Моделирование распределенных научных вычислительных процессов посредством применения технологии рабочих процессов

© Нестеренко А.К.
alexn@ccas.ru

[*] Бездушный А.А.
lexa971@mail.ru

Сысоев Т.М.
tim@ccas.ru

Бездушный А.Н.
bezdushn@ccas.ru

[**] Ярощук И.О.
yaroshchuk@poi.dvo.ru

Вычислительный центр им. А.А. Дородницына РАН
[*] Московский Физико-Технический Институт
[**] Тихоокеанский Океанографический Институт ДВО РАН

Аннотация

В работе рассматриваются проблемы и задачи, возникающие при моделировании сложных распределенных научных вычислительных процессов. Анализируются соответствующие требования к системе исполнения научных рабочих процессов, описывается архитектура одной из реализаций такой системы. Решение демонстрируется в применении к задаче автоматизации исследований гидрофизических процессов.

1 Введение

Рабочие процессы привлекают большое внимание со стороны разработчиков информационных систем, потому что автоматизация исполнения рабочих процессов [13], протекающих в различных областях деятельности человека, позволяет:

- Повысить контролируемость и предсказуемость таких процессов.
- Ускорить процесс взаимодействия с пользователями и приложениями.
- Посредством развитых средств мониторинга собирать статистику выполнения рабочих процессов для их последующей оптимизации.
- Разрабатывать гибкие языки декларации маршрута рабочих процессов.
- Избавить исполнителя от выполнения множества рутинных типовых задач, а курьеров – от необходимости играть роль «транспортного механизма» рабочих процессов.
- Выполнять эффективную параллельную работу.
- Привлекать к этапам выполнения процесса только специалистов с необходимым уровнем квалификации за счет гибкой политики ролей пользователей.

На данный момент существует целый ряд систем управления потоками работ. Большая часть усилий разработчиков программного обеспечения в последнее время направлена на исследование рабочих процессов в бизнес среде. Хотя бизнес процессы и заслуживают того внимания, которое им уделяется, существует еще один класс рабочих процессов, наиболее часто встречающихся при решении сложных научных вычислительных задач [1,2,3]. Процессы данного класса получили название *научных потоков работ*. К основным отличиям бизнес процессов и научных вычислений можно отнести следующее:

- Бизнес процессы ориентированы по большей части на сложные потоки управления, в то время как в научных процессах большее внимание уделяется потокам данных.
- Научные процессы предъявляют высокие требования к средствам преобразования данных, так как экспериментальные данные представлены в сильно различающихся форматах.
- В научных процессах особо важная роль отводится поддержке средств моделирования и анализа полученных данных.
- Интеграция разнородных данных в соответствии с семантическим описанием их структуры является неотъемлемой частью научного вычислительного процесса.
- Научные процессы тесно взаимодействуют с человеческими ресурсами, которые могут «вмешиваться» практически в любое состояние,

Труды 7[ой] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005, Ярославль, Россия, 2005.

например, изменяя контекст, вводя дополнительные параметры.
- Важной задачей научных вычислительных процессов является хранение промежуточных результатов вычислений для последующих обработки и анализа.

Термин *научные потоки работ* описывает набор структурированных *действий* и вычислений, которые возникают при решении научной проблемы. Данный тип рабочих процессов особенно важен с точки зрения автоматизации повседневной деятельности научных сотрудников. Действительно, вот несколько причин, по которым *научные потоки работ* представляют особый интерес исследователям в данной области:
- Научные сотрудники составляют значительный процент пользователей информационных и вычислительных сетевых ресурсов.
- Наука всегда была требовательной к объемам вычислений, обрабатываемой информации. Ученые больше не могут обходиться при проведении повседневных экспериментов без помощи разнообразных измерительных и вычислительных систем, данных смежных областей.
- Научные потоки работ могут стать основными составляющими успеха при автоматизации гетерогенных вычислительных процессов. Потоки работ, позволяющие выполнять параллельные вычисления, асинхронное взаимодействие с пользователями и внешними системами, обработку исключительных ситуаций дают ученым возможность полноценно использовать комплексные вычислительные процессы для решения сложных научных проблем, не прибегая к низкоуровневому программированию.

