Architecture solutions for the metadata extraction toolkit, taking into account the built-in privacy extracts

Vladimir I. Budzko
Full Member of Russian Cryptography Academy, Dr.Sc.(Eng.)
Deputy Director for Research, Institute of Informatics Problems of Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences,
Professor of National Research Nuclear University «MEPhI»

Vadim I. Korolev
Professor, Dr.Sc.(Eng.)
Leading Research Fellow of Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Professor of Financial University under the Government of the Russian Federation

Dmitry A. Melnikov
Associate Professor, Ph.D.(Eng.)
Leading Research Fellow of Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Associate Professor of National Research Nuclear University «MEPhI»

Victor G. Belenkov
Ph.D.(Eng.)
Leading Research Fellow of Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences

Peter A. Keyer
Senior Research Fellow of Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Moscow, 119333, Vavilov str. 44/2
vbudzko|dmelnikov|pkeyer|vkorolev|vbelenkov@ipiran.ru

1 Built-in privacy extracts as information security metadata

The basic task of ensuring information security in automated data processing systems is to manage the delimitation of access to information related to certain types of secrets and regulated by access by law. Management of access control implies the existence of a legal field for access relations between information objects and user entities. This problem is most difficult to solve for data intensive systems.

Data intensive system - automated information systems that provide data analysis and management, information processing and solving functional research and applied problems in various fields with intensive use of data (Data Intensive Domains - DID). These areas are associated with the prevailing global trend of creating massive data collections and providing the possibility of their joint use in solving research and decision-making tasks. The main sources of information resources for data intensive systems are arrays of collections of unstructured data [1]. And the problem of identification of information objects of limited access remains unresolved [2]. Therefore, it is necessary in the system to provide the ability to retrieve and identify information objects of limited access, to form a legal field. This feature is inherently similar to solving the problem of extracting knowledge from content, and its solution is associated with the idea of using built-in privacy extracts.

In order to solve this problem, the authors formulated in [3, 4, 5] a system of interrelated concepts:

- *An artifact* – a content or a piece of content that is isolated in a certain way that is considered as a whole together with the complex of credentials accompanying it and the container. The container contains content and a set of credentials accompanying it.

- *Protected entity* – specific naming of areas of activity, objects belonging to them, their indicators / characteristics / types of information, mentions of which or specific information are subject to protection in specific areas of activity (hereinafter - the entity).
- *Confidentiality extract (CE)* - the minimum selection (for example, a word or other attributes) from the content area, the result of extracting from the artifact an indicator of the availability of information that requires protection and assignment to a certain level of confidentiality in specific areas of activity. In fact, the data included in the privacy extract related to the protected information is the metadata of the special artifact– its security tokens.
- *Input content of data intensive systems - information from various data sources, may have various forms of presentation.*
- *Output content of data intensive systems* - texts, tables, spreadsheets, charts, graphs, drawings, situational situation on the map, highlighting regions on the map and other forms of displaying information according to the results of the request.

As the input content of data intensive systems can be considered such types as:

- relational databases (single and multidimensional);
- non-relational databases;
- files generated by office software and user applications, including files containing unstructured and structured texts, tables, charts, graphs, drawings, scanned images, photographs, etc .;
- artifacts in web formats, including in the form of sites, pages, etc .;
- files with topographic data, including digital terrain model (DTM) data;
- files with satellite data (optical and radar), for example, weather data, earth surface, etc .;
- artifacts of content management systems (for example, EMC Documentum, EMC Documentum eRoom, FileNet P8, FileNet Document & Image Services, Interwoven TeamSite, Lotus Notes / Domino, Microsoft SharePoint, OpenText LiveLink;
- email artifacts;
- social network artifacts (e.g. FaceBook, LinkedIn, Twitter);
- artifacts of photo hosting and video hosting (for example: Flickr, Yuotube), blogs, etc .;
- artifacts with machine data (for example, in the form of logs of transaction systems, clickstream logs, tickers, trades, traffic information, data from sensors and tools, coming on-line or off-line, including from systems with a large number of transactions
- streaming data: audio and video data, telemetry data, instrument and sensor data, for example, received from climate stations;
- mixed artifacts.

