Управление знаниями и формирование связанных данных в Статкомитете СНГ*

Юрий Михайлович Акаткин[а)],
Елена Донатовна Ясиновская[а)],
Андрей Владимирович Шилин[б)],
Михаил Геннадиевич Бич[б)]
[а)] Российский экономический университет имени Г.В. Плеханова, г. Москва, Россия;
[б)] ООО «Электронное проектирование», г. Москва, Россия

В данной статье представлены решения, реализованные Статкомитетом СНГ в информационных системах управления знаниями, подготовки связанных данных и «умных» (богатых смыслом, семантикой) метаданных в составе строящегося датахаба СНГ. На основе анализа международного опыта и проведения собственных многолетних исследований была сформулирована цель работы – повышение эффективности и потенциала использования статистических данных за счет обеспечения однозначной содержательной интерпретации данных, в том числе в информационных системах потребителей. Для достижения этой цели авторы предложили новые подходы и технологии к построению системы управления знаниями на базе семантической сети, позволившие связать семантические модели, интерпретируемые машинами, с человекочитаемыми представлениями знаний. Решение задачи систематизации знаний о статистической методологии играет ключевую роль для повышения потенциала использования связанных данных и обеспечения совместной обработки статистических данных. Предложенный методический и технологический подход направлен на формирование контекста предметной области, который используется для разработки связанных данных и генерации «умных» метаданных, а также обеспечивает новые возможности для работы потребителей со статистическими данными и метаданными – их интерпретации, содержательного анализа, сопоставления и совместной обработки. Наряду с описанием рабочего цикла систем в статье проведен содержательный анализ проблем гармонизации статистической терминологии, выявленных в результате практической работы с доменом «Статистика труда». Отдельное внимание уделено роли экспертного сообщества в развитии системы управления знаниями.

Ключевые слова: система управления знаниями, экспертное сообщество, глоссарии, гармонизация, онтологии, семантические активы, семантическое обогащение данных, связанные открытые статистические данные, «умные» метаданные, семантическая интероперабельность.
JEL: C82, C87, J21, M15.
doi: https://doi.org/10.34023/2313-6383-2024-31-3-80-90.

Для цитирования: Акаткин Ю.М., Ясиновская Е.Д., Шилин А.В., Бич М.Г. Управление знаниями и формирование связанных данных в Статкомитете СНГ. Вопросы статистики. 2024;31(3):80–90.

Knowledge Management and Linked Data Generation in the CIS Statistics Committee*

Yuri M. Akatkin[a)],
Elena D. Yasinovskaya[a)],
Andrew V. Shilin[b)],
Mikhail G. Bich[b)]
[a)] Plekhanov Russian University of Economics, Moscow, Russia;
[b)] Electronic Design LLC, Moscow, Russia

This article presents the actions implemented by the Interstate Statistical Committee of the CIS (CIS-STAT) in knowledge management information systems, preparation of linked data and «smart» (semantically rich) metadata as part of the CIS data hub that is under construction. Based on the analysis of international experience and after conducting their own long-term research, the authors set out the purpose behind the work – to increase the efficiency and potential of using statistical data by ensuring an unambiguous and meaningful data interpretation, including in consumer information systems. To reach this goal, the authors proposed new approaches and technologies for building a knowledge management system based on the semantic network, which made it possible to link machine-interpretable semantic

* Статья подготовлена в развитие доклада на Международном форуме производителей и пользователей статистики (г. Санкт-Петербург, 12–14 сентября 2023 г.), представленного Статкомитетом СНГ.

The article was prepared as a follow-up to the report presented by the CIS-Stat at the International Forum of Statistics Producers and Users (St. Petersburg, September 12–14, 2023).

models with human-readable knowledge representations. Addressing the objective of organizing knowledge about statistical methodology is a key to increasing the potential for using linked data and enabling collaborative processing of statistical data. The proposed methodological and technological approach is aimed at contextualizing a subject area used to develop linked data and generate «smart» metadata. It also provides new opportunities for consumers to work with statistical data and metadata – their interpretation, meaningful analysis, comparison and joint processing. Along with a description of the systems operating cycle, the article provides a meaningful analysis of the issues of harmonizing statistical terminology, identified by practical work with the «Labor Statistics» domain. Special attention is paid to the role of the expert community in developing a knowledge management system.

Keywords: knowledge management system, expert community, glossaries, harmonization, ontologies, semantic assets, semantic data enrichment, linked open statistical data, smart metadata, semantic interoperability.

JEL: C82, C87, J21, M15.

doi: https://doi.org/10.34023/2313-6383-2024-31-3-80-90.

For citation: Akatkin Yu.M., Yasinovskaya E.D., Shilin A.V., Bich M.G. Knowledge Management and Linked Data Generation in the CIS Statistics Committee. *Voprosy Statistiki*. 2024;31(3):80–90. (In Russ.)

«Умные» метаданные и связанные открытые статистические данные – будущее мировой статистики

В современном мире цифровых экосистем и искусственного интеллекта хорошо организованные и понятные данные приобрели особое значение. Потребителям статистики важны качество и полнота метаданных, которые описывают данные. Работы по формированию метаданных ведутся давно, и поддерживающие их стандарты активно внедряются на практике как за рубежом, так и в странах СНГ. В последнее время акценты в подготовке метаданных смещаются со структуры данных на отражение их смысла. Для этого на основе систематизации знаний о статистической методологии и практики с помощью современных технологий формируется контекст предметной области, разрабатываются «умные»[1] метаданные (то есть метаданные, обогащенные смыслом – семантикой), а также связанные данные. Такой подход обеспечивает возможность понимания данных и метаданных людьми и их интерпретации информационными системами (ИС) [1].

Требования к описанию метаданных закреплены в целом ряде международных и национальных стандартов, активно развивающихся с начала 2000-х годов. Например, ГОСТ Р ИСО/МЭК 11179 «Информационная технология. Регистры метаданных»[2] является национальным стандартом Российской Федерации по управлению регистрами метаданных. В мировой статистике большим стимулом для развития метаданных стало широкое внедрение стандарта Statistical Data and Metadata eXchange (SDMX) на международном уровне [2]. На сегодняшний день стандарт SDMX позволил создать полноценную информационную среду, предоставляющую обучающие и методические материалы, инструментарий разработки, а также пространство для взаимодействия специалистов в области стандартизации статистических данных.

