American Journal of Operations Research, **2013, 3, 181-186**
http://dx.doi.org/10.4236/ajor.2013.31A017 Published Online January 2013 (http://www.scirp.org/journal/ajor)

CAS-Based Approach for Automatic Data Integration

Eli Rohn
Department of Information Systems Engineering, Ben-Gurion University, Beersheba, Israel
Email: elirohn@bgu.ac.il

Received November 29, 2012; revised December 30, 2012; accepted January 14, 2013

ABSTRACT

Research of automatic integration of structured and semi-structured data has not resulted in success over the past fifty years. No theory of data integration exists. It is unknown what the theoretical necessary requirements are, to fully support automatic data integration from autonomous heterogeneous data sources. Therefore, it is not possible to objectively evaluate if and how much new algorithms, techniques, and specifically Data Definition Languages, move towards meeting such theoretical requirements. To overcome the serious reverse salient the field and industry are in, it will be helpful if a data integration theory would be developed. This article proposes a new look at data integration by using complex adaptive systems principles to analyze current shortcomings and propose a direction that may lead to a data integration theory.

Keywords: Software Engineering; Data Definition Languages; Data Integration; Complex Adaptive Systems

1. Motivation and Introduction

Data integration is a pervasive challenge faced in applications that need to query across multiple autonomous and heterogeneous data sources. It is also a major challenge for companies experiencing mergers or acquisitions. Data integration is crucial in large enterprises that own a multitude of data sources, for progress in large-scale scientific projects, where data sets are being produced independently by multiple researchers, for better cooperation among government agencies, each with their own data sources, and in offering good search quality across the millions of data sources on the World-Wide Web.

Research of automatic integration of structured and semi-structured data has been largely unsuccessful over the past fifty years. No theory of data integration exists. It is unknown what the theoretical necessary requirements are to fully support automatic data integration from autonomous heterogeneous data sources. Therefore, it is not possible to objectively evaluate if and how much new algorithms, techniques, and specifically Data Definition Languages (DDLs), move towards meeting the requirements of automatic data integration. Nor is it possible to suggest a better algorithm, technique or DDL that might advance the state-of-the-art of automatic data integration, because the requirements do not exist.

2. Data Structures

Data structures are used to organize and represent related facts to ultimately satisfy a goal. Computerized data structures are constructed using a given syntax. The syntax is usually referred to as the Data Definition Language (DDL). A DDL specifies how to organize and interconnect related elementary pieces of data into useable structures. DDLs come in three types: "structured", (e.g., COBOL, SQL) "semi-structured", (e.g., web pages, Word documents) and "unstructured" (e.g., images, voice). Data structures can differ on three aspects: their structure (which also implies level of details), field or tag names, and syntax used to define the data structure.

DDLs are used to codify messages to be sent or received by computerized systems or their components. Scores of DDLs have been developed over the years. Examples of DDL include Cobol's structured File Description (FD) section; delimited flat files such as Comma Separated Values (CSV) and Data Interchange File Format (DIFF) for data exchange; Structured Query Language (SQL) for relational databases; Extensible Markup Language (XML) for semi-structured data and metadata; ontologies expressed in a variety of DDLs such as Resource Description Framework (RDF) and Web Ontology Language (OWL). Standards such as EDI and SWIFT also define data. EDI (Electronic Data Interchange) established a "common language" for exchanging business-related transactions via the creation and enforcement of a standard. The Society for Worldwide Interbank Financial Telecommunication (SWIFT) created a "common language" for exchanging monetary related transactions via the creation and strict enforcement of standards.