The types of presentation of the output content of data intensive systems can be artifacts: texts, tables, spreadsheets, charts, graphs, drawings, situational situations on the map, highlighting regions on the map and other types, as well as mixed-type content.

The concept of a *container* is introduced. Such objects as a message, file, folder, database, or its structural element, website, portal, web page, etc. can be considered as a container.

Filling CE can be presented in artifacts explicitly, or obtained by methods of extracting knowledge from artifacts. In this case, semantic processing based on various models can be used semantic and conceptual networks, frames, etc.

From the above definition, CE make it possible to determine or form a unique correspondence between artifacts extracted from information resources at the input / output of a DID system or in the process of processing information by a system and their level of confidentiality. In this case, the level of confidentiality refers to the assignment of information included in the artifact to the corresponding type of secrets, defined by law and requiring appropriate decisions to protect this information. Moreover, when establishing compliance in any case, the level of confidentiality must comply with legal requirements. In accordance with the regulatory framework of the Russian Federation, the level of confidentiality is determined either at the state level (in relation to information constituting a state secret) or by the owner of the information in accordance with its policy for regulating information relations, which should not contradict the regulatory framework (in relation to commercial information, scientific information). A separate segment of information resources related to confidential data is personal data.

Confidentiality extracts are attributes of artifacts and can be presented in them either explicitly in the artifact information array, or can be obtained from the artifact context by methods of extracting knowledge about the availability of information constituting a particular secret.

Thus, the procedure for determining the privacy extract in each case is associated with its extraction from the artifact.

2 General architectural construction of the metadata extraction tool complex

The formation of C1 is carried out taking into account three factors:

- the context corresponding to the areas of activity to which the owners and users of information belong;
- the context corresponding to the objects or systems to which the data relates;
- contextual privacy requirements.

The main requirement for the architectural construction of a tool complex for data extraction is that CE processing should be carried out on dedicated instrumental hardware-software complexes (HSC-CE). HSC-CE isolation is necessary because determining the level of access to information in data intensive systems for a particular artifact in the general case depends on the information of this artifact that is supposed to be processed in the system. The level of access to this information can be above the acceptable level of confidentiality for the system. Based on this, HSC-CE should pre-process artifacts coming from data sources, as well as post-process the results of query execution before transmitting these results to the user.

Moreover, each artifact accepted by the system before its substantial processing is transferred to the system after HSC-CE.

Based on the results of processing the artifact, HSC-CE transmits a notification to the system, which may include two types of content. The first type reflects a positive result and contains an anonymized technological identifier of the artifact, as well as the level of tolerance of the artifact and the area of activity to which it belongs. The second type of notification contains an information security incident code (ISIC) associated with a violation of an indicator of an EC clearance level.

Similar procedures are implemented in the process when user requests data. The results of each user request, performed by the DID system before providing them to the user, are transmitted to HSC-CE to determine their level of access or their general availability. Based on the results, notifications of similar content are generated: on the level of admission to the result or the ISIC` incident code.

The place of HSC-CE as a structural component of the system in the technological process of data processing is illustrated by the scheme shown in Figure 1.

