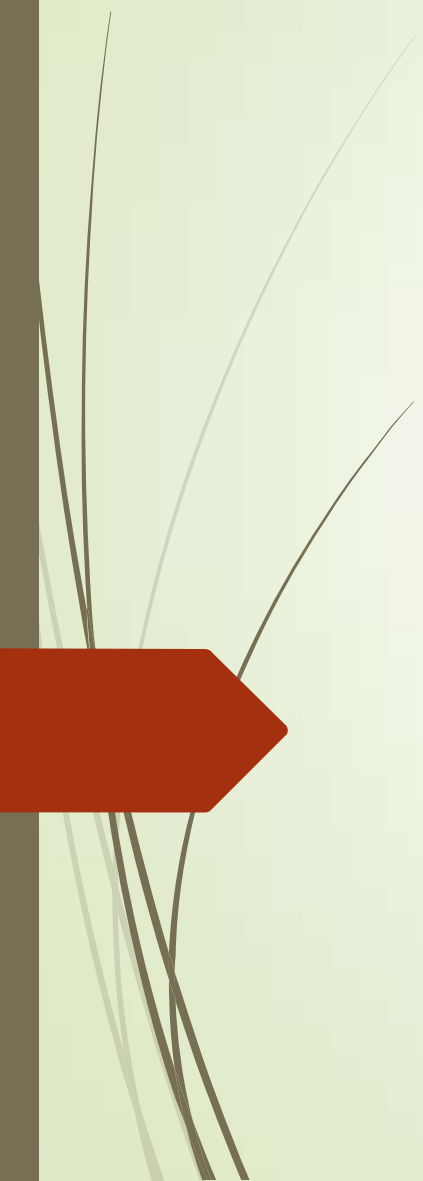




Поддержка повторного использования спецификаций потоков работ за счет обеспечения их независимости от конкретных коллекций данных и сервисов

Брюхов Д.О., Вовченко А.Е., Калиниченко Л.А.

Институт Проблем Информатики РАН



План

- Проблемы организации исследований в науках с интенсивным использованием данных (НИИД)
- Среды для публикации и повторного использования потоков работ
- Подход для решения проблем повторного использования за счет обеспечения их независимости от конкретных коллекций данных и сервисов
- Пример применения подхода на основе задачи определения вторичных стандартов

Введение

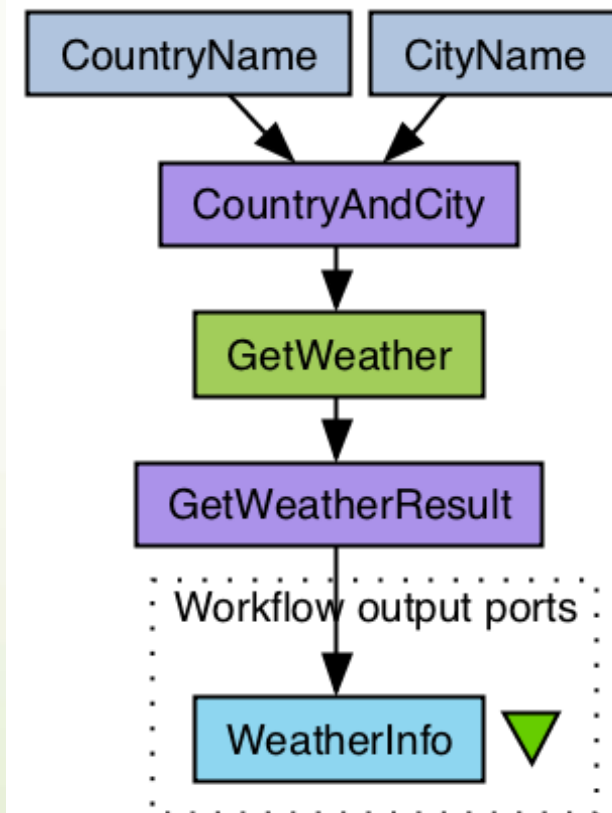
- Науки с интенсивным использованием данных (НИИД) развиваются в рамках новой парадигмы научных исследований, согласно которой новые знания образуются в результате анализа разнообразных данных, накопленных в результате проведения измерений, наблюдений, моделирования, вычислений.
- Данные становятся ключевым источником получения знаний в НИИД.
- Объем, разнообразие и качество накапливаемых данных быстро растет.
- Системы с интенсивным использованием данных имеют существенное пересечение с быстро развиваемой областью, именуемой «Big Data».

Проблемы организации исследований в НИИД

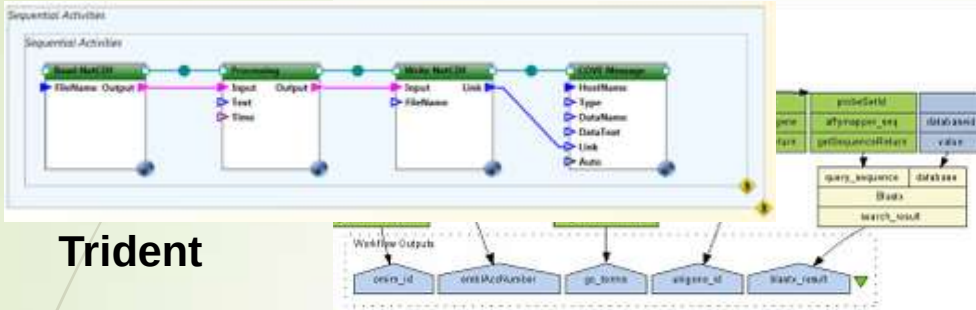
- Создаваемые в НИИД методы анализа данных и алгоритмы решения задач как правило ориентированы на конкретные коллекции данных, находящиеся в поле зрения конкретных ученых в конкретный момент. Из-за этого отсутствует возможность повторного использования таких методов, алгоритмов и их реализаций над другими данными, в других коллективах НИИД.
- Отсутствует практика накопления и повторного использования методов анализа данных, алгоритмов решения задач и их реализаций в научном сообществе НИИД.
- В НИИД отсутствует практика формирования ИТ-базированных, согласованных в сообществах концептуальных определений научных областей (включающих их структуру, понятия, спецификации методов, задач, и пр.).

