Связанные статистические данные: актуальность и перспективы

Юрий Михайлович Акаткин[а)],
Константин Эмильевич Лайкам [а), б)],
Елена Донатовна Ясиновская[а)]
[а)] Российский экономический университет имени Г.В. Плеханова, г. Москва, Россия;
[б)] Федеральная служба государственной статистики, г. Москва, Россия

В данной статье после развернутой аргументации актуальности проведенного исследования рассмотрены перспективы внедрения концепции связанных статистических данных, формируемых в рамках единого информационного пространства, обеспечивающего эффективное производство, распространение и повторное использование статистических и административных данных. Реализация этой качественно новой концепции на основе технологических новаций, предпринимаемая в целях более полного удовлетворения быстро возрастающих потребностей пользователей - ключевая задача цифровой трансформации, определенная Правительством Российской Федерации в области официальной статистики. Большая часть открытых данных связана со статистикой: демографическими, экономическими и социальными показателями. Их описание и представление в виде связанных данных могло бы стать важной основой для ускорения социально-экономического развития страны путем создания новых общественно значимых государственных, муниципальных, некоммерческих и коммерческих услуг/продуктов.

В статистике связанные открытые данные (Linked Open Statistical Data, LOSD) позволяют выполнять анализ на основе скоординированной, интегрированной информационной базы как альтернативы использованию разрозненных и часто противоречивых наборов данных. Национальные статистические службы и государственные органы целого ряда стран, а также международные организации уже перешли на парадигму связанных данных. Авторы статьи рассматривают преимущества этого подхода, а также практику его применения в международных проектах.

Приведены примеры и лучший опыт создания связанных открытых статистических данных в публикациях и стратегических документах в рамках Европейской статистической системы. Показано, что развитие связанных статистических данных сдерживается отсутствием доступных онтологий и стандартов - расширений, необходимых для обеспечения требований к классификации различных концептов в статистике и управлению ими. Проведенный в статье анализ проектов и инициатив отражает возможности и перспективы решения данной проблемы в сфере государственной статистики. Сформулированные авторами рекомендации основаны как на анализе международной практики, так и на результатах собственного опыта разработок в рамках научно-исследовательского проекта «Центр семантической интеграции».

Ключевые слова: связанные статистические данные, цифровое государственное управление, статистика, датацентричность, онтологии в статистике.
JEL: C81, C82, D80, M40, O10.
doi: https://doi.org/10.34023/2313-6383-2020-27-2-5-16.

Для цитирования: Акаткин Ю.М., Лайкам К.Э., Ясиновская Е.Д. Связанные статистические данные: актуальность и перспективы. Вопросы статистики. 2020;27(2):5-16.

Linked Open Statistical Data: Relevance and Prospects

Yuri M. Akatkin[a)],
Konstantin E. Laykam[a), b)],
Elena D. Yasinovskaya[a)]
[a)] Plekhanov Russian University of Economics, Moscow, Russia;
[b)] Federal State Statistics Service, Moscow, Russia

After a detailed argumentation of the study's relevance, this article discusses the prospects for introducing the concept of linked open statistics produced within the framework of a single information environment that ensures efficient production, dissemination, and reuse of statistical and administrative data. The implementation of this qualitatively new concept based on technological innovations and aimed to meet rapidly growing user demands is a key task of digital transformation, defined by the Government of the Russian Federation in the field

of official statistics. The major part of open data concerns statistics such as demographic, economic and social indicators. Describing and presenting them in the form of linked open statistics sets an important background for accelerating socio-economic development by introducing new socially significant state, municipal, non-commercial and commercial services/products.

Linked Open Statistical Data (LOSD) allows performing analysis based on a coordinated, integrated information environment as an alternative to using disparate and often controversial data sets. National statistical institutes and government bodies in many countries, together with international organizations, have already chosen the paradigm of linked open statistics. The authors discuss the advantages of this approach, as well as its practical application in international projects.

The article presents the examples and best practices of linked open statistics in a number of publications and strategic documents within the European Statistical System. It also shows the constraints of the linked open statistics development due to the lack of accessible ontologies and standards - the extensions necessary to meet the requirements for classification and management of various concepts in statistics domain. The analysis of projects and initiatives carried out in the article reflects the possibilities and prospects of solving this problem in the field of state statistics. The authors formulate a set of recommendations based both on the analysis of international practice and on the results of their own development experience within the research project «Center of Semantic Integration».

Keywords: Linked Statistical Data, digital government, statistics, data centricity, statistics ontology.
JEL: C81, C82, D80, M40, O10.
doi: https://doi.org/10.34023/2313-6383-2020-27-2-5-16.

For citation: 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.)

Введение

Международные, государственные и частные организации все чаще открывают свои данные для повторного (англ. re-used) использования [1-3]. Большая часть открытых данных связана со статистикой: демографическими (например, данные переписи), экономическими и социальными показателями (например, количество новых предприятий, уровень безработицы) [4-6]. Открытые многомерные статистические данные сегодня составляют важную основу для ускорения социально-экономического развития путем создания новых общественно значимых государственных, муниципальных, некоммерческих и коммерческих услуг/продуктов [7, 8].

Правительством Российской Федерации в 2019 г. были утверждены Концепция создания цифровой аналитической платформы предоставления статистических данных[1] и Концепция создания и функционирования национальной системы управления данными[2], а также Национальная программа «Цифровая экономика Российской Федерации»[3] и федеральный проект «Цифровое государственное управление»[4]. Федеральная служба государственной статистики (Росстат) в целях реализации этих программных документов осуществляет работы по созданию государственной информационной системы «Цифровая аналитическая платформа предоставления статистических данных».

Создание цифровой платформы направлено на достижение следующих целей:

- формирование и использование единого информационного пространства, обеспечивающего эффективную реализацию процесса производства статистической информации в контуре цифрового государственного управления;
- однократное предоставление первичных статистических данных и их многократное использование;
- снижение затрат на сбор, хранение, обработку и распространение статистической информации;
- снижение отчетной нагрузки на респондентов;
- повышение эффективности процесса распространения официальной статистической информации и степени удовлетворенности пользователей предоставленными данными.

