УДК 330:46; 311:42

Пугачова М.В., к.т.н., с.н.с.,
Науково-проектний інститут статистичних технологій НТК статистичних досліджень

МЕТАДАНІ В СИСТЕМАХ СТАТИСТИЧНОГО МОНІТОРИНГУ

Надано визначення понять статистичних метаданих у системах статистичного моніторингу та статистичної метаінформаційної системи, визначені категорії користувачів статистичних метаданих та джерела отримання статистичних метаданих

Вступ. Відомо, що в останнє десятиріччя проблемам створення інформації про статистичні дані, а також її зберіганню та управлінню нею приділяється значна увага. Проведено низку міжнародних конференцій, присвячених цим питанням, міжнародні організації постійно видають документи та керівництва (див., наприклад, [1-5]).

У літературі наводиться визначення статистичних метаданих через такі основні положення про них [1, 6-8]:

1. Метадані – це фізичне представлення метаінформації (метазнань) – аналогічно тому, що дані – представлення інформації (знань).
2. Метадані забезпечують інформацію про дані: про процеси їх створення та використання.
3. Метадані – це дані, що необхідні для належного виробництва та використання даних, про які вони повідомляють.
4. Дані забезпечують інформацію про об'єкти в реальній об'єктній системі, а метадані – про інформаційну систему – метаоб'єкти. Подібно реальним об'єктам, метаоб'єкти мають властивості і пов'язані один з одним об'єктними відношеннями.
5. Даними та метаданими керують інформаційні системи.

Можна зауважити, що три перші твердження можна розглядати як три "проекції" тривимірного визначення метаданих:

- **синтаксичної** (орієнтованої на представлення);
- **семантичної** (орієнтованої на зміст);
- **прагматичної** (орієнтованої на ціль).

У першому пункті термін "знання" використовується як альтернативний терміну "інформація", хоча, на наш погляд, це не зовсім коректно. Взагалі, "знання" повинно інтерпретуватись як поняття, що "ширше" і "глибше", ніж "інформація". Поняття інформації часто фокусується на фактичному знанні (в протилежність знанню, що має характер визначень, правил, законів, алгоритмів тощо). Подібно "знанню" термін "інформація" має на увазі інтерпретацію людським мозком, але "знання" має на увазі подальше "усвоювання" або "розуміння".

Постановка проблеми. На цей час у вітчизняній літературі існує багато визначень метаданих, але майже не приділялось уваги метаданим, що використовуються у статистичних інформаційних системах (СІС). А ті визначення, що надаються у закордонній літературі, відрізняються великою різноманітністю. На нашу думку, слід надати більш просте та зрозуміле поняття тим метаданим, що використовуються у статистичних інформаційних системах та системах статистичного моніторингу, а для цього необхідно вирішити цілу низку супутніх питань: хто є користувачами статистичних даних та статистичних метаданих, які їх потреби та наскільки поглиблену інформацію про статистичні метадані повинні отримувати різні категорії користувачів. Слід також визначити функції, які виконує база статичних метаданих у СІС та системах статистичного моніторингу (ССМ).

Таким чином, **метою дослідження** є визначення місця статистичних метаданих у системах статистичного моніторингу, видів

статистичних метаданих та категорій їх користувачів і виробників, а також надання для прикладу опису статистичних метаданих для ССМ ділової активності підприємств.

Викладення основного матеріалу дослідження. Статистичні метадані у системах статистичного моніторингу – це дані, що необхідні для належного виробництва та використання статистичних даних. Вони описують статистичні дані та – до деякого ступеня – процеси та інструменти, які залучені до виробництва та використання статистичних даних. Тобто, статистичні метадані – це дані про статистичні дані.

Для роботи із статистичними метаданими повинні існувати інструменти, за допомогою яких вони залучаються до процесів обробки та використання статистичних даних. Для цього можна ввести поняття статистичної метаінформаційної системи, тобто такої системи, що використовує і виробляє статистичні метадані, і здійснює свої задачі шляхом виконання таких функцій як "збирання статистичних метаданих", "обробка статистичних метаданих", "зберігання статистичних метаданих" і "розповсюдження статистичних метаданих".

Метаінформаційна система може бути активною або пасивною. Активна метаінформаційна система фізично з'єднана з інформаційною системою, яка містить дані, про які повідомляють метадані в метаінформаційной системі. Пасивна метаінформаційна система містить тільки посилання на дані, а не дані безпосередньо.