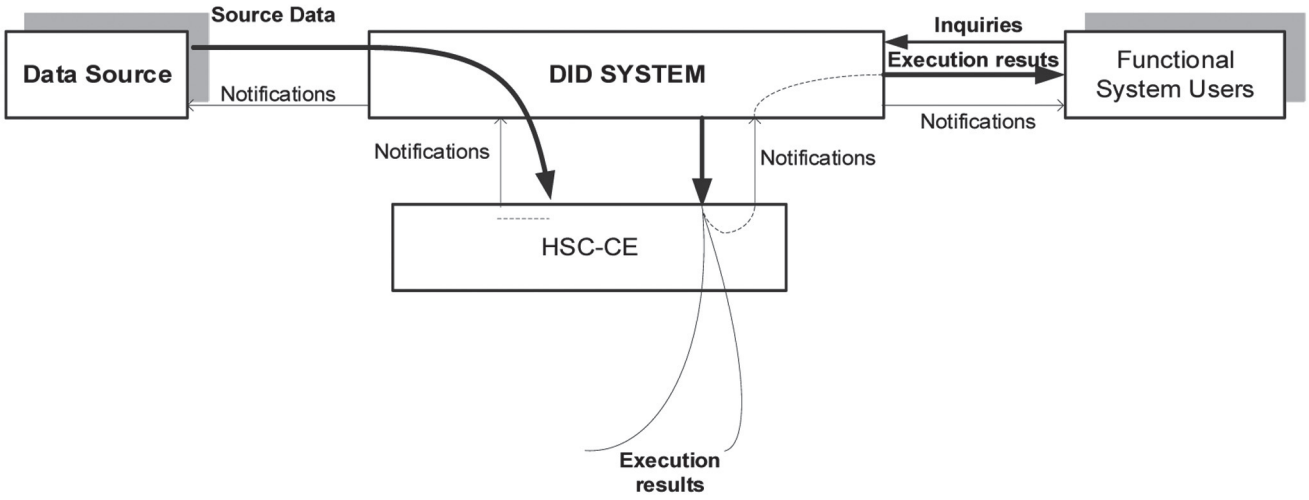

Figure 1. - Schematic diagram of providing information security through HSC-CE in the technological process of processing large amounts of data

3 The formation of confidentiality extracts and the determination of the level of confidentiality in HSC-CE based on text analysis

The sequence of actions for the formation of confidentiality extracts and determining the level of confidentiality based on the analysis of the text of the material includes the following basic procedures:

1) reduction of the Artifact received from one of the data sources or prepared to provide the user with the system in the form of a single material, including heterogeneous fragments (for example, texts, tables, forms, etc.);
2) bringing the Artifact to a single form of presentation of information;
3) bringing the Artifact to a single format for representing data in the system;
4) the definition of the positions of naming Entities in the Artifact or its fragment;
5) reduction of the Artifact or its fragment to a single naming of Entities;
6) replacement of personal pronouns with nouns of Entities;
7) determination of the assignment of naming / mentioning of specific objects to naming / mentioning of specific areas of activity;
8) determination of the assignment of naming / mentioning of specific indicators / characteristics / types of information to naming / mentioning of specific objects or to naming of specific areas of activity;
9) verification of ownership and the formation of a list of Entities, references to which are subject to protection;
10) the formation of a list of Entities, specific information about which is subject to protection;

11) determination of the attribution of texts to naming / mentioning of specific indicators / characteristics / types of information;
12) clarification of confidentiality levels of information on indicators / characteristics / types of information;
13) determination of confidentiality levels of information about objects and fields of activity;
14) clarification of confidentiality levels of information about objects;
15) clarification of confidentiality levels of information on areas of activity;
16) final clarification of confidentiality levels of information on areas of activity;
17) determining the level of confidentiality of the Artifact / its fragment;
18) the formation of privacy token (PT) for Artifact / its fragment.

In essence, the content and sequence of these procedures determine the HSC-CE functioning algorithm for extracting special-purpose metadata in the form of confidentiality extracts. A description of the integrated algorithm for the functioning of the HSC-CE is given in Table 1

Таблица 1.