Зачем нужны потоки работ?

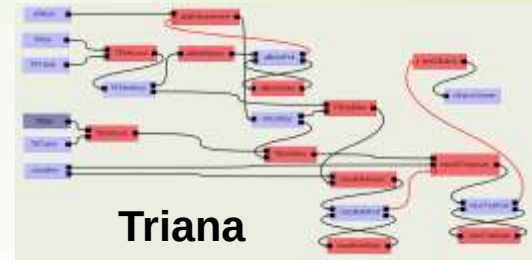
- Потоки работ как понятия (1996, Workflow Management Coalition):
 - «Частичная или полная автоматизация бизнес-процессов, в ходе которой документы, информация или задачи передаются от одного этапа обработки к другому для осуществления операции, в соответствии с набором процедурных правил»
- Система управления потока работ
 - «система, которая описывает поток работ (по сути, бизнес-процесс), создает его и управляет им при помощи программного обеспечения, которое способно интерпретировать описание процесса, взаимодействовать с участниками потока работ и при необходимости вызывать соответствующие программные приложения и инструментальные средства»
- Научные потоки работ
 - полезная парадигма для описания, управления и обмена при поддержке сложного научного анализа в НИИД



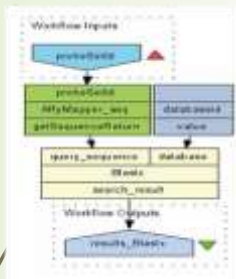
Системы управление потоками работ



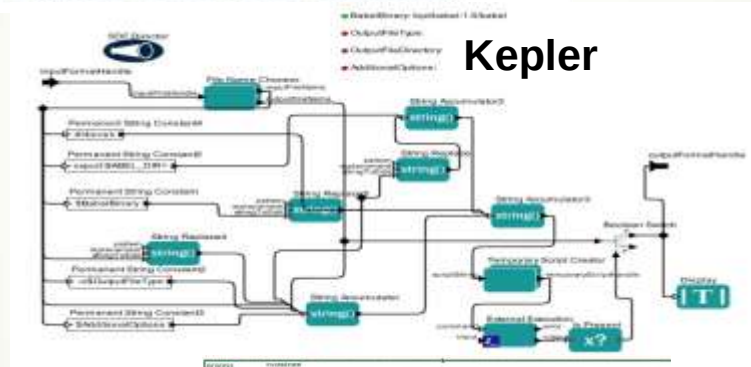
Trident



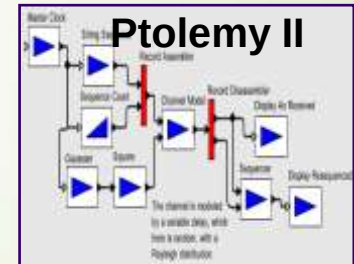
Triana



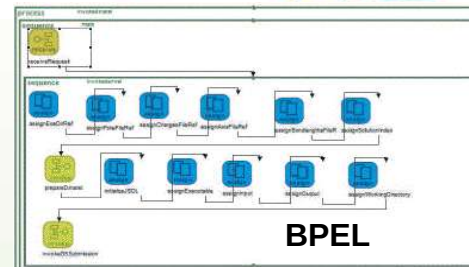
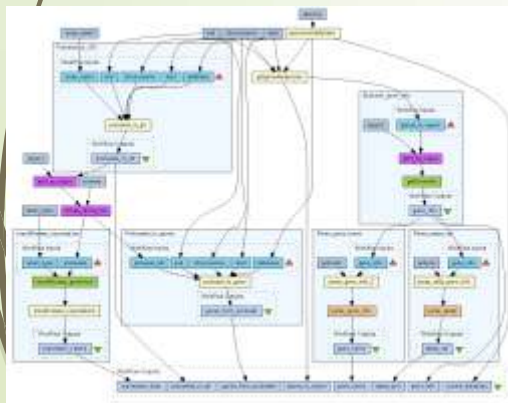
Taverna