Специфика научных вычислительных процессов определяет ряд задач, которые должна решать система управления потоками работ:
- Решение научной проблемы, как правило, сопряжено с обращением к ряду вычислительных и измерительных систем. Вычисления включают в себя большое количество этапов по преобразованию и обработке данных, а так же этапы обычной верификации и валидации данных на входе и выходе математических алгоритмов.
- Должны обрабатываться несоответствия в форматах данных между ресурсами хранилищ научной информации и аналитическими средствами с конвертированием данных в случае необходимости к требуемому представлению.
- При возникновении исключительных ситуаций должен выполняться «семантический откат» выполненных в ходе процесса изменений, так как обычный откат изменений в хранилищах данных зачастую невозможен в связи с большой продолжительностью по времени некоторых этапов вычислительного процесса.
- Многие научные вычислительные процессы могут продолжаться длительный период времени (недели и даже месяца), что накладывает дополнительные требования к надежности и отказоустойчивости системы управления потоками работ.
- Научные потоки работ так же могут привлекать человеческие ресурсы. Это особенно актуально на ранних стадиях работы вычислительного процесса. Роли участвующих в процессе людей должны быть четко определены для обеспечения эффективного взаимодействия процесса и наиболее подходящего специалиста.
- Вычислительная среда гетерогенна. Она включает в себя средства от суперкомпьютеров и специализированных измерительных систем до рабочих станций. Это обеспечивает дополнительную сложность задачи поддержки и управления вычислительными процессами.

На текущий момент большая часть действий, составляющих научные вычислительные процессы, уже выполняется учеными-экспериментаторами. Однако, при их автоматизации с помощью научных потоков работ, повышается эффективность обеспечения вычислений, которая определяется такими факторами, как: выразительные языки описания рабочих процессов, эффективные средства их исполнения и мониторинг выполнения этапов вычислительного процесса.

2 Проблемы, возникающие на пути решения задачи автоматизации научных вычислительных процессов

Научные потоки работ выходят за рамки обычных бизнес процессов. Для практического использования научных рабочих процессов должен быть решен ряд возникающих при этом проблем. Первая категория таких задач относится к рабочим процессам в целом. Она включает такие задачи, как:
- декларативное определение потоков управления и данных;
- семантическая обработка исключений;
- принятие решений человеком по отдельным этапам вычислений;
- управление ролями участников процесса и динамическое изменение этих ролей;
- автоматическое исполнение и мониторинг рабочих процессов;
- координация и синхронизация с другими научными и бизнес процессами.

Вторая категория задач характерна для научных потоков работ и включает требования, выполнение которых необходимо при проведении научных вычислений, но которые не могут быть полностью адресованы традиционным системам управления потоками работ:

- возможность взаимодействия с большим количеством аналитических средств, не только с хранилищами данных;
- функционирование в различных вычислительных средах, включая суперкомпьютеры;
- широкие возможности по преобразованию данных;
- наглядное визуальное представление описания и состояний исполняющегося потока работ с поддержкой вывода графической информации.

На основании приведенного сравнения требований к функциональности стандартных и научных рабочих процессов можно выделить список общих требований к системе исполнения потоков работ для возможности моделирования сложных научных вычислений:

- наличие выразительных средств декларации потоков работ;
- обеспечение прозрачного доступа к научным данным и вычислительным сервисам;
- возможность построения композиций вычислительных процессов;
- масштабируемость: многие научные потоки работ оперируют большими объемами данных и/или нуждаются в высокоскоростном доступе к вычислительным системам;
- асинхронное взаимодействие: долгоживущие потоки работ должны иметь возможность исполнения в фоновом режиме на удаленном сервере без необходимости сохранения постоянной связи с клиентом;
- эффективные механизмы обработки исключительных ситуаций;
- взаимодействие с пользователями: многие научные потоки работ нуждаются в принятии решений человеком на различных этапах выполнения;
- динамический выбор вычислительных систем, удовлетворяющих потребностям процесса, для взаимодействия;
- эффективные механизмы преобразования данных в различных форматах.

