ARPN Journal of Systems and Software

http://www.scientific-journals.org

Information Integration Redefined

[1]C.Punitha Devi, [2]V. Prasanna Venkatesan, [3]G. Shanmugasundaram
[1]Department of Computer Science and Engineering, [2,3]Department of Banking Technology, Pondicherry University
[1]Punitha_c@yahoo.com, [2]Prasanna_v@yahoo.com, [3]sundar_gss2004@yahoo.co.in

ABSTRACT

A great challenge across business today is to achieve visibility into all the information produced within and outside the organization, fitting it to the context depending on the current scenario and for initiating effective action with the latest information. Information integration is one technique that helps in achieving this. Each organizations need for accessing information, list of input sources and in which form the outcome to be are all details contributing to the issues. Hence a need for a precise representation as to what information integration state is obvious. This paper gives the survey on various definitions stated for information integration which has been classified and discussed over the completeness of them analyzing on the three aspects input, output and process. Here the incompleteness of the stated definitions is explained in terms of where they limit. With these findings we present a comprehensive characterization of information integration with an effort to fit in all aspects i.e. meaning, distinctness and completeness of information integration leading to further research.

Keywords: *Information Integration, Information Integration definitions, Information Integration categorization, Information Integration characterization.*

1. INTRODUCTION

Availability of information in present day is enormous and there exists no suitable methodology or technology to acquire the needed information, increasing the risk of information overload, directing to improper usage of information and poor decision–making. This huge information is distributed across enterprises, organizations and geographical conditions. Enterprise IT problems are tied to the "islands of information" caused by many legacy architectures distributed across geographies, business units and multiple subsidiaries. Getting access to all these information sources to obtain insight and better decisions is a challenge for all enterprise or organizations.

2. INFORMATION INTEGRATION

Information integration has been conceptualized in many ways through different persons and on different contexts [3][4][5][7][9]. On studying and analyzing the definitions, it is implied that information integration revolves on three elements i.e. input, process and output. Each definition's perspectives differ from one another and lead to stressing upon the elements. Not all definitions indicate in complete the three elements. These definitions are listed here and they have been classified into four types based on where the significance of the definition is noted as to its constituents input parameters or output parameters or the process or goal centric. Information integration as per the study reveals that explanations or statements could be focused on several aspects but more emphasize is thrown to input sources. How could it be said that focus is towards any one constituents input or output or process?

For eg,

Information integration systems provide facilities that support access to heterogeneous information sources in a way that isolates users from differences in the formats, locations and facilities of those sources. [1]

The input, process and output aspects of the definition are

Input: Heterogeneous information sources with difference in formats, locations and facilities

Process: Support access to

Output: Isolate users

It is seen that the input sources are elaborated and given a wide understanding while compared to the other aspects. The other part of process and output are given a thin lining. The 'support' that's been talked about the process does not actually convey the real or exact meaning, the same way the output 'isolate users'. Both seem to be ambiguous as to what sort of support, what is the methodology behind this word support, targeted towards which user, to what extent this isolation has to be considered. By analyzing the definition explanation or a broad scope is placed on the input parameters

Likewise each definition has been keenly analyzed and categorized as below.

2.1 Input Centric Definitions

Definitions that are biased towards the input sources of information integration with a mere mentioning or an abstract notation about process or output are stated here.

- According to [2] the input sources have been stressed upon by specifying it as information in despite of its organization, structure, naming convention, encode, semantic and so on, whereas output is stated in simple as a single interface and process has not been specified at all except for the word information integration.
- As per [13] the concentration is towards input sources, which has been stated as distributed heterogeneous information and information

processing resources, while the process part is mentioned as assembling, and the output being a coherent view.

- Here as per [9] the input sources has been emphasized by regarding it as various applications or systems like data management systems, content management systems, data warehouses, and other enterprise applications. Whereas the process has been given as combine and the output as a common platform, where both aspects have been stated with a thin lining.

