



**IntechOpen**

# Linked Open Data

Applications, Trends and Future Developments

*Edited by Kingsley Okoye*





---

# Linked Open Data - Applications, Trends and Future Developments

*Edited by Kingsley Okoye*

Published in London, United Kingdom

---



## IntechOpen





*Supporting open minds since 2005*





Linked Open Data – Applications, Trends and Future Developments

<http://dx.doi.org/10.5772/intechopen.80197>

Edited by Kingsley Okoye

#### Contributors

Shishir Kumar, Anju Shukla, Harikesh Singh, Julthep Nandakwang, Prabhas Chongstitvatana, Jung-Ran Park, Andrew Brenza, Chang Kuo-Chi, Monday Osagie Adenomon, Kingsley Okoye, Lori Richards, Kai-Chun Chu, Hsiao-Chuan Wang, Yuh-Chung Lin, Tsui-Lien Hsu, Yu-Wen Zhou

© The Editor(s) and the Author(s) 2020

The rights of the editor(s) and the author(s) have been asserted in accordance with the Copyright, Designs and Patents Act 1988. All rights to the book as a whole are reserved by INTECHOPEN LIMITED. The book as a whole (compilation) cannot be reproduced, distributed or used for commercial or non-commercial purposes without INTECHOPEN LIMITED's written permission. Enquiries concerning the use of the book should be directed to INTECHOPEN LIMITED rights and permissions department ([permissions@intechopen.com](mailto:permissions@intechopen.com)).

Violations are liable to prosecution under the governing Copyright Law.



Individual chapters of this publication are distributed under the terms of the Creative Commons Attribution 3.0 Unported License which permits commercial use, distribution and reproduction of the individual chapters, provided the original author(s) and source publication are appropriately acknowledged. If so indicated, certain images may not be included under the Creative Commons license. In such cases users will need to obtain permission from the license holder to reproduce the material. More details and guidelines concerning content reuse and adaptation can be found at <http://www.intechopen.com/copyright-policy.html>.

#### Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2020 by IntechOpen

IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 5 Princes Gate Court, London, SW7 2QJ, United Kingdom

Printed in Croatia

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from [orders@intechopen.com](mailto:orders@intechopen.com)

Linked Open Data – Applications, Trends and Future Developments

Edited by Kingsley Okoye

p. cm.

Print ISBN 978-1-83962-671-5

Online ISBN 978-1-83962-672-2

eBook (PDF) ISBN 978-1-83962-673-9

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

**5,100+**

Open access books available

**126,000+**

International authors and editors

**145M+**

Downloads

**156**

Countries delivered to

Our authors are among the  
**Top 1%**

most cited scientists

**12.2%**

Contributors from top 500 universities



**WEB OF SCIENCE™**

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)







# Meet the editor



Kingsley Okoye received his PhD in Software Engineering from the University of East London, UK in 2017. He also completed an MSc in Technology Management in 2011 and a BSc in Computer Science in 2007. He is an MIET member at the Institution of Engineering and Technology UK, a Graduate Member of IEEE, a member of IEEE SMCS Technical Committee (TC) on Soft Computing, and Data Architect at Tecnológico de Monterrey. He is a devoted researcher in industry and academia in both hardware and software fields of computing. He serves as guest editor, editorial board member and reviewer in reputable journals and conferences. His research interests include process mining, semantic web technologies, learning analytics, computer education, data management, internet applications and ontologies.



# Contents

<b>Preface</b>	<b>XIII</b>
<b>Chapter 1</b> BIBFRAME Linked Data: A Conceptual Study on the Prevailing Content Standards and Data Model <i>by Jung-Ran Park, Andrew Brenza and Lori Richards</i>	<b>1</b>
<b>Chapter 2</b> TULIP: A Five-Star Table and List - From Machine-Readable to Machine-Understandable Systems <i>by Julthep Nandakwang and Prabhas Chongstitvatana</i>	<b>19</b>
<b>Chapter 3</b> Linked Open Data: State-of-the-Art Mechanisms and Conceptual Framework <i>by Kingsley Okoye</i>	<b>41</b>
<b>Chapter 4</b> Analysis of Effective Load Balancing Techniques in Distributed Environment <i>by Anju Shukla, Shishir Kumar and Harikesh Singh</i>	<b>61</b>
<b>Chapter 5</b> Study on IoT and Big Data Analysis of 12” 7 nm Advanced Furnace Process Exhaust Gas Leakage <i>by Kuo-Chi Chang, Kai-Chun Chu, Hsiao-Chuan Wang, Yuh-Chung Lin, Tsui-Lien Hsu and Yu-Wen Zhou</i>	<b>79</b>
<b>Chapter 6</b> Financial Time Series Analysis via Backtesting Approach <i>by Monday Osagie Adenomon</i>	<b>99</b>



# Preface

Today, modern tools and techniques for the collection and analysis of data in all fields of science and technology are proving to be more complex. The growing complexities are evidenced by the need for a more generalized and standardized description (integration) of the various data sources and formats to allow for the flexible exploration of different data types. Theoretically, the challenge has been how to create automated systems capable of providing an understandable format for the different datasets, as well as making the derived formats and/or standards applicable across the different platforms. Over the past few decades, one of the recent technologies that have proved indispensable in this area is the Linked Open Data (LOD). The LOD systems consist of a number of machine-readable datasets with Resource Description Framework (RDF) triples that are useful in describing data classes and the underlying properties. Moreover, indications from early research note that one of the problems with such existing data or information processing systems is the need for not just representing the data (or information) in formats that can be easily understood by humans, but also for building the intelligent systems that process the information that they contain or support. In other words, “machine-understandable” systems. By machine-understandable system, we assume that the extracted information or models are either semantically labeled (annotated) to ease the analysis process, or represented in a formal structure (ontology) that allows a computer (the reasoning engine) to infer new facts by making use of the underlying relations. Indeed, the main idea for such data or information processing systems or those aspects of aggregating the data and computing the hierarchy of several process elements; is that they should not only be machine-readable but also machine-understandable. An adequate knowledge-base system is perceived to be, on the one hand, understandable by people, and on the other hand understandable by the machines. As devices become smarter and produce data about themselves, it has become increasingly important for data scientists to take advantage of more powerful tools and/or data integration techniques to help provide a common standard for information dissemination across the different platforms. To this end, the content of this book demonstrates that technologies such as the semantic web, machine learning, deep learning, natural language processing, internet of things, knowledge graph, process mining, and artificial intelligence, etc. which encompasses the wider spectrum of the LOD are of paramount importance. Therefore, this book presents two main drivers for the LOD technologies as follows: (i) encoding knowledge about specific data and process domains, and (ii) advanced reasoning and analysis of the big datasets at a more conceptual level.

This book intends to provide the reader with a comprehensive overview of the latest developments within the LOD framework and the benefits of the supported methods – ranging from the semantics-aware techniques that exploit knowledge kept in big data to improved data reasoning (big analysis) beyond the possibilities offered by most traditional data mining techniques. Fundamentally, the book covers the entire spectrum of “Linked Open Data - Applications, Trends and Future Developments”. It consists of six chapters selected after a rigorous review

by both the academic editor, reviewers, and the IntechOpen book editorial team. Technically, each of the individual chapters provides a comprehensive conceptualisation of the LOD framework and its main application components. The authors of the different chapters are reputable scholars and researchers from across the world with wide areas of research interests. Ranging from the computational fields of computer science, data science, information science, software engineering, knowledge graph and library linked data, internet of things, and semantic web technologies to the engineering and manufacturing fields of process modelling, process intelligence, and then to knowledge and data management, e-commerce and financial analytics, and educational innovation. Fundamentally, the rich contents of this book, conveyed by the authors through the various chapters, explore the different topics of interest in relation to LOD, ranging from the latest in LOD clouds and systems, to research problems and challenges with LOD, and then its application in real-time or real-world settings. This includes detailed subjective knowledge of the gaps in the existing literature, suitable methodologies applied to address the identified gaps, design and development of LOD frameworks, and case studies.

To this end, *Chapter 1* looks at the extent to which the Bibliographic Framework Initiative (BIBFRAME) has been used to integrate library data from the silos of online catalogues, and then discusses some of the challenges that need to be addressed in order to optimize the potential capabilities that the BIBFRAME model holds. *Chapter 2* discusses the several attempts within the Linked Data research area to transform table and list datasets into machine-readable formats by proposing a data model named TULIP. The method focused on transforming tables and lists into RDF format while thoroughly maintaining their essential configurations and for the future development of the Semantic Web. *Chapter 3* presents the latest mechanisms and conceptual framework of the LOD by proposing a Semantic-Based Linked Open Data Framework (SBLODF) that integrates the different elements or entities in information systems or models with semantics (metadata descriptions) to produce explicit and implicit information based on the user's search queries. The SBLODF framework is a machine-readable and machine-understandable system that proves to be useful in encoding knowledge about different process domains and representation of the discovered information or models at a more conceptual level. *Chapter 4* discusses the need and issues that involves the analysis of effective load balancing techniques in a distributed environment. It looks at the heterogeneous nature of distributed computing, interoperability, fault occurrence, resource selection, and task scheduling for performance optimization of web resources through various balancing algorithms and scheduling methods, and then provides a concise narrative of the problems encountered and dimensions for future extension. *Chapter 5* is on the study of the internet of things (IoT) and big data analysis for advanced processes. It discusses a semiconductor process for manufacturing that is set on the cloud database for big data analysis and decision-making through a continuous monitoring system. *Chapter 6* assesses the implication of the backtesting approach in financial time series analysis when choosing a reliable Generalized Auto-Regressive Conditional Heteroscedastic (GARCH) model for analysing stock returns in a case study of financial institution settings.

Resourcefully, this book is a reference and educational book targeted to be beneficial for data scientists, software developers, semantic web engineers, information system designers, process managers, teachers, and researchers, and general consumers in application and implementation of the LOD framework

and research in the various contexts. With this book, the editor and authors focused on bridging the practical and theoretical gap in the methodological use and commercial application of the LOD concepts in computer science and engineering.

I would like to thank IntechOpen publishers, and the publishing team especially the book project Process Manager, Ms. Jasna Božić, for the professional commitment to ensuring the successful completion of this book. Most importantly, I would like to specially thank the authors for their dedicated hard work, wonderful research, and informative contributions that form the rich content of this book.

**Dr. Kingsley Okoye**  
Data Architect,  
Tecnologico de Monterrey,  
Monterrey, Mexico





# BIBFRAME Linked Data: A Conceptual Study on the Prevailing Content Standards and Data Model

*Jung-Ran Park, Andrew Brenza and Lori Richards*

## Abstract

The BIBFRAME model is designed with a high degree of flexibility in that it can accommodate any number of existing models as well as models yet to be developed within the Web environment. The model's flexibility is intended to foster extensibility. This study discusses the relationship of BIBFRAME to the prevailing content standards and models employed by cultural heritage institutions across museums, archives, libraries, historical societies, and community centers or those in the process of being adopted by cultural heritage institutions. This is to determine the degree to which BIBFRAME, as it is currently understood, can be a viable and extensible framework for bibliographic description and exchange in the Web environment. We highlight the areas of compatibility as well as areas of incompatibility. BIBFRAME holds the promise of freeing library data from the silos of online catalogs permitting library data to interact with data both within and outside the library community. We discuss some of the challenges that need to be addressed in order to optimize the potential capabilities that the BIBFRAME model holds.