Большая работа в области метаданных и управления знаниями в международной статистике проделана Евростатом, реализовавшим свою базу данных[3] с применением SDMX. Опубликованные в ней наборы статистических данных сопровождаются как краткими, так и расширенными метаданными. Однако проведенный анализ практики применения SDMX показал некоторые ограничения стандарта, существенно влияющие на возможность машинной интерпретации распространяемых статистических данных:

1. Поскольку SDMX имеет объектную информационную модель, поддержка всего многообразия связей между сущностями метаданных является весьма сложной.

2. Краткие метаданные (Data Structure Definition, DSD) описывают только структуру и кодировку наборов данных. При публикации данных в форматах, отличных от XML, утрачивается связь между DSD и данными. Эта связь, отображенная в визуальных интерфейсах, не может интерпретироваться ИС.

[1] Smart Metadata Manifesto, Cosmos 2024: Conference on Smart Metadata for Official Statistics, 2024 edition. URL: http://cosmos-conference.org/2024/.

[2] URL: https://docs.cntd.ru/document/1200087954?ysclid=ls4uo5jjl5959347510.

[3] Eurostat. Database. URL: https://ec.europa.eu/eurostat/data/database.

3. В то же время расширенные метаданные (в соответствии с MetaData Structure Definition) – это результат большой методической работы. Для каждого вида статистических работ формируется объемный документ из 19 разделов. Однако расширенные метаданные распространяются отдельно от данных и не предназначены для машинной обработки. При этом большинство разделов – текстовые и практически не содержат ссылок на связанные понятия и показатели.

Как результат, несмотря на фундаментальный подход, реализуемый Евростатом, наблюдаются разрывы в метаданных: в разных информационных ресурсах они не всегда соответствуют друг другу. Так, зачастую термины и определения глоссария статистической энциклопедии Евростата[4] могут отличаться от тех терминов и определений, которые используются при подготовке расширенных метаданных. Пример таких расхождений представлен в таблице 1.

Таблица 1

Сравнение определений термина «Безработный» в различных ресурсах Евростата

Расширенные метаданные, описывающие показатели Обследования рабочей силы[5]	Статистическая энциклопедия Евростата «Statistics Explained»[6]
Безработный (Unemployed)	
Безработными считаются все лица в возрасте от 15 до 74 лет (от 16 до 74 лет в Эстонии, Италии и Великобритании), которые не были трудоустроены в течение учетной недели, активно искали работу в течение последних четырех недель и были готовы приступить к работе немедленно или в течение двух недель	Человек в возрасте от 15 до 74 лет; не трудоустроен в течение учетной недели в соответствии с определением занятости; в настоящее время доступен для работы, то есть доступен для оплачиваемой работы или самостоятельной занятости до истечения двух недель, следующих за контрольной неделей; активно ищет работу, то есть либо осуществлял деятельность в течение четырехнедельного периода, заканчивающегося учетной неделей, в поисках оплачиваемой работы или самостоятельной занятости, либо нашел работу, которую можно начать в течение периода не более трех месяцев с конца учетной недели

Осознание этих проблем привело статистические международные организации к развитию связанных открытых статистических данных (СОСД) [3] и «умных» метаданных, в конструкции которых данные и метаданные неразрывно связаны. Решения, реализуемые в этом направлении, базируются на применении стандарта RDF Data Cube Vocabulary [4], сочетающего преимущества SDMX и связанных данных.

Связанные данные, разработанные в соответствии с принципами Semantic Web [5], содержат необходимые семантические связи с описывающими их понятиями, словарями и онтологиями [6]. А технологии Semantic Web обеспечивают среду, в которой приложения могут запрашивать данные и управлять ими, формировать интерфейсы и делать логические выводы с учетом семантических связей. Поэтому Консорциум Всемирной паутины (World Wide Web Consortium, или W3C) рекомендует связанные данные в качестве наиболее эффективного способа открытия данных в Интернете. Однако, как показывают исследования в области создания связанных данных, успех их использования решающим образом зависит от того, насколько полно семантические модели отражают контекст предметной области. При использовании минимального набора моделей и инструментов (например, связка schema.org и JSON-LD) получаются связанные данные, удовлетворяющие заявленным стандартам. Но их бедная семантика, по сути, не дает возможности реализовать весь потенциал Semantic Web.

Зачастую понимание и использование данных потребителями, не являющимися экспертами в данной области, а также их интерпретация информационными системами затруднены из-за отсутствия формализованных знаний о предметной области и машиночитаемых данных, дополненных семантикой для понимания смысла набора данных [1]. Бедность семантики приводит к тому, что связанные данные трудно обнаружить в Интернете по их описаниям и также однозначно связать с понятиями предметной области. Поэтому многие из уже опубликованных наборов не готовы для объединения и совместной обработки [7].

Построение богатой семантики реализуется за счет «множества свойств, описывающих ресурс, включая библиографическую информацию, аннотации, относящиеся к предметной области, ссылки на взаимосвязанные источники данных,

[4] Eurostat. Statistics Explained. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Main_Page.
[5] URL: https://ec.europa.eu/eurostat/cache/metadata/en/lfsi_esms.htm.
[6] URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Unemployment.

а также повторное использование существующей терминологии и поддержка языковых версий» [1] и позволяет использовать метаданные различными способами как людям, так и машинам [8].

Сделанные заключения созвучны с новым направлением (треком) развития, продвигаемым Группой высокого уровня ЕЭК ООН по модернизации официальной статистики[7]. Речь идет о богатых, или «умных» метаданных[8], которые не только стандартны (то есть их можно понять и использовать везде) и активны (позволяют реализовать управление статистическими процессами), но и делают данные находимыми (F), доступными (A), интероперабельными (I) и пригодными для повторного использования (R) [9]. Ключевым способом реализации принципов FAIR считается применение семантических технологий не только для распространения данных, но и для формализации знаний в виде семантических моделей – стандартных словарей, глоссариев, тезаурусов, онтологий [10].