3 Архитектура системы управления научными потоками работ

В данном разделе описывается архитектура разрабатываемой системы автоматизации исполнения научных потоков работ. Данное решение следует ряду WEB-стандартов для поддержки открытой модульной архитектуры. На следующей диаграмме приведена компонентная структура системы исполнения научных рабочих процессов:

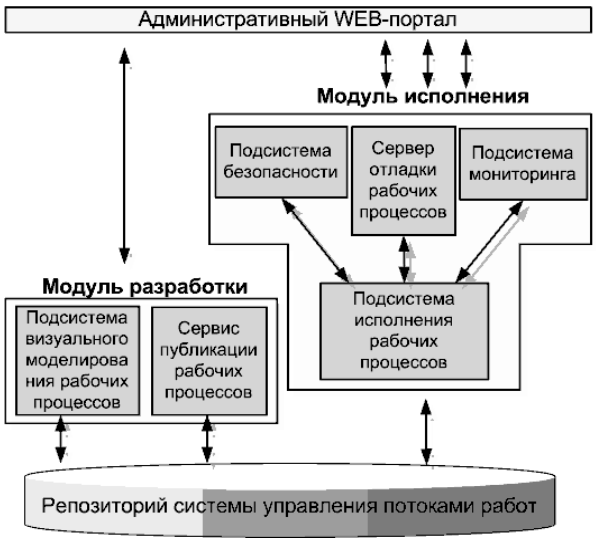

Рис. 1. Компонентная схема системы исполнения рабочих процессов

В качестве языка описания автоматизированных рабочих процессов в системе используется язык BPEL4WS (Business Process Execution Language For WEB-Services[4]), как наиболее выразительный язык, позволяющий описывать как блочные, так и графовые потоки работ. Вычислительные сервисы, сервисы преобразования данных и другие участники рабочего процесса представлены WEB-сервисами, следующими архитектуре WSA и стандартам WSDL (Web Services Description Language[5]) и SOAP (Simple Object Access Protocol[6]). Язык описания композиций WEB-сервисов BPEL4WS поддерживает основные конструкции, необходимые для эффективного описания вычислительных процессов:

- параллельное вычисление с поддержкой графовых структур;
- обработка исключительных ситуаций с возможностью «семантической» компенсации контекстов в случае долгоживущих вычислительных транзакций;
- преобразование данных;
- асинхронное взаимодействие с поддержкой механизмов корреляции сообщений;
- возможность динамического выбора участников процесса;
- представление рабочего процесса в виде WEB-сервиса (создание композиций вычислительных процессов).

Для возможности визуального моделирования описаний научных процессов реализованы средства моделирования BPEL4WS-документов с поддержкой средств синтаксической и структурной верификации, а так же механизмов управления уровнями детализации редактируемых описаний:

