УДК 004.9
МРНТИ 20.51.19

DATA INTEGRATION METHODS IN INFORMATION SYSTEMS

SADIRMEKOVA ZH.[1], TUSSUPOV J.[1], SAMBETBAYEVA M.[1], ALTYNBEKOVA ZH.[2]

[1]Eurasian National University named after L.N. Gumilev
[2]Taraz innovation and Humanities University

Annotation: *In accordance with the purpose of the study, the first task is to analyze the existing solutions in the field of data integration into a single information space.it is also necessary to note the features, advantages and disadvantages of the integration system architectures used. This will allow us to identify problems related to the structural and semantic heterogeneity of data in information systems, as well as formulate an approach to combining information systems. Currently, world leaders in software development are presenting their ready-made solutions for data integration. Their consideration will complement the study of theoretical approaches and will identify the areas that should be developed and used in the construction of integrated systems. To solve the problem of semantic heterogeneity of information in the integration of information systems, it is proposed to use ontologies of the subject area. The problem of integration of information systems data in this paper is considered as a problem of integration of information from heterogeneous information systems and resources (repositories) in order to ensure a unified representation of data. In order to correctly integrate heterogeneous information systems, it is necessary to find out the commonality and differences of the ontologies underlying them, as well as to agree on heterogeneous ontological specifications and further implement information transformation. As a result, heterogeneous information systems work together in the context of the subject area of the problem at a semantically significant level. Analysis of the state of research on the harmonization of ontologies shows that the existing methods are mostly informal and based on the subjective opinion of a human expert. So far, this has not been studied in sufficient depth, mainly for special cases.*

Key words: *information systems, repository, data integration, ontology*

МЕТОДЫ ИНТЕГРАЦИИ ДАННЫХ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Аннотация: *В соответствии с целью исследования первоочередной задачей является аналитический обзор существующих решений в области интеграции данных в единое информационное пространство, также необходимо отметить особенности, преимущества и недостатки применяемых архитектур интегрирующих систем. Это позволит выделить проблемы, связанные со структурной и семантической неоднородностью данных информационных систем, а также сформулировать подход к объединению информационных систем. В настоящее время мировые лидеры по разработке программного обеспечения представляют свои готовые решения по интеграции данных. Их рассмотрение дополнит исследование теоретических подходов и позволит выявить те направления, которые следует развивать и использовать при построении интегрируемых систем. Для решения проблемы семантической неоднородности информации при интеграции информационных систем предлагается использовать онтологии предметной области. Задача интеграции данных информационных систем в настоящей работе рассматривается как задача интеграции информации из разнородных информационных систем и ресурсов (репозиториев) с целью обеспечения единого представления данных. С целью корректной интеграции неоднородных информационных систем необходимо выяснить общность и различия*

онтологий, лежащих в их основе, а также согласовать неоднородные онтологические спецификации и далее осуществлять преобразование информации. Как следствие, обеспечивается совместная работа неоднородных информационных систем в контексте предметной области задачи на семантически значимом уровне. Анализ состояния исследования по согласованию онтологий показывает, что существующие методы, в основном, неформальны и основываются на субъективном мнении человека эксперта. До сих пор это исследовано недостаточно глубоко, преимущественно для частных случаев.

Ключевые слова: *информационные системы, репозиторий, интеграция данных, онтология*