С учетом накопленного международного опыта и многолетних исследований авторами реализован новый способ построения систем управления знаний на основе семантической сети, которая обеспечивает связь знаний, представленных в виде семантических моделей, интерпретируемых машинами, с визуальными представлениями, понятными людям. Для повышения потенциала использования связанных данных и обеспечения совместной обработки статистических данных с применением этого способа необходимо:

– сформировать контекст предметной области статистики;

– предоставить целостную (неразрывную) среду интерпретации данных для потребителей (людей и информационных систем);

– формировать умные метаданные и связанные статистические данные, обогащенные смыслом (семантикой).

Данный подход был реализован при разработке прототипов Системы управления знаниями в области статистической методологии (СУЗ) и Системы подготовки и распространения связанных открытых статистических данных (СПР СОСД) Статкомитета СНГ, представленных в этой статье.

На начальном этапе проекта разработана Концепция подготовки и распространения связанных статистических данных Статкомитета СНГ, которая базируется на многолетних исследованиях авторов [11–13] и учитывает существующий международный опыт в области применения семантических методов для обеспечения качества, сопоставимости и эффективной совместной обработки статистических данных. Концепция предусматривает создание СУЗ и СПР СОСД как составной части строящегося датахаба Статкомитета СНГ (см. рис. 1).

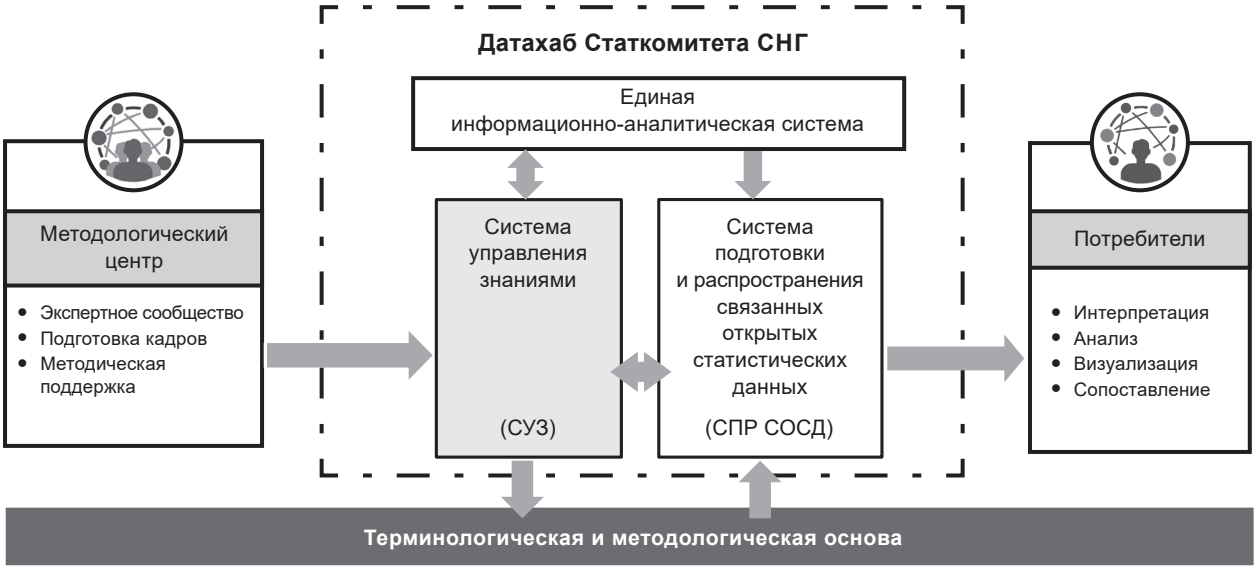

Рис. 1. **Прототипы СУЗ и СПР СОСД в датахабе Статкомитета СНГ**

[7] UNECE High-level Group for the Modernisation of Official Statistics, HLG-MOS Project proposal for 2023, 2023. URL: https://unece.org/sites/default/files/2022-12/HLG-MOS%20ProjectProposal%202023_LinkingClassifications.pdf.

[8] Smart Metadata Manifesto, Cosmos 2024: Conference on Smart Metadata for Official Statistics. 2024 edition. URL: http://cosmos-conference.org/2024/.

Важнейшей задачей экспертного сообщества, решаемой в датахабе Статкомитета СНГ, является формирование в СУЗ терминологической и методологической основы, составляющей контекст предметной области статистики. Этот контекст используется для разработки связанных данных и генерации «умных» метаданных, что обеспечивает новые возможности для работы потребителей со статистическими данными и метаданными – их интерпретации, содержательного анализа, сопоставления и совместной обработки.

Знания в СУЗ организованы по статистическим доменам в разрезе стран. На первом этапе работа была построена на базе домена «Статистика труда». Краткое описание основных разделов СУЗ представлено в таблице 2.

