг состояние метаданных,
– агрегация метаданных.

Создание широкого спектра взаимосвязанных объектов метаданных. Например, в рамках проекта SeaDataNet создается пять объектов метаданных; в ЦБМД ЕСИМО выделено более двадцати объектов. Основными объектами метаданных являются сведения о технологиях, информационных ресурсах – массивах и базах данных, наблюдательных сетях, форматах хранения данных, организациях, наблюдательных платформах, методах, проектах, параметрах, др. Выделение отдельных объектов метаданных позволяет значительно уменьшить дублирование информации в различных объектах, упростить поддержку актуальности данных и более полно представлять информацию при визуализации любого объекта метаданных.

В течение многих лет основным объектом метаданных считалось описание массивов и баз данных. К сожалению, наличие этого объекта не позволяет правильно связать весь спектр объектов метаданных, относящихся к тому или другому массиву данных. Поэтому предлагается главным объектом метаданных считать **технологию обработки** тех или иных данных. Технология продуцирует или использует данные при их сборе, обработке или распространении. С технологией всегда связаны один или несколько массивов данных (входной, выходной массивы, может быть несколько массивов на входе или на выходе). Все эти массивы должны быть представлены в метаданных. Каждое описание массива должно быть связано с такими объектами метаданных, как организация, формат, программные средства, методы, др. Каждый массив может продуцировать один или несколько информационных ресурсов. Каждый информационный ресурс должен наследовать описание массива данных, из которого он получен.

Важным моментом является правильный выбор единицы описания данных. Единица описания – это совокупность параметров, непересекающаяся с другими совокупностями, данные по которым хранятся в БД и массивах. Атрибутами, которые определяют единицу описания, являются

- уровень обработанности (агрегации) данных – первичные данные наблюдений, обобщенные данные, диагностические и прогностические данные;
- тип системы хранения данных – СУБД, система структурированных файлов данных, система объектных файлов данных, программные приложения;
- пространственное разрешение (точка, сетка, профиль, объект);
- временное разрешение (случайное, ежемесячное и др.);
- место хранения (организация);
- методы (приборы) наблюдений.

Технологии предлагается классифицировать по этапам обработки. То есть выделяются технологии для производства наблюдений; сбора данных; обработки, накопления и хранения; и распространение данных.

Базовый набор объектов метаданных включает сведения о: технологиях обработки данных, наблюдательных сетях, массивах и базах данных, форматах данных, платформах наблюдений (судах, прибрежных станциях, буях, спутниках), проектах, моделях, программных средствах, инструментах измерений, методах, библиографических источниках, организациях, рейсах научно-исследовательских судов, персонах, должностях, параметрах, терминах, интерфейсах, морских картах, выпускаемой информационной продукции, временных рядах измерений, данных в узлах сетки, объектных файлах, профилях.