3. Data Integration Approaches

Lack of a data integration theory has not discouraged continuing efforts to create a data integration panacea. However, lacking fundamental understanding of failures root causes, none of these technologies was able to escape, circumvent or overcome an invisible reverse salient. Many are poor in thorough analysis, often devoid of sound mathematical foundations, and littered with short lived solutions [1]. This has led to what the Gartner Group calls the "hype cycle"—a model of the relative maturity of technologies in a certain domain [2,3]. Data integration solutions end up in some form of an electronic graveyard upon reaching the "disillusion stage" in Gartner's hype cycle. Such was the fate of the Metadatabase project [4-6], the STRUDEL project [7], the NIMBLE solution [8], XOP (XML-binary Optimized Packaging), the Ozone project [9], OIL (Ontology Inference Layer) [10,11], DAML (DARPA Agent Markup Language) [12], DAML + OIL [11,13], CICA (Context Inspired Component Architecture), XVIF (XML Validation Interoperability Framework) and others. Usage of these decade old approaches has not stopped, even though researchers recognize that data integration techniques do not work without significant human intervention or the demand to use a specific data structure, making it a defacto standard. For example, the Quality-Aware Service-oriented data Integration (QASI) project relies on available "data networks [that] consists of autonomous systems—data peers—which may have both data exchange relationships and virtual mappings between each other" [14]. However, if these do exist and if they are accessible to QASI, they merely relay on aforementioned decade old techniques that cannot be fully automated. In fact, the authors state that data marketplaces aid their customers by providing online schema browsing tools and schema documentation. The astute reader understands that such tools are for humans to use, in the quest of integrating data. Another example of a recent project is an incident information management framework based on data integration, data mining, and multi-criteria decision making [15]. It too references data integration techniques that are two decades old, none of which produced a data integration solution that doesn't require substantial human intervention. With these grim results, it is worth taking the risk and look at DDL engineering geared towards data integration from an entirely different perspective. This may lead to new workable insights. Hence, we look at DDL engineering using principles of Complex Adaptive Systems CAS [16-18].

4. Relevant CAS Attributes

Analysis of DDLs using CAS theory should be done by using relevant CAS characteristics: variety, tension and entropy. To those we should add a control mechanism referred to as "regulator". Each is explained hereafter.

4.1. Variety

In its simplest form, given a set of elements, its variety is the number of distinguishable elements. Thus the set {wnbcbbccbscnn} has a variety of four letters {,b,c,n,w}. It may be more conveniently measured by the logarithm of this number. If the logarithm is taken to base 2, the unit is the bit. E.g., $\mathrm{Log}_2(4) = 2$. For a given system, variety is the number of meaningful different states and disturbances that a system has.

Disturbances are irregular inputs or system states outside normal values or boundaries. To handle them without breaking down, a system needs to have a sifting and response mechanism. Inputs that are ignored (filtered) are irrelevant to the system. Filtering is a part of the regulator. The remaining inputs need to be dealt with, using a regulator that generates a proper response. That is, the irregularity is mapped into the system, because it helps the system to achieve its goals. If more than one response is possible, the regulator should use the one that best meets the system's goals.

4.2. Tension

Tension in physical systems can be expressed as interaction of parts in a mechanical system, measurable in some units of energy. Suspension bridges provide a prime example of tension. The tension on their cables is supposed to preserve the relation between the state of the bridge and some aspects of its environment. A CAS degree of sensitivity towards its environment is tension; it preserves, at least temporarily, the relation between the inner state of the CAS and some aspects of its environment. The mapping corresponds "closely with the current conception of 'information', viewed as the process of selection of a variety has meaning" [20].

4.3. Entropy

Von Bertalanffy demonstrated the physical equivalence of thermodynamics entropy and information theory entropy [21]. A complexion is any specific set of choices out of all the possible sets, made by each element. The number of complexions in an arrangement is the number of possible alternatives one can choose from. This is equivalent to ensemble of variety in information theory [22], and the entropy measure thus can be used.

4.4. Law of Requisite Variety

If a system aims to successfully adapt, achieve or survive, it requires a certain amount of flexibility. That amount of flexibility has to be proportional to the variety that the system must contend with. By analogy, in a chess game a pawn has a limited number of responses to a threat compared to the queen's. That is, the queen's variety of moves is far greater than the pawn's and therefore can better adapt to the threat in order to assure its survival and decrease its player's chances of losing. This necessary flexibility is known as Ashby's Law of Requisite Variety (LRV) [16].

5. Data Integration and CAS Principles

Automatic data integration from autonomous and heterogeneous sources is viewed here as a transition from a closed to an open system, which is in essence an adaptive information processing system. Such systems do not simply engage in interchange of data or information with its environment. This interchange is an essential factor underlying the system's viability, continuity, and its ability to change further. Therefore, the suitability of mechanisms, such as DDLs, for data integration should be analyzed using a matching paradigm, namely CAS.

Data integration is a goal oriented process of combining data residing at different sources. Since the data is represented using non-identical data structures, the integration requires correct mapping of data elements from one or more source data structure to the destination data structure. Such mapping is far from trivial, as shown by Batini, Lenzerini *et al*. [20], Hunter and Liu [21], and others [22,23].