Таблица 2

Основные разделы СУЗ

Раздел	Краткое описание
Библиотека	Формируется для каждого статистического домена. Разработана классификация по типам материалов. Кроме того, дополнительно используется сортировка по источникам (международные организации или страновые), а также по году выпуска документа
Методология	Наиболее важные материалы из библиотеки публикуются в структурированном виде в специализированном разделе «Методология», который формируется как по доменам, так и по странам. При структурировании материалов гипертекстовая разметка используется для связывания терминов и документов
Глоссарий	По результатам анализа и структурирования загруженных материалов формируется глоссарий домена. Глоссарий в виде семантической модели распространяется в машиночитаемом формате. Каждый термин глоссария описан по единому шаблону. Термины глоссария связаны между собой с помощью гипертекстовой разметки. В целях сопоставления международной терминологии в СУЗ поддержаны механизмы мультиязычности. Для терминов глоссария домена «Статистика труда» приведены также определения на английском языке из международных статистических стандартов
Гармонизация	СУЗ поддерживает процедуры гармонизации терминологии для устранения противоречий в определениях терминов из различных источников. В специализированном шаблоне консолидируются все варианты, выявленные на этапе анализа, формируется предложение по гармонизации, а затем фиксируется решение. Страница гармонизации такого термина сохраняется для отслеживания истории создания терминологической статьи и указывается, как источник термина
Модели	Спроектированные в СПР СОСД семантические модели справочников публикуются на страницах СУЗ. Для этого разработан специализированный шаблон, в котором наряду с описанием представлена структура справочника и его содержимое. При моделировании элементы справочника связываются со ссылками на их эквиваленты во внешних информационных ресурсах (например, Geonames, DBpedia и др.). Установление такого соответствия позволяет обеспечить сопоставимость с данными, описанными на основе этих ресурсов
Показатели и наборы СОСД	В разделе публикуются описания показателей и наборы СОСД, отражающие их значения. Формализованное описание показателя, на основе которого формируются «умные» метаданные для Единой информационно-аналитической системы Статкомитета СНГ (ЕИАС), имеет гипертекстовую разметку, выделяющую связанные термины и документы, а также содержит такие разделы как: 1) сведения о регламентирующих документах (ссылки на опубликованные в СУЗ материалы); 2) методика сбора данных (с гипертекстовой разметкой связанными терминами и документами); 3) система классификаций (ссылки на связанные семантические модели – справочники). Наборы СОСД распространяются в машиночитаемом формате и опубликованы в СУЗ в специализированном шаблоне, содержащем формализованное описание набора и визуальные представления (интерактивные таблицы). В описаниях используется гипертекстовая разметка для выделения связанных терминов, справочников, документов
Обзоры	Информационно-аналитические обзоры включают следующие блоки: 1) выявленные тенденции – блок текстов, сгенерированных на основе анализа набора СОСД; 2) экспертное мнение – блок для комментариев эксперта с гипертекстовой разметкой; 3) таблицы и графики – блок для публикации визуальных компонентов, отображающих СОСД в табличном виде или в виде графиков и диаграмм. Таблицы и графики нумеруются автоматически, если их несколько. Описания таблиц и графиков также могут быть размечены связанными терминами или другими структурными элементами

Для результативности применения «умных» метаданных и СОСД необходима открытая семантически богатая среда их интерпретации. Комплекс систем СУЗ и СПР СОСД формирует единую непротиворечивую терминологическую и методическую основу для разработки богатых семантических моделей, а затем обеспечивает возможность их использования для подготовки, распространения и интерпретации связанных данных и «умных» метаданных. Важным принципом, лежащим в основе предложенных методов и инструментов, является обеспечение совместной работы ИТ-специалистов и экспертов-статистиков.

Общая логика работы комплекса систем представлена на рис. 2. Основной рабочий цикл из семи нижеперечисленных блоков разделен между СУЗ (блоки работ 1, 2 и 7) и СПР СОСД (блоки 3–6):

1. Сбор и систематизация методологических документов (создание электронной библиотеки), структурирование, HTML-разметка (связанными терминами и документами) и размещение наиболее важных документов в специализированном разделе «Методология».

2. Разработка глоссариев (формирование терминологических статей) и описаний показателей

на основе анализа методологических документов, а затем генерация соответствующих семантических активов[9] (СА).

3. Каталогизация сгенерированных в СУЗ семантических активов в каталоге СА СПР СОСД.

4. Разработка и каталогизация в СПР СОСД необходимых СА, справочников и онтологий статистических доменов в соответствии с семантическими стандартами.

5. Загрузка наборов данных из хранилища ЕИАС. Трансформация наборов в стандарт RDF Data Cube, семантическое обогащение и каталогизация в СПР СОСД.

6. Визуализация и валидация каталогизированных семантических моделей и наборов СОСД.

7. Конструирование с учетом структуры OLAP-кубов ЕИАС «умных» метаданных, которые передаются для размещения на информационных панелях.