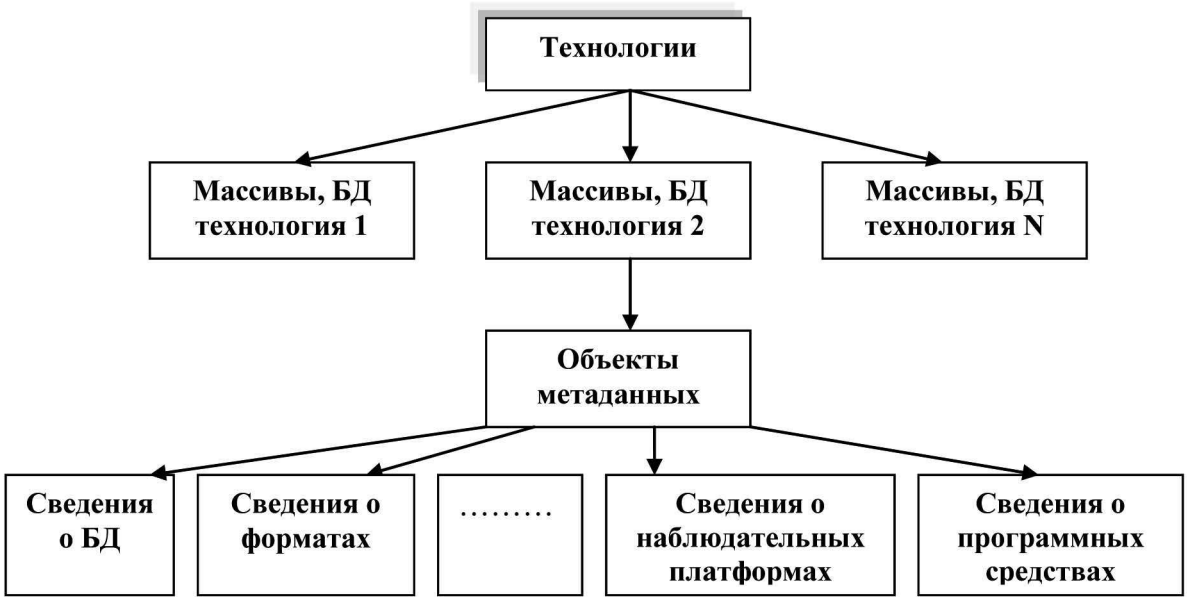

Рис. 1. Схема организации связей объектов метаданных

Разные объекты метаданных связаны с одним или несколькими объектами метаданных. Например, описание проекта связано со сведениями об организациях, экспертах, платформах, а описание массивов и баз данных - со сведениями об организациях, сетях наблюдений, форматах, проектах, экспертах, программных средствах, наблюдательных платформах, методах, приборах, нормативно-методических документах, терминах.

Рис. 2. Модель базы метаданных

Для описания объектов метаданных можно рекомендовать выделить следующие блоки (записи) для описания одинаковых комплексов атрибутов в различных объектах метаданных: иерархия структуры и содержания данных; временное обобщение (срок, сутки, месяц, год, другое); вертикальное обобщение (стандартные уровни, глубина/высота и т.п.);

классификаторы (общие коды); географические характеристики (координаты, названия морских районов, стран, субъектов федерации, городов, портов и т.п.); сведения о формате транспортного файла; принадлежность пользователя к группе, роль пользователя, тип аутентификации пользователя; дополнительная информация (ссылки, ключевые слова).

Метаданные должны быть организованы по централизованно – распределенной схеме. То есть управление метаданными должно быть централизованно в рамках одной предметной области, а хранение, особенно для одинаковых объектов метаданных (рейсы НИС, описания массивов данных, проектов) – распределено между центрами. Для организации обмена метаданными можно использовать язык XML, структуры некоторых объектов метаданных разработаны в проекте Sea Data Net для систем EDMED, CSR, EDMERP, CDI. Каждая организация должна представлять метаданные на своих сайтах в стандартизованных форматах, а национальные и международные центры данных могут периодически собирать эти XML файлы и объединять их в соответствующие объекты метаданных.

При ведении метаданных должен использоваться единый словарь параметров, включающий следующие атрибуты: название, единицы измерения, формат хранения параметра, используемый классификатор и метод определения параметра, реферат, другое. Словарь должен использоваться в различных объектах метаданных (описание массива данных, информационного ресурса, прибора, программных средств, др.), при интеграции данных и прикладной обработке.

При создании метаданных должны использоваться общие классификаторы и коды (из стандарта ИСО 19115, Межправительственной океанографической комиссии ЮНЕСКО, Всемирной метеорологической организации, Международного гидрографического бюро). При этом любой центр или пользователь волен использовать локальный классификатор, но при этом в метаданных должно быть указание на его использование. Тогда при обмене метаданными может быть произведено маппирование используемых кодов.

Важным моментом управления метаданными является мониторинг состояния метаданных, т.е. автоматизированный контроль ошибок ввода, оценка полноты заполнения полей и всех описаний объектов метаданных, получение информации о вкладе центров в пополнение метаданных.

Накопленные метаданные уже сейчас позволяют сделать выводы о состоянии перевода в цифровую форму данных о природной среде, поэтому в технологии ведения метаданных необходимо включение средств **агрегации сведений о данных, находящихся в базах метаданных** (получение количества массивов данных по организациям, видам наблюдений, наблюдательным платформам; расчет количества логических единиц сбора данных – станций, экспедиций, наблюдательных платформ, профилей, параметров за период наблюдений, для географического района, организации – хранителя данных и т.п.).