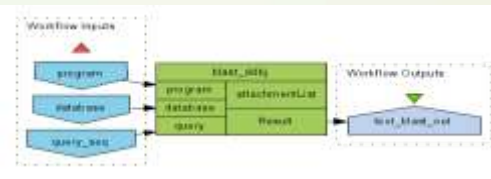
Kepler



Ptolemy II

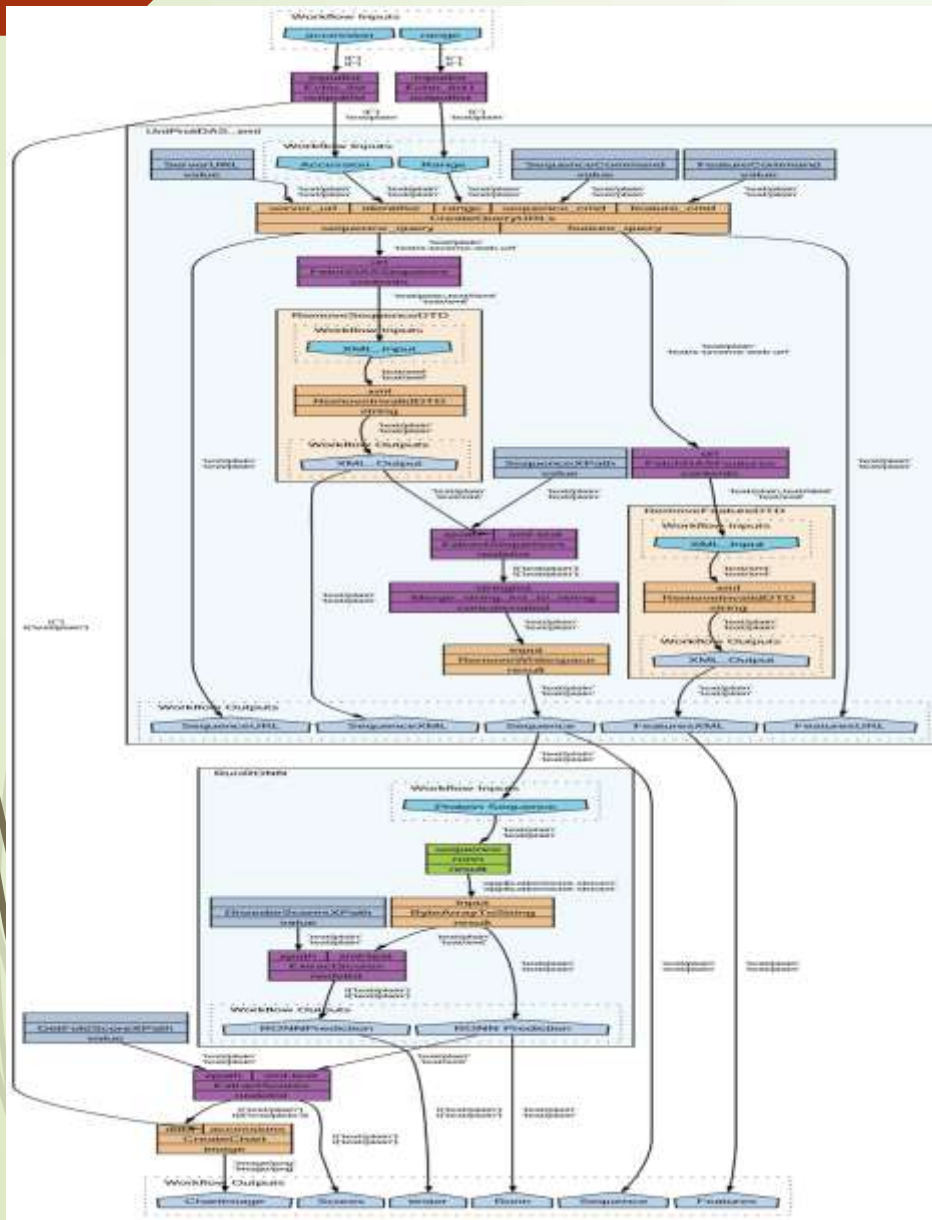


BPEL



BioExtract

Система управления потоками работ Taverna



- модель потоков данных (data flow model)
- поток работ состоит из сервисов, представляющих собой программные компоненты, и направленных связей между ними, выражающих зависимости по данным
- поддержка широкого набора как локальных, так и удаленных сервисов в различных областях науки
- поддержка вложенных потоков работ

Среда myExperiment

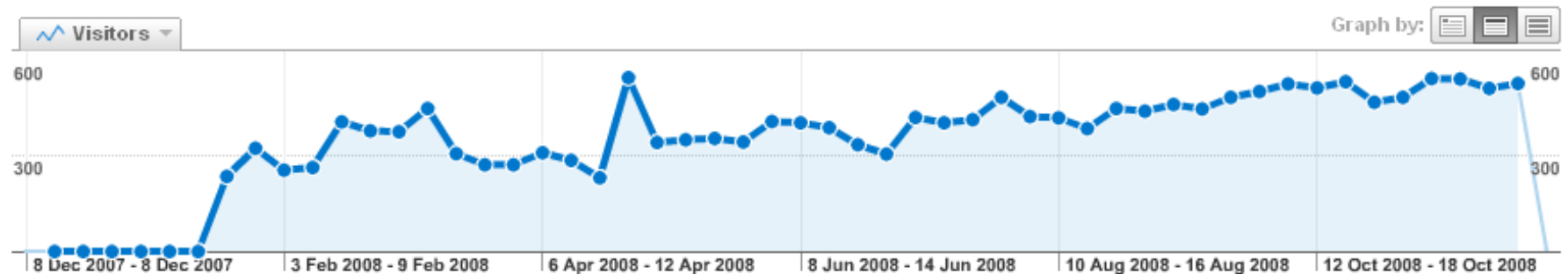
- Крупнейший социально ориентированный репозиторий научных потоков работ
- Позволяет ученым
 - находить интересующие их потоки работ
 - воспроизводить результаты этих потоков работ
 - повторно использовать существующие потоки работ для решения задач
- Позволяет создавать и присоединяться к группам для совместных исследований

The screenshot displays the myExperiment web interface. At the top, the logo 'my experiment' is visible alongside links for 'About', 'Mailing List', and 'Publications'. A navigation bar includes 'Home', 'Users', 'Groups', and 'Workflows'. Below this, a breadcrumb trail reads 'Home » Workflows » Success-Abandonment-Classification'. The main content area features the title 'Workflow Entry: Success-Abandonment-Classification' and metadata: 'Created at: 06/02/08 @ 14:35:41' and 'Last updated: 02/07/08 @ 17:15:'. A list of links follows: 'License', 'Credits (2)', 'Attributions (0)', 'Tags (6)', 'Featured in Packs (1)', 'Ratings (1)', 'Attributed Citations (0)', 'Version History', 'Reviews (0)', and 'Comments (0)'. A yellow banner indicates 'Version 3 (latest) (of 3)' with a 'View version: 3 (latest)' dropdown. Below this, it states 'Version created on: 06/02/08 @ 14:35:41 by: Andrea Wiggins' and 'Last edited on: 02/07/08 @ 17:15:25 by: Andrea Wiggins'. The workflow details show 'Title: Success-Abandonment-Classification' and 'Type: Taverna 1'. A 'Preview' section includes a thumbnail of a complex workflow diagram with the text '(Click on the image to get the full size)'.