Слід підкреслити, що метадані, накопичені в рамках тільки однієї статистичної системи, являють собою великі масиви інформації, які необхідно поповнювати, оновлювати, зберігати, одним словом, з якими потрібно працювати. Для цього і створюються системи статистичних метаданих (статистичні метаінформаційні системи), що функціонують всередині статистичних інформаційних систем. Управління статистичними метаданими – це складна та багатовимірна задача, якою займається ціла низка фахівців у різних організаціях та галузях знань, як у статистичних офісах, так і в університетах та інших наукових закладах, до того ж організоване активне міжнародне співробітництво, прикладом якого може бути ціла низка проектів Євросоюзу, таких як AMRADS, MetaNet, METAWARE, COSMOS [5].

Таким чином, статистична метаінформаційна система повинна бути інструментом, що дає можливість статистичній організації ефективно виконувати свою головну задачу – виробництво офіційних статистичних даних. Тобто управління всіма фазами цього виробництва (збирання, зберігання, оцінка та розповсюдження даних), кожна з яких повинна бути забезпечена необхідними метаданими. Мається на увазі виконання таких функцій:

1. Планування, проектування, впровадження та оцінювання статистичних виробничих процесів.
2. Управління методологічною діяльністю. Використання когерентних (узгоджених, логічно пов'язаних) метаданих у статистичній методології є дуже важливим.
3. Управління взаємодією з користувачами статистичних даних та інформації. Полегшення зворотного зв'язку.
4. Покращення корисності статистичних даних та метаданих для користувачів.
5. Розповсюдження статистичної інформації. Користувачам потрібні надійні метадані для пошуку, навігації та інтерпретації статистичних даних.
6. Покращення якості статистичних даних. Оцінювання якості статистичних даних є однією з найважливіших цілей статистичної діяльності. Система статистичних метаданих повинна мати релевантну множину метаданих для всіх відомих критеріїв оцінки якості статистичної інформації (див., наприклад, [9]).
7. Управління джерелами статистичних даних і взаємодія з постачальниками даних (респондентами).
8. Покращення інтеграції статистичної інформаційної системи з іншими національними інформаційними системами. Наразі зростають потреби у використанні

адміністративних даних для статистичних цілей. Це зобов'язує до кращої інтеграції та розподілу метаданих серед статистиків та державної адміністрації для гарантування когерентності та відповідності інформації, призначеної для обміну.

9. Покращення інтеграції статистичної інформаційної системи з інформаційними системами міжнародних організацій, які вимагають поєднання своїх власних метаданих з метаданими національних статистичних офісів з метою зробити статистичну інформацію більш порівняною та сумісною.

10. Управління, уніфікація та стандартизація виробничих процесів всередині статистичного офісу.

11. Створення бази знань щодо процесів у СІС. Це дозволяє також розподілити знання серед працівників таким чином, щоби мінімізувати риски, пов'язані з їх міграцією.

12. Покращення адміністрування СІС.

13. Сприяння підвищенню дохідності СІС для статистичного офісу.

14. Уніфікація концепції статистичної методології як рушійної сили для кращої комунікації та розуміння між менеджерами, проектувальниками, галузевими статистиками, методологами, респондентами та користувачами СІС.

Очевидними користувачами статистичних метаданих є, з одного боку, виробники статистичних даних, а з іншого – користувачі статистичних даних. Чому ці дві категорії людей нуждаються в статистичних метаданих?

Користувач статистичних даних під час вирішення будь-якої своєї проблеми або питання визначає статистичні дані, що, за його думкою, потенційно необхідні для її розв'язання. Потім він ідентифікує їх, розшукує, аналізує, інтерпретує, а після цього, можливо, повторює частини процедур пошуку та аналізу спочатку. Кожний з наведених кроків вимагає деякої додаткової інформації про статистичні дані, тобто статистичних метаданих, а їх необхідна глибина буде, в першу чергу, залежати від наявних у користувача знань. Тобто, різні категорії користувачів мають різні вимоги до статистичних метаданих.

У табл. 1 наведено деякі умовні категорії користувачів та цілей, для яких вони використовують статистичні дані.