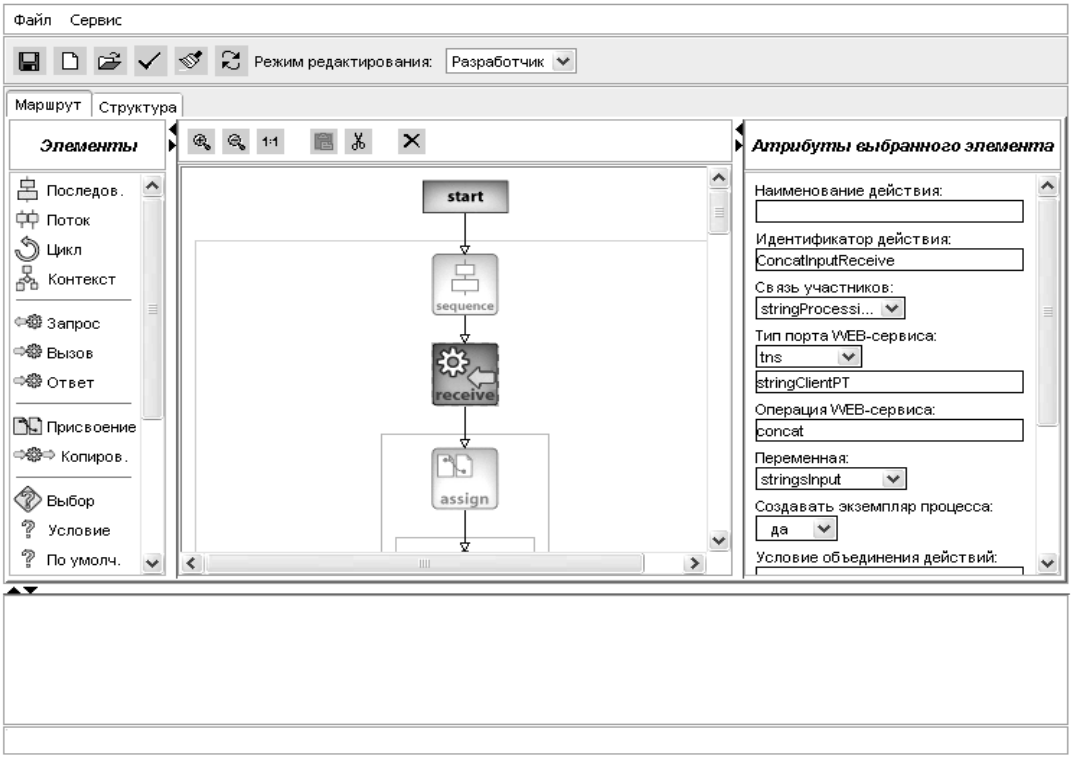

Рис. 2. Средства визуального моделирования описаний рабочих процессов

Регистрация новых описаний рабочих процессов и их размещение в репозитории системы управления осуществляется посредством административного WEB-сервиса, после чего информация о зарегистрированном процессе и списке активных экземплярах доступна для просмотра и модификации.

Динамический запуск и исполнение потоков работ осуществляется через визуальное приложение, оформленное в виде Java-апплета и взаимодействующее с сервером исполнения по специальному XML-протоколу. Данный клиентский интерфейс позволяет визуализировать процесс исполнения потока работ.

Разработанная система исполнения рабочих процессов имеет развитые средства визуальной отладки потоков работ. При этом реализована клиент-серверная архитектура, состоящая из сервера отладки, визуального клиента и XML-протокола для их взаимодействия, позволяющая выполнять полноценную отладку рабочих процессов в реальном времени.

Для возможности протоколирования хода исполнения рабочего процесса предназначен модуль журналирования состояний действий и исключительных ситуаций, возникающих на маршруте рабочего процесса. Доступ к журналу сообщений осуществляется через WEB-интерфейс.

Управление безопасностью доступа к операциям рабочего процесса осуществляется посредством подключаемых модулей безопасности, с помощью которых можно реализовать любую схему управления доступом. Для аутентификации и авторизации пользователя внешними системами, к которым производится обращение на маршруте потока работ, поддерживается возможность распространения контекста безопасности в заголовках SOAP-сообщений, передаваемых внешним вычислительным сервисам.

4 Применение технологии научных потоков работ к решению задачи гидрофизических исследований и мониторинга

В этом разделе рассматривается задача применения технологии рабочих процессов к решению задачи автоматизации исследований гидрофизических процессов.

В России, в рамках подпрограммы "Исследование природы Мирового океана" федеральной целевой программы "Мировой океан", проводятся исследования, направленные на решение научных, социально-экономических, правовых, экологических, институциональных проблем управления прибрежными территориями. Важную роль в контексте глобальной программы играют научные организации, ведущие комплексные исследования региональных акваторий. Изучение гидрофизических процессов, развивающихся на шельфах морей Дальнего Востока, является необходимой частью общих исследований и дает ценный вклад в общую

информационную базу. Для систематизации сведений и накопления знаний разрабатываются системы, формирующие общее информационное пространство. Развитие подобных систем направлено на решение задач комплексной обработки, анализа и интерпретации данных.