АҚПАРАТТЫҚ ЖҮЙЕЛЕРДЕ ДЕРЕКТЕРДІ БІРІКТІРУ ӘДІСТЕРІ

Аңдатпа: *Зерттеу мақсатына сәйкес бірінші кезектегі міндет деректерді бірыңғай ақпараттық кеңістікке біріктіру саласындағы қолданыстағы шешімдерді талдамалы шолу болып табылады, яғни интеграциялайтын жүйелердің қолданылатын архитектураларының ерекшеліктерін, артықшылықтары мен кемшіліктерін атап өту қажет. Бұл Ақпараттық жүйелер деректерінің құрылымдық және семантикалық біртекті еместігімен байланысты проблемаларды бөліп көрсетуге, сондай-ақ ақпараттық жүйелерді біріктіруге көзқарасты қалыптастыруға мүмкіндік береді. Қазіргі уақытта бағдарламалық қамтамасыз етуді әзірлеу бойынша әлемдік көшбасшылар деректерді интеграциялау бойынша өздерінің дайын шешімдерін ұсынады. Оларды қарау теориялық тәсілдерді зерттеуді толықтырады және интеграцияланатын жүйелерді құру кезінде дамыту және пайдалану қажет бағыттарды анықтайды. Ақпараттық жүйелерді интеграциялау кезінде ақпараттың семантикалық біркелкі емес проблемасын шешу үшін пәндік саланың онтологиясын қолдану ұсынылады. Осы жұмыста ақпараттық жүйелердің деректерін интеграциялау міндеті деректерді бірыңғай ұсынуды қамтамасыз ету мақсатында әр текті Ақпараттық жүйелер мен ресурстардан (репозиторийлерден) ақпаратты интеграциялау міндеті ретінде қарастырылады. Біртекті емес Ақпараттық жүйелерді дұрыс интеграциялау мақсатында олардың негізінде жатқан онтологиялардың ортақтығы мен айырмашылықтарын анықтау, сонымен қатар біртекті емес онтологиялық ерекшеліктерді келісу және одан әрі ақпаратты түрлендіруді жүзеге асыру қажет. Нәтижесінде семантикалық мәнді деңгейде тапсырманың пәндік саласы контекстінде біртекті емес ақпараттық жүйелердің бірлескен жұмысы қамтамасыз етіледі. Онтологиялардың келісімі бойынша зерттеудің жай-күйін талдау қолданыстағы әдістер негізінен бейресми және сарапшының адамның субъективті пікіріне негізделгенін көрсетеді. Жеке жағдайлар үшін бұл әлі күнге дейін тереңірек зерттелмегенін байқадық.*

Түйінді сөздер: *ақпараттық жүйелер, репозитория, интеграциялық деректер, онтология*

Introduction

Data integration in information systems is understood as providing a single unified interface for accessing a certain set of, generally speaking, heterogeneous independent data sources [1, 2]. thus, for the user, the information resources of the entire set of integrated sources are represented by: as a new single source. The system that provides the user with these features is called a data integration system.

The data integration system frees users from the need to know which data sources they use other than the integrated one, what the properties of these sources are, and how to access them. Integrated data sources can be traditional database systems that support various data models (relational, object, object-relational, graph, etc.), a variety of legacy systems, repositories, websites, and structured data files. Providing access to data from multiple sources through a single interface means that it actually supports the representation of a set of data from multiple independent sources in terms of a single data model. Finally, it is important to note that the composition of multiple sources may be pre-

defined or dynamically updated, and data sources may have unchanged or updated content.

The development of methods for integrating information resources is one of the most pressing problems in the field of information systems (is). It has attracted particular attention in recent years. However, the problem of data integration is not new. The first steps in this area date back to the mid-70s, when the development of distributed database systems began and when, thanks largely to the ANSI/X3/SPARC report [3], a clearer understanding of the multi-level architecture of database systems, data models as a tool for modeling reality, and the display of data models was formed. This was mainly about supporting the global schema for a set of local databases operating in different nodes of the network under DBMS management, which support the same or, in General, different data models. Later, a somewhat more General form of this task was associated with the creation of multibases and federated databases, data warehouses, various repositories of information resources, as well as web applications. In recent years, the development of electronic libraries (Digital Libraries) has been widely developed in many countries) [4-7] problems of integration of heterogeneous data have become a key role, and there is also a problem of integration of text information resources from various independent sources.

The multidimensional nature of the problem

The problem of data integration is extremely multidimensional and diverse. The complexity and nature of the methods used to solve it significantly depend on the level of integration that needs to be provided, the properties of individual data sources and the entire set of sources as a whole, and the required integration methods.