Метаданные должны более широко использоваться при прикладной обработке данных, т.е. приложения должны активно использовать один или несколько объектов метаданных для предоставления пользователю необходимой информации об источниках данных, методах определения параметров среды и т.д. Например, в приложении «Мониторинг наблюдательных сетей», кроме состояния наблюдательных сетей по ведомствам, регионам, можно получить сведения о наблюдательных платформах, приборах, экспертах, нормативно методических документах, в этой области. Кроме того, метаданные используются в электронно-справочных климатических пособиях, при прикладной обработке

данных. При этом для отражения метаданных, их поиска и представления в пространстве (расположение наблюдательных платформ, изученность наблюдениями, др.) должны использоваться геоинформационные системы.

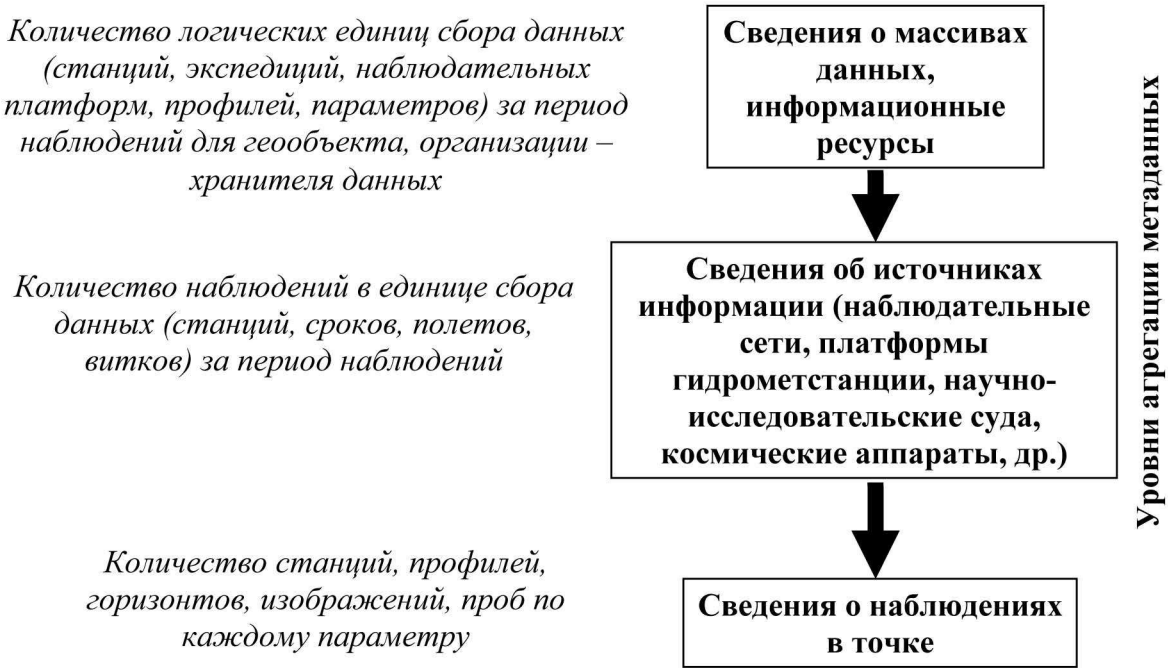

Рис. 3. Схема агрегации метаданных

Агрегированные метаданные необходимы для оперативного отслеживания потоков информации для подготовки информационно – аналитических материалов, принятия решений по планированию развития информационных ресурсов, баз данных, организации разработки новых таблиц, схем, подсхем, приложений с целью улучшения информационного обеспечения морской деятельности; получения метаданных о количественных характеристиках баз данных, их пополнении в любой момент из любого места.