В течение длительного периода на стационарном гидрофизическом полигоне ТОИ ДВО РАН «мыс Шульца» проводятся комплексные океанологические наблюдения. Полученные сведения описывают разнообразные гидрофизические процессы в прибрежной области и являются основой для дальнейшего изучения шельфовой зоны Японского моря.

В настоящее время накоплен и продолжает поступать разнообразный материал экспериментальных измерений, результатов обработки и интерпретации. По результатам натурных данных гидроакустических измерений проводилось численное моделирование эволюции внутренних волн и распространения звука в шельфовой области [7,8,11,12]. Соответствующая схема исследований представлена на следующей диаграмме:

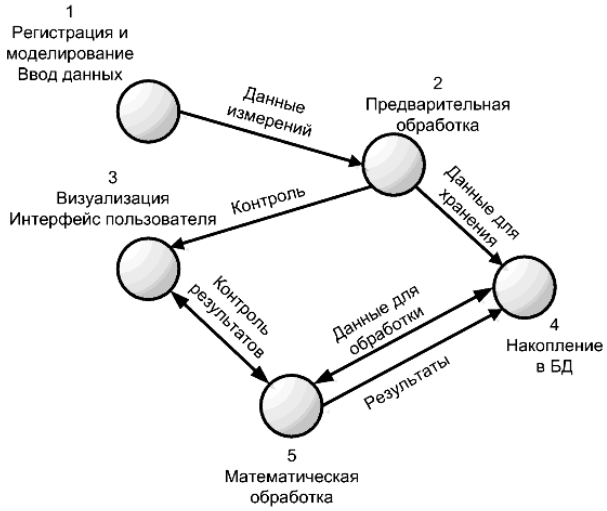

Рис. 3. Базовая схема мониторинга

Модульность системы отражает тот факт, что в процессе мониторинга задействованы достаточно большие вычислительные ресурсы: несколько компьютеров, аппаратно-регистрационные комплексы (буи, несущие датчики температуры, давления и др.). Программный комплекс состоит из системы ввода и предварительной обработки, модулей визуализации и численного анализа, системы хранения. Все системы могут работать как вместе, так и отдельными сегментами, синхронно или с разрывом по времени. Тем не менее, все операции являются компонентами единого процесса мониторинга, отражающего эволюцию реального физического объекта.

В соответствии с приведенной схемой процесс гидрофизического мониторинга условно разбивается на три основных этапа:

- Снятие показаний приборов.
- Предварительная обработка информации и ее размещение в хранилище данных.
- Использование полученного массива данных в качестве параметров математических алгоритмов, размещение результатов работы алгоритмов в хранилище информации.

Ввиду этого процесс обработки экспериментальных данных может быть представлен в виде композиции трех потоков работ.

4.1 Процесс сбора экспериментальных данных

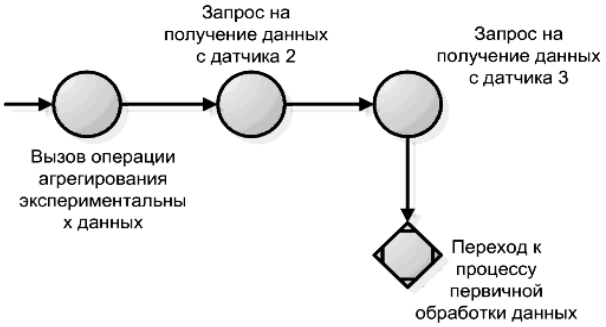

Рис. 4. Пример процесса последовательного сбора экспериментальных данных

Каждая из групп экспериментальных установок управляется отдельным микропроцессором [9,11] и представлена снаружи Web-сервером, поддерживающим взаимодействие по протоколу SOAP. Для снятия показаний с приборов в рамках рабочего процесса происходит последовательное обращение к методам таких Web-сервисов. Полученный таким образом массив информации является входным параметром процесса первичной обработки данных.