Data integration systems can integrate data at the physical, logical, and semantic levels. Integrating data at the physical level is theoretically the simplest task and is reduced to converting data from various sources into the required unified format for their physical representation. Data integration at the logical level allows access to data contained in different sources in terms of a single global schema that describes their joint representation, taking into account the structural and possibly behavioral (when using object models) properties of the data. Semantic properties of data are not taken into account. Data integration at the semantic level provides support for a unified representation of data taking into account their semantic properties in the context of a single domain ontology. Data sources can have various properties that are essential for choosing data integration methods – they can support the representation of data in terms of a particular data model, they can be static or dynamic, and so on. The set of integrated data sources can be homogeneous or heterogeneous with respect to the characteristics corresponding to the integration level used. As for how to integrate data, there are two possible approaches – virtual or actual (materialized) representation of integrated data. The first approach creates an access mechanism that generates data in the required view directly from data sources when processing a user request. The full materialized representation of integrated data in terms of a single user interface is not supported. The virtual approach is most often used when using frequently updated data sources. On the contrary, the second approach creates a complete materialized representation of integrated data at the integration stage, which is alienated from the original sources and coexists with them. This is the data representation that is used for processing user requests. This approach is used, in particular, in data warehouses.

Basic tools

The main tools used to ensure the integration of information resources include data converters that integrate data models, data model display mechanisms, object adapters (Wrappers), intermediaries (Mediators), ontological specifications, tools for integrating schemas and integrating ontological specifications [8, 9], as well as an architecture that provides interaction of tools used in a specific resource integration system. An example of a developed infrastructure

that provides semantic integration of data using a set of these tools is discussed in [10].

Semantic data integration tools

The most common approach to semantic data integration is based on the use of semantic intermediaries [11]. Intermediaries support unified meta descriptions of integrated data sources. As a rule, semantic intermediaries are developed for a specific narrow subject area. Intermediary mechanisms are based on ontological specifications of sources. An integrated ontology of used sources is created for the intermediary. Such systems also require an integrating data model with advanced data semantics modeling capabilities. Research methods of semantic data integration are devoted, in particular, to the work [5, 7, 12, 13, 14]. In recent years, there have been a number of publications devoted to solving the problem of semantic integration of data from multiple sources, in which it is proposed to use the descriptive logic apparatus, embodied in the OWL ontology description language, to represent the global schema in the data integration system. Many works are devoted to this topic, in particular [15-16].

To ensure semantic integration, it is necessary to analyze not only the data and its structure, but also the analysis of semantic information about the subject area. Semantic integration means that it is possible to establish a correspondence between the meanings of information system units.

Unfortunately, at the moment, there are no clearly defined methods for solving the problem of semantic integration, but there are some projects that try to implement this approach in one way or another.

Examples of some of these projects:

• eCulture – a search engine that allows you to simultaneously search in several collections of cultural heritage institutions. Working by the transference of the collections in the metadata, linking collection objects as instances of classes using publicly available dictionaries, thereby creating a large graph;

• IPISAR – the project explores the dissemination, study and rational use of cultural heritage. The project offers a number of ideas that may simplify the integration of information. The project has developed an application "Pescador", which will store cataloged data in stable repositories;

• SWHi – an ontology developed for an electronic library from the point of view when the main data sources in the repository are described by metadata. This metadata is displayed and stored in an ontology that is based on the schema ontology. To enrich the ontology, they also add new, related information from selected web documents.

Using the ontological approach as a basis for data integration

In the works [17- 26] ontologies were used to solve the problem of data integration.

As noted in [27-28], the ontology – based approach is used in various problem areas, from knowledge representation to information integration. Ontology is used in knowledge management systems to formally describe the simulated part of the world in the form of a dictionary of terms shared by specialists in the selected subject area. Based on this common dictionary, various sources of knowledge can be integrated. Thus, using a common dictionary, it is possible to understand and compare different information systems [23].

The ontology can be used at the stages of development and operation of an information system.

Ontology is defined both as a specification for the conceptualization of a subject area and as a means of representing the semantics of information units [29, 30]. As noted in [31], " domain ontology can be used to describe information objects, their properties, and relationships." Information about these objects can be stored in different sources, and to see the full picture, it is necessary to build an information model [32-33].

One of the advantages of using ontology for IP integration is the availability of software that supports ontological analysis. There are a number of tools (Ontolingua, OntoEdit, OilEd, WebOnto, ODE) that support editing, documenting, visualizing, importing and exporting ontologies,

as well as combining and comparing them [34-38]. Such tools are used for both designing and analyzing the ontology, performing typical operations.