В настоящее время Росстат использует SDMX (международный XML формат обмена статистическими данными) в Единой межведомственной информационно-статистической

[1] Утверждена распоряжением Правительства Российской Федерации от 17 декабря 2019 г. № 3074-р. URL: http://static.government.ru/media/files/4YejV8mvcCSeGWTg2kXprmthtNbWyfrU.pdf.

[2] Утверждена распоряжением Правительства Российской Федерации от 3 июня 2019 г. № 1189-р. URL: http://static.government.ru/media/files/jYh27VIwiZs44qa0IXJlZCa3uu7qqLzl.pdf.

[3] Паспорт национальной программы утвержден решением президиума Совета при Президенте Российской Федерации по стратегическому развитию и национальным проектам 24 декабря 2018 г. URL: http://government.ru/info/35568/.

[4] URL: https://digital.gov.ru/ru/activity/directions/882/.

системе (ЕМИСС), а также применяет стандарт Data Documentation Initiative (DDI)[5] при распространении итогов различных выборочных обследований населения.

В целях создания единой методологической и структурной основы для построения интегрированной системы статистических ресурсов авторами была предложена модель статистического показателя, обеспечивающая возможность его унифицированного описания для использования в межведомственных информационных ресурсах [9, 10]. Ряд положений этого подхода нашел отражение в ЕМИСС.

Применение объектных моделей, к которым относятся SDMX и DDI, обеспечивает улучшение понимания данных пользователями и повышение интероперабельности статистических информационных систем. В то же время использование объектного подхода не позволяет преодолеть высокую фрагментацию информационного пространства. Объектные модели имеют существенные ограничения по уровню глубины и сложности, их трудно наращивать и связывать между собой, они не формируют многомерных структур понятий, не отражают вариативность отношений и взаимосвязей, существенно важных для представления концептов реального мира. Именно поэтому публикация открытых статистических данных (Open Statistical Data - OSD) автоматически не дает явных преимуществ [11] - их повторное эффективное использование затруднено, что объясняется в первую очередь фрагментарностью среды OSD, разрозненностью данных и отсутствием возможности их содержательной интерпретации [12] .

На решение этой проблемы направлена совокупность технологий семантической сети (Semantic Web - SW), таких как RDF, OWL, SKOS, SPARQL и др. В семантической сети данные представляются в стандартном виде, с учетом отношений (связей) между ними, которые создаются в соответствии с принципами SW[6]. Такое семантическое аннотирование данных позволяет не только человеку, но и компьютеру однозначно определять их содержательную интерпретацию с использованием семантических моделей (онтологий, тезаурусов, глоссариев и словарей), которые не имеют ограничений по сложности, связанности и вариативности. Коллекции взаимосвязанных наборов данных называют также *связанными данными* (Linked Data - LD)[7]. Публикация LD облегчает поиск и интеграцию данных [13], а технологии SW обеспечивают среду, в которой приложения могут запрашивать данные и управлять ими, формировать интерфейсы и делать выводы с учетом семантических связей. В 2017 г. Консорциум всемирной паутины (World Wide Web Consortium - W3C) рекомендовал связанные данные в качестве наиболее эффективного способа открытия данных в Интернете[8].

В статистике связанные открытые данные (Linked Open Statistical Data - LOSD) позволяют выполнять комплексный анализ разрозненных и изолированных наборов данных [14-16]. В результате многие национальные статистические службы и государственные органы (например, правительство Шотландии, правительство Фландрии, Национальный институт статистики Италии) уже сейчас активно используют парадигму связанных данных для публикации статистических показателей в Интернете [17, 18]. В этом направлении было предложено множество стандартных словарей (например, QB, SKOS, XKOS) и ведется разработка необходимых семантических моделей (например, в проекте LOD2[9]) [19].

LOSD в Европейской статистической системе

В рамках Европейской статистической системы (European Statistical System - ESS) создана сеть LOSD ESSnet[10] для сбора и анализа лучшего опыта публикации связанных открытых статистических данных как внутри статистических организаций (национальные статистические службы, Евростат), так и за их пределами (например, программа европейской интероперабельности - ISAI).

В LOSD ESSnet входят четыре национальные статистические службы: Болгарии (координатор проекта), Франции, Ирландии, Италии. В конце 2017 г. стартовал проект, направленный на решение следующих задач:

[5] URL: https://ddialliance.org/Specification/DDI-Lifecycle/3.2/.
[6] Berners-Lee T. Linked Data - Design Issues. W3C. URL: https://www.w3.org/DesignIssues/LinkedData.html.
[7] URL: https://www.w3.org/standards/semanticweb/data.
[8] URL: https://www.w3.org/TR/dwbp/.
[9] URL: http://aksw.org/Projects/LOD2.html.
[10] URL: https://ec.europa.eu/eurostat/cros/content/essnet-linked-open-statistics_en.

- изучение опыта национальных статистических систем, публикующих статистические данные в виде LOSD;

- обеспечение возможности пользователей легко взаимодействовать со статистической информацией, представленной в виде связанных открытых данных;

- подготовка национальных статистических служб к публикации LOSD в Европейской статистической системе - ESS;

- разработка рекомендаций о дальнейшем развитии ESS в части получения и использования LOSD [20].

Первоначальные связанные открытые статистические данные были опубликованы в апреле и октябре 2018 г., а окончательные результаты получены в апреле 2019 г. Проект показал, что среди национальных статистических служб, которые имеют опыт публикации LOSD, существует общее понимание преимуществ связанных данных. К прямым преимуществам относятся: 1) более гибкие средства распространения данных; 2) расширенные возможности анализа данных, в том числе полученных из различных наборов данных; а также 3) возможность связывания с другими источниками (например, в рамках национальной статистической системы); при этом 4) информация о происхождении данных сохраняется. Косвенные преимущества заключаются в том, что проекты LOD способствуют обеспечению внутренней согласованности данных и метаданных, усиливают роль национальных статистических служб в разработке стандартов и стимулируют их партнерские отношения.

