



# Концепция и архитектура тематического интеллектуального научного интернет-ресурса

**Ю.А. Загорулько, Г. Б. Загорулько, В. К. Шестаков,  
И. С. Кононенко**

*Институт систем информатики имени А.П. Ершова СО РАН  
Новосибирск, Россия*



К настоящему времени накоплено громадное количество информации по различным областям знаний, при этом значительная ее часть представлена непосредственно в сети Интернет, но **проблема эффективного обеспечения научного сообщества информацией по интересующим его тематикам** пока не имеет удовлетворительного решения.

Это объясняется тем, что современные информационные системы используют довольно ограниченный набор методов представления, поиска и интерпретации информации.

Нерешенной остается и **проблема удобного доступа к средствам обработки информации**. Даже уже реализованные и представленные в Интернет методы обработки информации остаются недоступными широкому кругу пользователей из-за отсутствия содержательной информации о них.



Для информационной поддержки научной и производственной деятельности в определенной области знаний предлагается строить **тематические интеллектуальные научные интернет-ресурсы (ИНИР)**.

**Тематический ИНИР** представляет собой доступную через Интернет информационную систему, обеспечивающую

- (1) систематизацию и интеграцию** научных знаний и информационных ресурсов определенной области знаний,
- (2) содержательный эффективный доступ** к ним (поиск и навигацию) и средствам их интеллектуальной обработки,
- (3) поддерживающую их использование** при решении различных научных и производственных задач **посредством предоставления соответствующих интерфейсов и сервисов.**



ИНИР базируется на формализмах **онтологий** и **семантических сетей**.

**Онтология** наряду с описанием моделируемой области знаний содержит соотнесенное с ним описание структуры и типологии интегрируемых информационных ресурсов и методов интеллектуальной обработки содержащихся в них данных и знаний.

**Семантическая сеть** играет роль интеллектуального хранилища данных, в котором накапливается информация о релевантных научных информационных ресурсах и web-сервисах, реализующих методы обработки содержащейся в них информации.

На основе онтологии и семантической сети организуются **навигация** по информационному пространству ИНИР и **содержательный поиск**.



Формально система знаний ИНИР описывается четверкой:

$$KS = \langle O, Th, SN, IRs \rangle, \text{ где}$$

*O* – онтология ИНИР,

*Th* – тезаурус области знаний ИНИР,

*SN* – семантическая сеть, служащая для представления информационного содержания (контента) ИНИР,

*IRs* – информационные ресурсы и средства их интеллектуальной обработки, интегрируемые в ИНИР,





**Онтология области знаний ИНИР** строится на основе двух базовых онтологий – *онтологии научной деятельности* и *онтологии научного знания*.

**Онтология научной деятельности** включает классы понятий, относящиеся к организации научной и исследовательской деятельности, такие как *Персона, Организация, Событие, Деятельность, Проект, Публикация* и др.

**Онтология научного знания** фиксирует основные содержательные структуры, используемые для построения онтологий конкретных областей знаний: *Раздел науки, Метод исследования, Объект исследования, Научный результат* и др.



Основным классом **онтологии научных интернет-ресурсов** является класс *Информационный ресурс*, который служит для описания, представленных в сети Интернет информационных ресурсов.

Набор атрибутов и связей класса ***Информационный ресурс*** основан на стандарте Dublin core [8]. Его атрибутами являются: *название ресурса, язык ресурса, тематика ресурса, тип доступа к ресурсу* и т.п.

Объекты этого класса могут быть связаны семантическими отношениями с другими информационными объектами, представляющими в контенте ИНИР *организации, персоны, публикации, события, разделы науки* и т.д.



**Онтология задач и методов** включает:

- **описания задач**, на решение которых нацелен ИНИР, и **методов их решения**;

- **описания web-сервисов**, реализующих методы обработки информации, содержащейся в интегрируемых в ИНИР информационных ресурсах.

Описания интегрируемых web-сервисов базируются на онтологии OWL-S, предназначенной для описания семантических web-сервисов (Semantic Web Services) и представленной на языке OWL.

Обычный web-сервис представляет собой программную систему, описание функциональности которой представлено в сети Интернет в виде интерфейса, заданного в пригодном для машинной обработки формате (WSDL).



Семантический web-сервис предоставляет потенциальным пользователям не только описание своего интерфейса в терминах типов входных и выходных данных, но и описание своей семантики, т.е. того, что сервис делает, его предметной области, ограничений на область применения и качество сервиса и т.п. Причем все его свойства, «способности» (функциональность) и интерфейсы кодируются в однозначной поддающейся машиной обработке форме,

Наличие семантического описания у таких сервисов обеспечивает не только реализацию их поиска и корректного использования (исполнения), но и возможность композиции из них новых сервисов с целью получения функциональности, требуемой для решения пользовательских задач.



**Наличие у семантических web-сервисов содержательных описаний создает предпосылки и для их успешной интеграции в ИНИР.**

При этом будет обеспечиваться **содержательный доступ** к ним не только для программных агентов, но и **для человека**, желающего найти необходимые для решения его задач средства интеллектуальной обработки информации.

Возможна также интеграция в ИНИР и обычных (не семантических) web-сервисов. Для этого такие сервисы предварительно снабжаются семантическими описаниями, составленными с использованием понятий онтологии OWL-S и онтологии области знаний ИНИР.



**Тезаурус** служит для описания смысла понятий (терминов), используемых в моделируемой области знаний.

При этом тезаурус позволяет задавать смысл понятий не только с помощью определений, но и посредством соотнесения их с другими понятиями, используя для этого не только семантические отношения, задающие иерархии понятий, но и отношения синонимии и ассоциации. Благодаря этому тезаурус может применяться как при обработке пользовательских запросов, так и при поиске и аннотировании информационных ресурсов, интегрируемых в ИНИР.