Stage	Name of procedure	Procedure content
1.	Bringing an Artifact to a single material view	Bringing the Artifact and its attributes, received from the ID or prepared to be provided to users, in the form of a single material, including fragments that are heterogeneous in form (texts, tables, forms, etc.).
2.	Bringing the artifact to a single form of information presentation	Transformation of recognizable artifact information objects in graphic forms, spreadsheets, etc. into a text submission form. Standard tools for recognizing graphic information objects, exporting data from spreadsheets, reformatting, etc. are used. There may be distortions associated with poor recognition of texts in graphical form, with errors in converting spreadsheets to text forms, with reformatting errors, etc.
3.	Bringing the Artifact to a single format for presenting content (data) in the system	Bringing the results of processing the Artifact at the previous stages to a single format for presenting the content of the Artifact, unified in HSC-CE (HTML, PDF, DOC, etc.).
4.	Definition of Entity naming positions in the Artifact or its fragment	Formation of a table of positions of naming Entities in the Artifact. For each Entity, its current and initial naming, position in the source text and in the text formed at the previous stage are indicated. To name the Entity, determined taking into account the distances between words, the positions for each of its words are indicated. For object naming, positions are reserved to indicate the naming of the areas of activity to which these objects belong. For the naming of indicators / characteristics / types of information, the positions of the naming of objects and areas of activity to which they relate are reserved.
5.	Bringing an Artifact or its fragment to a single naming of Entities	Exclusion of ambiguity in naming the same Entity. To avoid ambiguity, the Artifact searches for additional data using information from the Protected Information (PI) Networks and Presentation tables, taking into account the homonymy (context) of naming Entities, as well as their designations, synonyms, abbreviations, codes, epithets, descriptive expressions that uniquely indicate the naming object, their various spellings, spellings with typical mistakes. The unified naming convention is implemented. If the Presentation tables for objects of different fields of activity have the same expressions for the Entity, a naming group is formed with an indication for each naming of the field of activity to which this naming refers. If the Presentation tables for different indicators / characteristics of different objects and areas of activity have the same expressions for the Entity, a naming group is formed with an indication for each naming of the object and / or area of activity to which this naming refers. During transformations, distortions are possible due to errors in machine semantic analysis of texts.
6.	Replacing personal pronouns with nouns of Entities	Exclusion of ambiguity in naming the same Entity. As a result of processing the request and / or in its sections, a search is made for personal pronouns and semantic analysis of the text to determine their relationship to one of the nouns or to a group of words. They are replaced by the corresponding noun or group of words (basic naming of Entities). The output is the result of processing the Artifact in a unified format with personal pronouns replaced with the corresponding nouns or groups of words.

7.	Definition of assignment of naming/mentioning of specific objects to naming/mentioning of specific areas of activity	An exception to the ambiguity in assigning object naming to naming of areas of activity. Using the methods of semantic analysis, the identification of object naming belongs to one of the areas of activity for which information is subject to protection. Moreover, uniform nouns of Entities are used instead of personal pronouns. The table of entities naming entries for object naming indicates the areas of activity to which these objects belong. In object naming groups, one element remains. The input of the implementation of this step receives the results of processing the previous step (table of positions for naming entities) and the results of processing in steps 5 and 6; an updated table of positions for naming entities is formed at the output.
8.	Definition of assigning naming / mentioning of specific indicators / characteristics / types of information to naming / mentioning of specific objects or to naming of specific areas of activity	The elimination of ambiguity in attributing the names of indicators / characteristics / types of information to the names of specific objects and areas of activity. It is determined whether the naming of indicators / characteristics / types of information belongs to the naming of objects and areas of activity for which information is subject to protection. We use the methods of semantic analysis. Uniform nouns of Entities are used instead of personal pronouns. The naming of objects of areas of activity and areas of activity to which these indicators / characteristics / types of information are related is indicated in the table of positions of entity naming for naming indicators / characteristics / types of information. One element remains in the naming groups of indicators / characteristics / types of information. The input of the implementation of this step receives the results of processing the previous step (table of positions for naming entities) and the results of processing in steps 5 and 6; an updated table of positions for naming entities is formed at the output.
9.	Verification of ownership and compilation of a list of Entities, references to which are subject to protection	Formation of CE Entities, any information about which is subject to protection. The presence of naming Entities is checked for which references are to be protected (the corresponding vertex in the PI Network has a sub vertex “DUMMY”) using the information present in the Network of Protected Information (PI Network) and the table of positions of entity naming. Entities level of confidentiality is determined using the information present in the PI Network. The table CE of the Access Object begins to be formed. It contains the naming of Entities taking into account their affiliation (indicators / characteristics / types of information - to objects or areas of activity, objects - to areas of activity) and their level of confidentiality. The CE of the Artifact is formed at the output of this step.
10.	Formation of a list of Entities, specific information about which is subject to protection	Formation of CE indicators / characteristics / types of information, specific information about which is subject to protection. The level of naming entities confidentiality is determined using the information present on the PI Network and their table of positions for for indicators / characteristics / types of information. The formation of the CE Artifact table continues in terms of naming indicators / characteristics / types of information, taking into account their and their level of confidentiality. An updated CE of the Artifact is formed at the exit of the stage.
11.	Definition of assignment of texts to naming / reference to specific indicators / characteristics / types of information	Determination of the presence and exclusion of ambiguity in assigning information to the names of indicators / characteristics / types of information. The text is determined to belong to the names of indicators / characteristics / types of information for which information is subject to protection, using the methods of semantic analysis. The processing results at step 5 are used when working with the unified naming of Entities that replace naming in which the distance between words does not exceed the specified value. When working with the unified naming of Entities that replace naming in which the distance between words does not exceed the specified value, the processing results are used at step 5. As a result of processing the table of entity naming positions for naming indicators / characteristics / types of information, the text positions in the Artifact are indicated that relate to these indicators / characteristics / types of information. The input of the implementation of this step receives the results of the processing of the previous step (table of positions for naming entities) and the results of processing at stages 5 and 6; an updated table of positions for naming entities is formed at the output.