DDLs to date not come with a robust regulation mechanism that satisfies LRV [16,23,24]. Hence, data integration approaches do not yield tension unless humans intervene in the mapping process and invest mental energy to keep the relations from falling apart when a data source changes its data structure. Such failures are due to the absence of a regulator that can successfully overcome the existing semantic heterogeneity, which in turn is a manifestation of the theoretically infinite variety that exists in the environment.

"Resolution" of semantic heterogeneity, in CAS terminology, is an attempt to constraint the variety. Each proposed data integration method reviewed in the literature earlier is a form of a regulator, in the sense explained by Ashby [16] and by Casti [25]. For example, an EDI implementation requires skilled personnel and specialized software to map data from the organization's internal data formats to EDI and vice-versa. Without such mapping, the system's weakness grows. The mapping requires (mental) energy supplied by humans whose mission is to reduce such weakness, hopefully eliminating it. Correct mapping among data elements creates (or sustains) the tension mentioned above. All EDI implementations rely upon industry consensus, implemented as standards. This is a form of a regulator in Ashby's sense. It is noted that adaptation to a changing environment is not a characteristic of EDI systems. It is also noted that social (business) agreements provide an Ashby regulation mechanism. Both cases exemplify how reduction of semantic heterogeneity is achieved.

Many attempts to constrain variety and address semantic heterogeneity have been proposed, including ontology based solutions. Elements of a given data structure are mapped to an ontology for aiding with semantic resolution tasks. However, ontologies themselves are not constrained. Thousands of competing ontologies have been built and published in recent years [26-30]. Further, several non-compatible ontology mechanisms have been proposed and put to use [10-13]. From a CAS perspective, these were attempts to build a regulator. As explained earlier, they too failed to deliver the promise, since they did not manage to control semantic variety and relations variety. These solutions also have the risk of creating a circular reference (**Figure 1**), thus not leading to a resolution and not meeting their proposed goals.

Two successive articles [31,32] performed an extensive literature review and analysis of DDLs from the 1950's to date. They conclude that existing DDLs are indistinguishable when their variety is analyzed and that all DDLs have almost identical expressive power in terms of information bits. The main culprit is the level of natural language built into DDLs, referred to as semantic heterogeneity in contemporary literature.

It is vital to recognize that from a CAS perspective, data integration is in essence the creation of a meaning preserving mapping, or relation, between an ensemble and its external constrained variety. Such mapping preserves the meaning of the variety vis-à-vis information systems, whose goal is to integrate at least some external data. Mappings, or relations, that last for the duration that they are needed, are held together by tension. In symbols-mediated CAS, tension can be measured by formal meaning preservation requirements [33]. The level of organization created by a specific set of relations (tension) out of all the possible sets (complexion) is measurable by using entropy.

6. Desired Attributes

A data definition language that is designed for automatic data integration of heterogeneous sources should satisfy some CAS characteristics: ability to selectively map to the variety presented by a system's environment; autonomously maintain the mapping as long as needed; ability to add, remove or update its own elements and relations dynamically. This requires the ability to build a regulator that has at least the same variety it needs to regulate, because it will have to satisfy LRV. In CAS terminology, the DDL should be able to "make sense" of relevant variety in its environment by means of some mediator (regulator), create tension and sustain it (preserve