Data integration is an information support system for scientific and educational activities.

In our case, the ontology of the subject area is used as a conceptual scheme of the information system for supporting scientific and educational activities (ISNOD). The advantage of this approach is not only that the user interface is based on a high-level semantic data model, but also the possibility of reasoning in terms of ontology, which serves as a conceptual model.

The information system for supporting research and educational activities stores information about employees and their publications, conferences and projects that were attended by researchers, as well as information about organizations associated with specific scientific projects, various types of scientific publications, and so on.

Information objects describe the main classes of entities scientific information space, such as Organization, Person, Scientific activity, Publication, Research activities, Training course, division of science, Expertise, Geographic location, conference proceedings, etc., as well as the relationships between them.

A repository of digital objects was created for the rapid dissemination of research results and providing access to them. Which was implemented based on the well-known DSpace system using the PostgreSQL DBMS and the Apache Tomcat server. Repository data is structured in hierarchical sections and collections. To date, 4 sections have been created ("Publication", "Collection of conference materials", "Working programs of training courses", "Competence") and more than 400 collections. To collect information, repositories use the OAI-PMH metadata exchange Protocol, the Dublin Core metadata representation format, and the XML file format. For processing metadata from XML format, there is an import parser to external information systems (Fig. 1). In our case, the domain ontology was used to solve the problem of data integration. An important advantage of this approach is the ability to use existing top-level ontologies to describe extracted metadata.

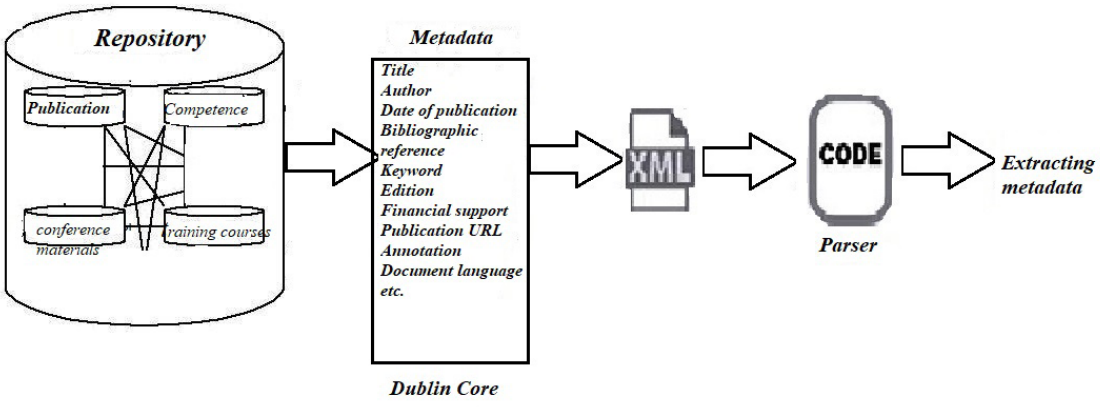

Fig.1 – The process of extracting metadata

One of the promising methods for solving the problem of semantic integration is the method based on the use of metadata [39]. Metadata is a secondary information resource. And despite the fact that the integration of resources within information systems is possible, as a rule, only at the metadata level, the main interest is undoubtedly the primary resource. Therefore, any schemas must contain links to the primary resource. If there is such a link in the metadata record, extracting the primary information becomes a technical task for the organization of user interfaces.

Thus, the scenario for accessing primary information in the ISNOD looks like this:

1. The user makes metadata search.

2. the User views the found entries and selects the ones that are needed.

3. In the selected records, the user clicks a link to view the primary source

Conclusion

A study of data integration problems has been conducted, and it is shown that finding new solutions to these problems is an urgent task. The analysis of modern tools and technologies for ensuring structural and semantic interoperability is carried out. A method for ensuring semantic interoperability based on the use of metadata and the ontology of the subject area is considered. It is concluded that to create a new integrating system within the framework of this dissertation work, it is necessary to use this approach based on the ontology of the subject area. This will support a unified representation of data, taking into account their semantic properties. Integration of data based on ontologies, adequately describes their semantic features. The on-demand integration approach using ontologies solves the problem of information integration, it lacks some of the disadvantages that are inherent in other technical methods, and provides the ability to develop applications that work with information at the semantic level.