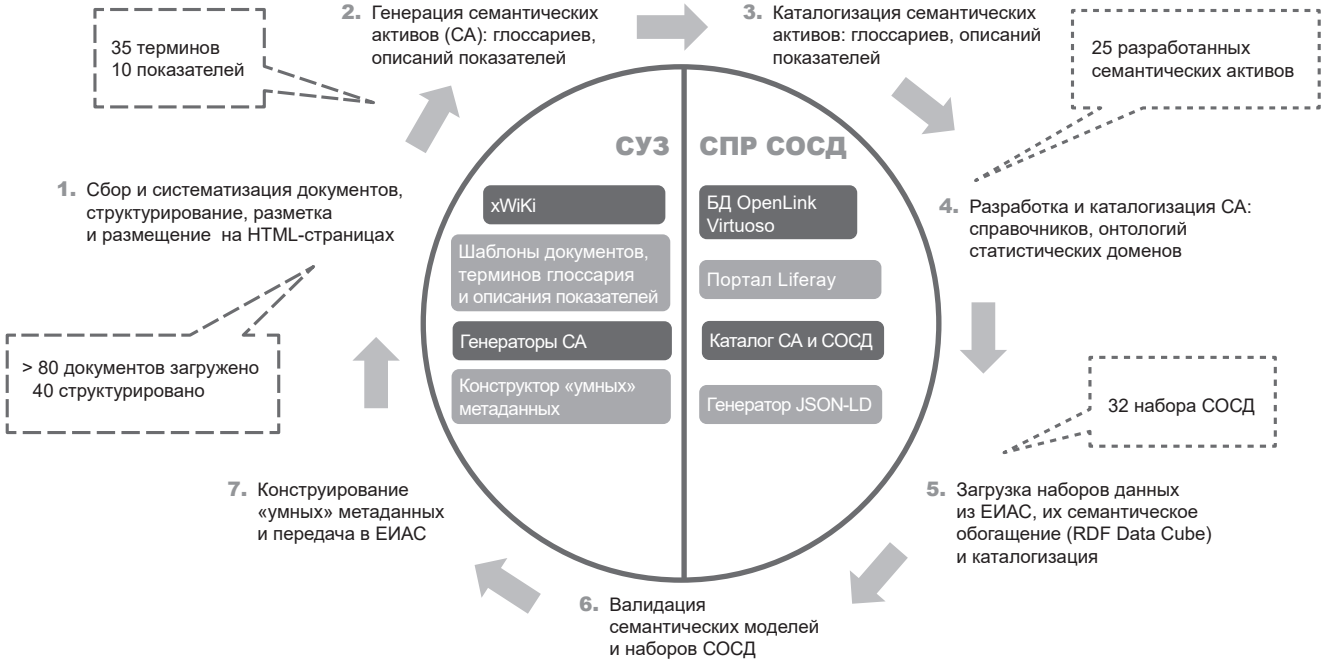

Рис. 2. **Рабочий цикл СУЗ и СПР СОСД**

Библиотека домена «Статистика труда» в СУЗ уже включает более 80 документов, из которых 40 структурированы и размещены на страницах xWiKi с использованием гипертекстовой разметки для связывания с терминами глоссария и упомянутыми документами. Каталог наборов СОСД содержит 32 семантически обогащенных набора СОСД домена «Статистика труда».

Для проведения экспертной оценки результатов семантического моделирования применяются инструменты визуализации семантических моделей и наборов СОСД. Визуальные интерфейсы отображают результат автоматического преобразования метаданных в человекочитаемое представление – интерактивные таблицы, деревья и графы. Как показано для иллюстрации на рис. 3, семантически обогащенный набор СОСД содержит персистентные (устойчивые) гиперссылки на элементы среды интерпретации, что позволяет валидировать их консистентность.

СУЗ Статкомитета СНГ построена на основе расширения xWiKi для использования семантических технологий. В ней поддерживаются шаблоны публикации документов, терминов глоссария и описания показателей, которые используются для наполнения СУЗ при участии экспертов предметной области и обеспечивают человекочитаемое представление знаний, зафиксированных в СА. Для автоматизации формирования семантических моделей (активов) и «умных» метаданных разработаны соответствующие генераторы и конструкторы.

[9] Под семантическими активами понимаются подготовленные для многократного использования описания данных: 1) метаданные, например, XML- и RDF-схемы; 2) общие модели данных; 3) онтологии; 4) тезаурусы; 5) справочные данные, например, списки кодов, таксономии, словари, глоссарии. СА публикуются как открытые стандарты данных.

Рис. 3. **Пример визуализации набора СОСД**

Система подготовки и распространения СОСД построена с использованием технологий OpenLink Virtuoso и Liferay. Основу системы составляют каталоги СА и СОСД, разработанные в соответствии со стандартами ADMS и DCAT. В качестве внешнего средства семантического моделирования используется Protégé. Разработанный генератор JSON-LD обеспечивает подготовку семантически богатых описаний наборов СОСД для индексации и поиска в соответствии с требованиями центра поиска Google.

Таким образом, в прототипах СУЗ и СОСД реализован полный цикл от загрузки материалов до подготовки «умных» метаданных и их публикации в ЕИАС, а также семантически богатых СОСД. Полученный опыт и апробированные подходы могут быть использованы не только в статистике, но и в других предметных областях.

Применение механизмов гармонизации для решения конфликтов терминологии и классификации

Проведенный в процессе наполнения СУЗ анализ методических документов и опубликованных межотраслевых глоссариев и словарей в домене «Статистика труда» выявил проблему несогласованности терминологии как на международном, так и на национальном уровнях стран СНГ. Эта проблема требует особого обсуждения. Для примера на рис. 4 продемонстрированы разные определения термина «Рабочая сила», зафиксированные в документах различного уровня (МОТ, СНГ, Россия).

Аналогичная проблема возникла при анализе англоязычных ресурсов (глоссариев международных организаций, статистических стандартов, академических словарей). 10 из 27 терминов в глоссарии СУЗ, относящихся к статистике рабочей силы, имеют разные определения в различных авторитетных источниках. Термины с наибольшей вариативностью источников представлены в таблице 3.

Проблема, по мнению авторов, усугубляется тем, что международные документы, например, резолюции МОТ, как правило, не включают специальный раздел по терминологии. С другой стороны, при формировании таких разделов как «Понятия и определения» в страновых документах зачастую используется вариант термина, предложенный составителями без опоры на единое терминологическое пространство. Поэтому терминологические (семантические) конфликты существуют не только на уровне документов разных стран, но и в различных документах одной страны (или СНГ в целом).

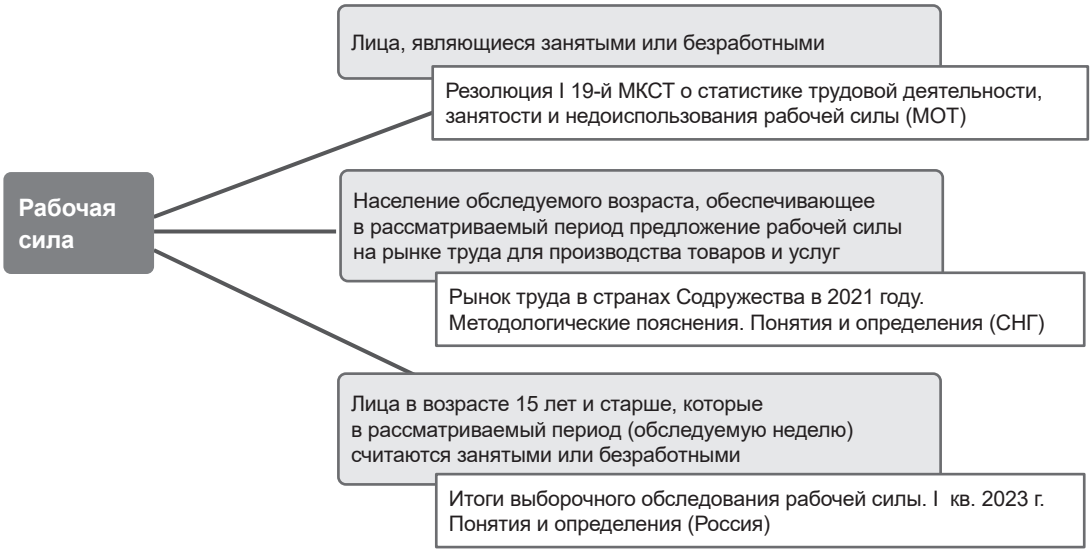

Рис. 4. **Варианты определений термина «Рабочая сила» на международном, наднациональном и национальном уровнях**