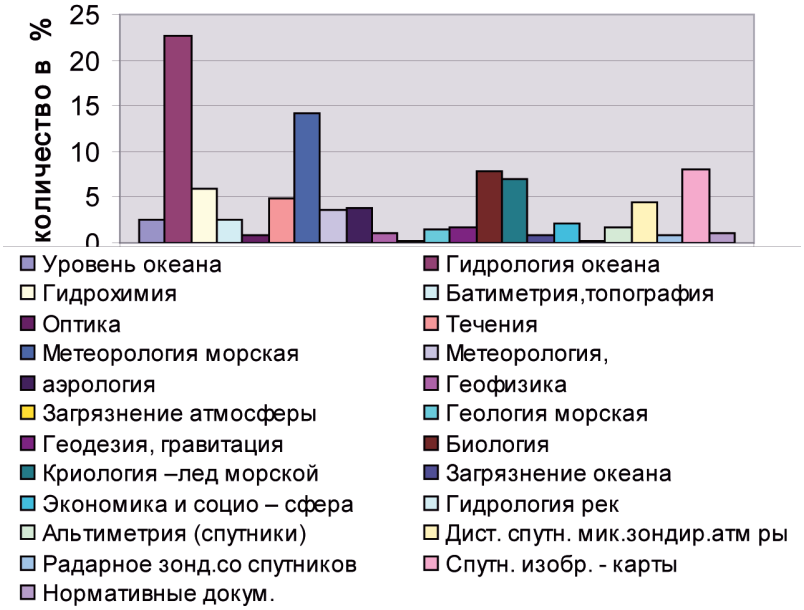

Рис. 4. Пример агрегированной информации «Количество баз данных по видам наблюдений» [8]

Таблица 1. Состояние метаданных ЕСИМО

Объекты метаданных	***Количество экземпляров***
Сведения о массивах и базах данных	
Рейсы НИС	34344
Морские прибрежные станции и посты	797
Информационные ресурсы	1450
Сведения об организациях	4072
Сведения о судах	21404
Сети наблюдений	47
Приборы и измерительные комплексы	492
Проекты и программы исследований Мирового океана	268
Массивы и БД	1030
Эксперты	386
Интернет-ссылки	520
Программные средства	205
Форматы	157
Методы	33
Термины	1316
Параметры	1716
Кодификаторы	363

Заключение

Предлагаемые подходы по развитию метаданных использованы при реализации ЦБМД ЕСИМО (http://data.oceaninfo.ru/meta/), портала Международного полярного года (http://mpg-info.ru) и могут быть использованы во многих предметных областях, в т.ч. и при хранении спутниковой информации. Пример удаленного ввода метаданных можно найти по адресу http://data.oceaninfo.ru/inf/index.jsp.

Совместное использование метаданных из нескольких источников для широкого спектра объектов метаданных в распределенной архитектуре – это стратегическое направление развития метаданных.

Литература

1. *Алексеенко Е.А., Вязилов Е.Д., Рогачев А.Е.* Проектирование БД справочной океанографической информации. - М. Гидрометеоиздат. 1986. – 40 с.
2. *Хохлов Ю.Е., Арнаутов С.А.* Обзор форматов метаданных. - Институт развития информационного общества. 2005. [Электронный ресурс]. Режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/methodology/md_rev/md_intro/md_example.
3. *Муралидхар Прабхакаран.* Управление метаданными в корпорации. [Электронный ресурс]. 15.07.2005. Перевод: Intersoft Lab. Режим доступа: *http://www.iso.ru/journal/articles/416.html*
4. *Marine* Community Profile of ISO 19115. Version 1.1. 2006-05-01. – 55 с.
5. *Еремеев В.Н., Суворов А.М., Владимиров В.Л. и др.* Каталогизация данных океанологических наблюдений на Украине. - Севастополь. Препринт. Национальный научно-технический совет по проблемам Мирового океана, НАМИТ при Кабинете Министров Украины, Комиссия по проблемам Мирового океана НАН Украины, Морской Гидрофизический Институт НАН Украины. 1995. - 78 с.
6. *Вязилов Е.Д.* Метаданные как основа управления глобальными и локальными базами данных // Журнал «Новости ЕСИМО». 2001. Вып.7. -11 с.

7. *Вязилов Е.Д.* Консолидация метаданных в области наук об окружающей среде // Журнал "Вычислительные технологии" Т. 10, Спецвыпуск. СВ-Томск, 2005 С.30-38.
8. *Вязилова Н.А.* Агрегированные характеристики некоторых объектов метаданных на портале ЕСИМО // Новости ЕСИМО. 2005. Вып. 23. 12 с.

The modern approaches to metadata development

A.E.Kobelev, E.D.Vyazilov

State Establishment "All-Russian Research Institute of Hydrometeorological Information – World Data Centre"
249035, Obninsk, Koroleva Str, 6
E-mail: vjaz@meteo.ru

It is presenting the metadata appointment. There are systemic, thematic, interfaces metadata and process metadata. It is selected the problems of working metadata systems. It is presenting the places of возникновения и and using of metadata. It is describing the approaches for metadata development (the creating of centrally– distributing scheme metadata, the using of content management systems for remote metadata input, monitoring of metadata statement, metadata aggregation, wide users application of metadata, data bases administrations, implicating systems, etc.).

Keywords: classification, origin place, metadata using, approaches, creating methods, metadata aggregation.