Таблиця 1. Користувачі статистичних даних та їх цілі

Хто?	*Для чого?*
Уряд	Оцінювання, планування, прийняття управлінських рішень
Компанії (підприємства)	Прийняття ділових рішень
Науково-дослідні організації	Аналіз та пояснення реальних явищ, використання для наукових досліджень
Навчальні заклади	Використання у навчальних і наукових процесах
Широкий загал Політичні діячі Преса	Участь у демократичних процесах

Поняття "виробництво статистичних даних" охоплює цілий цикл життя статистичного обстеження або статистичної інформаційної системи, включаючи проектування, виконання, контроль, підтримку та оцінку. Тому до виробників статистичних даних згідно з [10, 11] віднесено

– проектувальників статистичних обстежень і статистичних інформаційних систем: статистиків предметних галузей, статистиків-методологів, спеціалістів з інформаційних систем;

– постачальників вхідних даних, наприклад, респондентів;

– статистиків виробництва, тобто тих, хто безпосередньо проводить обстеження.

Також, на нашу думку, до цієї групи користувачів слід віднести і керівників

(менеджерів) різного рівня, що працюють у системі статистики і мають безпосереднє відношення до процесів виробництва статистичної інформації.

Кожна з цих категорій виробників статистичних даних має свої типові потреби в метаданих. Наприклад, проектувальник статистичного обстеження повинен знати про потреби користувачів, як подібні обстеження проводились раніше, як вони проводяться у цей час іншими статистичними службами. Постачальник даних для статистичного обстеження зацікавлений в отриманні знань про цілі обстеження та про витрати і вигоди від участі у ньому. Статистику виробництва необхідно мати контрольні списки, реєстри та іншу документацію системи виробництва, щоби, з одного боку, завжди знати як виконати всі виробничі етапи належним чином, а з іншого боку, – навчати, при необхідності, нових співробітників. При оцінці статистичної інформаційної системи необхідні метадані про функціонування системи, включаючи зворотну інформацію від користувачів.

Таким чином, всіх користувачів статистичних метаданих можна розподілити на категорії, як це показано на рис. 1.

Зовнішньому користувачу, що працює із статистичними даними, необхідні метадані трьох типів: описові, орієнтовані на процеси та глобальні метадані. Описові метадані надають інформацію про зміст, сенс, точність, доступність, надійність статистичних даних, тобто про виміри якості даних. Проте, описові метадані не завжди достатні, щоби надати користувачеві адекватне розуміння даних, іноді потрібна інформація про те, як здійснюється виробничий процес отримання, оброблення, зберігання та узагальнення даних. Цю інформацію надають орієнтовані на процеси метадані. Користувач, який шукає статистичні дані, що, можливо, допоможуть у вирішенні його питання або проблеми, потребує глобальні метадані, які охоплюють багато статистичних обстежень, щоби мати можливість ідентифікувати і визначати місцезнаходження можливо релевантних статистичних даних. Прикладами таких глобальних даних можуть бути такі:

– описи – більш або менш формалізовані – всієї доступної статистики та реєстрів спостережень;

– добре структуровані та інформативні каталоги – по можливості, глобальні – специфікована доступна статистика та реєстри спостережень разом із посиланнями на більш детальні описи даних;

– індекси до каталогів;

– словник (довідник) для підтримки процесу визначення запитів, який би забезпечував визначення та трактування запитань як більш широке, так і більш вузьке, та надавав пов'язані з ними терміни.

Під час аналізу статистичних метаданих, отриманих у результаті пошуку, користувачу теж будуть потрібні метадані, що відносяться до категорії загального знання (тобто глобальні метадані), подібно тому, як необхідні керівництва, що описують різні методи статистичного аналізу.

Виробнику статистичних метаданих необхідні такі їх види:

– такі ж самі, як для користувачів (для того, щоби бути в змозі надати їм допомогу);

– детальна інформація про процес виробництва (щоби пам'ятати, як правильно виконувати виробничі стадії, та для навчання нових співробітників);

– зворотна інформація від користувачів та постачальників вхідних даних, щоби покращити процес виробництва даних (цей тип метаданих, природно, повинен бути представлений за категоріями користувачів та видами послуг, що їм надаються. Запити про статистичні дані, що не можуть бути задовільнені у наявному проекті обстеження, можна реєструвати як інформацію для удосконалення обстеження у майбутньому);

– глобальні метадані для проектування / перепроектування процесів виробництва (наприклад, загальні методологічні знання, що містяться в енциклопедіях, експертних системах, керівництвах);