**Keywords:** linked data, functional requirements for bibliographic records (FRBR), resource description and access (RDA), semantic web, machine readable cataloging (MARC)

## 1. Introduction

Over the last several decades, the library community has been faced with the challenge of remaining relevant as an authoritative source of bibliographic data within the larger networked environment of the Web. This relevance has particularly been tested by what a number of information professionals see as the library community's reliance on resource description such as Machine Readable Cataloging (MARC), which do not fully support the establishment of relationships between resources across the Web at large nor optimize library data for machine readability. As a result, the vast majority of bibliographic data held in libraries has been locked in library catalogs, which, although automated, essentially function as electronic equivalents of the physical card catalogs of a hundred years ago [1].

However, due to the rapidly changing technology environment, there is now the opportunity for the library community to expose the data created by cataloging

and metadata professionals and to establish interconnections to related resources across the Web [2]. Newer technologies, such as developed by the World Wide Web Consortium's (W3C) linked open data (LOD) initiative under the banner of the Semantic Web, offer libraries the potential to permit library data to be read and indexed by major online search engines, enhancing user access to authoritative sources of bibliographic data, as has been the library community's historic role to create. As the World Wide Web Consortium defines it, the Semantic Web "is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [3]. In other words, the Semantic Web is a method whereby those who are creating content on the Web can markup this content with specific types of metadata in such a way that machines, meaning Web browsers and other applications, can better understand it and use it in novel ways.

Already a number of prominent libraries have developed projects that have published library data that are in compliance with Semantic Web principles, including the Swedish National Library, the French National Library (BnF), the British Library, the Spanish National Library, the German National Library as well as the OCLC [2]. Additionally, implementation of Semantic Web technologies like W3C's Resource Description Framework (RDF) within the library community holds the potential for enriching user experience by permitting users to explore the diverse interconnections between resources through optimizing the machine readability of library data. Lastly, by altering the cataloging process to conform to LOD standards, libraries are afforded the opportunity to reduce cataloging costs through a reduction in duplicate cataloging efforts and to better leverage existing bibliographic data produced elsewhere.

In response to these challenges and opportunities, the Library of Congress (LOC) has developed a high-level model of bibliographic description called the Bibliographic Framework Initiative or BIBFRAME, which aims not only to replace MARC but to provide a framework for optimizing library data within the networked environment. BIBFRAME is essentially an entity-relationship model which uses the Web as architecture and a Resource Description Framework/Extensible Markup Language (RDF/XML) serialization for the description of bibliographic resources. It involves a radical reconceptualization of bibliographic description, eliminating the static, bibliographic record as the product of cataloging in favor of a series of machine readable statements that result in a graph of interconnected entities.

The purpose of this paper will be to examine the development of BIBFRAME through a comprehensive review of relevant literature. We will begin with an overview of BIBFRAME by LOC, outlining the history and structure of the model [in Section 2]. We will then examine the relationship of BIBFRAME to other relevant bibliographic models and content standards including MARC [in Section 3.1], Functional Requirements for Bibliographic Records (FRBR) [in Section 3.2], Resource Description and Access (RDA) [in Section 3.3], and Semantic Web [in Section 3.4]. We will highlight areas of compatibility as well as areas of incompatibility when known. Then, we will end the paper with some concluding remarks.

## **2. History and overview of BIBFRAME**

Officially established in 2011 by the Library of Congress, the Bibliographic Framework Initiative, or BIBFRAME, is a high-level model designed to facilitate the bibliographic description of information resources as well as the exchange of bibliographic data in the networked environment. In 2012 the Library of Congress contracted Zepheria, a consulting firm that specializes in the deployment of semantic

web technologies, to assist with the development of the model. In addition to its work with the Library of Congress, Zepheria has also played, in partnership with Google, Yahoo, and Bing, a key role in the development of Schema.org, a common set of web developer metadata schemas designed to describe websites in support of the indexing efforts of the Internet's major search engines. Over its brief history, BIBFRAME has produced and published a vocabulary for the model, a number of discussion papers related to the vocabulary or other aspects of BIBFRAME implementation, and tools for data conversion.

In its essence, BIBFRAME is an entity-relation model similar to the model put forth in the Functional Requirements for Bibliographic Description. As such, it consists of entities and attributes designed for the description of resources typically managed by cultural heritage institutions. As a result of this entity-relation model, BIBFRAME emphasizes its focus on capturing data elements relevant to bibliographic description, such as title, author, publisher, etc., instead of the creation of complete bibliographic records, which has historically been the focus of the library community. In this way, BIBFRAME establishes a framework for bibliographic description that clearly separates information related to the intellectual contents of resources from their physical properties.

Within this entity-relation model, BIBFRAME is further modeled within RDF/XML in order to bring the model in-line with Semantic Web principles. The use of RDF/XML allows users of the model to identify entities and to describe the relationships between them more clearly and completely. Moreover, it permits these relationships be processed more easily by machines, making library data more conducive to the Web environment. In other words, it allows library data to be found more easily by Internet search engines and, by extension, users. At the heart of this development is the use of Universal Resource Identifiers, or URIs, to name entities and data values, instead of text strings. Thus, the entire BIBFRAME vocabulary of entities and properties has been rendered in URI form.

In summary, BIBFRAME utilizes Web architecture for the description, maintenance, and exchange of bibliographic data in order to accomplish three primary goals [4]:

1. Differentiate clearly between conceptual content and its physical manifestation(s) (e.g., works and instances).
2. Focus on unambiguously identifying information entities.
3. Leverage and expose relationships between and among entities.

## **2.1 The BIBFRAME model**

The newest BIBFRAME model, version 2.0, consists of three core class entities [5, 6]. These are defined below:

- Work: “a resource reflecting a conceptual essence of the cataloged resource” [5]
- Instance: “a material embodiment of a work” [5]
- Item: “an actual copy (physical or electronic) of an instance” [5].

As these entities and their definitions make clear, BIBFRAME, like FRBR, separates the intellectual content of a resource (creative work) from its physical

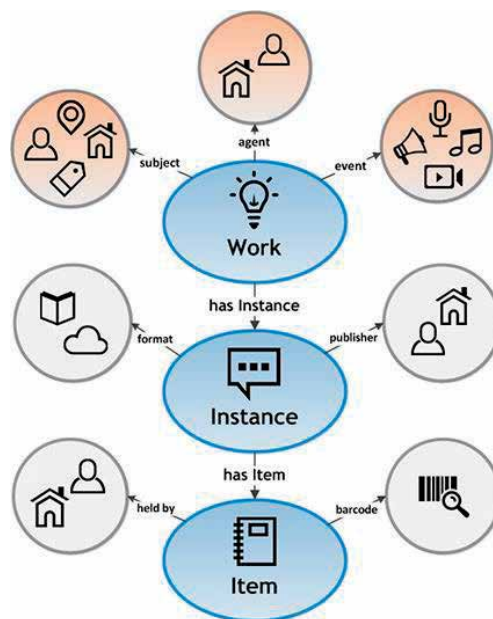
realization (instance). However, instead of FRBR's four entity classes (work, expression, manifestation, and item), BIBFRAME models only three. Thus, although BIBFRAME and FRBR are conceptually related, it appears that BIBFRAME has simplified the number of entity classes required for bibliographic description.

Below (**Figure 1**) is a graphical depiction of the BIBFRAME model that highlights the relationships between these core entities.

While presenting the evolution of the latest version of BIBFRAME 2.0 from the previous version, McCallum reports the participation of vendors in linked data: "Another major step is now beginning to happen as the vendors who supply many of the services in the community have started to explore linked data, and they are the community's essential innovators" [7, p. 84].

BIBFRAME offers a significant amount of flexibility with resource description. However, per the BIBFRAME documentation, other relationships can also be described. Namely, works can be related to works, instances to instances, works to instances, and instances to works [8]. Beyond the main classes of entities, BIBFRAME also includes a number of properties that are related to each entity. For instance, the creative Work class contains properties that, as one researcher notes, reflect traditional bibliographic elements such as title, creator, language, etc. [9] as well as specific resource Work types that can be used to increase the granularity of a work's description. These properties include resource-type concepts like audio, text, and movingimage.

The instance class contains properties which serve to describe the physical "embodiment" of resources. These properties include terms that overlap with those of the work class such as title and creator, as well as those that describe the aspects of a resource at the manifestation level, such as publisher [9]. Although there is overlap in terminology between the work and instance class, the modeling of these properties in RDF/XML serves as a means to disambiguate terms with the same name through the assignment of a specific URI. Thus, despite identical text names, the use of URIs serves to identify properties within their specific classes.



**Figure 1.**  
Graphical depiction of BIBFRAME model [5].

To put it plainly, BIBFRAME attempts to be content standard and model agnostic. Its framework is intended to be flexible enough to accommodate existing models (FRBR, MARC, etc.) and content standards (RDA, VRA, DACS) as well as models and standards that have yet to be developed. Thus, it appears that BIBFRAME appears to be poised to provide the library community with a new model of bibliographic description and exchange that takes full advantage of the Web as architecture. Furthermore, the model also promises to make library data more visible on the Web, not only to the benefit of users looking for library resources but also for re-use in contexts outside of the library community. Finally, it appears that BIBFRAME will permit the full description of relationships between and among resources, enhancing user experience of library information.

## **2.2 BIBFRAME profiles**

It is worth noting that the high degree of flexibility and extensibility built into the model comes with a cost. The under-specification of the model, which is what lends it flexibility, means that there are no built in mechanisms within the model or its RDF schemas that guide and constrain the generation of BIBFRAME data [10]. Nevertheless, the initiative proposes the use of BIBFRAME profiles to address this issue. A BIBFRAME profile can be understood as “a document, or set of documents, that puts a Profile (e.g. local cataloging practices) into a broader context of functional requirements, domain models, guidelines on syntax and usage, and possibly data formats” [10]. In other words, a BIBFRAME Profile serves as a kind template for the generation of BIBFRAME descriptions through the establishment of metadata structure and value constraints. BIBFRAME data can be validated against relevant profiles in order to ensure conformance to an established metadata structure.

However, it should be noted that BIBFRAME profiles exist externally to the model and must be developed within the context of local needs and practices, likely within an application used by cataloguers to capture bibliographic data. In other words, a BIBFRAME profile matches the metadata structures needed within a given context. As long as the overall structure of the data conforms to the BIBFRAME model, then that data should remain interoperable on the Web. Thus, it appears that the initiative is attempting to balance the need for a flexible structure within the model itself and the need to contain that flexibility within a viable framework that can produce consistent and reliable data at the local level.

The study in [11] compares locally created Dublin Core metadata scheme-based application profiles from a number of institutions and digital projects (n = 8). The results of the study present the commonalities and variations of locally developed application profiles and shed light on the effects of resource type and subject domain on naming conventions. The experiences and lessons drawn from the implementation processes of locally developed metadata application profiles are invaluable in the sense that they offer insights and efficient mechanisms for metadata planning and reuse. Thus, the study may shed light on the development of BIBFRAME application profiles in local practice settings.

## **3. Relationship of BIBFRAME to prevailing content standards and models**

It is the intention of the BIBFRAME initiative to design the model in such a way that it not only can serve as the standard encoding and interchange format of bibliographic data within the library community but also to be a model for integrating

library data within the Web environment more generally. As such, the model is designed with a high degree of flexibility in the hope that it can accommodate any number of existing models as well as models yet to be developed. Put simply, the model's flexibility is intended to foster extensibility. The following sections will discuss the relationship of BIBFRAME to the prevailing content standards and models employed by cultural heritage institutions, or those in the process of being adopted by cultural heritage institutions, in an effort to determine the degree to which BIBFRAME, as it is currently understood, can be a viable and extensible framework for bibliographic description and exchange in the Web environment.

### **3.1 Machine readable cataloging (MARC)**

BIBFRAME is intended to replace MARC as the encoding and exchange format for the bibliographic data produced by the library community. But why? What is it about MARC's design that requires the format to be replaced?