Несмотря на то, что связанные данные - это область, которая для национальных статистических служб в основном все еще остается экспериментальной, достигнуто общее понимание необходимости использовать преимущества LOSD для скоординированной разработки последующих шагов на уровне ESS. В целях дальнейшей демонстрации осуществимости и преимуществ LOSD планируется использовать и конкретные результаты различных экспериментальных проектов.

В разработанной в рамках реализации проекта стратегии определены следующие приоритеты:

- необходимо наращивать потенциал на уровне национальных статистических служб и ESS посредством обучения кадров, совместных пилотных проектов и сотрудничества между междисциплинарными командами;

- общие подходы к LOD и процессы управления ими должны разрабатываться совместно и внедряться в существующие структуры ESS, Евростата и национальных статистических служб, поскольку управление LOD является ключевым элементом;

- несмотря на то, что сегодня существует множество технологий, основным преимуществом ESS должен стать набор стандартных инструментов и руководств, направленный на достижение эффективной работы и надежности;

- ESS следует систематически поддерживать связь с разработчиками стандартов за пределами ЕС (например, Австралии, Японии) [21].

Онтологии в статистике

Одновременно с развитием связанных статистических данных исследователи пришли к выводу о том, что для обеспечения требований к классификации различных концептов в статистике и управлению ими необходимы значительные расширения. Например, в SKOS - одном из часто используемых в LOSD базовом словаре - стандартные отношения («расширяет», «сужает», «связан с»[11]) описывают связи, традиционные для тезаурусов, но они недостаточны для описания статистических классификаторов, которым зачастую свойственны более формально определенные иерархические отношения, например «наследование» или «разделение» (целое/части). Кроме того, иерархии статистических классификаторов структурированы в соответствии с уровнями, отражающими более детальные представления о рассматриваемой предметной области. В то же время управление статистическими концептами требует использования ассоциаций, которые являются более конкретными, чем общие типа «связан с», поскольку необходимо определять как причинно-следственные, так и временные отношения. Для снятия этих ограничений в 2013 г. ООН, ОБСЕ и Евростат предложили использовать расширение XKOS (eXtended Knowledge Organization System), которое позволяет представить более подробные описания, необходимые для управления статис-

[11] «broader than», «narrower than», and «related to».

тическими классификаторами, расширяя существующие определения объектов и отношений классов SKOS. XKOS разработан не только на основе потребностей статистического сообщества, но и с учетом требований терминологических стандартов, в частности ИСО 704: 2009 (ISO 704) и ИСО 1087-1: 2000 (ISO 1087), которые определяют конструкции и отношения, необходимые для управления концептами и полного описания статистических классификаторов [22]. Другие варианты расширения SKOS и применения расширенного состава внешних словарей и моделей представлены также в проекте открытых данных Японии[12].

Онтологии уже на протяжении многих лет успешно используются в менее формализованном пространстве Semantic Web, обеспечивая формальное именование, определение и описание концептов предметной области, а также отношения между этими концептами. В официальной статистике также существует ряд моделей, словарей или других семантических моделей, но они, как правило, не являются формально выраженными или согласованными друг с другом.

Большая работа по развитию онтологий в государственных данных, в том числе в домене статистики, проводится в Великобритании, активно реализующей цифровую трансформацию государственного управления[13]. На ранних стадиях публикации связанных государственных данных было определено, что практически все они принимают форму таблиц или многомерных кубов. При этом зачастую необходимо обращаться к организациям (ведомствам, подведомственным структурам и т. д.) не только для того, чтобы описать их структуру, но и для того, чтобы иметь возможность связывать данные с теми организациями, которые их собирают, лицензируют и публикуют.

Процесс развития государственных данных, направленный на достижение пятого уровня 5-звездочной модели зрелости, предложенной Тимом Бернерсом-Ли[14] (все государственные данные открыты, связаны и опубликованы в машиночитаемом формате, предоставляя потребителю контекст), сдерживается отсутствием доступных онтологий или стандартов. Поэтому при поддержке Национального архива Великобритании была поставлена задача разработать онтологии для представления кубов данных (статистика, измерения, расходы) и для представления организаций, а затем встроить их в международные стандарты.

В рамках этой программы был разработан так называемый Словарь кубов данных - Data Cube Vocabulary, который в 2014 г. был рекомендован W3C[15] и широко используется в различных областях. В основе словаря лежит концепт «набор данных», который представляет собой совокупность наблюдений. Наблюдения организованы по набору измерений (например, время, географический регион), и каждое наблюдение имеет одно или несколько связанных измерений (например, структура населения или качество воздуха). Для надежной интерпретации измерений может использоваться другая информация, такая как единицы измерения или используемый процесс измерения; эти аннотации называются атрибутами. Также была разработана и принята в качестве рекомендации W3C Онтология организации (Organization Ontology)[16], позволяющая публиковать и связывать информацию об организациях и их структуре.

В то же время совместно с органом по стандартизации электронного правительства Великобритании (LeGSB iStandUK) велась работа над созданием Онтологии расходов (Payment Ontology)[17] для описания данных о расходах правительства на местном уровне. Сложность моделирования информации о расходах заключается в том, что у каждого местного органа управления существует свой способ организации анализа собственных расходов: различные способы группировки и классификации, а также установление связей как с бюджетом, так и с услугами. Применение онтологий допускает использование открытого (расширяемого) набора схем анализа, каждая из которых описывает себя и организована с использованием подхода SKOS. В основе онтологии лежит Data Cube Vocabulary. Это позволило, с одной стороны, гибко публиковать результаты анализа, сохраняя то значение (семантику), которое необходимо органам власти на местах, а с другой - обеспечило возможность их сопоставле-

[12] URL: http://data.e-stat.go.jp/lodw/en/outline/example#1-3-7.
[13] URL: https://ukparliament.github.io/ontologies/.
[14] Berners-Lee T. 5-Star Open Data. URL: http://5stardata.info/en/.
[15] The RDF Data Cube Vocabulary. URL: https://www.w3.org/TR/vocab-data-cube/.
[16] Organization Ontology: ORG. URL: https://www.w3.org/TR/vocab-org/.
[17] URL: https://www.epimorphics.com/casestudy/data-gov-uk/.