2.2 Process Centric Definitions

As per the explanation stated in the previous subsection, here the definitions which focus on the process of information integration are listed. Each of the definition is analyzed below.

- Here [5] definition emphasize on the process involved for integration by specifying it as integrating data from multiple sources without loading them into data warehouse by giving abstract information about input and output aspects as data from multiple sources and providing tools respectively.

2.3 Output Centric Definitions

There exists some definitions that emphasize on output aspects but those definitions also stress on process aspect or input aspect of information integration, hence such definitions have been categorized under hybrid type.

2.4 Hybrid Definitions

Some definitions give broader description of one or more aspects of the input, output or process, such are explained below.

- As per [10], the process is elaborated here by stating it as bringing together physically or logically data with the help of data warehousing tools. Another aspect that is elaborated here is the output, by indicating that the applications or users are made to use all data either directly or indirectly.
- [12] talks about information integration by specifying details on all aspects through stating the input as diverse forms of business information, the integration process as a set of process namely coherent search, access, replication, transformation, and analysis, and the output aspect to provide real time read and write access, enabling data transformation, data interchange, data placement for quality aspects of performance, currency, availability to meet business needs.
- According to [1] the emphasize is given on the input by explaining it as data from existing heterogeneous, distributed sources, and also on the output by giving it as illusion of single database or system. The process aspect has not been considered mush but only to be mentioned as integration facilities.
- Information from disparate heterogeneous sources often with no appropriate common schema needs to be synthesized in a flexible, transparent and intelligent way in order to respond to the demands of a query thus enabling a more informed decision by the user or application program.[11]

2.5 Generic Definitions

This categorization specifies the definitions that seem to be not specific towards any aspects of information integration except to just convey what it is.

- [3] States it as a category of middleware that provide access to data giving an inference of a single database.
- [7] Gives an abstract definition by stating it as combining information from various sources to give a unified format.
- [12] Treat it as a solution to manage voluminous and diverse data transparently.

2.6 Goal Centric Definitions

The definitions stated here refer to the intention of information integration without any specification on the input or output or the process.

- [2] Indicates the goal as combining multiple information sources
- [4] The goal is addressed as to enable new applications that require information from several sources to be built quickly.
- [8] Here the goal has been specified with more explanation when compared to the previous ones, as to make available an integrated and coherent view of data stored various heterogeneous information sources.

The goal specified by [11] is to focus on what the user requires by providing a uniform interface to information sources without any regard of how or where to find the information.

The definitions however could be visualized with accordance to the categorization in the table below. Table 1 would give a very precise, readable and easily implied version of the categorization.

ARPN Journal of Systems and Software

http://www.scientific-journals.org

Table 1: Categorization of information integration definitions

Sl No.	Input parameters	Process	Output characteristics	Definition centered on
1	Heterogeneous information sources with difference in formats, locations and facilities	Support access to	Isolate users	Input
2	Data from existing heterogeneous distributed sources	Integration	Illusion of a single integrated information system	Input and output
3	Information regarding its organization, structure, country, semantics and so on	Viewing	Single interface	Input
4	Applications	Access data, middleware	Single database	Generic
5	Data from multiple sources	View – integrate without first load into central Data Warehouse	Tools	Process
6	Different sources	Combining information	Unified format	Generic
7	Distributed heterogeneous information process sources from inter organization and collaborative service provision.	Assembling	Coherent view	Input
8	DBMS, Content management system, Data warehouse, EAI	Combine core elements	Common platform	Input
9	Complimentary data	Brought together physically or logically	Applications to make use of all relevant and indirect data	Process and Output
10	Diverse forms of business information across organization	Integrate, transform data for business analysis and data interchange	Provide real time access, and data placement for performance, currency and availability	Input, Process and Output
11	Voluminous and diverse data	Manage transparently	-	Hybrid
12	Disparate heterogeneous sources , no common schema	Synthesize with flexibility, transparency and intelligence	Respond to a query for informed business decision by user or application program	Hybrid

3. PROPOSED CHARACTERIZATION OF INFORMATION INTEGRATION

The table 1 gives the input and output parameters along with the process of the mentioned definitions. The categorization shown above clearly indicates that the definition is biased towards any one of the constituents input, output, process, or goal. The ultimate meaning underlying information integration has been addressed in all definitions but in different and incomplete forms.