First of all, the design of MARC can perhaps be best understood as an exchange format which emphasizes the display of bibliographic information about specific library holdings within electronic catalogs. As a result of this emphasis, MARC records can be conceived as aggregates of information that include descriptions of both the conceptual essence of resources as well as aspects of their physicality [4]. These aggregates are realized in the cataloging process through the application of content standards such as AACR2 and now RDA and are captured, for the most part, in a series of tagged literals or tagged text strings. Ultimately, the overarching structure of MARC records and the content rules used to realize them serve as means to display bibliographic data in much the same way as the physical card catalogs which were its predecessor [1]. MARC's design has served the library community well over the years and has, as the Library of Congress points out in their introductory paper on the BIBFRAME model, allowed librarians to accomplish three important bibliographic tasks [4]:

1. To capture information about the intellectual essence of resources
2. To capture information on the physical aspects of resources
3. To capture information about the management of resources such as control numbers and record handling codes

However, within the current context of the Web environment coupled with the increased processing capabilities of modern computers and applications, MARC's design presents the library community with a number of structural difficulties that limit the potential uses of bibliographic data. First of all, MARC's reliance on the use of literals as identifiers for resources and the elements that compose bibliographic records limits the ability of machines to process MARC information [4]. As a result, variations or equivalences of literals are difficult for machines to parse. Secondly, MARC does not separate information regarding the intellectual content of a resource and its physical carrier clearly enough [4]. Even with adjustments to MARC, such as those included in RDA, an FRBR-based content standard that makes a clearer distinction between the content and carrier, the very format of MARC will not allow machines to utilize it fully [12]. Thirdly, the structure of MARC records, although information rich, are poor at expressing relationships between bibliographic elements in ways that machines can easily understand [13]. Again, even with adjustments to MARC, such as MARC/XML, a serialization intended to increase the machine readability of MARC records, the use of content standards



like AACR2 which were developed primarily with display issues in mind prevents the processing of MARC data significantly [14]. Ultimately, this means that library data is unable to interact with the vast majority of computer applications automatically, limiting the exposure of bibliographic data on the Web, preventing the rich relationships between data elements from being realized and effectively hiding bibliographic information from online users.

BIBFRAME is designed to address these issues. To begin, as one researcher notes, BIBFRAME is not only designed to replace MARC as an encoding and exchange format but to offer a complete re-conception of bibliographic description itself, one that is in-line with the capabilities of the Web environment [15]. BIBFRAME accomplishes this in a number of ways. First, BIBFRAME replaces the idea of the catalog record with the notion that a resource is defined by a discrete series of bibliographic elements. These elements clearly distinguish between the intellectual content of a resource, its physical carrier, and the various entities responsible for its production. Freed from the record as a bundle of data elements, the individual elements are better able to interact in computer applications, and the cataloguer is better able to describe relationships between elements. Secondly, text strings or literals are replaced by URIs or Universal Resource Identifiers. By using URIs to identify bibliographic elements and their values, machines are better able to process the bibliographic information and to utilize the relationships described between them. These two elements, when built upon a Web-based architecture and serialized in RDF/XML, permit BIBFRAME bibliographic data to interact more freely on the Web.

However, despite these changes and the claim that it is standard agnostic, the BIBFRAME initiative also claims that BIBFRAME will be backwards compatible with MARC, meaning that MARC will be mapped to BIBFRAME in such a way that MARC data can be automatically converted to BIBFRAME data without loss of information. Indeed, the BIBFRAME initiative has already developed tools that are available on its website which can translate MARC data into BIBFRAME 2.0 (Figure 2) [16]. As the relationship between MARC elements and BIBFRAME entities may be complex, may even be many-to-many, as one researcher notes [17], the success of such a mapping remains to be seen.

The screenshot shows the 'BIBFRAME Comparison Tool' interface. At the top, it says 'Compare MARCXML to BIBFRAME2'. Below this is a search bar with the identifier '8226' and a 'Search' button. The main area is split into two columns: 'MARC' on the left and 'BIBFRAME (Turtle)' on the right. The MARC column displays a record for 'Snoopy on wheels / by Terry Flanagan.' with various fields like 010, 020, 030, 040, 042, 050, 060, 070, 080, 090, 100, 245, 260, 300, 400, 500, 600, 700, 800, 900, and 970. The BIBFRAME column shows the corresponding RDF/XML representation of this record, including URIs for the title, author, and subject, and a list of identifiers.

**Figure 2.** Screenshot of the BIBFRAME comparison service results page showing MARC data (left) and BIBFRAME RDF/XML data (right) for Terry Flanagan's Snoopy on wheels.

### 3.2 Functional requirements for bibliographic records (FRBR)

Published in 1998 by the International Federation of Library Associations (IFLA), the final draft of the Functional Requirements for Bibliographic Records provided a radical re-conception of bibliographic description. In essence, FRBR is an entity-relation model which is composed of four primary classes (work, expression, manifestation, and item) that separate the intellectual content of resources from various aspects of their physical properties, resulting in a new emphasis on the component pieces of bibliographic data rather than the bibliographic record as a whole [15]. As BIBFRAME, with its three primary entity classes (work and instance and tem), is related, at least superficially to FRBR, and considering the likelihood of FRBR's international acceptance as the standard model of bibliographic description, it is useful to compare the two models to determine the degree of compatibility and potential interoperability.

At least on the surface, BIBFRAME and FRBR appear to be closely related. Both models employ the entity-relation approach to bibliographic description and divide the bibliographic record into component pieces which are attached as attributes to entities. As noted, FRBR defines four primary entities for bibliographic description. These are as follows:

- **Work:** “a distinct intellectual or artistic creation” [18]. As such, a work is abstract, pertaining to the intellectual content of a resource as separate from its physical existence. For example, Shakespeare’s *Romeo and Juliet* is a work apart from all of the various editions (print and electronic), performances, and films that have embodied it.
- **Expression:** “the intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation, sound, image, object, movement, etc., or any combination of such forms” [18]. For example, the English text of *Romeo and Juliet*, as separate from the various ways it is presented in different editions is an expression of the work.
- **Manifestation:** “the physical embodiment of an expression of a work” [18]. For example, the 1998 Signet Classics edition of *Romeo and Juliet* is a manifestation. In other words, when the expression of a work takes on a physical form, as text, film, sound recording, etc., it becomes a manifestation.
- **Item:** “a single exemplar of a manifestation” [18]. For example, an item is a single copy of the 1998 Signet Classics edition of *Romeo and Juliet*.

As can be seen, the FRBR main entities represent a hierarchical movement from abstraction to specificity of a particular information resource [17]. In a similar fashion, BIBFRAME is constructed of entities in a hierarchical fashion, but instead of FRBR's four levels, BIBFRAME defines three [4]:

1. **Work:** “a resource reflecting a conceptual essence of the cataloged resource”
2. **Instance:** “a material embodiment of a work”
3. **Item:** “an actual copy (physical or electronic) of an instance”

Thus, although BIBFRAME only uses three main entity classes, there is still the same movement from abstraction to specificity as represented in the FRBR

hierarchy. Nevertheless, the lack of conformance to the FRBR hierarchy has resulted in much discussion, and, perhaps, even some confusion about how BIBFRAME relates to FRBR. For instance, there appears to be some disagreement in the literature regarding the exact relationship between BIBFRAME and FRBR entities, especially with regard to how the BIBFRAME entities may represent confluences of FRBR entities. Although a number of researchers espouse a correspondence between the BIBFRAME work entity and the FRBR entities work and expression [13, 15, 16, 19], at least one researcher sees a correspondence only between BIBFRAME Work and FRBR Work [20]. Similarly, it appears that most researchers see a correspondence between BIBFRAME instance and FRBR manifestation entities [13, 15, 19], while others see a correspondence between BIBFRAME instance and FRBR manifestation and expression [20].

Perhaps some of the difficulty of mapping BIBFRAME to FRBR lies in the basic ambiguity of the meaning of the respective concepts. For instance, as is noted by IFLA, the FRBR concept of work is an abstraction, meaning that it is hard to define its “precise boundaries” and that the divisions between works and between works and expressions may in fact be culturally dependent [18]. Furthermore, as other researchers have noted, efforts at operationalizing the concept of work have led to at least two different conceptions of the concept. For instance, some have argued that a work can be conceived as the intellectual content of an endeavor with no “assumptions about how it is physically realized,” while, from a different point of view, a work can be conceived as the sum of all common attributes (author, title, etc.) from a set of manifestations [17]. Perhaps complicating the matter is fact that neither BIBFRAME’s nor FRBR’s hierarchy constitutes a definable bibliographic whole. For instance, although FRBR’s entities are organized hierarchically, and are often pictured within a box, there is no single concept to which this hierarchy relates [19]. The need for a kind of super-entity has been noted well in the literature [19]. It would seem that these questions regarding FRBR are equally applicable to BIBFRAME since BIBFRAME does not include a super-entity that encapsulates the work and instance entities. Thus, it appears that there may still be some serious conceptual difficulties that need to be overcome if BIBFRAME, as an entity-relation model, is to be a viable framework for bibliographic description.

Nevertheless, because BIBFRAME appears to be a simplified version of FRBR, perhaps some of the conceptual difficulties regarding FRBR will not negatively affect BIBFRAME as much. For instance, perhaps BIBFRAME’s conflation of FRBR’s work and expression concepts is useful since it is sometimes difficult to determine the boundaries between a work and its expression. However, since the BIBFRAME initiative has suggested that its model is agnostic, meaning that it can be applied to any model, it must be able to be mapped clearly to other models if it is to foster interoperability. Yet, as one researcher notes, to make the model completely agnostic may be unrealistic, since to be perfectly interoperable, both models require almost equivalent semantics and granularity, a situation which would suggest the redundancy of one of the models [2]. This does not seem to be the case between FRBR and BIBFRAME, which means that the initiative may need to re-examine the possibilities of BIBFRAME working with other models.

### **3.3 Resource description and access (RDA)**

BIBFRAME is designed to be content standard agnostic, meaning that the model does not include requirements or specifications for the use of any particular content standard for bibliographic description. In fact, per the initiative, BIBFRAME is intentionally underspecified so that any content standard may be applied successfully within the context of the model, including those that have yet to be developed [4].

Thus, this intentional under-specification is designed to maximize the extensibility of the model and to help ensure its usefulness in a wide range of extant and future information management contexts and use scenarios, as well as for the widest variety of current and future resource types [4].

However, since the BIBFRAME initiative has positioned the model to be the replacement for MARC as the primary method of bibliographic description and data exchange between libraries, the initiative is doing more than simply ensuring the openness of the model to accommodate RDA and other content standards. Per the initiative, the designers are planning on taking an active look at the elements in RDA and other content standards, including the *Anglo-American Cataloging Rules, Second Edition* (AACR2). As a number of researchers have noted, it appears that BIBFRAME is also being designed to specifically accommodate RDA [1, 13, 20], which suggests that this particular content standard may be playing a stronger role in the design of the model than may have been suggested initially. As BIBFRAME is still under development, it remains to be seen exactly to what degree RDA plays a role in the design of the model and what effects this might have on the model's extensibility.

Nevertheless, BIBFRAME designers suggest that the use of profiles will be another way to accommodate a variety of content standards within the model. A BIBFRAME profile is "a document, or set of documents, that puts a Profile (e.g., local cataloging practices) into a broader context of functional requirements, domain models, guidelines on syntax and usage, and possibly data formats" [10]. According to the initiative, such profiles can be used to define constraints in the creation of BIBFRAME records such as those required by any content standard, including RDA.