ния и проведения комплексного анализа данных различных ведомств.

Национальный институт статистики Италии (Italian National Institute of Statistics - Istat) также сообщает об использовании онтологий для целей интеграции и распространения данных. Они основываются на парадигме управления данными на базе онтологий (OBDM), предложенной для интеграции нескольких разнородных источников данных. Этот опыт был применен на портале связанных открытых данных Istat [23]. Онтологии использовались и в ходе проведения переписи населения. На сегодняшний день опубликованы:

- Онтология «Местоположение» - итальянский профиль приложения для базового европейского словаря «Местоположение»[18], который отражает основные характеристики адреса;

- Онтология «Население и домашнее хозяйство» - итальянский профиль приложения, который описывает людей, их место жительства, место рождения и домашнее хозяйство;

- Универсальная онтология - версия Istat для профилирования онтологии GSIM Общей информационной статистической модели (Generic Statistic Information Model, GSIM), разрабатываемой ЕЭК ООН[19].

GSIM содержит объекты, которые определяют информацию о реальном мире («информационные объекты»), и включает данные и метаданные (такие как классификации), а также правила и параметры, необходимые для запуска процессов управления (например, правила редактирования данных). GSIM идентифицирует около 110 информационных объектов, которые объединены в четыре группы верхнего уровня:

• Группа «*Бизнес*» используется для определения планов и статистических программ, а также процессов, выполняемых для реализации этих программ. Она включает в себя определение статистических требований, бизнес-процессов, составляющих статистические программы и их оценку.

• Группа «*Обмен*» применяется для каталогизации информации, которая поступает в статистическую организацию и выходит из нее через каналы обмена. Она включает в себя объекты, которые описывают сбор и распространение информации.

• Группа «*Понятия*» служит основой для определения семантики данных, обеспечивая понимание того, какие данные измеряются.

• Группа «*Структура*» используется для описания и определения терминов, применяемых в отношении информации и ее структуры [24].

В ЕЭК ООН внедрение онтологий ведется группой высокого уровня по модернизации официальной статистики (High-Level Group for the Modernisation of Official Statistics - HLG-MOS[20]). В ней выделена группа поддержки стандартов[21], сформированная для того, чтобы найти способы разработки, улучшения, интеграции, продвижения стандартов, необходимых для модернизации статистики, и содействовать их внедрению. Она несет операционную ответственность за поддержание и развитие Общей модели деятельности для статистических организаций (GAMSO)[22], Общей модели статистических бизнес-процессов (GSBPM)[23], Общей модели статистической информации (GSIM) и документации Общей архитектуры производства статистики (CSPA)[24]. Группа ведет работу по следующим направлениям:

- связывание GSBPM и GSIM;

- согласование комплексных процессов GSBPM с GAMSO;

- создание базовой онтологии для официальной статистики;

- глоссарий метаданных.

Создание Базовой Онтологии для официальной статистики (Core Ontology for Official Statistics - COOS) направлено на решение проблемы разнородности и фрагментарности существующих семантических моделей в статистике. Реализация этого мероприятия началась в ноябре 2018 г.

Задачами COOS являются:

- обеспечить формальное представление (используется OWL) базовых понятий, присутствующих в основных моделях, особенно тех, которые не имеют формального моделирования (GSPBM, GAMSO);

- определить отношения между этими понятиями, в частности между теми, которые принадлежат к разным моделям;

[18] URL: https://joinup.ec.europa.eu/solution/core-location-vocabulary.
[19] URL: https://www.istat.it/en/ontology.
[20] URL: https://statswiki.unece.org/pages/viewpage.action?pageId=187891840.
[21] URL: https://statswiki.unece.org/display/hlgbas/Modernisation+Groups.
[22] URL: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2015/12-Generic_Activity_Model_for_Statistical_Organisations_HLG.pdf.
[23] URL: https://ec.europa.eu/eurostat/cros/content/gsbpm-generic-statistical-business-process-model-theme_en.
[24] URL: https://ec.europa.eu/eurostat/cros/system/files/07a._validation_services_developed_in_the_ess_-_cspa.pdf.

- предложить отношения между статистическими понятиями и объектами, определенными в других онтологиях.

Косвенной, но важной целью является также создание сообщества статистиков, заинтересованных в разработке онтологий. Команда COOS начала работать в феврале 2019 г. посредством виртуальных собраний на основе специального хранилища GitHub с целью создания первой версии COOS и ее последующего представления на семинаре HLG-MOS [25].

Собственный опыт разработок в области LOSD

В 2016 г. в Российском экономическом университете имени Г.В. Плеханова (РЭУ им. Г.В. Плеханова) стартовал научно-исследовательский проект - «Центр семантической интеграции» (ЦСИ)[25], направленный на достижение следующих целей:

- исследование и апробация современных подходов к управлению семантическими активами, разработке методов и инструментов семантической интеграции на основе принципов MDA (архитектура, управляемая моделью) и повторного использования семантических активов для обеспечения семантической интероперабельности;

- создание платформы коллективной работы, которая обеспечит возможность каталогизации и управления семантическими активами, предоставит инструментарий моделирования информационных систем и сервисов информационного обмена, а также перевода открытых данных в связанные данные для сопоставления, анализа и визуализации.