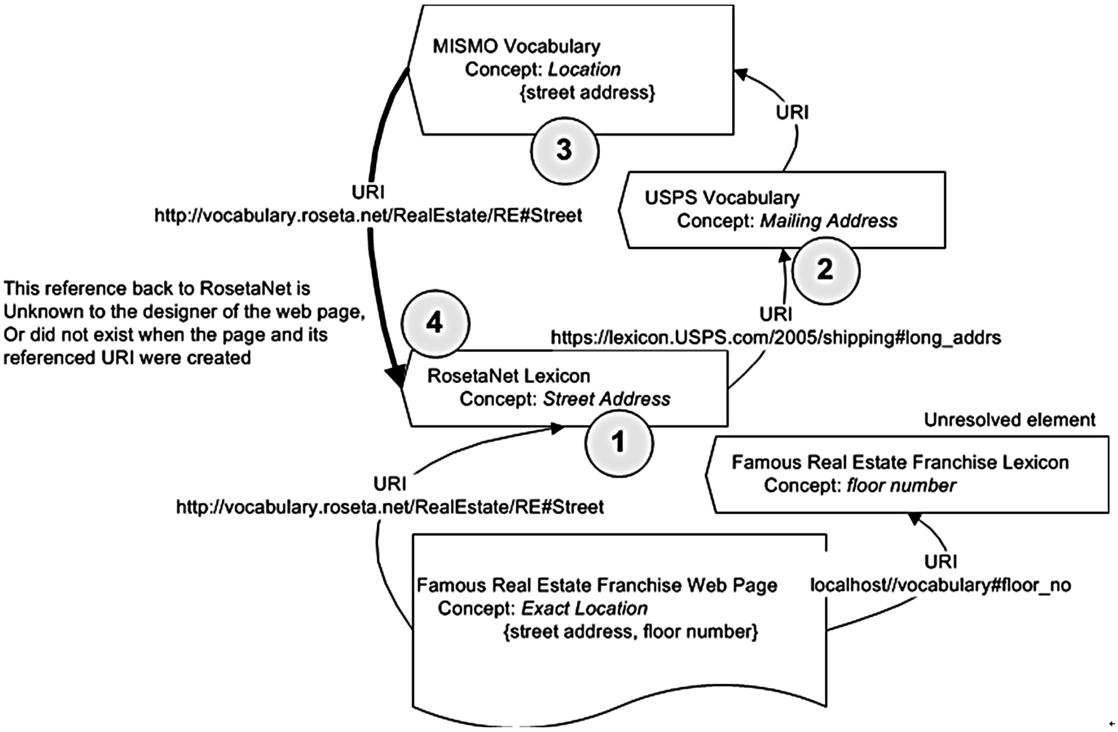
MISMO Vocabulary
Concept: Location
{street address}
3
URI
USPS Vocabulary
Concept: Mailing Address
2
URI
http://vocabulary.roseta.net/RealEstate/RE#Street
This reference back to RosetaNet is
Unknown to the designer of the web page,
Or did not exist when the page and its
referenced URI were created
4
URI
https://lexicon.USPS.com/2005/shipping#long_addrs
RosetaNet Lexicon
Concept: Street Address
1
Unresolved element
Famous Real Estate Franchise Lexicon
Concept: floor number
URI
http://vocabulary.roseta.net/RealEstate/RE#Street
URI
localhost//vocabulary#floor_no
Famous Real Estate Franchise Web Page
Concept: Exact Location
{street address, floor number}

Figure 1. Circular reference example.

meaning). To "make sense" it should have a perfect disambiguation mechanism, or be built on foundations with no ambiguity and therefore exhibit perfect (or maximum) entropy, as defined in information theory [22]. Therefore, it is required to build a regulator that is a canonical control system [25,34], a regulator that is completely reachable and completely observable.

To be more precise, a DDL designed for automatic data integration of heterogeneous sources may be viewed as an entity D including an automatic intelligent control system con(D) with a phase (or state) space sp(D) having an associated finite set of relations rel(D) satisfying the following properties:

1) The control system con(D) is completely observable and completely reachable [34]. That is, all elements of sp(D) can be compared, corrected and integrated, when they are identified with heterogeneous data sources.

2) The phase space is denumerable and generated by a finite set of rules or operations from a small finite set of atoms or building blocks.

3) The set of relations is finite

4) Disable to dynamically modify both sp(D) and rel(D).

5) All of the dynamical data processes of D are performed with maximum entropy [22].

6) D is universal for a large class of data sources S in the sense that it is capable of creating and maintaining and inverting an injective (one-to-one), meaning preserving mapping (homomorphism or monomorphism): φ: A → sp(D) for any A in the class S.

Creating a CAS based DDL requires the presence of a regulator that has at least the same variety it needs to regulate, such that it satisfies the law of requisite variety. In CAS terminology, the DDL should be able to "make sense" of some of at least some of the variety in the environment by means of some mediator (regulator), create tension and sustain it (preserve meaning). To "make sense", this regulator should have a perfect disambiguation mechanism; alternatively, is could be build on foundations with no ambiguity and therefore perfect entropy ($H = 1$).

This research advocates lowering expectations, and limit the regulator to handling an enumerable set. That is, variety that can be generated using some simple rules of derivation. Such a regulator should be a canonical control system as described hereto.