REFERENCES

1. Levy A.Y. Logic-Based Techniques in Data Integration. Logic-based Techniques in Data Integration. In: Logic Based Artificial Intelligence. Edited by J. Minker. Kluwer Publishers, 2000.
2. Manolescu I., Florescu D., Kossman D. Answering XML Queries over Heterogeneous Data Sources. Proc. Of the 27th VLDB Conference, Roma, Italy, 2001.
3. ANSI/X3/SPARC Study Group on Data Base Management Systems Interim Report. FDT Bulletin, 7 (2), 1975, pp. 1-140.
4. Bezduchny A.N., Zhizhchenko A.B., Kulagin M.V., Serebryakov V.A. An integrated system of information resources of the Russian Academy of Sciences and technology for developing digital libraries. Programming, 2000, 4, p. 3-14.
5. Kalinichenko L.A., Kolchanov N.A., Podkolodny N.L. Problems of creating a subject intermediary for the integration of molecular genetic information resources. The Second All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Electronic Collections", Protvino, 2000, p. 174-184.
6. Digital Libraries Initiative. NSF. http://www.dli2.nsf.gov/.
7. Kalinichenko L.A., Briukhov D.O., Skvortsov N.A., Zakharov V.N. Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections. The Second All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Electronic Collections", Protvino, 2000, p. 78-90.
8. [8] Briukhov D.O., Shumilov S.S. Ontology Specification and Integration Facilities in a Semantic Interoperation Framework. Proc. of the Second Intern. Workshop ADBIS'95. Moscow, 1995. - P. 195-200.
9. Jean-Christophe R. Pazzagli, Suzanne M. Embury. Bottom-up Integration of Ontologies in Database Context. KRDB, 1998, 7.1-7.7.
10. Zh.B. Sadirmekova, O.L.Zhizhimov, D.A. Tussupov, M.A. Sambetbayeva//Requirements for information system to support scientific and educational activities// CEUR Workshop Proceedings (DICR-2019), //Novosibirsk, Russia, 2019. 44-47pp.
11. Wiederhold G. Mediators in the Architecture of Future Information Systems. IEEE Computer 25:3, pp. 38-49, 1992.