4.2 Процесс первичной обработки информации и размещение ее в хранилище данных

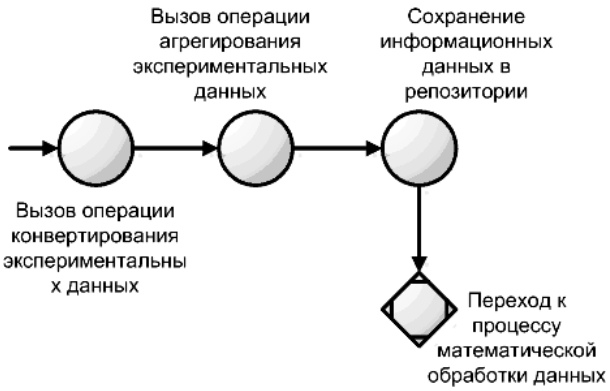

Рис. 5. Пример простого процесса первичной обработки экспериментальных данных

На данном этапе происходит обращение рабочего процесса к Web-сервисам конвертирования, а затем агрегирования данных, в

задачу которых входит преобразование форматов данных, получаемых с различных измерительных устройств к каноническому формату и их агрегирование [10] в соответствии с каноническим описанием (метаданными) [12]. После этого процесс обращается к Web-сервису импорта-экспорта хранилища экспериментальных данных для загрузки обработанной информации. Дальнейшая процедура анализа полученных данных – математическая обработка.

4.3 Процесс математической обработки

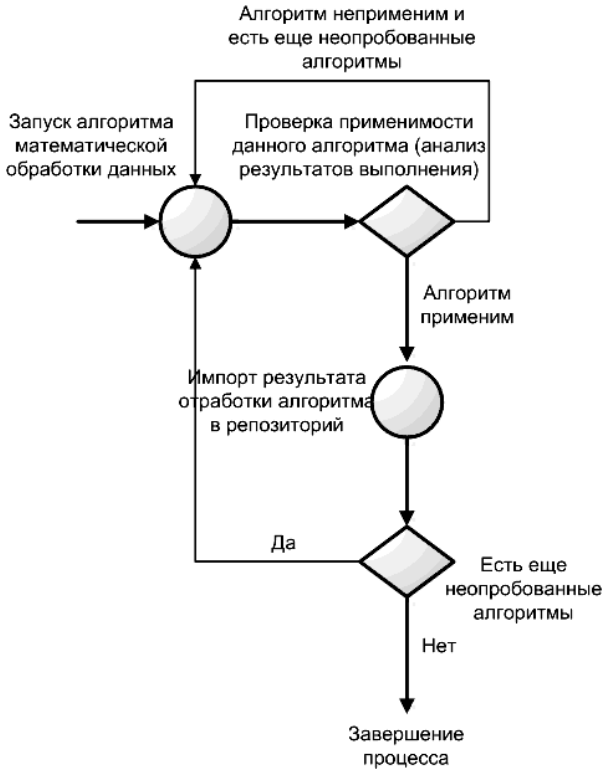

Рис. 6. Типовая схема процесса математической обработки экспериментальных данных

Данный процесс производит обращения к подключаемым Web-сервисам, реализующим различные алгоритмы математической обработки экспериментальных данных. При этом на этапе контроля применимости данного алгоритма процесс может привлекать людей для формирования экспертной оценки. В случае успешной отработки алгоритма, происходит сохранение результата обработки в хранилище данных посредством обращения к сервису импорта. Число подобных итерации равно числу доступных математических алгоритмов. При этом сам процесс выполнения такого алгоритма может также быть представлен в виде отдельного потока работ, выполняя, например:

- Обращение к дополнительным источникам данных для получения статистической информации, необходимой для выполнения расчетов [10].
- Использование внешних вычислительных сервисов общего характера для решения конкретных математических уравнений.
- Обработку исключительных ситуаций.
- Привлечение экспертов к отдельным этапам работы алгоритма.

5 Заключение

Таким образом, весь процесс гидрофизического мониторинга укладывается в применение набора стандартных технологий, предлагаемых потоками работ, в значительной степени автоматизирующих данный процесс и делающих его более гибко определяемым и управляемым.