Considering the input of the definitions, even though the focus is on the information sources, each one differ , one address it as various databases, another as applications, some as data and content databases, while others address more deeply as to sources across organizations. There exists no commonality among them. There arises an ambiguity as to what really is the input, either any one of the aspects mentioned above or all or a combination of some. This selection of input sources is dependent over what? Hence such questions leave these definitions to be incomplete.

Analyzing the output in the definitions, the most important aspect is how the output has to be represented, or communicated and in which format. This

ARPN Journal of Systems and Software

http://www.scientific-journals.org

is unclear as one say it be a coherent view, one address it to be tool, one specify it as a unified format and the list goes on. Not much is specified as to who it is intended to, the purpose or the output form.

The same could be said about in the process involved. Even though the process here is to integrate, how actually is the integration done. As it is mentioned in terms of assemble, combine, integrate, transform, support access to, it leads to confusion as to what are the processes involved in integration and what sort of mapping or actual transformation that is needed.

Some definitions do exist which address all the aspects in their definition but when look in to them, one could say that they are not complete. Emphasize or explanation about the three aspects has been stated [11,12] but there lies a gap in case of what the actual process is along with in addition to how the process could be done and which form the output should be rather than being an answer to a query.

This explanation stated above has clearly stated that there exists no proper characterization for information integration which had initiated this study to propose a complete characterization for information integration thereby indicating the research initiatives in it. The proposed characterization is:

Information integration is the process of combining information modeling patterns with process patterns to gain visibility and federated access to distributed heterogeneous information sources at real time, thereby to provide information on demand through information delivery patterns for achieving better business insight.

4. DISCUSSION

To check for the conceptualization of a matter one normally answers the three questions of what it means, whether it addresses all constituents, and how exactly it is stated. Here information integration's definitions as portrayed in the previous section are analyzed with three dimensions of meaning, distinctness and completeness (exactness). These dimensions specify in common the necessities of any definition or conceptualization. By meaning it is to check whether it portrays the concept in the right sense and intention, distinctness refers to the addressing of its constituents in the concept and representation whereas exactness gives the check as to all the elements of the concept are stated precisely. The table below has been designed with the intention of portraying the extent to which these definitions stated above conveys the three constituents of a definition.

The explanation for the indicators are explained below

Table 2: Expression of components of various definitions

Definitions	Meaning	Distinctness	Completeness
a	--	+-	--
b	+-	--	--
c	+-	+-	--
d	+-	++	--
e	+-	+-	--
f	--	--	--
g	++	--	--
h	++	++	--
i	+-	++	--
j	++	+-	+-
k	--	+-	--
l	++	+-	+-
Proposed Characterization	++	++	++

-- The criteria is not mentioned at all

-+ It is just mentioned and does not satisfy the requirements at the basic level itself

+- The requirements are addressed partially

++ The requirements satisfy the criteria completely

The outcome of this study is presented in table 2, which has taken its input source as table 1 and supporting information stated above to analyze and to have produced such an output. This outcome gives a collective view of all what researchers wanted to convey on defining information integration and where actually they had lacked and thereby paved way for the proposed characterization specified in the previous section which give a comprehensive view on information integration by filling out the gaps in the related work.

5. CONCLUSION

Information integration builds on the solid foundation of existing data management solutions. However data management solutions do not provide the real information that would help in informed decision making for bringing about better business insight. This paper has brought out the various definitions that have been stated by various researchers by addressing their characteristics and their limits. This gives way for a need of a complete characterization of information integration in line with the context, which has been considered as the main outcome of this paper. The discussion part states clearly the limitations of the definitions while compared with the requirements of a definition. The proposed characterization specifies many issues that give way for another round of research in information integration with varied perspective.