Среды для публикации и повторного использования потоков работ

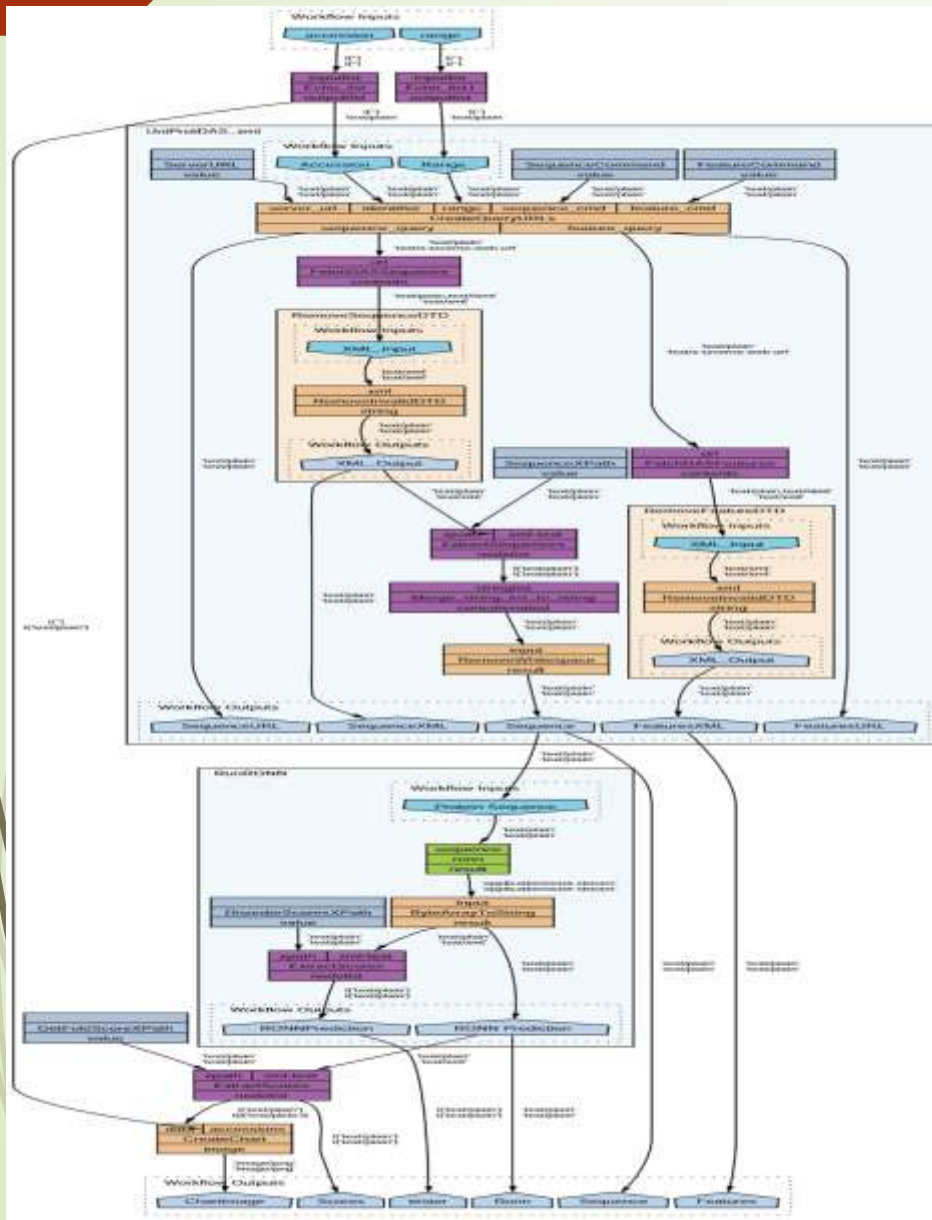
- myExperiment - среда для совместных исследований, в которой ученые могут публиковать потоки работ для решения задач
 - содержит более 2000 потоков работ
 - позволяет публиковать потоки работ в различных системах управления потоками работ (таких, как Taverna, Galaxy, Trident)
- ER-flow предоставляет ученым программную поддержку для создания, обмена и запуска потоков работ в различных системах управления потоками работ
 - поддерживает ASKALON, Galaxy, GWES, Kepler, LONI Pipeline, MOTEUR, Pegasus, P-GRADE, ProActive, Triana, Taverna, WS-PGRADE
- crowdLabs – система для поддержки обмена, визуализации и анализа потоков работ для системы управления потоками работ VisTrails