Таблица 3

Базовые термины домена «Статистика труда» и источники, в которых они определены по-разному

Термин	Источники определений
Безработные / Persons in unemployment / Unemployed	ILO, Eurostat Glossary, OECD, BLS US, UK DATASERVICE, Cambridge Dictionary, IATE, Eurostat Metadata
Волонтеры/Volunteer	ILO, EURO-LEX, Cambridge Dictionary
Занятость/Employment	ILO, BLS US, Cambridge Dictionary, IATE, DBPedia
Учетный период / Reference Period	ILO, EURO-LEX, IATE
Лица, не входящие в состав рабочей силы / Persons outside the labour force	ILO, Eurostat Glossary, OECD, BLS US, Britannica, IATE, DBPedia
Потенциальная рабочая сила / Potential labour force / Entrants	ILO, Eurostat Glossary, OECD, UK DATASERVICE, Cambridge Dictionary, IATE, DBPedia
Рабочее место / Job	ILO, OECD, BLS US, EUROLEX, Cambridge Dictionary
Трудовая деятельность / Work	ILO, OECD, Cambridge Dictionary, DBPedia

Для устранения конфликтов терминологии в СУЗ сформирован раздел «Гармонизация», в котором генерируется специальная страница (см. рис. 5) с отображением всех возможных вариантов термина, выявленных на этапе анализа материалов, фиксирующая:

– различия в наименовании (например, «занятые лица» и «занятое население – это лица»);

– различия в определениях (при их наличии);

– различия в пояснениях (при их наличии).

Для каждого варианта указывается ссылка на источник. Он может быть размещен как в библиотеке загруженных материалов, так и опубликован в разделе «Методология» в структурированном виде. Эксперты принимают решение по гармонизации термина, в котором по результатам обсуждения могут быть зафиксированы наименование, определение, пояснение, а также альтернативные названия. Для терминов глоссария, созданных в результате гармонизации, в качестве источника указывается страница гармонизации. Это позволяет отследить весь процесс появления новой версии термина.

Большое значение для гармонизации терминологии на пространстве СНГ имеет мультиязычность. Модель глоссария СУЗ поддерживает множество языковых версий, что позволяет реализовать работу с материалами, опубликованными на языках стран Содружества. Для подготовки языковой версии важно использовать только релевантные и актуальные материалы, в которых указаны точное название термина, его определение и, по возможности, пояснение.

Для английской версии приоритетными являются тексты международных методических материалов (например, резолюций МОТ) и международных глоссариев. При наполнении языковой версии преимущество имеет текст зарубежного источника, а перевод целесообразно приводить только для терминов, не имеющих международных аналогов (например, термин «баланс трудовых ресурсов»).

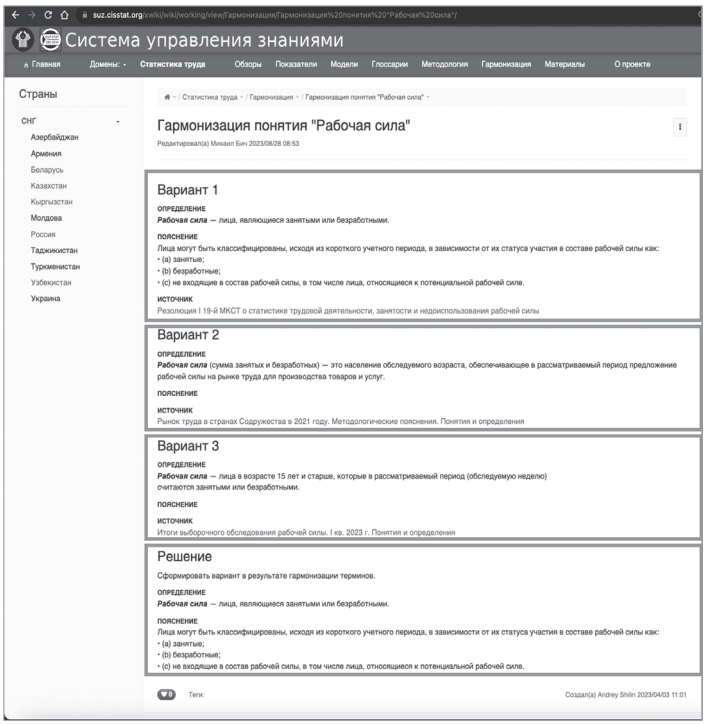

Рис. 5. **Страница гармонизации термина «Рабочая сила»**

В дальнейшем при наполнении СУЗ примененные для терминов механизмы будут также использоваться для гармонизации справочников и классификаций.

Заключение

Представленные в статье методы и технологии систематизации и структурирования информации, реализованные в СУЗ и СПР СОСД, позволили на примере домена «Статистика труда» одновременно решить несколько весьма актуальных задач:

– консолидировать знания предметной области путем формирования единой электронной библиотеки материалов в области статистической методологии, их структурирования, выделения связанных понятий и документов, а также проведения семантического моделирования;

– заложить основу единого терминологического пространства в виде глоссариев доменов, опубликованных на страницах СУЗ, и семантических моделей, распространяемых в машиночитаемом формате;

– сформировать целостную (неразрывную) среду интерпретации данных для потребителей в целях подготовки, распространения и применения «умных» (обогащенных семантикой) метаданных и связанных данных.

Предложен принципиально новый подход к построению систем управления знаниями. СУЗ Статкомитета СНГ, базирующаяся на семантической сети, обеспечивает связь семантических моделей, интерпретируемых внешними информационными системами, и человекочитаемых визуальных представлений, доступных широкому кругу потребителей. Впервые в СНГ на примере домена «Статистика труда» разработано более 30 наборов связанных открытых данных по 10 статистическим показателям.