As other researchers have noted, RDA may not have gone far enough in distinguishing the content from the carrier of information resources [1, 14]. This potential fundamental flaw in the content standard may pose further difficulties in mapping RDA to BIBFRAME. Such difficulties are presented in the study [21] which shows the uneven mapping between existing RDA classes and BIBFRAME 2.0— particularly the RDA Expression class. The study demonstrates many-to-many relationships in the mapping between RDA and BIBFRAME. Nevertheless, as BIBFRAME is in a relatively early stage of development, the nature and magnitude of these difficulties remain to be seen.

### **3.4 Semantic web**

The current Web environment is structured in such a way that machines, and thus users, are unable to take full advantage of the links that are established among and between resources. In other words, the Web is an environment composed of Web pages and hypertext links that do not describe the nature of the links that connect pages together nor the nature of the data (content) contained in Web pages. In other words, as many researchers note, the current web is a "Web of Documents" versus a "Web of Data" [22, 23]. As a result, current search mechanisms, such as the major search engines, are limited in their ability to utilize information on the Web, relying almost solely on harvesting algorithms to index the content of Web pages and then to match this indexed information against the search terms entered by users. While, as one researcher notes, this method has served the Web well, permitting users to locate needed resources within the vast sea of online information, it lacks the ability to lead users to related content, even when complex and intelligent relevancy algorithms are employed [14]. Furthermore, within the context of the library community, it means that most library data remains relatively difficult to locate online and relatively static with regard to other online resources relevant to

library holdings. In other words, library data, in its current form, remains in the proverbial silo of its online catalogs.

However, through the employment of Semantic Web technologies, there is the potential to expand the uses of library data in the Web environment and thereby to enhance user experience of this data. As is commonly the current case on the Web, a typical hyperlink connects resources but the nature of the connection remains unexplained. However, through the use of Semantic Web and Linked Data principles, such as the use of URIs to identify resources and the embedding of URIs in RDF statements, the nature of these connections can be exposed. In this scenario, a hyperlink can then be defined in almost any way that the user can imagine, indicating the link points to a reference, an author, a subject an authority, etc. Machines can then use this data to “infer” other resources that have been described similarly, such as resources with the same subject heading as the one in question, and permit users to explore these relationships more readily.

At the heart of the Semantic Web are four principles that Tim Berners-Lee, inventor of the World Wide Web and founder and director of the W3C, set forth in his paper entitled “Linked Data” [24]. These principles define the nature of Linked Data as it can be implemented in the current Web environment. Furthermore, they serve as a framework and guide for those interested in making their Web content viable within the Semantic Web, as some conformance to a standard model is required for successful implementation. These principles are as follows:

1. Use URIs as names for things [24].
2. Use Hypertext Transfer Protocol (HTTP) URIs so that people can look up those names [24].
3. When someone looks up a URI, provide useful information, using the standards (RDF\*, SPARQL) [24].
4. Include links to other URIs, so that they can discover more things [24].

Perhaps most significantly, the conception of Linked Data requires the use of URIs to identify resources or, more specifically, the data elements of resources (Principle 1). In other words, as was mentioned in the discussion on MARC above, the use of text strings to identify resources makes machine processing difficult. The shift to URIs as identifiers means that machines can better understand the identity of resources, especially if they are known by different names or to disambiguate different resources known by the same name. Furthermore, the shift to URIs also signals the shift in understanding in regards to the nature of information resources as described in the above FRBR section. It emphasizes the identification of discrete data elements within information resources versus the identification of the resource as a whole. In other words, it emphasizes the atomization of resources into their relevant components.

Principle 2 emphasizes the need for a common schema for the definition of URIs. Since HTTP is already the foundation of data transfer on the Web and since it appears to be serving its function well, Berners-Lee suggests that using this common protocol for the definition of URIs will increase the usefulness of data described in Semantic Web compliant ways. Furthermore, as the BIBFRAME initiative notes, these URI schemes should not be obscure, even if they are represented in HTTP, in order to facilitate data interaction and reuse [4].

Principle 3 emphasizes the need for a common framework for the exchange of information described with URIs. Typically this means the use of RDF for the

modeling of data, which, as the BIBFRAME initiative notes, is the most common framework within the LOD community [4]. As a conceptual framework for representing resources on the Web [15], RDF can be understood as a kind of syntax for structuring data in such a way that it fosters the machine readability of that data through the use of URIs and the delineation of relationships between data elements. RDF is typically rendered in XML, but other languages, such as N3, Turtle, and N-Triples, are also used [22]. In its basic format RDF consists of statements, called triples, which, like sentences, contain subjects, predicates, and objects. A basic RDF statement might read as “Book A (subject)—Written By (predicate)—Author A (object),” where Book A, Written By, and Author A are all identified by URIs, with the possible exception of the object, which could be populated with a text string [22]. The power of this model is that the type of relationships between resources (Book A and Author A) is defined (Written By). **Figure 3** illustrates this statement graphically. Thus, as a result of delineating relationships between data elements, tools called “reasoners” can make inferences about the data [19].

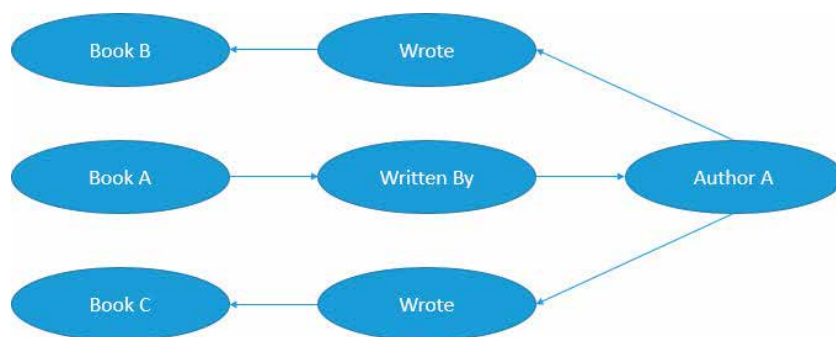
A reasoner is a software application that can make logical inferences based on a set of statements, or axioms, provided to it through queries. Although there are many query languages that can be used to access and manipulate data modeled in RDF, the SPARQL Protocol and RDF Query Language (SPARQL) has emerged as the most popular [23]. For instance, a reasoner, beginning with a SPARQL query to a database that contained the above RDF statement, could use that statement to make inferences about other books written by Author A and present those to users without the user specifically querying the system to do so (**Figure 4**). Furthermore, there are no restrictions on the number of RDF triples that can be created for a particular resource, which fosters the development of rich data graphs, or the decentralized interconnections between data elements, within the Web environment. Although RDF is not a data format, but a model for representing data elements on the Web, it has been serialized in a number of ways. For instance, BIBFRAME has been modeled in RDF/XML, but other languages, like N-Triples, ATOM, and JSON, also exist. Although BIBFRAME has been modeled in RDF/XML, the Initiative claims that any data format that conforms to the standard model of URIs embedded in triples should be compliant with the BIBFRAME model [4].

Principle 4 encourages broad use of the connections established through the first three principles [4]. Thus, data that has been described in conformance with the above principles can be considered Linked Data and Semantic Web compliant. However, if the URIs expose, point to, or otherwise include information that is made freely available for reuse on the Web, such as through a Creative Commons license, this data can be considered Linked Open Data, not just Linked Data.

As stated earlier, a number of prominent libraries have published library data in compliance with Semantic Web principles [2]. Even though these projects are not BIBFRAME projects, they are generally in-line with FRBR principles of bibliographic description. It is worth examining the degree to which the model conforms to the current understanding of Linked Data and the Semantic Web. To begin, BIBFRAME has defined URIs for all BIBFRAME entities and properties within the BIBFRAME namespace. This is particularly important as some properties that belong to different classes have identical names. The use of URIs serves as a clear



**Figure 3.**  
*Graphical depiction of a basic RDF statement.*



**Figure 4.**  
*Graphical depiction of a reasoner using RDF statements to infer additional resources.*

means to disambiguate these properties. Secondly, as has been noted, BIBFRAME has been modeled in RDF/XML [25].

In addition to these two factors, the BIBFRAME model, like FRBR, deconstructs bibliographic records into their component pieces through the entity-relation conception of bibliographic description. Taken together, these elements suggest that BIBFRAME conforms well to the current understanding of Linked Data and the Semantic Web. Furthermore, even though the initiative has rendered the model in RDF/XML, BIBFRAME is also designed to be compliant with other data formats which conform to the structured use of URIs within syntax of triples statements. Thus, it also appears that BIBFRAME is, at least in principle, poised to integrate library data with other data produced within contexts outside the library community. This aspect too suggests that BIBFRAME is Semantic Web friendly.

#### 4. Discussion

There are challenges that may hinder the widespread adoption of BIBFRAME within the library community. In addition to the modeling difficulties and potential conceptual misalignment of BIBFRAME in relation to MARC, FRBR, RDA, Linked Data, and RDF, there are difficulties posed by complex resource types such as audiovisual materials, manuscript, and serial publications [26]. Additionally, although MARC is in essence an exchange format for bibliographic data, it has become so intertwined with the content standards applied to it, first AACR2 and now RDA; this union of the two may further entrench it within the library community. Without consensus regarding the fate of MARC, it may be difficult to persuade MARC's adherents, even if BIBFRAME proves to offer more capabilities to catalogers.

There may be significant conceptual difficulties with mapping RDA to BIBFRAME. For instance, RDA was developed within the context of the FRBR entity-relationship model. As such, RDA separates resources into FRBR's four main entity classes: Work, Expression, Manifestation and Item. However, as has already been noted, BIBFRAME's main entity classes do not align with FRBR's classes in an exact manner [20]. This lack of alignment may make the mapping between RDA and BIBFRAME difficult.

Although it appears that BIBFRAME conforms to current conceptions of Linked Data and the Semantic Web, there are still a number of issues worth considering. First, since the usefulness of the relationships delineated through the RDF triples depends on the quality and stability of the resources to which they are linked, the BIBFRAME initiative will have to determine the degree to which it will maintain



its own controlled vocabularies and ontologies versus relying on others to do so. Ontologies suitable for the Linked Data environment are taxonomies and thesauri that meet the W3C Web Ontology Language (OWL) standard [22]. For example, the Library of Congress Subject Headings modeled in the Simple Knowledge Organization System (SKOS) framework is an OWL-compliant ontology.

The existence of high-quality, stable ontologies is particularly a relevant concern with regard to the use and reuse of Linked Open Data resources. For instance, as one researcher notes, many LOD ontologies and vocabularies are developed in the context of research projects, which means that for a particular moment they may be up-to-date, accurate, and in compliance with current standards, though it does not ensure continued governance and maintenance [12]. Thus, the reliance on such vocabularies could present the threat of obsolescence should governing bodies discontinue their activities. Thus, it appears that BIBFRAME will need to assess the stability of ontologies and vocabularies, such as those for resource type, and determine if it is better to develop and maintain its own within the BIBFRAME namespace or to link to resources outside the initiative.

Secondly, although BIBFRAME claims that the model should be interoperable with any serialization using triples and URIs, the fact that the initiative has serialized the model in RDF/XML may be a limitation. In other words, because the initiative has limited its serialization within a single framework, it may discourage implementation in other formats. As one researcher notes, it may be better for the initiative to provide potential implementers with examples from a number of possible serializations in order to demonstrate the model's flexibility, extensibility, and potential for interoperability [2].

Thirdly, there may be difficulties with viably implementing the BIBFRAME model which are rooted in the nature of RDF itself. As the study in [19] notes in their comparison of BIBFRAME, FRBR, and RDA, there is nothing in RDF that prevents people from making nonsensical RDF triples. In other words, there are no validation mechanisms for the creation of RDF statements, as there are for well-formed XML or HTML documents. While, as the researchers note, BIBFRAME has proposed the use of profiles in order to establish content rules and constraints on the creation of BIBFRAME records, these do not prevent potential difficulties with the integration of BIBFRAME data elements with data elements modeled in other frameworks such as FRBR.