REFERENCES

[1] N.W. Paton,C.A. Goble, S. Bechhofer, Knowledge based information integration systems, Information and Software Technology 42 (2000) 299–312

[2] Feng beiming, Ma Manfu, Gou Heping, An Architecture to integrate distributed information integration, ACM, Eighth International

Conference Grid and Cooperative Computing, 45 – 49, 2009

[3] Mukesh Mohania and Manish Bhide, New Trends in Information Integration, IBM Almaden Research Lab, Proceedings of the 2nd international conference on Ubiquitous information management and communication, ACM, 2008

[4] Laura Haas, Beauty and the Beast: The Theory and Practice of Information Integration ICDT 2007 , Vol. 4353 (2006), pp. 28-43

[5] Alon Halevy, Anand Rajaraman, Joann Ordille, Data Integration in Teenage Years, VLDB `06, ACM 1-59593-385-9/06/09

[6] Alon Y. Halevy, Naveen Ashishy, Dina Bittonz, Michael Careyx, Denise Draper, Jeff Pollockk, Arnon Rosenthal, Vishal Sikkayy , "Enterprise Information Integration: Successes, Challenges and Controversies", ACM SIGMOD, June 2005.

[7] Philip A. Bernstein, Laura Haas, Information Integration in the Enterprise, Communications of the ACM,Volume 51 Issue 9, September 2008

[8] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi, Riccardo Rosati Information Integration: Conceptual Modeling and Reasoning Support, Proceedings-3rd IFCIS, International Conference on Cooperative Information Systems, Pgs 280-289, Aug 1998.

[9] M. A. Roth,D. C. Wolfson, J. C. Kleewein, C. J.Nelin, Information integration: A new generation of information technology, IBM SYSTEMS JOURNAL, pgs 563-577,VOL 41, NO 4, 2002

[10] A. D. Jhingran,N. Mattos,H. Pirahesh, Information integration: A research agenda, IBM SYSTEMS JOURNAL, pgs 555-562, VOL 41, NO 4, 2002

[11] Yannis Dimopoulos and Antonis Kakas, Information Integration and Computational Logic, Computing Research Repository - CORR , vol. cs.AI/0106, 2001

[12] www.Ibm.com.db2.ii.doc

[13] Patrick C. K. Hung, Dickson K. W. Chiu, Developing Workflow-based Information Integration (WII) with Exception Support in a Web Services Environment, Proceedings of the 37th Hawaii International Conference on System Sciences – 2004.

BIOGRAPHY

C. Punitha Devi obtained her B.sc in Computer Science (1996) and M.C.A (1999) from Bharathidasan University and M.Tech in Computer Science & Engineering (2007) from Pondicherry University. Currently she is pursuing PhD in Department of Computer Science and Engineering, Pondicherry University. She is having 10 years of teaching experience. She has published one book and papers in national and international journals/ conferences. Her research area includes Service Oriented Architecture and Software Engineering.

Dr. V. Prasanna Venkatesan is currently an Associate Professor, Department of Banking Technology, Pondicherry University. He earned his B.Sc in Physics (1986) from Arignar Anna Arts College, karaikal. He received his M.C.A (1989) from Pondicherry Engineer-ing College, M.Tech in Computer Science & Engineering (1995) from Pondicherry University and Ph.D in Computer Science & Engineer-ing (2008) from Pondicherry University. He is having more than 20 years of teaching experience. He has published 3 books and papers in national and international journals/ conferences. His research area includes Software Architecture, Banking Technology, Object Oriented Modeling and Design, Smart Banking.

G. Shanmugasundaram obtained his B.Tech in Information Technology (2005) from BCET, Pondicherry University. He received his M.Tech in Computer Science and Engineering (2008) from SMVEC, Pondicherry University. Currently he is pursuing PhD in Department of Banking Technology; Pondicherry University. He is having 2 years of teaching experience and 1 year in software development. His research area includes Service Oriented Architecture and Web Technologies.