Ценный практический опыт совместной работы ИТ-специалистов и экспертов-статистиков получен в процессе наполнения СУЗ. Формализовано представление контекста предметной области за счет применения специализированных шаблонов при формировании библиотеки материалов, публикации структурированных документов, подготовки терминов-кандидатов, а также поддержки процессов гармонизации. Знания экспертов о предметной области перенесены ИТ-специалистами в семантические модели (онтологии, тезаурусы и глоссарии), которые обеспечивают возможность их интерпретации использования информационными системами потребителей.

Статкомитет СНГ уделяет значительное внимание формированию экспертного сообщества. Для наполнения СУЗ разработана методология загрузки и классификации методических и ана-

литических материалов, а также подготовлены соответствующие инструменты. Первым шагом для создания такого сообщества является сформированная Статкомитетом СНГ экспертная группа, которая в 2024 г. ведет наполнение СУЗ по семи доменам статистики: система национальных счетов, цены, сельское хозяйство, промышленность, уровень жизни, заработная плата, туризм.

Большой потенциал имеет расширение экспертного сообщества для обсуждения ключевых вопросов в области методологии, терминологии, гармонизации за счет привлечения экспертов стран СНГ. Их совместная работа может быть организована в рамках международных семинаров, а полученный результат и принятые решения в виде модельных документов – фиксироваться в СУЗ и учитываться при подготовке СОСД и «умных» метаданных.

Перспективным направлением представляется применение отработанного подхода для повышения эффективности мониторинга при реализации государственных и наднациональных программ.

Литература

1. **Abgaz Y.** et al. Towards a Comprehensive Assessment of the Quality and Richness of the European a Metadata of Food-Related Images. 2020. 1st International Workshop on Artificial Intelligence for Historical Image Enrichment and Access (AI4HI-2020). Paris: ELRA. 2020. P. 29–33. doi: https://doi.org/10.13140/RG.2.2.29753.39521.

2. **Stahl R., Staab P.** History of SDMX. Measuring the Data Universe. Data Integration Using Statistical Data and Metadata Exchange. Springer Cham. 2018. P. 73–83. doi: https://doi.org/10.1007/978-3-319-76989-9_11.

3. **Kalampokis E., Zeginis D., Tarabanis K.** On Modeling Linked Open Statistical Data // Journal of Web Semantics. 2019. Vol. 55. P. 56–68. URL: https://www.sciencedirect.com/science/article/pii/S1570826818300544?via%3Dihub.

4. **Escobar P.** et al. Adding Value to Linked Open Data Using a Multidimensional Model Approach Based on the RDF Data Cube Vocabulary // Computer Standards & Interfaces. 2019. Vol. 68. doi: https://doi.org/10.1016/j.csi.2019.103378.

5. **Bizer C., Heath T., Berners-Lee T.** Linked Data: The Story So Far // Semantic Services, Interoperability and Web Applications: Emerging Concepts. IGI Global. 2009. P. 205–227. doi: https://doi.org/10.4018/978-1-60960-593-3.

6. **Feitosa D.** et al. A Systematic Review on the Use of Best Practices for Publishing Linked Data // Online Information Review. 2018. Vol. 19. No. 1. P. 107–123. doi: https://doi.org/10.1108/OIR-11-2016-0322.

7. **Akatkin Yu.** et al. The Challenges of Linked Open Data Semantic Enrichment, Discovery, and Dissemination // Physics of Particles and Nuclei. Pleiades Publishing. 2024. Vol. 55. P. 538–549. doi: https://doi.org/10.1134/S106377962403002X.

8. **Zaveri A.** et al. Quality Assessment for Linked Data: A Survey // Semantic Web. 2016. Vol. 7. No. 1. P. 63–93. doi: https://doi.org/10.3233/SW-150175.

9. **Wilkinson M.** et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship // Scientific Data. 2016. Vol. 3. Article 160018. URL: https://www.nature.com/articles/sdata201618.

10. **Amdouni E., Bouazzouni S., Jonquet C.** O'FAIRe Makes You an Offer: Metadata-Based Automatic FAIRness Assessment for Ontologies and Semantic Resources // International Journal of Metadata, Semantics and Ontologies. 2022. Vol. 16. Iss. 1. P. 16–46. doi: https://doi.org/10.1504/IJMSO.2022.131133.

11. **Akatkin Y., Yasinovskaya E.** Data-Driven Government in Russia: Linked Open Data Challenges, Opportunities, Solutions // Electronic Governance and Open Society: Challenges in Eurasia (EGOSE 2020). Communications in Computer and Information Science. Springer Cham. 2020. Vol. 1349. P. 245–257. URL: https://www.springerprofessional.de/en/data-driven-government-in-russia-linked-open-data-challenges-opp/18742244.

12. **Akatkin Yu., Laikam K., Yasinovskaya E.** The Concept and the Roadmap to Linked Open Statistical Data in the Russian Federation // Electronic Governance and Open Society: Challenges in Eurasia (EGOSE 2021) / Communications in Computer and Information Science. Springer Cham. 2022. Vol. 1529. P. 62–76. URL: https://link.springer.com/chapter/10.1007/978-3-031-04238-6_6.

13. **Акаткин Ю.М., Лайкам К.Э., Ясиновская Е.Д.** Связанные статистические данные: актуальность и перспективы. Вопросы статистики. 2020. Т. 27. № 2. С. 5–16. doi: https://doi.org/10.34023/2313-6383-2020-27-2-5-16.

Информация об авторах

Акаткин Юрий Михайлович – канд. экон. наук, заведующий научной лабораторией «Семантического анализа и интеграции», Российский экономический университет имени Г.В. Плеханова. 117997, г. Москва, Стремянный пер., д. 36. E-mail: u.akatkin@semanticpro.org. ORCID: https://orcid.org/0000-0001-6659-0961.