В 2018 г. в РЭУ им. Г.В. Плеханова была организована научная лаборатория (НЛ) «Семантической интеграции и анализа», которая развивает созданный макет ЦСИ и ведет на его базе работу по ряду направлений:

- формализация знаний на основе создания, распространения и повторного использования семантических моделей;

- обеспечение семантической интероперабельности для построения экосистем цифровых отраслей;

- развитие комплексного анализа данных статистики и мониторинга за счет использования семантических связей.

В рамках НЛ «Семантической интеграции и анализа» проводятся исследования и апробация различных методов и инструментов, предназначенных для создания, распространения, публикации и повторного использования семантических моделей, а также обогащения данных семантическими связями, например используемых в проекте LOD2 Statistical Workbench (см. рис. 1).

В области статистики в начале 2019 г. сотрудниками лаборатории с использованием макета

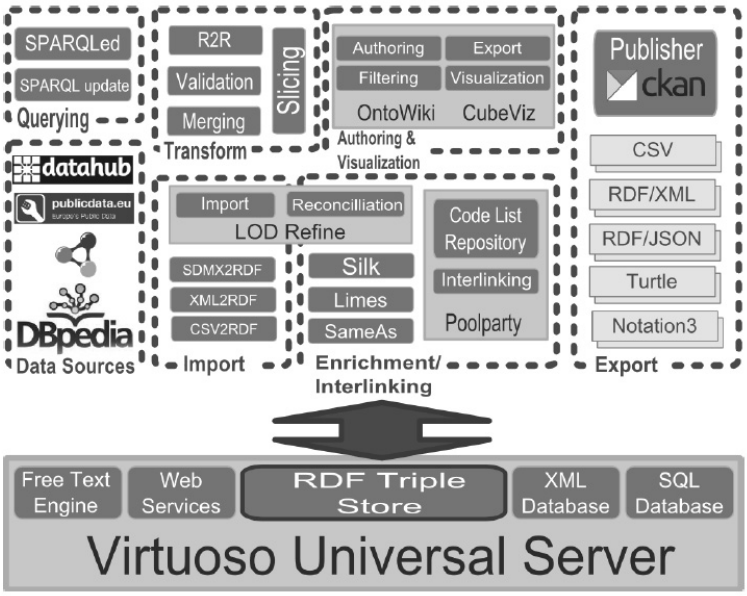

Рис. 1. **Пример архитектуры приложения LOD2 Statistical Workbench**

Источник: [26].

[25] URL: http://csi.semanticpro.org/.

ЦСИ были проведены работы по апробации технологии перевода данных EuroStat (Population as a percentage of EU28 population) в LOD.

В ходе этих работ была выполнена загрузка схем данных в формате SDMX, полученных на сайте Евростата, а также загрузка самих данных в формате CSV. С использованием инструмента Open Refine[26] проведена загрузка CSV с преобразованием в RDF.

В процессе преобразования проводились разметка данных в структуре RDF Data Cube Vocabulary и назначение словарей (например, указывалось, что выбранное поле соответствует skos:prefLabel определенного словаря). Для установки связи между элементами наборов данных после загрузки использовался инструмент Silk[27]. На рис. 2 представлен пример визуализации данных.