Распространение MyExperiment



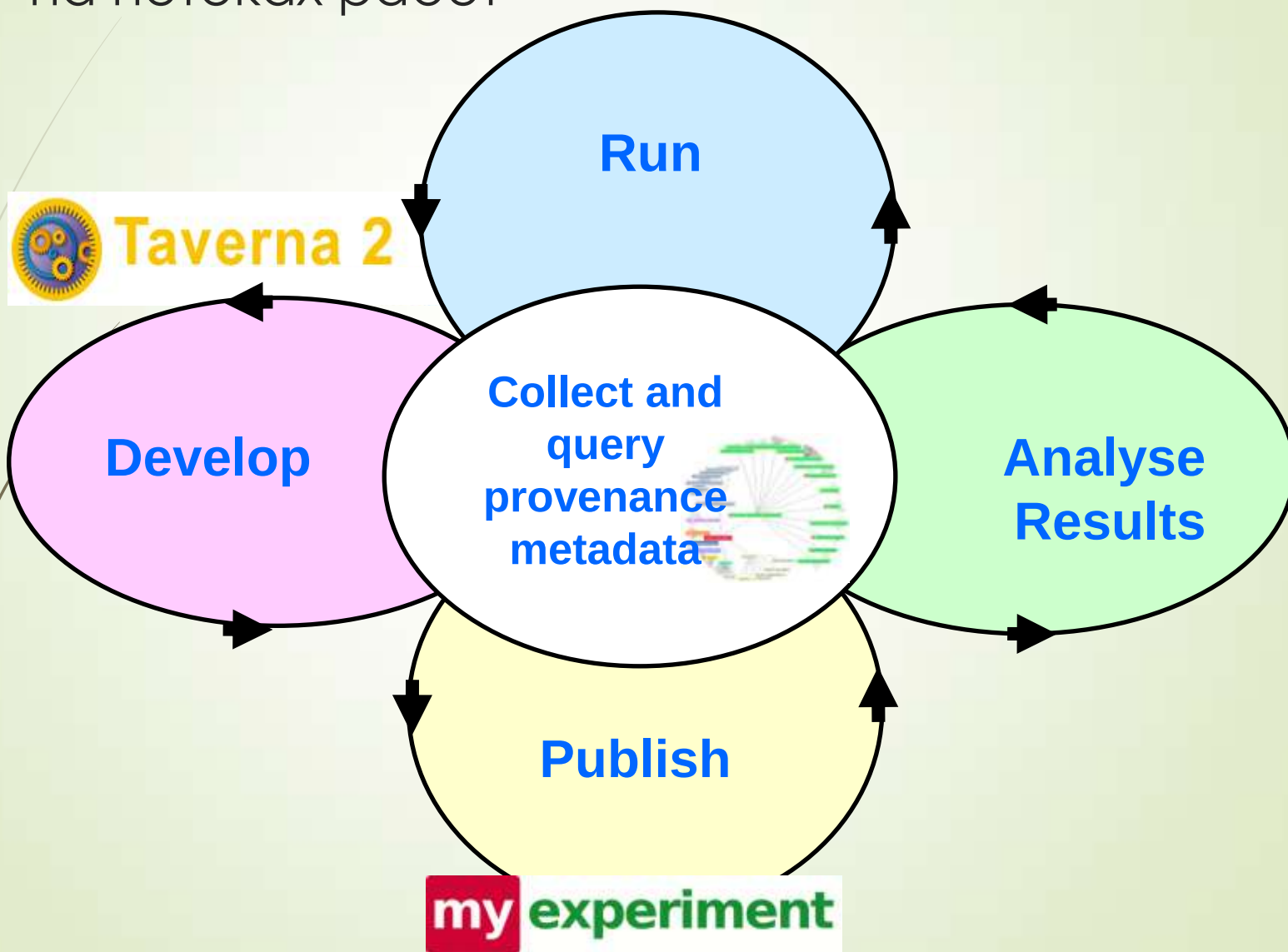
15,902 Absolute Unique Visitors

Система управления потоками работ Taverna



- модель потоков данных (data flow model)
- поток работ состоит из сервисов, представляющих собой программные компоненты, и направленных связей между ними, выражающих зависимости по данным
- поддержка широкого набора как локальных, так и удаленных сервисов в различных областях науки
- поддержка вложенных потоков работ

Жизненный цикл экспериментов, основанных на потоках работ





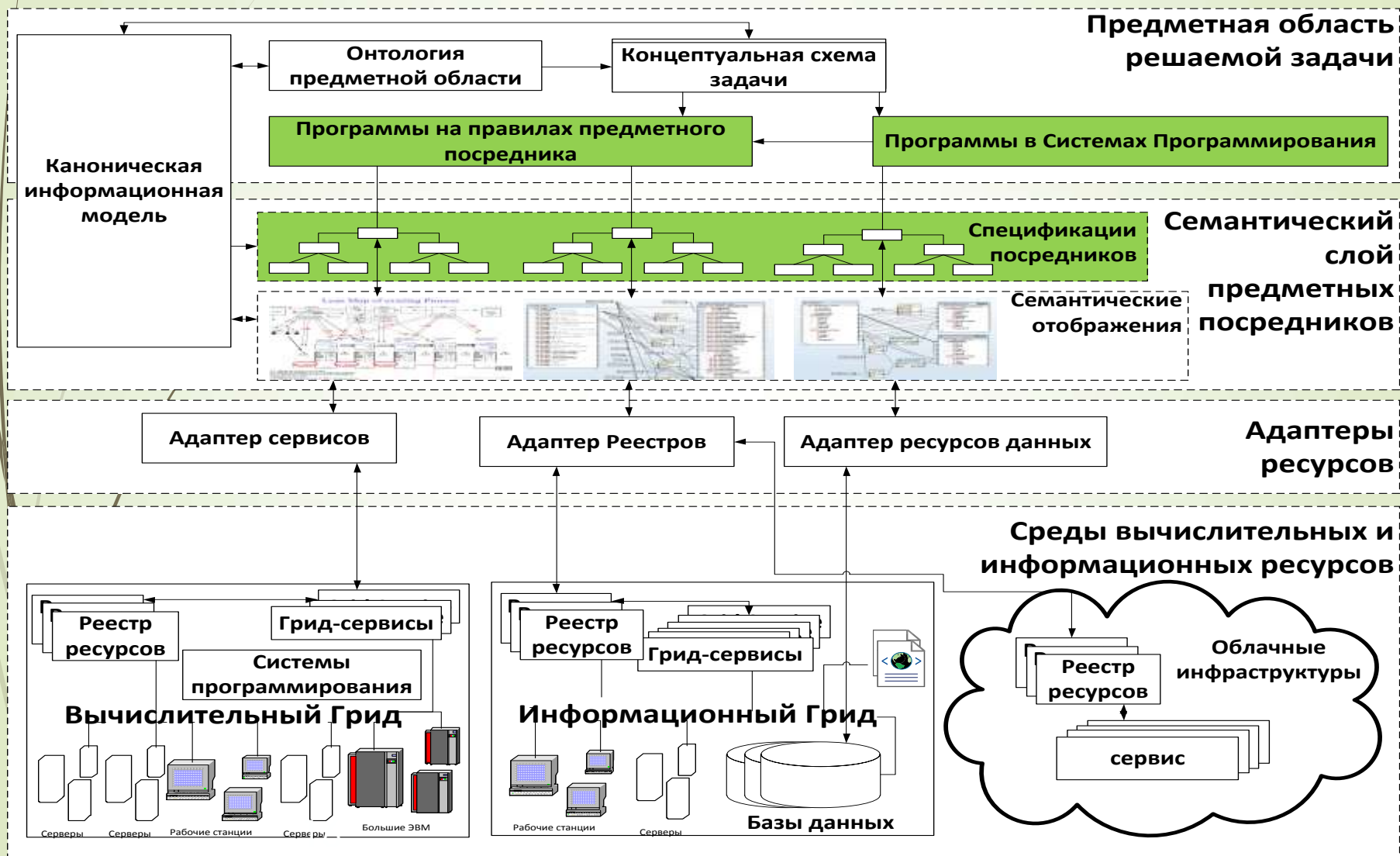
Проблемы повторного использования потоков работ в Taverna/myExperiment