На данном этапе работ над проектом выполняется прототипирование системы управления гидрофизическими наблюдениями на базе технологии научных рабочих процессов.

Научные потоки работ представляют собой интересный объект для исследования. Во-первых, научные вычислительные процессы имеют ключевые характеристики традиционных рабочих процессов, что делает их полигоном для обширного применения уже имеющихся наработок и исследований в этой области. Во-вторых, научные рабочие процессы достаточно сильно отличаются от бизнес процессов для того, чтобы являться объектом отдельных исследований, и тем самым приводят к появлению множества интересных задач, которые не возникают в результате анализа традиционных потоков работ. Процессы, аналогичные научным вычислениям, имеют место и в других применениях технологии рабочих процессов, что гарантирует актуальность рассматриваемых в данной статье проблем и их решений.

Литература

[1] Catherine Houstis, Spyros Lalis “A grid service-based infrastructure for accessing scientific collections: the case of the Arion system”, 2002.

[2] Bertram Ludäscher, Ilkay Altintas “Scientific Workflow Management and the Kepler System”, September 2004; revised March 2005. http://www.sdsc.edu/%7Eludaesch/Paper/kepler-swf.pdf

[3] Bertram Ludäscher, Kai Lin “Managing Scientific Data: From Data Integration to Scientific Workflows”, 2004. http://users.sdsc.edu/~ludaesch/Paper/gsa-sms.pdf

[4] Business Process Execution Language for Web Services Version 1.1. http://www-106.ibm.com/developerworks/library/ws-bpel/

[5] Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language. http://www.w3.org/TR/2004/WD-wsdl20-20040326/

[6] SOAP Version 1.2 Part 1: Messaging Framework. http://www.w3.org/TR/2003/REC-soap12-part1-20030624/
[7] Борисов О.В., Рутенко А.Н., Трофимов М.Ю. Пример гидроакустического мониторинга на шельфе Японского моря // Акустический журнал 1997, т 43.
[8] Борисов С.В., Рутенко А.Н., Коротченко Р.А. и др. Измерительно-регистрационный комплекс для акустико-гидрофизических исследований на шельфе и некоторые результаты его применения в натурных экспериментах.
[9] Коротченко Р.А., Трофимов М.Ю. Комплекс программ компьютерного моделирования гидрофизического полигона // Информатика в океанологии. ТОИ ДВО РАН, Владивосток, 1996. , с. 81-96.
[10] Нестеренко А.К., Сысоев Т.М., Бездушный А.А., Бездушный А.Н., Серебряков В.А. Интеграция распределенных данных на основе технологий Semantic Web и рабочих процессов. // Сборник докладов Шестой Всероссийской конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", Пущино, 2004.
[11] Коротченко Р.А., Бездушный А.Н., Ярощук И.О. Проект виртуального геофизического полигона на основе морской экспериментальной станции ДВО РАН "мыс Шульца"" // В кн. материалы докл. 3-й Всерос. симпозиума "Сейсмоакустика переходных зон". Владивосток: ДВГУ, 2003, с.163-165.
[12] Р.А. Коротченко, И.О. Ярощук, А.Н. Бездушный. Версия схемы метаданных экспериментальных исследований с приложением в гидроакустике // Электронный журнал, посвященный созданию и использованию электронных библиотек, том 7, выпуск 1. Москва: Институт развития информационного общества - 2004.
[13] А.К. Нестеренко, А.А. Бездушный, Т.М. Сысоев, А.Н. Бездушный. Возможности службы управления потоками работ по манипулированию ресурсами репозитория ИСИР // Сборник научных трудов X научно-практического семинара "Новые технологии в информационном обеспечении науки". Москва: 2003, с.206-231.

Modelling of the distributed scientific calculation processes with help of the workflow technology

Nesterenko A.K. Bezdushny A.A. Sysoev T.M. Bezdushny A.N. Yaroshchuk I.O.

This paper covers some important problems that take place while modeling complicated distributed scientific calculation processes. In the article appropriate requirements to the scientific workflow management system are analyzed and one of such systems architecture is presented. The solution is applied to the problem of the hydro physics researches automation.