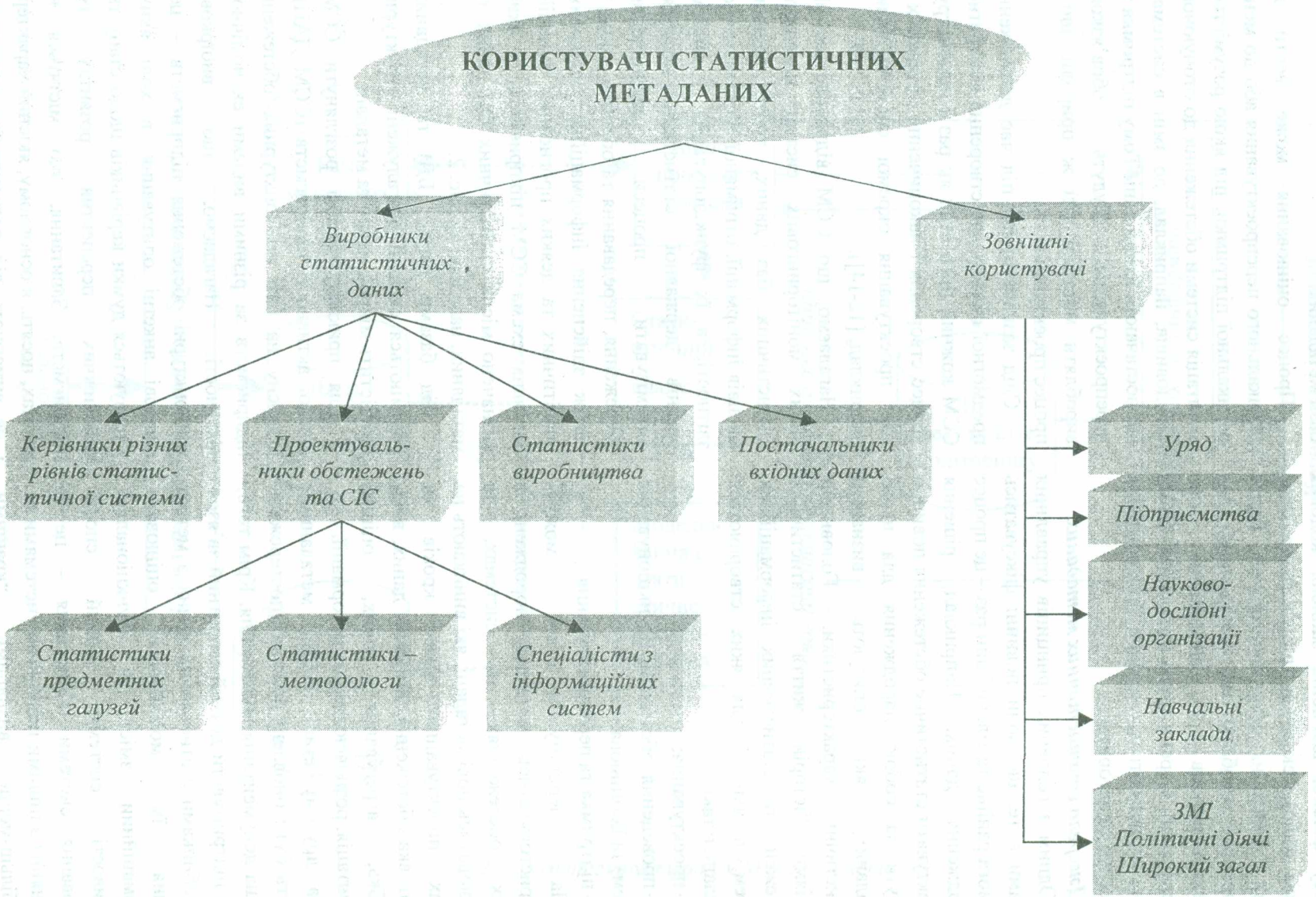

Рис. 1. Категорії користувачів статистичних метаданих

– проектувальникам статистичних інформаційних систем також необхідні метадані щодо форматів представлення даних (опису файлів, шаблонів записів, кодових описів тощо), текстових ярликів (наприклад, назв) і довільних текстових описів. Також важливими глобальними метаданими для проектувальників є інформація про класифікації. Окрема група метаданих, якими вони користуються, – опис статистичних процедур та алгоритмів.