One existing implementation of a canonical system is the symbolic language of chemistry. Admittedly, this system is not expressive enough for all the needs of contemporary chemists, but it is a working solution offering an intriguing idea for a DDL that supports automatic data integration and satisfies LRV. The periodic table offers unambiguous building blocks that by themselves have meaning for chemists and make some sense of the world. There are fewer than 100 elements in Mendeleev's periodic table, and they suffice to describe all known matter in our universe. Using a set of rather simple rules, one can combine two or more building blocks (atoms) to create new concepts, namely molecules, in chemistry. As long as the rules are understood and followed, a transmitter of information can create a new concept that doesn't exist up to that point and the recipient will be able to understand the concept using the same rules that created the new concept. For example, oxygen and hydrogen ("O" and "H" respectively) are two such building blocks, and each carries meaning. Their combination H_2O is validper the rules of chemistry. The new concept, that doesn't exist in the periodic table, is understood by anyone knowledgeable about chemistry. The same person will reject the concept $H_{2.5}O$ (2.5 particles of H) because the newly created complex concept violates the set of rules.

A challenge now exists—how to create an analogous DDL. It may require significant advancements in the field of formal knowledge representation before an attempt to build such a DDL becomes feasible.

7. Summary

Existing DDLs were designed for internal processing tasks and not for automatic data integration from autonomous heterogeneous sources. They are fundamentally ill-suited for this task. Wrapping them with layers of new syntaxes (e.g., XML, OWL, etc.) made the task look more like onion peeling—there is a need now to process more layers with new names whose essence is indistinguishable from the inner (older) layers, as illustrated in **Figure 2**. Hence, the added layers do not meet the LRV and therefore are futile. Rather than reduce semantic heterogeneity, they add to it. This shows that adding variety to a system is not always beneficial.

For automatic data integration of heterogeneous sources to be successful, there is a need to engineer DDLs and their necessary supporting mechanisms from the very beginning. The design must, as a minimum, satisfy all the aforementioned desired attributes, and contain a flexible regulator. CAS principles seem to be most promising, and should be evaluated further.

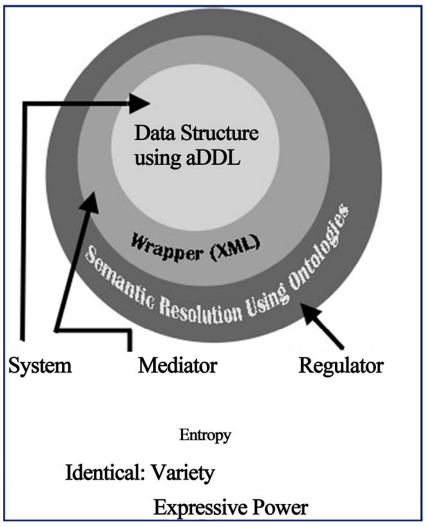

Figure 2. Insignificant layers.

REFERENCES

[1] R. E. Knox, "The XML Family of Standards: Four Years Later," Gartner Group, Stamford, 2001.

[2] R. E. Knox and C. Abrams, "Hype Cycle for XML Technologies for 2003," Gartner Group, Stamford, 2003.

[3] R. E. Knox, C. Abrams, T. Friedman, D. Feinberg, K. Harris-Ferrante and D. Logan, "Hype Cycle for XML Technologies, 2006," Gartner Group, Stamford, 2006.

[4] C. Hsu, "The Metadatabase Project at Rensselaer," *ACM SIGMOD Record*, Vol. 20, No. 4, 1991, pp. 83-90. doi:10.1145/141356.141394

[5] C. Shu, "Metadatabase: An Information Integration Theory and Reference Model," 2003. http://viu.eng.rpi.edu/mdb/iitrm.html

[6] J. Clark and S. DeRose, "XML Path Language (XPath) Version 1.0," W3C Recommendation, 1999. http://www.w3.org/TR/xpath/

[7] M. Fernandez, D. Florescu, J. Kang, A. Levy and D. Suciu, "STRUDEL: A Web Site Management System," *Proceedings of the* 1997 *ACM SIGMOD International Conference on Management of Data*, Tucson, 13-15 May 1997, pp. 549-552.

[8] D. Draper, A. Halevy and D. S. Weld, "The Nimble XML Data Integration System," Nimble Technology, Inc., San Jose, 2001.

[9] T. Lahiri, S. Abiteboul and J. Widom, "Ozone: Integrating Structured and Semistructured Data," Revised Papers from the 7th International Workshop on Database Programming Languages: Research Issues in Structured and Semistructured Database Programming, Springer-Verlag, Heidelberg, 2000.