However, perhaps the biggest threat to BIBFRAME as a mechanism to expose library data in a Semantic Web friendly way lies in the fact that, like the framework itself, the Semantic Web is still under development. For instance, as has been noted in the literature, understanding of what actually constitutes Linked Data is still under debate [19]. Since the very underpinning of the Semantic Web is still in flux, there is a possibility that any operationalization of the concept will change in the future. Thus, if the current methods for creating Linked Data alter significantly in the future, and if data described with current methods cannot be easily translated into the newer modes, then BIBFRAME Linked Data could potentially become obsolete, resulting in the relegation of library data to yet another, but different, silo.

This final point may also be exacerbated by the very fact that BIBFRAME is a model for the description of bibliographic data within the library community itself. For instance, as some researchers have noted, for data to be truly integrated in the Web, what is required is a common model for data description that includes not only bibliographic data but data of all types [2]. In other words, BIBFRAME, as a model for the description of bibliographic data, may not be intuitively understood by others outside the library community, which may result in a lack of implementation and difficulties with the integration of data embedded in other frameworks.

This is particularly important as BIBFRAME data is intended for use outside of the library community, especially with regard to the authority data such as controlled subject headings that have been the province of the library community for so long [2, 13]. Thus, while BIBFRAME holds the promise of freeing library data from the silos of online catalogs and to permit library data to interact with data both within and outside the library community, there may still be challenges to overcome in order to optimize these capabilities.

## 5. Conclusion

It is the intention of the BIBFRAME initiative to design the model in such a way that it not only can serve as the standard encoding and interchange format of bibliographic data within the library community but also be a model for integrating library data within the Web environment more generally. As such, the BIBFRAME model is designed with a high degree of flexibility that can accommodate any number of existing models as well as models yet to be developed within the Web environment. The model's flexibility is intended to foster extensibility.

However, regarding the model itself, there appears to be a significant need to consider the creation of a super-entity that would encapsulate the work and instance entities. With regard to the cataloging requirements for the description of complex resources such as audiovisual materials and serial publications, the creation of such a super-entity would solve a number of bibliographic description challenges. The existence of a super-entity would permit the description of resources and relationships that are currently difficult to model within the existing framework. Resources that do exhibit intellectual content or that are primarily event based would be easier to depict if such a super-entity was present.

BIBFRAME attempts to be content standard and model agnostic. Its framework is intended to be flexible enough to accommodate existing models. While increasing its extensibility, the framework may also result in an uncertainty of its application in specific cataloging contexts. This too may limit the willingness of the library community to invest in its adoption. Furthermore, even though BIBFRAME's potential for extensibility is intended to foster its adoption in a wide range of bibliographic contexts and to work equally well for divergent descriptive needs, its ability to accommodate most if not all modeling and content standards currently in use or yet to be invented may be optimistic. In this regard, BIBFRAME's ability to support widespread interoperability needs to be further addressed.

In this study we discussed the relationship of BIBFRAME to the prevailing content standards and models employed by cultural heritage institutions in order to determine the degree, to which BIBFRAME can be a viable and extensible framework for bibliographic description and exchange in the Web environment. Despite the promise of improved data management, sharing, and usage offered through the BIBFRAME model, there are various challenges that must be overcome for its adoption within the library community. However, if the initiative can overcome what will likely be significant challenges to the implementation of the model, BIBFRAME appears to be poised to become the next standard of bibliographic description and exchange for the library community and beyond. Furthermore, the model also promises to make library data more visible on the Web, not only to the benefit of users looking for library resources but also for reuse in contexts outside of the library community. Finally, it appears that BIBFRAME will permit the full description of relationships between and among resources, enhancing and enriching the user experience of library information.

### **Author details**

Jung-Ran Park\*, Andrew Brenza and Lori Richards  
The College of Computing and Informatics, Drexel University, Philadelphia, USA

\*Address all correspondence to: [jp365@drexel.edu](mailto:jp365@drexel.edu)

### **IntechOpen**

---

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Dean JW, Charles AC, Edward T. Coming to a library near you, via BIBFRAME. In: *The Library with the Lead Pipe*. 2013. Available from: <http://www.inthelibrarywiththeleadpipe.org/2013/charles-a-cutter-and-edward-tufte-coming-to-a-library-near-you-via-bibframe/> [Accessed: 02 June 2020]
- [2] Svensson LG. Are current bibliographic models suitable for integration on the web? *Information Standards Quarterly*. 2013;**25**(4):6-13
- [3] World Wide Web Consortium. The semantic web made easy [Internet]. Available from: <http://www.w3.org/RDF/Metalog/docs/sw-easy.html> [Accessed: 02 June 2020]
- [4] Library of Congress. Bibliographic Framework as a Web of data: Linked Data model and supporting services. 2012. Available from: <http://www.loc.gov/bibframe/pdf/marclid-report-11-21-2012.pdf> [Accessed: 02 June 2020]
- [5] Library of Congress. Overview of the BIBFRAME 2.0 model [Internet]. 2016. Available from: <http://www.loc.gov/bibframe/docs/bibframe2-model.html> [Accessed: 02 June 2020]
- [6] Library of Congress. BIBFRAME 2.0 Vocabulary [Internet]. 2016. Available from: <http://www.loc.gov/bibframe/docs/index.html> [Accessed: 02 June 2020]
- [7] McCallum S. BIBFRAME development. *JLIS.it*. 2017;**8**(3):71-85
- [8] Library of Congress. BIBFRAME relationships [Internet]. 2014. Available from: <http://www.loc.gov/bibframe/docs/bibframe-relationships.html> [Accessed: 02 June 2020]
- [9] Mitchell ET. Three case studies in linked open data. *Library Technology Reports*. 2013;**49**(5):26-43
- [10] Library of Congress. BIBFRAME profiles: Introduction and specification [Internet]. 2014. Available from: <http://www.loc.gov/bibframe/docs/bibframe-profiles.html> [Accessed: 02 June 2020]
- [11] Park J-r, Andrew B, Lu C. A comparative analysis of metadata best practices and guidelines: Issues and implications. *International Journal of Metadata, Semantics and Ontologies*. 2015;**10**(4):240-260
- [12] Fallgren N, Lauruhn M, Reynolds RR, Kaplan L. The missing link: The evolving current state of linked data for serials. *The Serials Librarian*. 2014;**66**(1-4):123-138
- [13] Kroeger A. The road to BIBFRAME: The evolution of the idea of bibliographic transition into a post-MARC future. *Cataloging & Classification Quarterly*. 2013;**51**:873-890
- [14] Breeding M. Linked data: The next big wave or another tech fad? *Computers and Libraries*. 2013;**33**(3):20-22
- [15] Ballegoie MV, Borie J. From record bound to boundless: FRBR, linked data, and new possibilities for serials cataloguing. *The Serials Librarian*. 2014;**66**(1-4):76-87
- [16] Library of Congress. MARC to BIBFRAME comparison viewer [Internet]. Available from: <http://id.loc.gov/tools/bibframe/compare-id/full-ttl> [Accessed: 02 June 2020]
- [17] Godby CJ. The relationship between BIBFRAME and OCLC's linked-data model of bibliographic description: a working paper [Internet]. 2013. Available from: <http://www.oclc.org/content/dam/research/publications/library/2013/2013-05.pdf> [Accessed: 02 June 2020]

[18] IFLA. Study group on the functional requirements for bibliographic records. In: *Functional Requirements for Bibliographic Records*. Munich: K.G. Saur Verlag; 1998

[19] Baker T, Coyle K, Petiya S. Multi-entity models of resource description in the semantic web: A comparison of FRBR, RDA, and BIBFRAME. *Library Hi Tech*. 2014;**32**(4):562-582

[20] Meehan TP. BIBFRAME. *Catalogue and Index*. 2014;**174**:43-52

[21] Taniguchi S. Examining BIBFRAME 2.0 from the viewpoint of RDA metadata schema. *Cataloging and Classification Quarterly*. 2017;**55**(6):387-412

[22] Yang S, Lee YY. Organizing bibliographic data with RDA: How far have we stridden towards the semantic web? In: Park JR, Howards L, editors. *New Directions in Information Organization*. Bingley, UK: Emerald Insight; 2013. pp. 3-27

[23] Yoose B, Perkins J. The linked open data landscape in libraries and beyond. *Journal of Library Metadata*. 2013;**13**(2/3):197-211

[24] Berners-Lee T. Linked data [Internet]. 2006. Available from: <http://www.w3.org/DesignIssues/LinkedData.html> [Accessed: 02 June 2020]

[25] Taniguchi S. Is BIBFRAME 2.0 a suitable schema for exchanging and sharing diverse descriptive metadata about bibliographic resources? *Cataloging and Classification Quarterly*. 2018;**56**(1):40-61

[26] Park J-R, Richards L, Brenza A. Benefits and challenges of BIBFRAME: Cataloging special format materials, implementation, and continuing educational resources. *Library Hi Tech*. 2019;**37**(3):549-565

# TULIP: A Five-Star Table and List - From Machine-Readable to Machine-Understandable Systems

*Julthep Nandakwang and Prabhas Chongstitvatana*

## Abstract

Currently, Linked Data is increasing at a rapid rate as the growth of the Web. Aside from new information that has been created exclusively as Semantic Web-ready, part of them comes from the transformation of existing structural data to be in the form of five-star open data. However, there are still many legacy data in structured and semi-structured form, for example, tables and lists, which are the principal format for human-readable, waiting for transformation. In this chapter, we discuss attempts in the research area to transform table and list data to make them machine-readable in various formats. Furthermore, our research proposes a novel method for transforming tables and lists into RDF format while maintaining their essential configurations thoroughly. And, it is possible to recreate their original form back informatively. We introduce a system named TULIP which embodied this conversion method as a tool for the future development of the Semantic Web. Our method is more flexible compared to other works. The TULIP data model contains complete information of the source; hence it can be projected into different views. This tool can be used to create a tremendous amount of data for the machine to be used at a broader scale.

**Keywords:** data labeling, knowledge discovery, knowledge representation, Linked Data, open data, semantic annotation, Semantic Web

## 1. Introduction

The web has evolved through many stages. Contents on the Web have been changed in both form and method. Searching for information on the Web with keywords is very limited. In the future, we will have a lot of information, causing the traditional search to be insufficient. To perform a better search, the semantic of information must be exploited. One way to represent such concept is to use Linked Data. The conversion of already abundant data into the Linked Open Data format will allow the intelligent search. This extension technology of the Web is called the Semantic Web.

This chapter will briefly introduce the Semantic Web and underlying technologies, including the fundamental element of the Semantic Web called Resource Description Framework (RDF). Currently, there are many different forms of information on the web. So, there is a lot of research related to the conversion of these

contents into a searchable format by the Semantic Web. This chapter discusses open data standards and the making of machine-understandable data.

Many forms of conversion have already been proposed by many research (we will review them in Section 3). However, the conversion of tables and lists is still problematic. We propose a novel method to convert tables and lists to five-star open data with the data model called TULIP. The following sections discuss TULIP vocabulary as well as brief examples of its application.

## 2. Background

Before mentioning the Semantic Web, it is useful to describe the development of the Web in a nutshell. Nova Spivack has explained Web 3.0, the latest development of the Web [1]. The first era is Web 1.0 which consist of contents that can rarely be changed, most of which are generated by research institutions and business organizations. The next era, Web 2.0, has contents that can be changed frequently, and most of them come from the users creating and updating their information, such as Weblog (blog), wiki, social networks, etc. Now we are in the era of Web 3.0 and Semantic Web. It focuses on linking the data between computers together and processing them directly by computers.