Джерела статистичних метаданих

Одним з головних принципів управління даними є те, що вони повинні фіксуватись якомога раніше та тільки один раз – це процес народження даних. Наприклад, рішення проектувати статистичне обстеження повинно тягнути за собою народження для нього метаданих, які описують визначені статистичні характеристики. Головними стадіями "історій життя" статистичних обстежень та статистичних інформаційних систем, в яких і для яких створюються метадані, є такі:

– проектування;

– проведення обстежень (функціонування систем) та їх оцінювання;

– підтримка та перепроектування.

На першому етапі можуть використовуватись метадані, породжені в інших обстеженнях та системах, та створюватись нові метадані, які виникають на етапах проектування таких кроків як підготовка обстеження, збирання даних, ввід, обробка, агрегування даних, оцінка, презентація, розповсюдження інформації.

На другому етапі новими метаданими можуть бути оновлені реєстри спостережень та інша документація обстеження. Крім того, треба підтримувати деякий зворотний зв'язок з виробниками статистичних даних з метою надання їм можливості оцінювати інформаційний зміст та функціональні можливості системи. Інший спосіб оцінювання системи обстеження – це її порівняння з іншими подібними системами та с більш-менш визнаними "кращими методами". Для полегшення таких порівнянь звичайною задачею для кожного проектування обстеження повинно бути внесення деяких метаданих до глобальної бази метаданих.

Процес оцінювання може вести до радикального перепроектування або до менш радикальної підтримки, під якою розуміється адаптація системи обстеження до оточуючого середовища, наприклад, до змін в системах, що постачають вхідні дані. Тому підтримка та перепроектування будуть обов'язково виробляти метàдані тієї ж природи, що і процес проектування.

Слід зазначити, що під час дослідження предметної області для створення конкретної ССМ кожний його етап як раз і демонструє процес створення і накопичення метаданих у ході проектування окремої системи (див., наприклад, [12-14]).

Нагадаємо, що ССМ відрізняються від інших моніторингових систем наявністю статистичних баз даних, тобто великих масивів інформації, отриманої із статистичної звітності, а їх функціонування у системі органів державної статистики дозволяє оптимізувати процеси накопичення, збереження, передавання та обробки даних, а також здійснення інформаційно-пошукових, аналітичних та деяких прогнозних функцій. Загальна схема ССМ підприємств, у який визначено місце статистичних баз даних та метаданих, наведена на рис. 2.

Під базами даних (БД) та метаданих розуміються складні розгалужені структури, що містить мікро-, макро та метадані.

Для прикладу можна розглянути ССМ ділової активності підприємств (ССМ ДАП), створену на базі кон'юнктурних обстежень підприємств за різними видами економічної діяльності. Нагадаємо, що вибіркові кон'юнктурні обстеження підприємств – це поштові анкетні опитування, в ході яких з'ясовуються думки керівників щодо стану та найближчих перспектив розвитку їх підприємств. Запитання, що містяться в анкетах, носять, в основному, якісний характер і не вимагають від респондентів залучення бухгалтерської звітності та іншої документації.

Рис. 2. Загальна структура ССМ підприємств

Метадані для ССМ ДАП, виходячи з викладеного вище, можна розподілити на категорії з наступним складом:

1) Описові метадані, тобто ті, що надають інформацію про зміст, сенс, статистичних даних та їх характеристик, про формати представлення даних, їх довільні текстові описи тощо. У ССМ ДАП – це така інформація:

– опис анкет обстежень різних видів економічної діяльності;

– опис показників з анкет обстежень;

– формати представлення даних у системі;

– опис основи для вибірок для різних секторів економіки;

– розміри вибірок;

– відсоток повернених анкет;

– періодичність проведення обстежень;

– метод отримання даних від респондентів;

– контактна інформація для користувачів.

2) Орієнтовані на процеси метадані у ССМ ДАП – це опис того, як здійснюються процеси вводу первинних даних, їх первинної обробки, аналізу, узагальнення та зберігання інформації. Такі метадані зберігаються у великих текстових файлах і являють собою структуровану текстову інформацію.

3) Глобальні метадані, тобто загальні методологічні знання, у ССМ ДАП містять

– методи побудови вибіркових сукупностей;

– реєстри спостережень (анкет);

– систему показників з анкет обстежень (каталог показників);

– методи розрахунків балансів, середніх (з урахуванням нулів та без них);

– опис розривів у рядах даних, пов'язаний із зміною методології, наприклад, з переходом на нові класифікації (КВЕД, КФВ та КОПФГ);

– методи розрахунків агрегованих показників;

– довідник для підтримки процесу визначення запитів;

– опис статистичних процедур та алгоритмів (наприклад, для імпутації даних або для рейтингування підприємств).