- зависимость спецификаций потоков работ от конкретных коллекций данных и/или сервисов
- если какой-либо из сервисов или база данных в настоящий момент недоступны, то весь поток работ не сможет быть выполнен
- недостаточность спецификаций потоков работ для их повторного использования

Концепции инфраструктур предметных посредников

- Посредники реализуют подход к решению задач, ориентированный на проблему
 - Формулируется концептуальная спецификация задачи, включающая базовые сущности и понятия предметной области, функции, процессы
 - Сущности и понятия предметной области, определенные таким образом, не зависят от существующих информационных ресурсов
- Программы для решения задачи формулируются в терминах предметной области
- Идентифицируются ресурсы, релевантные задаче
 - Релевантные задаче ресурсы регистрируются в предметных посредниках, задающих отображение схем ресурсов в концептуальную спецификацию.

Инфраструктура предметных посредников



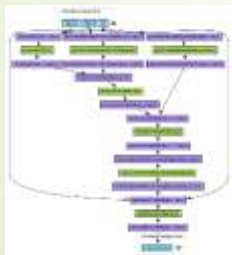
Обеспечению независимости потоков работ от данных

- Все сервисы в потоках работ Taverna определены в терминах конкретных сервисов и баз данных, что не позволяет задавать спецификации потоков работ независимо от конкретных ресурсов
- По сути, посредники представляют собой виртуальные базы данных, и в потоках работ Taverna их можно подключать аналогично обычным базам данных
 - посредством веб сервиса, созданного над посредником и реализующего интерфейс посредника
 - разработка специального плагина под средство разработки потоков работ Taverna Workbench
- Все доступные в Taverna ресурсы, используемые в качестве узлов в потоках работ, могут быть использованы также посредством посредников.

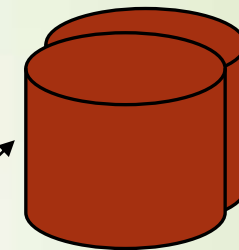
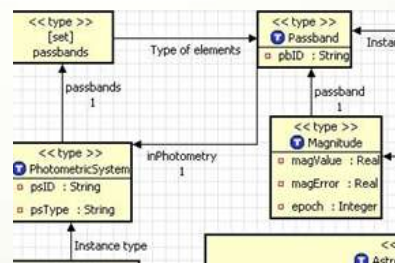
Подход для решения проблем повторного использования