Ясиновская Елена Донатовна – старший научный сотрудник научной лаборатории «Семантического анализа и интеграции», Российский экономический университет имени Г.В. Плеханова. 117997, г. Москва, Стремянный пер., д. 36. E-mail: elena@semanticpro.org. ORCID: https://orcid.org/0000-0001-8226-3549.

Шилин Андрей Владимирович – генеральный директор ООО «Электронное проектирование». 107023, г. Москва, Барабанный пер, д. 4. E-mail: a.shilin@e-projecting.ru. ORCID: https://orcid.org/0000-0001-5827-3923.

Бич Михаил Геннадиевич – канд. техн. наук, технический директор ООО «Электронное проектирование». 107023, г. Москва, Барабанный пер., д. 4. E-mail: misha@e-projecting.ru. ORCID: https://orcid.org/0000-0001-9380-8364.

References

1. **Abgaz Y.** et al. Towards a Comprehensive Assessment of the Quality and Richness of the European a Metadata of Food-Related Images [PowerPoint slides]. In: AI4HI-2020 *Virtual Workshop – 1st International Workshop on Artificial Intelligence for Historical Image Enrichment and Access*. Paris: ELRA; 2020. P. 29–33. Available from: https://doi.org/10.13140/RG.2.2.29753.39521.

2. **Stahl R., Staab P.** *History of SDMX. Measuring the Data Universe. Data Integration Using Statistical Data and Metadata Exchange*. Springer Cham; 2018. P. 73–83. Available from: https://doi.org/10.1007/978-3-319-76989-9_11.

3. **Kalampokis E., Zeginis D., Tarabanis K.** On Modeling Linked Open Statistical Data. *Journal of Web Semantics*. 2019;(55):56–68. Available from: https://www.sciencedirect.com/science/article/pii/S1570826818300544?via%3Dihub.

4. **Escobar P.** et al. Adding Value to Linked Open Data Using a Multidimensional Model Approach Based on the RDF Data Cube Vocabulary. *Computer Standards & Interfaces*. 2019;68. Available from: https://doi.org/10.1016/j.csi.2019.103378.

5. **Bizer C., Heath T., Berners-Lee T.** Linked Data: The Story So Far. In: Sheth A. (ed.) *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. IGI Global;2009. P. 205–227. Available from: https://doi.org/10.4018/978-1-60960-593-3.

6. **Feitosa D.** et al. A Systematic Review on the Use of Best Practices for Publishing Linked Data. Online Information Review. 2018;19(1):107–123. Available from: https://doi.org/10.1108/OIR-11-2016-0322.

7. **Akatkin Yu.** et al. The Challenges of Linked Open Data Semantic Enrichment, Discovery, and Dissemination. *Physics of Particles and Nuclei*. Pleiades Publishing. 2024;55:538–549. Available from: https://doi.org/10.1134/S106377962403002X.

8. **Zaveri A.** et al. Quality Assessment for Linked Data: A Survey. *Semantic Web*. 2016;7(1):63–93. Available from: https://doi.org/10.3233/SW-150175.

9. **Wilkinson M.** et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*. 2016;3:Article 160018. Available from: https://www.nature.com/articles/sdata201618.

10. **Amdouni E., Bouazzouni S., Jonquet C.** O'FAIRe Makes You an Offer: Metadata-Based Automatic FAIRness Assessment for Ontologies and Semantic Resources. *International Journal of Metadata, Semantics and Ontologies*. 2022;16(1):16–46. Available from: https://doi.org/10.1504/IJMSO.2022.131133.

11. **Akatkin Y., Yasinovskaya E.** Data-Driven Government in Russia: Linked Open Data Challenges, Opportunities, Solutions. In: Chugunov A. et al. (eds.) *Communications in Computer and Information Science: Proc. of the 7th International Conference, Electronic Governance and Open Society: Challenges in Eurasia (EGOSE 2020), St. Petersburg, Russia, November 18–19, 2020*. Springer Cham; 2020. P. 245–257. Available from: https://www.springerprofessional.de/en/data-driven-government-in-russia-linked-open-data-challenges-opp/18742244.

12. **Akatkin Yu., Laikam K., Yasinovskaya E.** The Concept and the Roadmap to Linked Open Statistical Data in the Russian Federation. In: Chugunov A.V. et al. (eds.) *Communications in Computer and Information Science: Proc. of the 8th International Conference, Electronic Governance and Open Society: Challenges in Eurasia (EGOSE 2021), Saint Petersburg, Russia, November 24–25, 2021*. Springer Cham; 2022. P. 62–76. Available from: https://link.springer.com/chapter/10.1007/978-3-031-04238-6_6.

13. **Akatkin Yu.M., Laykam K.E., Yasinovskaya E.D.** Linked Open Statistical Data: Relevance and Prospects. *Voprosy Statistiki*. 2020;27(2):5–16. (In Russ.) Available from: https://doi.org/10.34023/2313-6383-2020-27-2-5-16.

About the authors

Yuri M. Akatkin – Cand. Sci. (Econ.), Head, Research Laboratory of Semantic Analysis and Integration, Plekhanov Russian University of Economics. 36, Stremyanny Lane, Moscow, 117997, Russia. E-mail: u.akatkin@semanticpro.org. ORCID: https://orcid.org/0000-0001-6659-0961.

Elena D. Yasinovskaya – Senior Resercher, Research Laboratory of Semantic Analysis and Integration, Plekhanov Russian University of Economics. 36, Stremyanny Lane, Moscow, 117997, Russia. E-mail: elena@semanticpro.org. ORCID: https://orcid.org/0000-0001-8226-3549.

Andrew V. Shilin – General Director, Electronic Design LLC. 4, Barabannyj Lane, Moscow, 107023, Russia. E-mail: a.shilin@e-projecting.ru. ORCID: https://orcid.org/0000-0001-5827-3923.

Mikhail G. Bich – Cand. Sci. (Tech.), Technical Director, Electronic Design LLC. 4, Barabannyj Lane, Moscow, 107023, Russia. E-mail: misha@e-projecting.ru. ORCID: https://orcid.org/0000-0001-9380-8364.