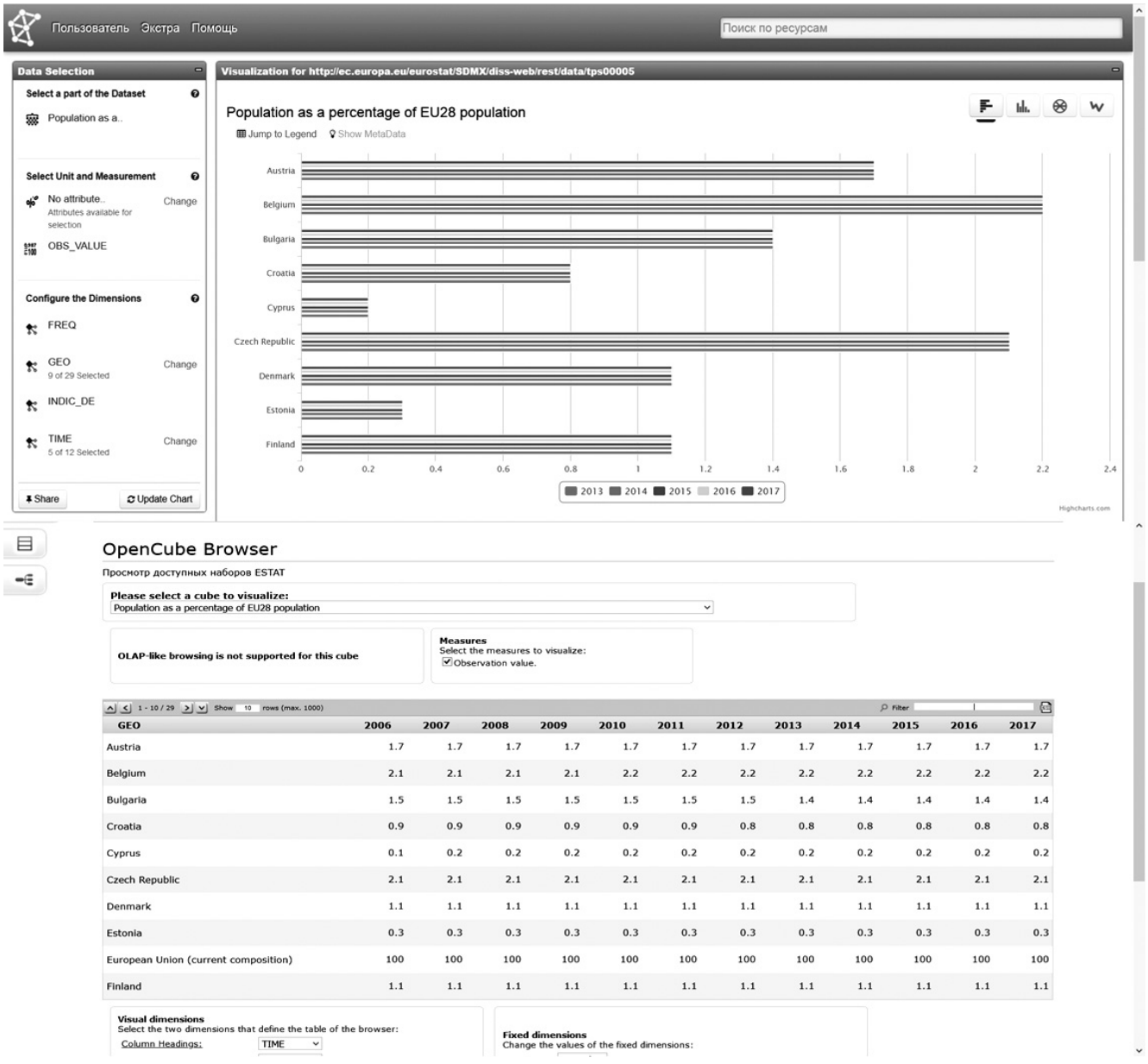

Рис. 2. **Пример визуализации загруженных данных Евростата**

[26] URL: https://openrefine.org/.
[27] URL: http://silkframework.org /.

Визуализация данных выполнялась путем применения двух инструментов: OntoWiki[28] для построения графиков и OpenCube Open Source Toolkit[29] для отображения данных в виде таблиц.

Заключение

Представленный в данной статье анализ публикаций и материалов, посвященных связанным статистическим данным, а также применению онтологий для формирования их семантических описаний, позволяет сделать вывод, что это направление исследований не только актуально, но и является одним из наиболее важных трендов развития международной статистики в цифровом обществе. Показано, что, несмотря на ряд концептуальных документов и множество научных исследований, практические работы по распространению LOSD пока еще носят фрагментарный характер. Наборы данных (например, Евростата), заявленные как связанные, к сожалению, зачастую недоступны, а реализованные в рамках академических исследований проекты значимых практических результатов пока не дали.

Следует признать, что создание и распространение семантических моделей - достаточно сложная, трудоемкая и многоаспектная работа, которая может быть выполнена лишь благодаря значительным совместным усилиям ИТ-специалистов и экспертов соответствующей предметной области. В то же время вложенные в «связывание» данных усилия вознаграждаются достижением нового уровня статистического анализа с применением средств визуализации и предоставлением возможности в полной мере учитывать знания и контекст, сформированные благодаря использованию семантического подхода.

Исходя из проведенного анализа и имеющегося опыта, мы предлагаем определить в качестве приоритетных следующие направления исследований и разработок:

- создание и распространение статистических онтологий и других семантических моделей статистических данных (инструменты и методы);

- инструменты и методы коллективной работы ИТ-специалистов и статистиков-методологов;

- средства визуализации, демонстрирующие преимущества связанных статистических данных;

- методы и инструменты управления семантическими моделями статистических данных.

Для российской статистики, на наш взгляд, наиболее актуальными (с учетом установленных национальных целей социально-экономического развития страны) и перспективными (с учетом накопленного разноплановой и значительного объема статистического материала) являются следующие предметные области: демография, уровень жизни (доходы населения, бедность, социальная помощь), рынок труда.

Литература

1. **Kalampokis E., Tambouris E., Tarabanis K**. A Classification Scheme for Open Government Data: Towards Linking Decentralised Data // International Journal of Web Engineering and Technology (IJWET). 2011. Vol. 6. No. 3. P. 266-285. URL: http://www.inderscience.com/offer.php?id=40725.

2. **Attard J.** et al. A Systematic Review of Open Government Data Initiatives // Government Information Quarterly. 2015. Vol. 32. Iss. 4. P. 399-418. doi: https://doi.org/10.1016/j.giq.2015.07.006.

3. **Zuiderwijk A., Janssen M.** Open Data Policies, Their Implementation and Impact: A Framework for Comparison // Government Information Quarterly. 2014. Vol. 31. Iss. 1. P. 17-19. doi: https://doi.org/10.1016/j.giq.2013.04.003.

4. Commission Notice - Guidelines on Recommended Standard Licences, Datasets and Charging for the Reuse of Documents // Official Journal of the European Union. C 240, 24.7.2014, p. 1-10. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014XC0724%2801%29.

5. **Capadisli S., Auer S., Ngonga Ngomo A.-C.** Linked SDMX Data // Semantic Web. 2015. Vol. 6. No. 2. P. 105-112. URL: http://semantic-web-journal.org/system/files/swj454.pdf.

6. **Cyganiak R., Hausenblas M., McCuirc E**. Official Statistics and the Practice of Data Fidelity // Wood D. (ed.) Linking Government Datahttps. New York: Springer-Verlag, 2011. P. 135-151. doi: https://doi.org/10.1007/978-1-4614-1767-5.

7. **Kalampokis E.** et al. Open Statistics: The Rise of a New Era for Open Data? // Scholl H. et al. (eds) Electronic Government. EGOV 2016. Lecture Notes in Computer Science, Vol. 9820. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-44421-5_3.

8. **Janssen M., van den Hoven J.** Big and Open Linked Data (BOLD) in Government: A Challenge to Transparency

[28] URL: http://ontowiki.net/.
[29] URL: http://opencube-toolkit.eu/.

and Privacy? // Government Information Quarterly. 2015. Vol. 32. Iss. 4. P. 363-368. doi: https://doi.org/10.1016/j.giq.2015.11.007.

9. **Акаткин Ю.М., Лайкам К.Э.** Методологические проблемы унифицированного описания статистических показателей для использования в межведомственных информационных ресурсах // Вопросы статистики. 2010. № 7. С. 3-9.

10. **Акаткин Ю.М., Лайкам К.Э.** О некоторых методических вопросах унифицированного описания статистического показателя // Вопросы статистики. 2011. № 7. С.11-19.