[10] S. Bechhofer, I. Horrocks, C. Goble and R. Stevens, "OilEd: A Reason-able Ontology Editor for the Semantic Web," In: C. A. Goble, D. L. McGuinness, R. Moller and P. F. Patel-Schneider, Eds., *Working Notes of the* 2001 *International Description Logics Workshop* (*DL*-2001), CEUR-WS.org, Stanford, 2001, pp.

[11] WEBONT, "W3C DAML+OIL Project," W3C Web-Ontology Working Group, 2009. http://www.w3.org/2001/sw/WebOnt/

[12] DAML.org, "DAML Ontology Library," US Government (DARPA), Arlington, 2004. http://www.daml.org/ontologies/

[13] M. Greaves, "2004 DAML Program Directions," DAML. org, Arlington, 2004.

[14] S. Dustdar, R. Pichler, V. Savenkov and H.-L. Truong, "Quality-Aware Service-Oriented Data Integration: Requirements, State of the Art and Open Challenges," *SIGMOD Record*, Vol. 41, No. 1, 2012, pp. 11-19. doi:10.1145/2206869.2206873

[15] Y. Peng, Y. Zhang, Y. Tang and S. Li, "An Incident In-

formation Management Framework Based on Data Integration, Data Mining, and Multi-Criteria Decision Making," *Decision Support Systems*, Vol. 51, No. 2, 2011, pp. 316-327. doi:10.1016/j.dss.2010.11.025

[16] R. W. Ashby, "An Introduction to Cybernetics," Chapman & Hall, London, 1956.

[17] Y. Bar-Yam, "Dynamics of Complex Systems," Westview Press, Boulder, 1997.

[18] W. Buckley, "Society—A Complex Adaptive System," Gordon and Breach Publishers, Amsterdam, 1998.

[19] N. Wiener, "Cybernetics or Control and Communication in the Animal and the Machine," MIT Press, Cambridge, 1948.

[20] W. Buckley, "Sociology and Modern Systems Theory," Prentice-Hall, Inc., Englewood Cliffs, 1967.

[21] R. C. Raymond, "Communications, Entropy, and Life," *American Scientist*, Vol. 38, No. 4, 1950, pp. 273-278.

[22] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Systems Technical Journal*, Vol. 27, No. 3, 1948, pp. 379-423.

[23] R. W. Ashby, "Adaptiveness and Equilibrium," *Journal of Mental Science*, Vol. 86, No. 5, 1940, pp. 478-484.

[24] R. W. Ashby, "The Nervous System as Physical Machine: With Special Reference to the Origin of Adaptive Behavior," *Mind*, Vol. 56, No. 221, 1947, pp. 44-59.

[25] J. L. Casti, "Canonical Models and the Law of Requisite Variety," *Journal of Optimization Theory and Applications*, Vol. 46, No. 4, 1985, pp. 455-459.

[26] MMI.ORG, "Marine Metadata Interoperability Ontology Registry and Repository," 2012.
http://mmisw.org/orr/#b

[27] OOR, "Open Ontology Repository," 2012.
http://ontolog.cim3.net

[28] OASIS, "OASIS Advances CAP and Emergency Data Exchange Language (EDXL) Specifications," 2006.
http://xml.coverpages.org/ni2005-09-08-a.html

[29] XML.GOV, "Registires," 2007.
http://xml.gov/registries.asp

[30] Protege, "Ontologies Registry," 2007.
http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library.

[31] E. Rohn, "Generational Analysis of Variety in Data Structures: Impact on Automatic Data Integration and on the Semantic Web," *Journal of Knowledge and Information Systems*, Vol. 24, No. 2, 2009, pp. 283-304.

[32] E. Rohn, "Generational Analysis of Tension and Entropy in Data Structures: Impact on Automatic Data Integration and on the Semantic Web," *Knowledge and Information Systems*, Vol. 28, No. 1, 2010, pp. 175-196.

[33] J. F. Sowa, "Worlds, Models and Descriptions," *Studia Logica*, Vol. 84, No. 2, 2006, pp. 323-360.

[34] E. D. Sontag, "Mathematical Control Theory, Deterministic Finite Dimensional Systems," 2nd Edition, Springer-Verlag, Heidelberg, 1998.