12. Kalinichenko L.A. SYNTHESIS: the language for description, design and programming of the heterogeneous interoperable information resource environment. Institute for Problems of Informatics, Russian Academy of Sciences. Moscow, 1993.
13. Wen-Syan Li, Chris Clifton. Semantic Integration in Heterogeneous Databases Using Neural Networks. Int. Conf. on VLDB, 1994, pp. 1-12.
14. Wiederhold G. Mediators in the Architecture of Future Information Systems. IEEE Computer 25:3, pp. 38-49, 1992.
15. Calvanese D., Giacomo G., Lembo D., Lenzerini M., Rosati R. Conceptual Modeling for Data Integration. http://www.inf.unibz.it/~calvanese/papers/calv-etal-book-mylopoulos-2009.pdf
16. Calvanese D., Giacomo G., Lembo D., Lenzerini M., Rosati R., Ruzzi M. Using OWL in Data Integration. Chapter 14. http://www.dis.uniroma1.it/~rosati/publications/Calvanese-etalSWIMBook-09.pdf
17. Adzhiev, A.S. Integration and loading of structured data in IS based on the ISIR platform / A.S. Adzhiev // Information support of science. New technologies: collection of scientific papers. - M., 2005. - p. 199−224.
18. Batrin, V.K. Development of a conceptual scheme (ontology) to ensure a unified semantics in an open system for integrating heterogeneous data / V.K. Batrovrin, M.R. Kogalovsky, A.S. Korolev, A.B. Petrov // Telematics'2006: materials of the All-Russian Scientific and Methodological Conference. St. Petersburg, June 2006 - St. Petersburg: because of St. Petersburg State University ITMO, 2006. - p. 90−91.
19. Heartless, A.N. The place of ontologies in a single integrated system of the RAS / A.N. Soulless, E.A. Gavrilova, V.A. Serebryakov, A.V. Shkotin // Modern technologies in the information support of science. URL: http://www.benran.ru/Magazin/cgi-bin/Sb_03/pr03.exe?!15 (accessed: 08/20/2011).
20. Volkova T.V. Improving the processes of information formation for university management on the basis of an integrated automated system: dissertation can. those. Sciences, Orenburg State University. Orenburg, .2008.
21. Gavrilova, T.A. Ontological approach to knowledge management in the development of corporate information systems / T.A. Gavrilova // News of artificial intelligence. - 2003. - No. 1 (55). - p. 24−30. [22] Gavrilova, T.A. Knowledge Base of Intelligent Systems / T.A. Gavrilova, V.F. Khoroshevsky //.− SPb .: Peter, 2002 .- p. 382
22. [23] Gladun, A.Ya. Ontologies in corporate systems / A.Ya. Gladun, Yu.V. Rogushina // Corporate systems. - 2006. - No. 1. - Part II.
23. Gribova, V.V. Automation of design, implementation and maintenance of the user interface based on the ontological approach: author. dis. Cand. those. Sciences / Gribova V.V. - Vladivostok, 2007.
24. Kleschev, A.S. Classification of properties of ontologies. Ontologies and their classification / A.S. Kleschev, E.A. Shalfeeva // Preprint. - Vladivostok: IAPU FEB RAS, 2005. - 19 p.
25. Lomov, P. A. Integration of semantically related information resources based on ontologies for effective information support of rational nature management / P. A. Lomov, M. G. Shishaev // Deep processing of mineral resources: a collection of materials from the IV school of young scientists and specialists "Balanced nature management." - Apatity: Publishing House of KSC RAS, 2008. - p. 243−247.
26. Guarino N. Formal ontology, conceptual analysis and knowledge presentation//International Journal of Human and Computer Studies, 43(5/6), pp. 625−640.
27. Guarino N. Formal Ontology in Information Systems, Proceedings of FOIS'98, Trento, Italy, 6−8 June 1998. Amsterdam, IOS Press, pp. 3−15.

28. Lande, D.V. Fundamentals of the integration of information flows: monograph / D.V. Lande //. − Kiev: Engineering, 2006 .- 240 p.
29. Gruber T.R. A Translation Approach to portable ontology specification//Knowledge Systems 92−7, Laboratory, Stanford University, Technical Report KSL.−1993.
30. Mikhailov I.S. Mathematical and software structural and semantic interoperability of information -odonic systems based on metamodels: dis. can those. sciences. MPEI, Moscow, 2008.
31. Travis, B. XML and SOAP: Programming for BizTalk Servers. The latest technology: transl. From English. / B. Travis. - M .: Publishing and trading house "Russian Edition", 2001. - 496 p.
32. Naykhanova, L.V. The use of fuzzy regulation methods in the combination of ontologies of the subject area [Text] / L.V. Naihanova // Software Products and Systems: Int. journal - Tver: NII TsPS. - 2008. - No. 2. - p. 41−44.
33. Nikonenko, A.A. Review of ontological knowledge bases / A.A. Nikonenko // Artificial Intelligence. - 2002. - No. 4. - p. 157–163.
34. Ontolingua [Электронныйресурс] URL: http://www.ksl.stanford.edu/software/ontolingua/ (date of the application: 14.05.2011).
35. OntoEdit: Collaborative ontology development for the Semantic Web. Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, D. Wenke // In Proc. of the Inter. Semantic Web Conference (ISWC 2002), Sardinia, Italia, June 2002.
36. Bechhofer, S. OilEd: A Reason-able Ontology Editor for the Semantic Web / S. Bechhofer, I. Horrocks, C. Goble, R. Stevens // Joint German/Austrian conf. on Artificial Intelligence. Lecture Notes in Artificial Intelligence LNAI 2174, Springer-Verlag, Berlin, 2001. − P.396−408.
37. WebOnto [Electronic resource] URL: http://webonto.open.ac.uk (accessed: 05/14/2011).
38. ODE, WebODE [Electronic resource] URL: delicias.dia.fi.upm.es/webODE/ (date of access: 05/14/2011).