12.	Clarification of confidentiality levels of information on indicators / characteristics / types of information	Determination of confidentiality levels by indicators / characteristics / types of information of a hierarchical nature. The level of confidentiality is determined as the maximum level of confidentiality. We use the information that is available on the PI network and CE Artifact tables for information about each indicator / characteristic / type of information that is hierarchical in nature. The input of the implementation of this step receives the processing results of step 11 (table of positions for naming entities), step 10 (table of CE Artifact), and an updated table of CE Artifact is generated at the output.
13.	Determination of confidentiality levels of information about objects and areas of activity	Determination of confidentiality levels of information about objects and areas of activity, which directly relate to indicators / characteristics / types of information identified at the previous stage. The level of confidentiality is determined as the maximum level of confidentiality of information that relates directly to indicators / characteristics for which specific information is available as a result of processing the request or its section. We use the information that is available on the PI network and CE Artifact tables, including information about each object or field of activity. The input of the implementation of this step receives the table of positions for naming entities and the processing results in step 12 (the table of the CE Artifact, an updated table of the CE Artifact is generated at the output.
14.	Refinement of confidentiality levels of information about objects	Determination of confidentiality levels of information about objects that are hierarchical. The level of confidentiality is determined as the maximum level of confidentiality of information relating directly to it of objects whose confidentiality levels were previously determined. We use the information that is available on the PI Network and CE Artifact tables, including information about each object that is hierarchical in nature. The input of the implementation of this stage receives the table of positions for naming entities and the processing results at step 13 (the table of the CE Artifact), an updated table of the CE Artifact is formed at the output.
15.	Clarification of confidentiality levels of information about areas of activity	Clarification of confidentiality levels of information about areas of activity to which objects directly relate, the confidentiality levels of information about which were determined at the previous stage. At this stage, the level of confidentiality is defined as the maximum level of confidentiality of information directly related to the objects, and taking into account the level of confidentiality of information about this area of activity, which was formed in step 13. We use the information that is available on the PI Network and CE Artifact tables, including information about each area of activity. The input of this step receives the table of positions for naming entities and the processing results at step 14 (the table of the CE Artifact), an updated table of the CE Artifact is generated at the output.
16.	Final clarification of confidentiality levels of information about areas of activity	Determination of confidentiality levels of information about areas of activity that are hierarchical in nature At this stage, the level of confidentiality of information is defined as the maximum of the levels of confidentiality of information directly related to areas of activity, the levels of confidentiality of which were previously determined. We use the information that is present on the PI Network and the tables of the CE Artifact, including information about each area of activity that is hierarchical in nature. The table of positions for naming entities and the results of processing at step 15 (the CE table of the Artifact) are input to the implementation of this step, the updated table of the CE Artifact is generated at the output.
17.	Determining the level of confidentiality of an Artifact / its fragment	Determination of confidentiality levels of information contained in the Artifact or its fragment. The level of confidentiality is defined as the maximum level of confidentiality of information about areas of activity for a given Artifact or fragment. The tables of the CE Artifact are used for the information contained in the Artifact or its fragment, standard tools for working with lists (search, retrieval, etc.), as well as tools for determining the largest value of a value that takes values from an ordered set. The output forms the level of confidentiality of the information contained in the Artifact or in its fragment.