Висновки та перспективи подальших досліджень:

1. У статті надано визначення статистичних метаданих та їх категорії, що дозволяє розподілити всю множину метаданих, які створюються для статистичних інформаційних систем або систем статистичного моніторингу, на окремі чітко визначені групи.

2. Представлено категорії користувачів статистичних метаданих в залежності від ролі, яку вони відіграють у процесах виробництва статистичних метаданих і користування ними як для виробничих цілей, так і для аналізу та прийняття рішень.

3. Наведено розподілені за категоріями статистичні метадані для системи статистичного моніторингу ділової активності підприємств, які, з подальшим розвитком цієї ССМ, будуть удосконалюватись та доповнюватись.

ЛІТЕРАТУРА:

1. Guidelines for the Modeling of Statistical Data and Metadata. Methodological Material // Conference of European Statisticians. United Nations: Geneva, 1995. – 29 p.

2. United Nations Statistical Commission and Economic Commission for Europe. Conference of European Statisticians. Guidelines for Statistical Metadata on the Internet (Statistical Standards and Studies). – Geneva, 2000. – No. 52. – 49 p. – http://www.unece.org/stats

3. SDMX Content-Oriented Guidelines: Metadata Common Vocabulary. – 164 p. – http://www.sdmx.org/

4. The IMF Metadata Repositories Project: An activity aligned with SDMX standards. Statistics Department of International Monetary Fund. – OECD Expert Group on Statistical Data and Metadata Exchange. April 1–2, 2004. – 13 p. – www. oecd.org

5. Meliskova J., Oakley G. Statistical Metadata System and Its Role in a Statistical Organization. Invited Paper on Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS). Geneva, 3-5 April 2006. – 46 p. – http://europa.eu.int/comm/eurostat

6. Malmborg E., Sundgren B. Integration of statistical information systems – theory and practice // Proceedings of the Seventh International Conference on Scientific and Statistical Database Management, University of Virginia, USA, September 1994. – IEEE Computer Society Press, 1994. – Pp. 78-99.

7. Sundgren B. "What metainformation should accompany statistical macrodata?" Report for the June 1991 Meeting of Working Party 9 of the OECD Industrial Committee as a basis for a discussion on the topic of Standards for Metadata in International Databases. – 63 p.

8. Sundgren, B. Statistical information systems in a modern society: roles, functions, and system designs. Invited paper for the Baltic Workshop on National Infrastructure Databases, Vilnius, Lithuania, May 1994. – 11 p.

9. Пугачова М.В. Забезпечення якості даних статистичного обстеження ділової активності підприємств // Статистика України. – 2005. – №4 (31). – С. 63-71.

10. Appel G. Metadata Driven Statistical Information Systems // Proceedings of the Statistical Metainformation Systems. Workshop in Luxemburg, February 1993. – Luxemburg: Eurostat, 1993. – Pp. 3-18.

11. Sundgren B. Advice Concerning the Strategic Decisions that Have to be Made by a Statistical Office when Developing and Implementing a Metadata Structure. – Stockholm: Statistics Sweden, 2002. – 27 p.

12. Пугачова М.В. Проблеми створення системи статистичного моніторингу ділової активності підприємств / Збірник наукових праць "Проблеми статистики". – Київ. – 2003. – № 5. – С. 118-126.

13. Концептуальні основи статистичного моніторингу. Монографія / Д.Д.Айстраханов, М.В.Пугачова, В.С.Степашко та ін.; за ред. М.В.Пугачової. – К.: ІВЦ Держкомстату України, 2003. – 343 с.

14. Пугачова М.В. Системи статистичного моніторингу в державній статистиці // Економіка: проблеми теорії і практики. Збірник наукових праць. Випуск 216: В 4 т. Том IV. – Дніпропетровськ: ДНУ, 2006. – С. 975-985.

ПУГАЧОВА Марина Володимирівна – кандидат технічних наук, старший науковий співробітник, директор Науково-проектного інституту статистичних технологій НТК статистичних досліджень

Наукові інтереси:

- кон'юнктурні обстеження підприємств за різними видами економічної діяльності;
- моніторингові та інші інформаційно-аналітичні системи