11. **Janssen M., Charalabidis Y., Zuiderwijk A.** Benefits, Adoption Barriers and Myths of Open Data and Open Government // Information System Management. 2012. Vol. 29. Iss. 4. P. 258-268. doi: https://doi.org/10.1080/10580530.2012.716740.

12. **Hassani H., Saporta G., Silva E.S.** Data Mining and Official Statistics: The Past, the Present, and the Future // Big Data. 2014. Vol. 2. No. 1. P. 31-43. doi: https://doi.org/10.1089/big.2013.0038.

13. **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.

14. **Abellro A.** et al. Fusion Cubes: Towards Self-Service Business Intelligence // International Journal of Data Warehousing and Mining. 2013. Vol. 9. Iss. 2. P. 66-88. doi: https://doi.org/10.4018/jdwm.2013040104.

15. **Kalampokis E., Tambouris E., Tarabanis K**. Linked open cube analytics systems: Potential and challenges // IEEE Intelligent Systems. 2016. Vol. 31. Iss. 5. P. 89-92. doi: https://doi.org/10.1109/MIS.2016.82.

16. **Bischof S.** et al. Enriching Integrated Statistical Open City Data by Combining Equational Knowledge and Missing Value Imputation // Journal of Web Semantics. 2018. Vol. 48. P. 22-47. doi: https://doi.org/10.1016/j.websem.2017.09.003.

17. **Klímek J.** et al. Publication and Usage of Official Czech Pension Statistics Linked Open Data // Journal of Web Semantics. 2018. Vol. 48. P. 1-21. doi: https://doi.org/10.1016/j.websem.2017.09.002.

18. **Chaniotaki E.** et al. Exploiting Linked Statistical Data in Public Administration: The Case of the Greek Ministry of Administrative Reconstruction. 23rd Americas Conference on Information Systems (AMCIS 2017), Boston, MA, USA, August 10-12, 2017. URL: https://aisel.aisnet.org/amcis2017/eGovernment/Presentations/15/.

19. **Kalampokis E., Zeginis D., Tarabanis K.** On modeling linked open statistical data // Journal of Web Semantics. 2019. Vol. 55. P. 56-68. doi: https://doi.org/10.1016/j.websem.2018.11.002.

20. LOSD Publication and Capacity Building. D1.1 Vision, Stakeholders and Business Case Definition. 15th April 2018. URL: https://ec.europa.eu/eurostat/cros/system/files/d1.1_vision_stakeholders_and_business_case_v3.pdf.

21. Eurostat. ESS Vision 2020: DIGICOM, Towards an ESS Strategy on (Linked) Open Data URL: https://ec.europa.eu/eurostat/cros/content/lod-strategic-documents_en.

22. **Cotton F.** et al. XKOS: An SKOS Extension for Statistical Classifications. Conference: ISI 2014, 59 World Statistics Congress, Hong Kong, China, August 2014. URL: https://www.researchgate.net/publication/280740700_XKOS_An_SKOS_Extension_for_Statistical_Classifications/stats.

23. **Aracri R.M.** et al. Using Ontologies for Official Statistics: The Istat Experience // Garrigys I., Wimmer M. (eds) Current Trends in Web Engineering. ICWE 2017. Lecture Notes in Computer Science, vol 10544. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-74433-9_15.

24. Generic Statistical Information Model (GSIM). High-level Workshop on Modernization of Official Statistics. Nizhni Novgorod, Russian Federation, 10-12 June 2014. URL: https://www.unece.org/fileadmin/DAM/stats/documents/fund.principles/2014/2-Generic_Statistical_Information_Model_EN.pdf.

25. **Cotton F.** Core Ontology for Official Statistics. Conference of European Statisticians. ModernStats World Workshop 2019. 26-28 June 2019, Geneva, Switzerland. URL: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2019/mtg2/MWW2019_COOS_Cotton_Abstract.pdf.

26. **Janev V.** et al. Supporting the Linked Data Publication Process with the LOD2 Statistical Workbench // Semantic Web - Interoperability, Usability, Applicability. Under review: submitted on 29.11.2013. URL: http://www.semantic-web-journal.net/system/files/swj591.pdf.

Информация об авторах

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

Лайкам Константин Эмильевич - д-р экон. наук, канд. техн. наук, заместитель руководителя Федеральной службы государственной статистики; директор НИИ «Институт современных экономических исследований», Российский экономический университет им. Г.В. Плеханова. 107450, г. Москва, ул. Мясницкая, 39, с. 1; 117997, г. Москва, Стремянный пер, 36. E-mail: laikam@gks.ru. ORCID: https://orcid.org/0000-0002-3205-1457.

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

References

1. **Kalampokis E., Tambouris E., Tarabanis K**. A Classification Scheme for Open Government Data: Towards Linking Decentralised Data. *International Journal of Web Engineering and Technology (IJWET)*. 2011;6(3):266-285. Available from: http://www.inderscience.com/offer.php?id=40725.

2. **Attard J.** et al. A Systematic Review of Open Government Data Initiatives. *Government Information Quarterly*. 2015;32(4):399-418. Available from: doi: https://doi.org/10.1016/j.giq.2015.07.006.

3. **Zuiderwijk A., Janssen M.** Open Data Policies, Their Implementation and Impact: A Framework for Comparison. *Government Information Quarterly*. 2014;31(1):17-19. Available from: doi: https://doi.org/10.1016/j.giq.2013.04.003.

4. Commission Notice - Guidelines on Recommended Standard Licences, Datasets and Charging for the Reuse of Documents. *Official Journal of the European Union*. C 240/1, 24.7.2014, p. 1-10. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014XC0724%2801%29.

5. **Capadisli S., Auer S., Ngonga Ngomo A.-C.** Linked SDMX Data. *Semantic Web*. 2015;6(2):105-112. Available from: http://semantic-web-journal.org/system/files/swj454.pdf.