### 2.1 Structured, semi-structured, and unstructured data

In terms of simplicity of data processing by computers, the data can be classified into three types:

1. Structured data is the data that has a definite structure, such as data contained in relational databases. This type of data can be directly processed.
2. Semi-structured data is the data that cannot be wholly identified for its structure, such as a table, list, chart, etc. Although humans can see these data as “structured” and can easily understand them, it is not possible for computers to manipulate these data directly because of uncertainty and ambiguity in terms of structure and meaning. It is necessary to convert them by means of various methods before further processing.
3. Unstructured data is the data that has no simple structure, such as text in the form of essays, pictures, audio, video, etc. They must be preprocessed by specific methods, such as natural language processing (NLP) and other methods to convert them into a format that can be manipulated by computers. This type of data has the highest uncertainty and ambiguity.

### 2.2 What is the semantic web

Indeed, the Semantic Web is not all new technology. Tim Berners-Lee, who invented the World Wide Web in 1990 [2], announced the concept of Semantic Web in 2001 in the *Scientific American* article [3]. Semantic Web is an extension of the Web that we currently use in which information is given well-defined meaning. In other words, Semantic Web is a Web of data that can be processed directly or indirectly and “understood” by computers. Steve Bratt, CEO of World Wide Web Consortium (W3C) [4], contrasts the World Wide Web which uses hyperlinks to link various resources between computers connected by the Internet and Semantic Web which uses relationship or “meaning” to link resources or “objects” together.

Each object in the Semantic Web is a part of a huge distributed database on the Internet which can be processed by computers, and results can be presented in a variety of formats as required by users.

In summary, Semantic Web is a technology that is based on the current technology and the Internet. It relies on a set of protocols at different levels that works together to create the distributed data structure on the Web in the form of relationships that linked together across the system through the Internet. An example of the benefits of the Semantic Web is to search for information about proteins that affect the treatment of Alzheimer's disease as currently being studied around the world. If searching using a regular search engine, it may reach about 223,000 documents around the Web. Many of these documents may not be relevant. However, if searching through the Semantic Web, the result is the list of 32 proteins from the Semantic Web of researchers sharing and exchanging information on the disease.<sup>1</sup>

### 2.3 Elements of the semantic web

As with other services on the Internet, most of which is the integration of standard or commonly used components. In the case of Semantic Web, it consists of various components such as Unicode, Uniform Resource Identifier (URI), Extensible Markup Language (XML), and other standards. Some frameworks have been developed, improved, or modified from the existing ones, such as the Resource Description Framework (RDF), RDF Schema (RDFS), Web Ontology Language (OWL), and SPARQL Protocol and RDF Query Language (SPARQL). In this chapter, we mainly focus on RDF and SPARQL.

Resource Description Framework (RDF) is the main structure for storing the smallest components of facts in the knowledge base linked within the Semantic Web. Basically, an RDF is a "sentence" that has three parts: a subject, a verb (or predicate), and an object. Both subject and object will be the identity, i.e., the name of the resource in the form of a URI (in the case of the latter, it can be literal or constant). A predicate (also in the form of URI) describes the relationship between them. These sentences are called RDF triples. The triples are linked together as a graph structure called the RDF graph, which is sometimes referred to as the semantic graph or knowledge graph.

### 2.4 What is linked data and five-star open data

It is said that Semantic Web, though simple, is still not being used extensively [5]. Linked Data is a set of guidelines for disclosing, sharing, and connecting pieces of information or knowledge on the Semantic Web using URI and RDF [6]. The Linked Open Data (LOD) project by Chris Bizer and Richard Cyganiak aims to expand the web with shared data by distributing open datasets in the RDF format on Semantic Web and creating the RDF links between these datasets [7]. A class of open data sharing level is defined as the number of stars (★) as follows:<sup>2</sup>

- ★ One-star level has the only requirement to make the information public in any data format.
- ★★ Two-star level has a provision that the disclosed information must be in a format that is not unstructured data, whether it is a proprietary format or not.

---

<sup>1</sup> [https://www.ted.com/talks/tim\\_berners\\_lee\\_the\\_next\\_web/transcript](https://www.ted.com/talks/tim_berners_lee_the_next_web/transcript)

<sup>2</sup> <https://5stardata.info/en/>



- ★★★Three-star level requires that the data must be in a structured form with an open standard format.
- ★★★★Four-star level determines that the data must be in an open standard in the Semantic Web format, such as RDF.
- ★★★★★Five-star level requires that the data must be linked to other open data to be a complete Linking Open Data.

### **3. Towards machine-understandable data**

There are many works related to transforming data from tables and lists to Linked Data. Some research involves extracting table-type data in various formats such as spreadsheets, relational databases, etc. and then converting them to RDF data. Those researches can be divided into groups as follows.

#### **3.1 Research related to the creation of facts into linked data**

There are many research works related to filling facts into Linked Data. The most discussed projects [8–11] are DBpedia, YAGO, Freebase, Wikidata, and OpenCyc. There are also several related researches which can be divided into groups as follows:

##### *3.1.1 Extracting facts from various parts of Wikipedia*

DBpedia is a joint research of the Free University of Berlin and Leipzig University in Germany [12]. The objective is to extract Wikipedia structured data such as infoboxes and categories including some unstructured data such as abstracts [13]. DBpedia supports Wikipedia information in many languages [14]. The goal of this project is to be the core to link other datasets of Linked Data together [15]. The result is a core that has more than 3 billion facts about 4.58 million topics which are divided into nearly 600 million facts from English Wikipedia articles, and the remainder is more than 2.5 billion facts from Wikipedia in other languages.

With the limitation of extracting data from tables, such as how to classify different types of tables and assigning names to them, DBpedia then chooses to extract only the structured data [13]. However, there are additional capabilities in the later version of the framework for extracting table data in Wikipedia, but it only extracts the table data as HTML tag block, not as the RDF triples that can be directly queried using SPARQL [12]. DBpedia has encouraged the development of algorithms to extract data from tables and lists in Wikipedia using the DBpedia framework by proposing a project in the Google Summer of Code (GSoC) from mid-2016 which continued to develop the project until 2017.<sup>3</sup> It has yet to publish the relevant academic work and has not yet been implemented in the latest version of DBpedia framework.

Isbell and Butler published a research paper created at HP's Digital Media Systems Laboratory [16]. They conducted a study of the conversion of data from Wikipedia structured infoboxes and has some parts that cover semi-structured and unstructured data.

YAGO is a research project from the Max Planck Institute for Informatics in Germany [17]. The objective is to extract structured data from Wikipedia categories by applying Synsets data from Princeton University's WordNet project. The project contains 120 million facts in 10 million topics.

---

<sup>3</sup> <https://wiki.dbpedia.org/blog/dbpedia-google-summer-code-2016> DBpedia @ GSoC 2016

BabelNet [18] and Multilingual Entity Taxonomy (MENTA) [19] extract facts from Wikipedia and WordNet as well as YAGO, but BabelNet and MENTA aimed at creating a multilingual knowledge base.

### 3.1.2 Manually recording facts into the knowledge base

Freebase of Metaweb Technologies [20] is a Web-based knowledge base where users share structured information directly through a Webpage specifically designed for recording and verifying information [21] (unlike DBpedia and YAGO, in which structured data was converted from Wikipedia.) After being acquired by Google in 2010, its data was transferred to Wikidata in 2014. Finally, in 2016, Freebase was closed, and it has been integrated into the Google Knowledge Graph. It is later being developed into Knowledge Vault: a Google research that aims to create an automated process to build the knowledge base directly from the Web [22].

In addition to the knowledge that users create in the system via the Web, Freebase also collects much information from Wikipedia [23] including Notable Names Database (NNDB), Fashion Model Directory (FMD), and MusicBrainz, in order to create a large amount of seed data. Before closing down, Freebase accumulated 2.4 billion facts in 44 million topics.

Wikidata is a project of the Wikimedia Foundation [24]. It is an open knowledge base, allowing users to manually record facts through a system designed to be easy to use, similar to Wikipedia. One interesting concept of Wikidata is the ability to keep the facts in conflict when it is not possible to conclude which fact is more accurate [25]. The “credibility” of information in Wikidata (including Wikipedia) does not focus on the “accuracy” of information more than the “provenance” of that information. For example, the population data of Mumbai is 12.5 million people, according to the Indian Bureau of Statistics but 20.5 million people when based on UN estimates. It is not the responsibility of the Wikidata community to find out what the truth is. Wikidata uses a straightforward way to store all information along with its source. The user has to choose which one to use. Currently, Wikidata has 30 million facts about 14 million topics. It can be seen that both DBpedia and Wikidata are the conversion of Wikipedia data into structured data using different methods [26]. However, some parts of Wikidata have been converted and incorporated into DBpedia Wikidata [27] and the ProFusion dataset [28].

Cyc is an extensive knowledgebase project by Douglas B. Lenat which started in 1984 [29]. The goal is to store a large number of facts and organize them automatically. OpenCyc is a smaller version of Cyc that reduces the size of the knowledge base and is publicly available [30]. However, OpenCyc was shut down in 2017, but ResearchCyc is still open for research studies [31].

### 3.1.3 Transform data from other formats to RDF

RDF123 by Han et al. [32] is a tool used to convert data in spreadsheet format to RDF format. Its concept can also be used to convert the table data into RDF. A survey paper [33] of the W3C RDB2RDF Incubator Group discusses many research projects that involve converting data from relational databases to RDF. Although this research does not mention the data conversion from the generic table, it can be applied to table conversion.

There are also many W3C recommendations by CSV on the Web Working Group<sup>4</sup> which discusses the conversion of data in the form of record sets in CSV format to other formats such as RDF or JSON.

---

<sup>4</sup> [https://www.w3.org/2013/csvw/wiki/Main\\_Page](https://www.w3.org/2013/csvw/wiki/Main_Page) CSV on the Web Working Group Wiki

### **3.2 Research related to the conversion of table and list to other formats**

There are several research works that involve converting table and list into other formats.

Yang and Luk [34, 35] discuss a thorough method for converting Web-based tables into key-value pair data and provide solutions to the problem of extracting data from the table in various cases.

The research of Pivk, Cimiano, and Sure [36] proposes a method to convert data from the Web-based table into F-logic (frame logic) which is a frame representation that can be applied to Semantic Web.

Table Analysis for Generating Ontologies (TANGO) is the research of Embley [37] and Tijerino et al. [38, 39]. The goal is to transform table data into ontology.

Table Extraction by Global Record Alignment (TEGRA) by Chu et al. [40] discusses the challenges of extracting structured data from Web-based table, in which in some case, a “table” that appear on a Webpage is not in HTML table format but it may be in HTML list or other arrangements.

DeExclerator is the research of Eberius et al. [41]. It is a framework for extracting structured data from HTML tables and spreadsheets.

Venetis et al. [42] solve the problem of dealing with semantics and ontologies by manually adding classes to column headers of the table without having to do schema matching. However, the user must have the skill to add this information.

WebTables [43, 44] is a project of Google Research to extract structured data from HTML tables on Webpages. They searched 14.1 billion HTML tables and found that only 154 million tables have sufficient quality to allow extraction of the structured data [45]. Most HTML tables on the web are used to define the layout of the webpage but are not used to present the data in the actual table format [46]. WebTables uses the classifier that is adjusted to focus on recall more than precision in order to filter the table from the Webpage as much as possible. It then selects only the table with a single-line header and ignores other more complicated tables. Later, this project has been developed into a system called Octopus [47] to help support the search engine more efficiently.