18.	Formation of CE Artifact / its fragment	The formation of the CE of a particular Artifact or its fragment. For the formation of CE tables are used CE Artifact. CE are a list of areas of activity, objects of areas of activity and their indicators / characteristics / types of information. The formation takes into account the belonging of objects to fields of activity, as well as the belonging of indicators / characteristics / types of information to objects and fields of activity. The hierarchical nesting of Entities is taken into account, indicating for each Entity the level of confidentiality of the information related to it and pointers to places in the text of the Artifact or its fragment. A table of positions for naming entities, the results of processing in step 16 (the table of the CE Artifact), steps 3 and 6 (the result of processing the Artifact) are input to this stage, the list of the CE of the Artifact is formed at the output. We use standard tools for working with lists (search, extraction, formation, etc.), as well as tools for searching and extracting fragments from text, determining the boundaries of sentences.

Literature

[1] Budzko V.I., Kalinichenko L.A., Stupnikov S.A., Vovchenko A.E., Bryukhov D.O., Kovalev D. Yu. Integration medium of large heterogeneous data collections. High Availability Systems. t. 10, No. 3, –M.: Radiotechnics, 2014. – P.3-19

[2] Miloslavskaya, N., Nikiforov, A., Budzko, V. Standardization of ensuring information security for big data technologies // Proceedings - 2018 IEEE 6th International Conference on Future Internet of Things and Cloud Workshops, W-FiCloud 2018, 8488175, c. 56-63, INSPEC Accession Number: 18150263, DOI: 10.1109/W-FiCloud.2018.00015, https://www.scopus.com/authid/detail.uri?authorId=56879039000

[3] Belenkov V.G., Borokhov S.V., Budzko V.I., Keyer P.A., Korolev V.I. Issues of ensuring information security of information systems that implement intensive use of data // Analytics and data management in areas with intensive use of data. Collection of scientific papers of the XIX International Conference DAMDID / RCDL'2017 (October 10-13, 2017, Moscow, Moscow State University, Russia) / Edited by L.A. Kalinichenko, J. Manolopoulos, N.A. Skvortsova, V.A. Sukhomlina. - Moscow: FIC IU RAS, 2017.S. 155-158. ISBN 978-5-519-60516-8 http://damdid2017.frccsc.ru/files/DAMDID_RCDL_2017_Proceedings.pdf

[4] Budzko V.I., Belenkov V.G., Korolev V.I. About one conceptual approach to information security in systems that implement DID // Information Technologies and Mathematical Modeling of Systems 2018. Proceedings of the international scientific and technical conference. - M.: Center for Information Technologies in Design of the Russian Academy of Sciences – FRC CSC RA, 2018. - P. 43 - 46. , ISBN: 978-5-6041390-8-0, https://elibrary.ru/item.asp?id=36725038

[5] Budzko V.I., Belenkov V.G., Korolev V.I. On the peculiarities of using tools and methods of information security in systems that implement DID // Information Technologies and Mathematical Modeling of Systems 2018. Proceedings of the international scientific and technical conference. - M.: Center for Information Technologies in Design of the Russian Academy of Sciences – FRC CSC RAS, 2018. - P. 47 - 51. ISBN: 978-5-6041390-8-0, https://elibrary.ru/item.asp?id=36725040