Taverna



Посредник



Базы данных

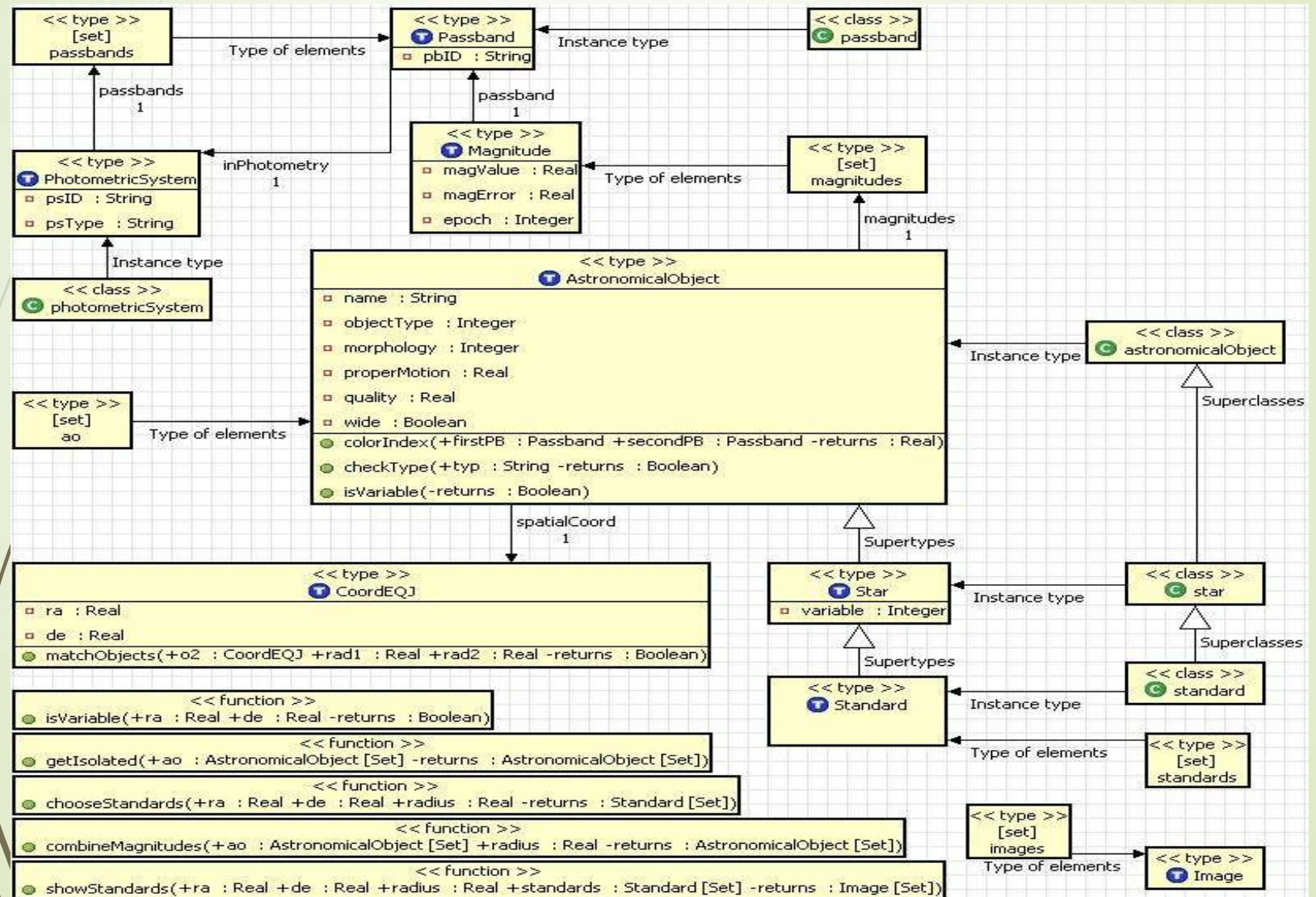


Сервисы

Задача определения вторичных стандартов

- Задача заключается в том, что по координатам площадки, требуется найти в ней звезды, удовлетворяющие ряду условий (не переменные, точечные, с хорошими изученными параметрами). Такие звезды называются «стандартами» и могут быть использованы для калибровки новых поступающих данных.
- Задача поставлена Институтом Космических Исследований РАН

Описание схемы посредника для задачи определения вторичных стандартов



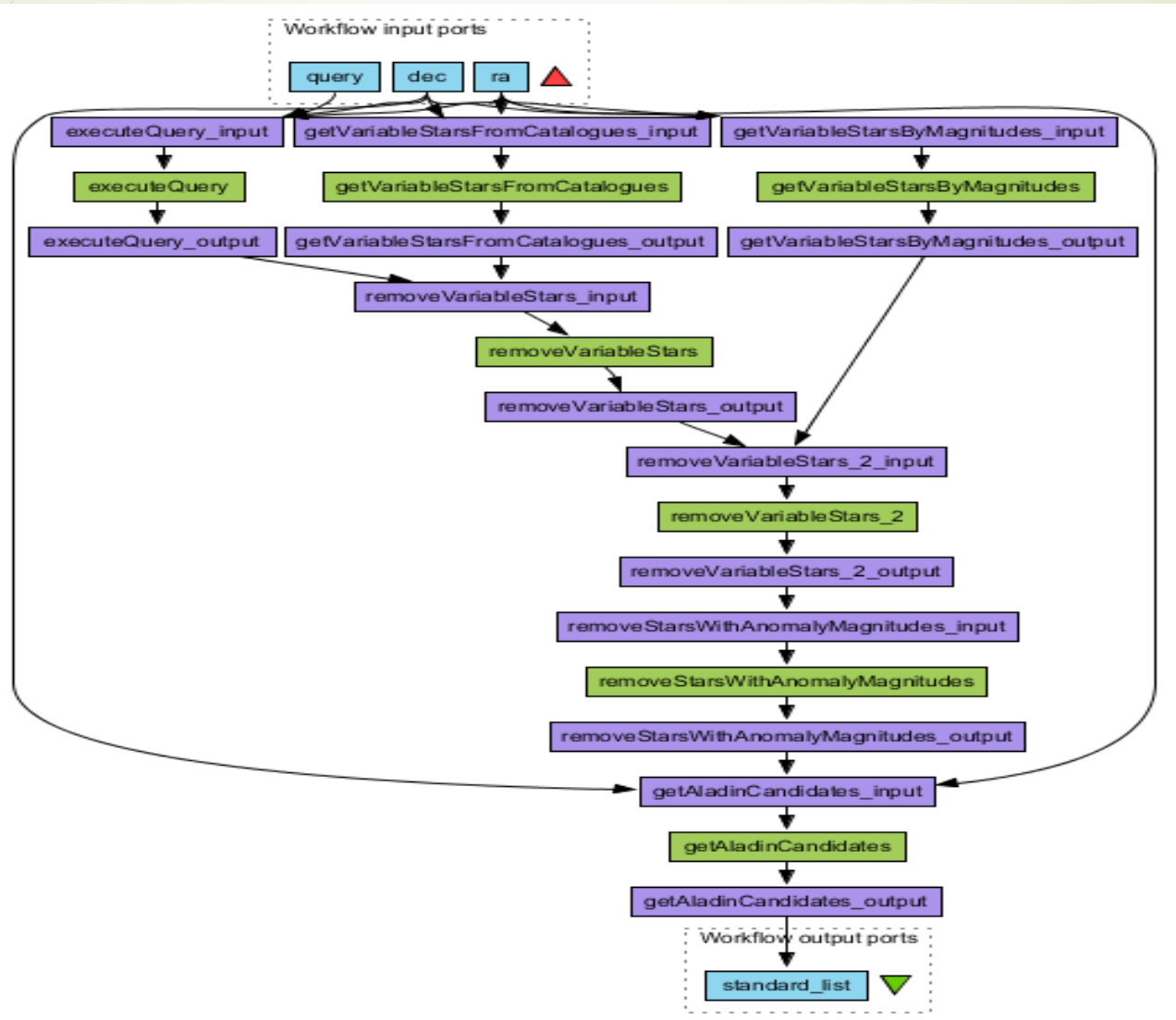
Программа решения задачи определения вторичных стандартов