Тезаурус ИНИР строится на основе ядра тезауруса, изначально включенного в систему знаний ИНИР. Ядро тезауруса содержит описание понятий базовых онтологий, включая описание терминов, с помощью которых они представляются в интернет-ресурсах.



Таким образом, система знаний ИНИР не только включает **формальное описание области знаний ИНИР**, задает **типологию релевантных ей информационных ресурсов** и решаемых в ней **задач и методов их решения**, описывает **смысл** используемых в этой области знаний **понятий**, но и обеспечивает эффективное **представление информации об интегрируемых в ИНИР информационных ресурсах и средствах их интеллектуальной обработки.**



# Архитектура ИНИР





## *Уровень доступа к информации*

<b>Представление запросов и результатов</b>	<b>Обеспечение навигации</b>
<b>Пользовательский интерфейс</b>	

**Первый уровень** обеспечивается **пользовательским интерфейсом** – программным компонентом, предоставляющим конечному пользователю содержательный доступ к контенту ИНИР при решении его задач.



Главными функциями пользовательского интерфейса являются **представление пользовательских запросов и результатов поиска и решений задач**, а также **обеспечение навигации** в информационном пространстве ИНИР.

Благодаря использованию онтологии и тезауруса пользовательский интерфейс позволяет **формулировать запросы в терминах моделируемой области знаний** и поддерживает **управляемую онтологией навигацию** по информационному пространству ИНИР.



## *Уровень обработки информации*

### Обеспечение информационных потоков

Модуль поиска в информационном пространстве ИНИР

Средства разработки и настройки базы знаний

Подсистема сбора онтологической информации об интернет-ресурсах

Редактор контента ИНИР

**Уровень обработки информации** обеспечивает все информационные потоки в ИНИР – от конструирования онтологии до обработки пользовательских запросов. Он включает **модуль поиска информации** в информационном пространстве ИНИР, **средства разработки/настройки базы знаний ИНИР** и **управления его контентом**, а также **подсистему сбора онтологической информации** (семантических метаданных) **об интернет-ресурсах**.



**Модуль поиска информации** в информационном пространстве ИНИР обеспечивает выполнение запросов на поиск информации, поступающих со стороны пользовательского интерфейса.

Настройка базы знаний выполняется с помощью средств **коллективной разработки и настройки базы знаний**, в состав которых входят редакторы онтологий и тезаурусов. Для управления контентом служит **редактор контента**, который позволяет создавать, редактировать и удалять информационные объекты и связи между ними.

Все редакторы реализованы как web-приложения, поэтому обеспечивают удаленную настройку ИНИР и поддержку его контента экспертами через Интернет.



**Подсистема сбора онтологической информации об интернет-ресурсах** выполняет поиск в Интернет информационных ресурсов, релевантных области знаний ИНИР, их семантический анализ и семантическое аннотирование на основе онтологии и тезауруса.

Кроме того, эта подсистема формирует по полученным аннотациям представление информации об интернет-ресурсах в контенте ИНИР в виде информационных объектов класса *Информационный ресурс* и их связей с другими информационными объектами. Благодаря этому объекты, соответствующие проаннотированным информационным ресурсам, могут использоваться при поиске информации и навигации по контенту ИНИР.



## *Базовый уровень*

### **Управление знаниями и данными**

**Система управления базами данных**

**Семантические технологии и web-сервисы**

Базовый уровень обеспечивает управление всеми данными и знаниями информационного пространства ИНИР. Он выполняет функции хранения и манипулирования данными (контентом ИНИР) и знаниями (онтологией и тезаурусом) с использованием средств стандартных СУБД (MySQL), технологий Semantic Web (OWL, RDF) и семантических web-сервисов (WSDL, OWL-S).

В частности, в качестве хранилища данных используется система Virtuoso (Open-Source Edition), обеспечивающая эффективную работу с большими объемами данных в RDF-формате.



## Заключение



- Представлен подход к построению тематических ИНИР, обеспечивающих систематизацию и интеграцию информационных ресурсов определенной области знаний и средств интеллектуальной обработки содержащейся в них информации, а также содержательный эффективный доступ к ним и их использование при решении различных задач.
- Основу ИНИР составляет онтология. Семантическая сеть, структура которой определяется онтологией, играет роль интеллектуального хранилища данных, в котором накапливается информация о релевантных научных информационных ресурсах и web-сервисах, реализующих интеллектуальную обработку содержащейся в них информации.
- Важное преимущество подхода: ИНИР позволяет упростить и ускорить доступ к затребованной пользователем информации и сократить время ее анализа благодаря аккумуляции описаний релевантных интернет-ресурсов и методов их обработки непосредственно в семантической сети ИНИР.



## Заключение



- Данный подход развивает технологию разработки порталов научных знаний, успешно применявшуюся при создании порталов знаний для ряда научных дисциплин.
- В отличие от порталов научных знаний ИНИР обеспечивает доступ не только к релевантным информационным ресурсам, но и средствам их интеллектуальной обработки (анализа), представленным в виде web-сервисов. Таким образом, ИНИР объединяет преимущества и порталов знаний, и мэшапов. От первых он наследует представление знаний и данных в виде онтологии и семантической сети, от вторых – интеграцию различных web-сервисов в одном приложении.
- Планируется создание ИНИР для осуществления информационной поддержки комплексных исследований в энергетике. При этом будут использованы накопленные в ИСЭМ СО РАН и представленные в электронном виде информационные ресурсы по энергетической тематике.



## Благодарности



*Работа выполнена при финансовой поддержке РФФИ (проект № 13-07-00422) и Президиума РАН (интеграционный проект СО РАН № 15/10 «Математические и методологические аспекты интеллектуальных информационных систем»)*