6. **Cyganiak R., Hausenblas M., McCuirc E**. Official Statistics and the Practice of Data Fidelity. In: Wood D. (ed.) *Linking Government Data*. New York: Springer-Verlag; 2011. P. 135-151. Available from: doi: https://doi.org/10.1007/978-1-4614-1767-5.

7. **Kalampokis E.** et al. Open Statistics: The Rise of a New Era for Open Data? In: Scholl H. et al. (eds) *Electronic Government. EGOV 2016. Lecture Notes in Computer Science, Vol 9820*. Springer, Cham. Available from: doi: https://doi.org/10.1007/978-3-319-44421-5_3.

8. **Janssen M., van den Hoven J.** Big and Open Linked Data (BOLD) in Government: A Challenge to Transparency and Privacy? *Government Information Quarterly*. 2015;32(4):363-368. Available from: doi: https://doi.org/10.1016/j.giq.2015.11.007.

9. **Akatkin Yu.M., Laikam K.E.** Methodological Problems of Unified Description of Statistical Indicators for Inter-Agencies Information Resources. *Voprosy Statistiki*. 2010;(7):3-9. (In Russ.)

10. **Akatkin Yu.M., Laikam K.E.** Selected Methodological Questions of Unified Description of Statistical Indicator. *Voprosy Statistiki*. 2011;(7):11-19. (In Russ.)

11. **Janssen M., Charalabidis Y., Zuiderwijk A.** Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information System Management*. 2012;29(4):258-268. Available from: doi: https://doi.org/10.1080/10580530.2012.716740.

12. **Hassani H., Saporta G., Silva E.S.** Data Mining and Official Statistics: The Past, the Present, and the Future. *Big Data*. 2014;2(1):31-43. Available from: doi: https://doi.org/10.1089/big.2013.0038.

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

14. **Abellrо A.** et al. Fusion Cubes: Towards Self-Service Business Intelligence. *International Journal of Data Warehousing and Mining*. 2013;9(2):66-88. Available from: doi: https://doi.org/10.4018/jdwm.2013040104.

15. **Kalampokis E., Tambouris E., Tarabanis K**. Linked open cube analytics systems: Potential and challenges. *IEEE Intelligent Systems*. 2016;31(5):89-92. Available from: doi: https://doi.org/10.1109/MIS.2016.82.

16. **Bischof S.** et al. Enriching Integrated Statistical Open City Data by Combining Equational Knowledge and Missing Value Imputation. *Journal of Web Semantics*. 2018;(48):22-47. Available from: doi: https://doi.org/10.1016/j.websem.2017.09.003.

17. **Klímek J.** et al. Publication and Usage of Official Czech Pension Statistics Linked Open Data. *Journal of Web Semantics*. 2018;(48):1-21. Available from: doi: https://doi.org/10.1016/j.websem.2017.09.002.

18. **Chaniotaki E.** et al. Exploiting Linked Statistical Data in Public Administration: The Case of the Greek Ministry of Administrative Reconstruction. In: *23rd Americas Conference on Information Systems (AMCIS 2017), Boston, MA, USA, August 10-12, 2017*. Available from: https://aisel.aisnet.org/amcis2017/eGovernment/Presentations/15/.

19. **Kalampokis E., Zeginis D., Tarabanis K.** On Modeling Linked Open Statistical Data. *Journal of Web Semantics*. 2019;(55):56-68. Available from: doi: https://doi.org/10.1016/j.websem.2018.11.002.

20. *LOSD Publication and Capacity Building. D1.1 Vision, Stakeholders and Business Case Definition*. 15th April 2018. Available from: https://ec.europa.eu/eurostat/cros/system/files/d1.1_vision_stakeholders_and_business_case_v3.pdf.

21. Eurostat. *ESS Vision 2020: DIGICOM, Towards an ESS Strategy on (Linked) Open Data*. Available from: https://ec.europa.eu/eurostat/cros/content/lod-strategic-documents_en.

22. **Cotton F.** et al. XKOS: An SKOS Extension for Statistical Classifications. In: *Conference: ISI 2014, 59 World Statistics Congress, Hong Kong, China, August 2014.* Available from: https://www.researchgate.net/publication/280740700_XKOS_An_SKOS_Extension_for_Statistical_Classifications/stats.

23. **Aracri R.M.** et al. Using Ontologies for Official Statistics: The Istat Experience. In: Garrigys I., Wimmer M. (eds) *Current Trends in Web Engineering. ICWE 2017. Lecture Notes in Computer Science, vol 10544. Springer, Cham*. Available from: doi: https://doi.org/10.1007/978-3-319-74433-9_15.

24. Generic Statistical Information Model (GSIM). In: *High-level Workshop on Modernization of Official Statistics. Nizhni Novgorod, Russian Federation, 10-12 June 2014*. Available from: https://www.unece.org/fileadmin/DAM/stats/documents/fund.principles/2014/2-Generic_Statistical_Information_Model_EN.pdf.

25. **Cotton F.** Core Ontology for Official Statistics. In: *Conference of European Statisticians. ModernStats World Workshop 2019. 26-28 June 2019, Geneva, Switzerland.* Available from: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2019/mtg2/MWW2019_COOS_Cotton_Abstract.pdf.

26. **Janev V.** et al. Supporting the Linked Data Publication Process with the LOD2 Statistical Workbench. *Semantic Web - Interoperability, Usability, Applicability.* Under review: submitted on 29.11.2013. Available from: http://www.semantic-web-journal.net/system/files/swj591.pdf.

About the authors

Yuri M. Akatkin - Cand. Sci. (Econ.), Head, 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.

Konstantin E. Laykam - Dr. Sci. (Econ.), Cand. Sci. (Tech.), Deputy Head, Federal State Statistics Service; Director, Institute of Modern Economic Researches, Plekhanov Russian University of Economics. 39, Myasnitskaya Str., Bldg.1, Moscow, 107450, Russia; 36, Stremyanny Lane, Moscow, 117997, Russia. E-mail: laikam@gks.ru. ORCID: https://orcid.org/0000-0002-3205-1457.

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