- среди всех астрономических объектов выбираются те, что попадают в указанную площадку
- отсеиваются неизолированные объекты.
- среди ранее выбранных объектов отсеиваются галактики, и выбираются звезды с очень малым собственным движением и качественными фотометрическими данными
- выбираются только переменные объекты
- выбираются переменные звезды из каталогов переменных звезд
- кросс-идентификация объектов из класса кандидатов в стандарты, и класса переменных звезд
- выбираются только те объекты, для которых не нашлось близко расположенного переменного объекта
- выбранные кандидаты в стандарты маркируются на изображение площадки гамма-всплеска

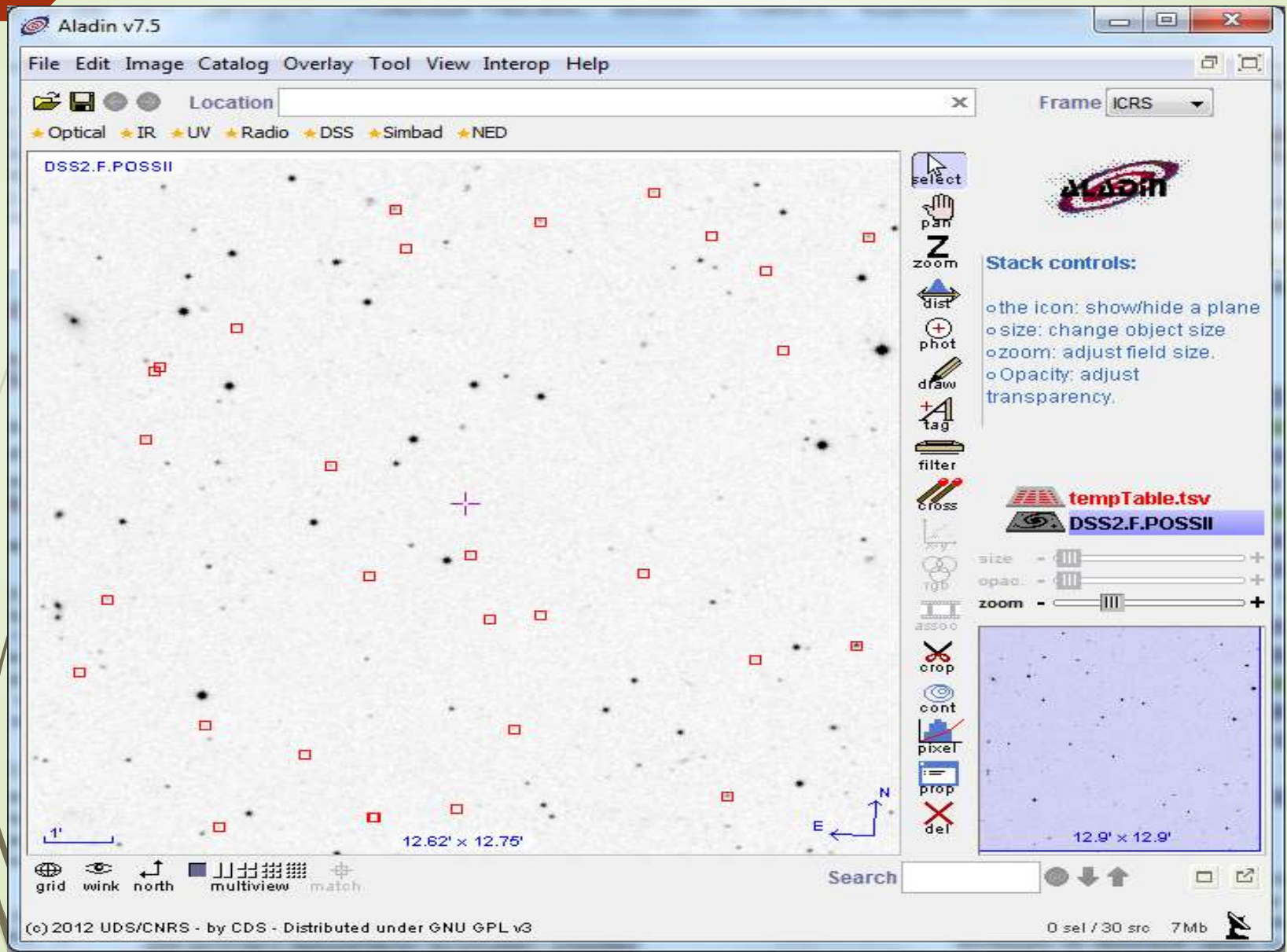
Описание Веб сервиса для доступа к предметному посреднику

- Для доступа к предметному посреднику решения задачи определения стандартов был разработан Веб сервис, включающий в себя следующие методы, реализующие описанные выше правила:
 - `executeQuery` – выполняет правило на языке формул посредника
 - `getVariableStarsFromCatalogues` - получает из посредника коллекцию переменных звезд в заданной области из каталогов переменных звезд
 - `getVariableStarsByMagnitudes` - получает из посредника коллекцию переменных звезд в заданной области, определяя переменная ли она по магнитудам
 - `removeVariableStars` - получает коллекцию стандартов, и коллекцию переменных. Из первой удаляются те объекты, которые содержатся во второй.
 - `getAladinCandidates` – по полученной коллекции объектов возвращает изображение, которое может быть открыто в программе Aladin

Поток работ решения задачи определения вторичных стандартов в среде Taverna



Результатом выполнения потока работ



Заключение

- Предлагаемый подход по встраиванию предметных посредников в среду организации исследований в НИИД позволяет упростить решение ряда проблем таких, как:
 - накопление методов анализа данных, алгоритмов решения задач и их реализаций в научном сообществе;
 - воспроизведение и повторное использование таких алгоритмов и методов;
 - формирование ИТ-базированных концептуальных определений научных областей;
 - использование методов и средств высокоуровневых декларативных определений методов анализа данных и алгоритмов решения задач в НИИД.
- Хотя статья рассматривает предлагаемый подход применительно к конкретной среде myExperiment и системе управления потоками работ Taverna, предлагаемый подход может быть аналогично использован в других средах с другими системами управления потоками работ.