At Google, Elmeleegy et al. [48] use WebTables to support a system called ListExtract to extract 100,000 lists from the web and then transform them into relational databases. Wong et al. [49] use 1000 machines to extract 10.1 billion tuples from 1 billion Webpages with parallel algorithms in less than 6 hours.

Fusion Tables [50] is a Google Research project designed to allow users to upload table data on the web for data analysis with various tools. It is currently available on Google Docs.

The Web Data Commons (WDC) [51] is a project to extract structured data from Common Crawl which is the largest webpage archive that is publicly available. A part of the WDC called Web Table Corpora only extracts structured data from HTML tables in the Common Crawl Web archive. Currently, Web Table Corpora has been available to download in two sets. The first set is the 2012 Corpus which extracts 147 million tables from 3.5 billion Webpages in 2012 Common Crawl. The second set is the 2015 Corpus which extracted 233 million tables from 1.78 billion webpages in July 2015 Common Crawl. The second set contains metadata about extracted tables, while this information is not reserved in the first set.

WDC Web Table Corpus has been used in many research. For example, it is used to measure the performance of schema matching approaches for various levels of table elements (such as table-to-class, row-to-instance, and attribute-to-property

---

<sup>4</sup> [https://www.w3.org/2013/csvw/wiki/Main\\_Page](https://www.w3.org/2013/csvw/wiki/Main_Page) CSV on the Web Working Group Wiki

matching) which previously used different datasets thus making it difficult to be compared [52].

The most similar work to our proposal is WikiTables [53] which is a tool to extract information from the tables in Wikipedia. It is used to discover new hidden facts. The result of this research is a set of 15 million tuples extracted from the Wikipedia tables.

### 3.3 Current “standard” representation of table and list

There are many ways to represent tables and lists in the standard data formats issued by many standard bodies such as:

- International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) standards
- Internet Engineering Task Force (IETF) Request for Comments (RFC)
- World Wide Web Consortium (W3C) Recommendations (REC)
- Internet Assigned Numbers Authority (IANA) Multipurpose Internet Mail Extensions (MIME) media types

We mention only the most common standard and format that are capable of table and list representation such as:

- Comma-separated values (CSV), i.e., RFC 4180 [54], and other delimiter-separated values (DSV) such as tab-separated values (TSV)<sup>5</sup>.
- Markup languages such as HTML table, i.e., RFC 1942 [55] (developed from USDOD SGML Table Model), and HTML list in HTML 2.0 RFC 1866 [56].
- Lightweight markup languages such as Wikitext table and list and other markdown languages. After many attempts to standardize various of them, they end up with RFC 7763 for the original syntax and RFC 7764 for other variants.
- Office spreadsheet and word processor table/list, e.g., OASIS OpenDocument Format (ODF) ISO/IEC 26300 and Microsoft Office Open XML (OOXML) ISO/IEC 29500. ISO/IEC also issues a comparison of both formats and guidelines for translation between them in ISO/IEC TR 29166.

## 4. TULIP: table/list interchangeable, unified, pivotal vocabulary

The main idea of TULIP is to transform the semi-structured data in the form of tables and lists, regardless of the source, to the structured data in the form of five-star open data as a set of RDF triples. Each triple contains only subject-predicate-object. The triples are connected to other triples and form a directed graph called the RDF graph. That is the principle of Linked Data, allowing Semantic Web

---

<sup>5</sup> <https://www.iana.org/assignments/media-types/text/tab-separated-values>

applications to consume TULIP's five-star open data in the same way as another Linked Data.

#### **4.1 TULIP in a *bud* shell**

TULIP is a set of RDF vocabulary (the completed TULIP specification is available at <http://purl.org/tulip/spec>) in the form of RDF Schema. It consists of a set of RDF properties and RDF classes that are used to define structures for data representation to completely store table and list data and preserve all of its original semantic structure. It includes basic properties needed for five-star open data such as identifiable, dereferenceable, etc., as well as the three unique properties of TULIP: interchangeable, unified, and pivotal.

##### *4.1.1 TULIP interchangeable property*

TULIP has the complete preservation property in order to preserve the semantic structure of the source table and list, such as the cell contents, table structure, column/row headers, list items, and hierarchy including some formats such as spanning cells, but not including decorative style, for example, typographic styles (bold/italics), fonts, colors, borders, backgrounds, etc. The original tables and lists can be recreated from the RDF triple set using TULIP vocabulary.

This is possible because TULIP has a set of properties and classes to store the content data and classes, such as tables, columns, rows, table cells, lists, list items, etc., as well as various characteristics such as table headers, spanning cells, enumerated lists, etc.

##### *4.1.2 TULIP unified property*

TULIP retains both the tables and lists in the same format as the hierarchical treelike structure. The structure is stored as a set of RDF triples which can represent directed graphs without order or precedence between sibling nodes. So, a set of additional RDF properties must be defined to mimic the hierarchical structures and precedence of nodes in the hierarchy. That resembles a feature of the RDF called the RDF Container; however, the TULIP has more specific features.

As TULIP retains the structure in this way, the output from a query can be projected to a new structure different from the original. It depends on how an application wants to present the information. TULIP data is stored with standard RDF properties such as `rdf:type` and `rdfs:label`. There are special RDF properties to model the treelike structure. The structure is overlaid on top of the standard RDF graph. So, Semantic Web applications that do not understand the TULIP schema can perform graph traversal and get all the contents from TULIP.

The advantage is that we can look at the data without paying attention to the origin of what type of data it came from. Instead, we can choose to look at it the way we want. For example, we can look at the data that originated from a table but think of it as if it comes from a list or vice versa. Otherwise, we may combine data in both formats. It is possible to show the data in entirely different formats, such as charts or diagrams. Therefore, if we want to create an infographic from TULIP data, we can create it dynamically and change its appearance freely in any form.

##### *4.1.3 TULIP pivotal property*

Pivotal properties or view manipulation can be done because TULIP has another type of structure modeled to mimic the multidimensional array on the RDF graph.

It can store data both in column-orient (columnar) and row-orient (row-based). This allows us to query specific content of all sizes and dimensions in a single query.

Moreover, with this model, we can apply the principles of data warehouse and online analytical processing (OLAP) operations, such as rolling up the entire data in the same group, drilling down to any layer, or even slicing to cut only some axes of multidimensional content including dicing, i.e., rotate to change the perspective which means that we can filter and pivot the view of the data in TULIP format any way we want.

One of the key concepts of TULIP is using an RDF feature called RDF collection, i.e., RDF list. Apply it as a one-dimensional array to store subscripts of each level in a multidimensional array by placing all subscripts as corresponding members of the RDF collection. Then put these collections to each node of TULIP. Access to each element of TULIP can be done by a SPARQL querying for its RDF collection items that match to the corresponding subscripts.

## 4.2 Creating five-star open data table and list with TULIP schema

Now, we will demonstrate how to create RDF triples using the TULIP schema to represent a simple table. We use the following small example table of three columns by three rows.

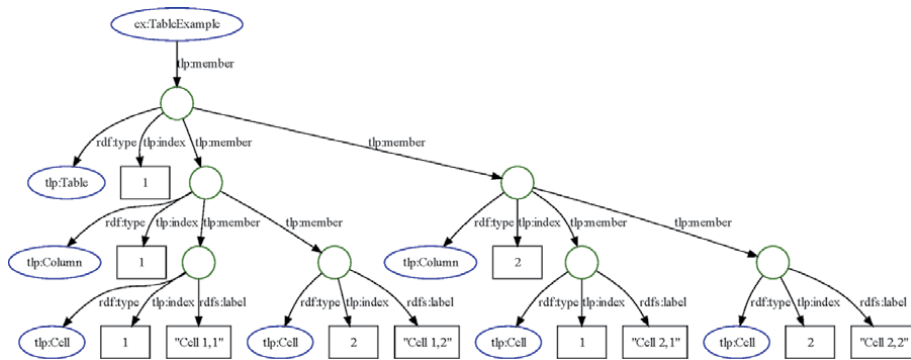
Cell Content 1,1	Cell Content 2,1	Cell Content 3,1
Cell Content 1,2	Cell Content 2,2	Cell Content 3,2
Cell Content 1,3	Cell Content 2,3	Cell Content 3,3

Because TULIP schema can represent the table in both column-oriented (column-major) or row-oriented (row-major), in this case, we represent a table with a column-major format. The sample data in the table cells is preceded by the corresponding column number, followed by the row number.

Excerpts of RDF triples used to represent the three-column by three-row table above using TULIP schema are shown in **Figure 1**.

```
ex:TableExample
  tlp:member _:Table1 .
  _:Table1 rdf:type tlp:Table ;
  tlp:index 1 ;
  tlp:member _:Col1, _:Col2, _:Col3 .
  _:Col1 rdf:type tlp:Column ;
  tlp:index 1 ;
  tlp:member _:Cell11, _:Cell12, _:Cell13 .
  _:Cell11 rdf:type tlp:Cell ;
  tlp:index 1 ;
  rdfs:label "Cell Content 1,1" .
  _:Cell12 rdf:type tlp:Cell ;
  tlp:index 2 ;
  rdfs:label "Cell Content 1,2" .
  ...
  _:Cell33 rdf:type tlp:Cell ;
  tlp:index 3 ;
  rdfs:label "Cell Content 3,3" .
```

**Figure 1.**  
*RDF triples of the example table represented by the TULIP schema.*



**Figure 2.**  
RDF graph of the example table using TULIP schema.

**Figure 2** shows these RDF triples in the RDF graph. (To make the figure more compact, we adjusted the table to a two-column by two-row dimension.)

Next, we will demonstrate how to create a simple five-star open data list using the TULIP schema by using the following example list.

- List Item 1
- List Item 2
  - List Item 2,1
  - List Item 2,2
    - i. List Item 2,2,1
- List Item 3

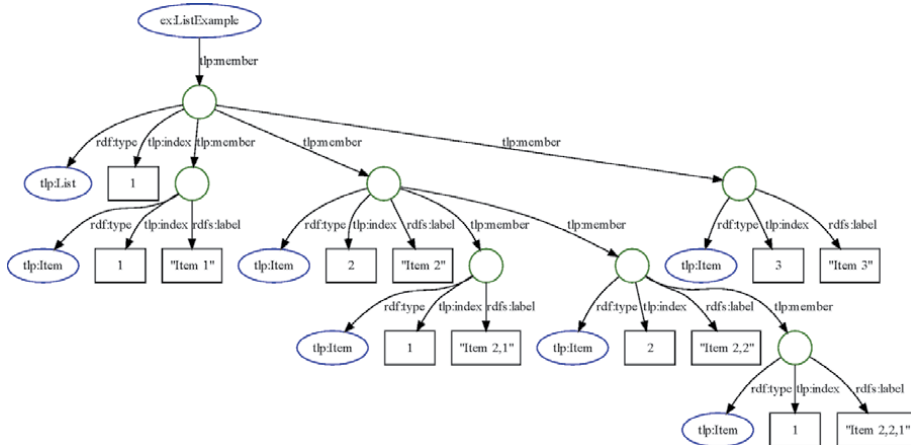
The RDF triples representing the above list using the TULIP schema are shown in **Figure 3**. The RDF graph of these RDF triples is shown in **Figure 4**.

```

ex:ListExample
  tp:member _:List1 .
  _:List1 rdf:type tlp:List ;
  tp:index 1 ;
  tp:member _:Item1, _:Item2, _:Item3 .
  _:Item1 rdf:type tlp:Item ;
  tp:index 1 ;
  rdfs:label "List Item 1" .
  _:Item2 rdf:type tlp:Item ;
  tp:index 2 ;
  rdfs:label "List Item 2" ;
  tp:member _:Item21, _:Item22 .
  ...
  _:Item22 rdf:type tlp:Item ;
  tp:index 2 ;
  rdfs:label "List Item 2,2" ;
  tp:member _:Item221 .
  _:Item221 rdf:type tlp:Item ;
  tp:index 1 ;
  rdfs:label "List Item 2,2,1".
    
```

```
_:item3 rdf:type tlp:Item ;
      tlp:index 3 ;
      rdfs:label "List Item 3" .
```

**Figure 3.**  
 RDF triples of the example list represented by the TULIP schema.



**Figure 4.**  
 RDF graph of the example list using TULIP schema.

### 4.3 Providing the way for direct data access

The RDF graphs, both in **Figures 2** and **4**, are already in a hierarchical structure and sequence with `tlp:index` (hereafter referred to as the “TULIP indexed member, TIM model”). It can be used to recreate the original table and list. This can be achieved by graph traversal. Furthermore, generic Semantic Web applications can access the data hierarchically. This is not much different from standard RDF containers. To access each arbitrary data, we have to indirectly access by performing graph traversal step by step until we reach the data we need. Writing SPARQL queries to perform this task is not easy. Therefore, we provide direct data access by assigning “position” to each data element. Similar to creating a multidimensional array index, we mimic this concept by using the feature of RDF called collections, i.e., RDF lists, to provide a way to access data directly in a single query without any nested match (hereafter referred to as the “TULIP index list, TIL model”). We use the `tlp:` element to point to the blank node of each item in the flat structure and create the `tlp:indexList` property for each node. The `tlp:indexList` property has its range, i.e., object in `rdf:List` class of sequence number according to the `tlp:index` in each node of the TIM model, and ends with zero (used to specify whether to separate only single element or group of all elements under the same parent node, more about this will be discussed later). If we take the example table and list to create the RDF triples using the TIL model, they will look like **Figure 5** and **Figure 6**.

```
ex:TableExample
  tlp:element _:Table1,
              _:Col1, _:Cell11, _:Cell12, _:Cell13,
```



```

        _:Col2, _:Cell21, _:Cell22, _:Cell23,
        _:Col3, _:Cell31, _:Cell32, _:Cell33 .
_:Table1 rdf:type tlp:Table ;
  tlp:indexList ( 1 0 ) .
_:Col1 rdf:type tlp:Column ;
  tlp:indexList ( 1 1 0 ) .
_:Cell11 rdf:type tlp:Cell ;
  tlp:indexList ( 1 1 1 0 ) ;
  rdfs:label "Cell Content 1,1" .
_:Cell12 rdf:type tlp:Cell ;
  tlp:indexList ( 1 1 2 0 ) ;
  rdfs:label "Cell Content 1,2" .
...
_:Cell33 rdf:type tlp:Cell ;
  tlp:indexList ( 1 3 3 0 ) ;
  rdfs:label "Cell Content 3,3" .

```

**Figure 5.**  
RDF triples of the example table represented by TIL model.

```

ex:ListExample
  tlp:element _:List1,
              _:Item1,
              _:Item2, _:Item21, _:Item22, _:Item221,
              _:Item3 .
_:List1 rdf:type tlp:List ;
  tlp:indexList ( 1 0 ) .
_:Item1 rdf:type tlp:Item ;
  tlp:indexList ( 1 1 0 ) ;
  rdfs:label "List Item 1" .
_:Item2 rdf:type tlp:Item ;
  tlp:indexList ( 1 2 0 ) ;
  rdfs:label "List Item 2" .
...
_:Item22 rdf:type tlp:Item ;
  tlp:indexList ( 1 2 2 0 ) ;
  rdfs:label "List Item 2,2" .
_:Item221 rdf:type tlp:Item ;
  tlp:indexList ( 1 2 2 1 0 ) ;
  rdfs:label "List Item 2,2,1" .
_:Item3 rdf:type tlp:Item ;
  tlp:indexList ( 1 3 0 ) ;
  rdfs:label "List Item 3" .

```

**Figure 6.**  
RDF triples of the example list represented by TIL model.

An example of the RDF graph for **Figure 6** (shows only the first five nodes) is in **Figure 7**.

Each of the `tlp:indexList` objects is the structure of the RDF collection, i.e., RDF list, which when extended, becomes an unbalanced binary tree structure, where leaf nodes are each member of the list, respectively. For example, the `tlp:indexList (1 2 0)` has a structure as **Figure 8**.

Access to each element of TULIP by SPARQL is achieved by querying its `tlp:indexList` that have RDF collection items matching the corresponding subscripts. The problem is that the current SPARQL specification has limited ability to directly handle the RDF collection (causing many attempts to create the alternatives to the



```

SELECT ?label
WHERE {
    ex:TableExample tlp:element ?elem .
    ?elem tlp:indexList/rdf:first 1 .
    ?elem tlp:indexList/rdf:rest/rdf:first 3 .
    ?elem rdfs:label ?label .
}

```

The results are as follows:

```

"Cell Content 3,1"
"Cell Content 3,2"
"Cell Content 3,3"

```

Alternatively, if we want the whole third row, we match the tlp:indexList with (1 ? 3) where “?” at the second subscript position is the tlp:Column level, which we will not filter. So we will get every column.

```

SELECT ?label
WHERE {
    ex:TableExample tlp:element ?elem .
    ?elem tlp:indexList/rdf:first 1 .
    ?elem tlp:indexList/rdf:rest/rdf:rest/rdf:first 3 .
    ?elem rdfs:label ?label .
}

```

The results are:

```

"Cell Content 1,3"
"Cell Content 2,3"
"Cell Content 3,3"

```

#### 4.4 Combining the advantages of both models

Both TULIP models, indexed member model and index list model, have different pros and cons. Users can choose either type as appropriate to their needs. Furthermore, they could create RDF triples using the merged model called the “TULIP hybrid, TH model,” which combines the advantages of both types in one structure. The idea is that we take the TIM model as the basis and then add TIL elements by inserting a tlp:indexList property into each blank node of TIM and adding all tlp:element to the primary resource. For example, when adding tlp:indexList and tlp:element to **Figure 1**, it will look like **Figure 9**.

```

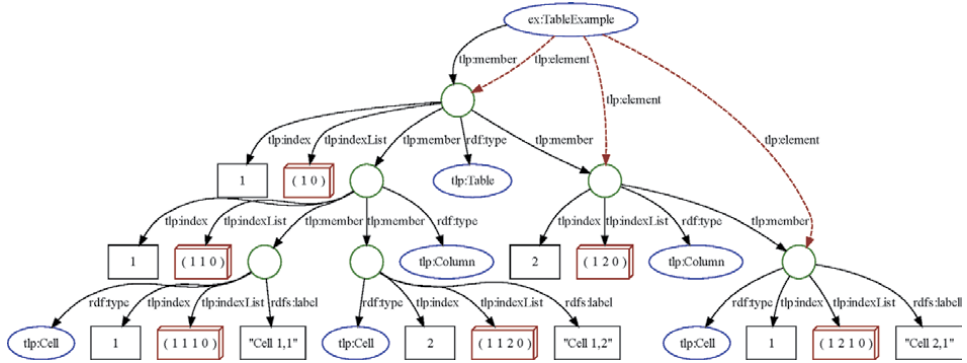
ex:TableExample
  tlp:element _:Table1,
    _:Col1, _:Cell11, _:Cell12, _:Cell13,
    _:Col2, _:Cell21, _:Cell22, _:Cell23,
    _:Col3, _:Cell31, _:Cell32, _:Cell33 ;
  tlp:member _:Table1 .
_:Table1 rdf:type tlp:Table ;
  tlp:index 1 ;
  tlp:indexList ( 1 0 ) ;
  tlp:member _:Col1, _:Col2, _:Col3 .
...
_:Cell33 rdf:type tlp:Cell ;
  tlp:index 3 ;
  tlp:indexList ( 1 3 3 0 ) ;
  rdfs:label "Cell Content 3,3" .

```

**Figure 9.** RDF triples of the example table represented by TH model.

When shown as RDF graph, it will look like **Figure 10**. (To make the graph more compact, we have resized the table to two columns by two rows and remove some nodes, and many edges of tlp:element have also been omitted. In fact, tlp:element will point to every blank node).

Likewise, the TH model RDF triples of the example list are in **Figure 11**. When shown as the RDF graph, it will look like **Figure 12** (showing only some tlp:element edges to make the graph more convenient and clear).

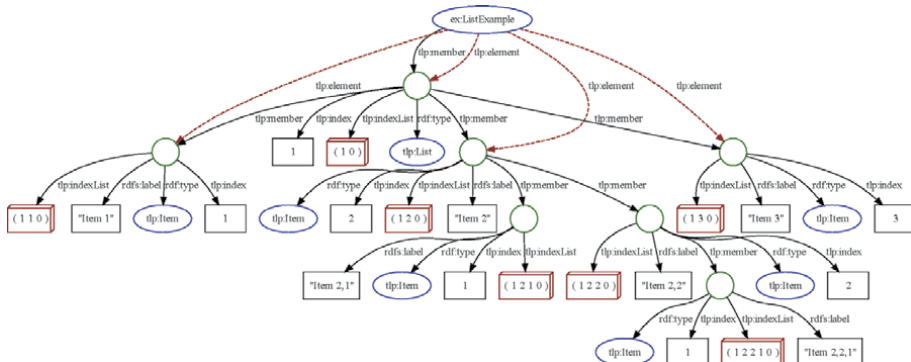


**Figure 10.**  
 RDF graph of the example table represented by TH model.

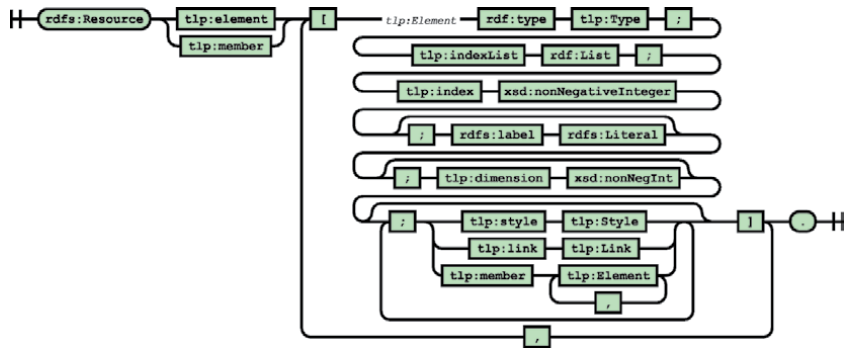
```

ex:ListExample
  tlp:element _:List1,
              _:Item1,
              _:Item2, _:Item21, _:Item22, _:Item221,
              _:Item3;
  tlp:member _:List1 .
_:List1 rdfs:type tlp:List;
  tlp:index 1;
  tlp:indexList ( 1 0 );
  tlp:member _:Item1, _:Item2, _:Item3 .
...
_:Item3 rdfs:type tlp:Item;
  tlp:index 3;
  tlp:indexList ( 1 3 0 );
  rdfs:label "List Item 3" .
    
```

**Figure 11.**  
 RDF triples of the example list represented by TH model.



**Figure 12.**  
 RDF graph of the example list represented by TH model.



**Figure 13.**  
Excerpt of TULIP vocabulary syntaxes *tlp:element* and *tlp:member*.

An excerpt of TULIP vocabulary syntax (detail specification can be found at <http://purl.org/tulip/spec>) for *tlp:element* and *tlp:member* is shown in **Figure 13**.

## 5. Experimental results and contributions

We have designed and implemented two reference libraries. The first library is a Python library that has functions to extract/transform Webpages and create TULIP datasets. The second library is a JavaScript library to query TULIP endpoint, consume its result sets, and manipulate them. The code is provided in the GitHub repositories at <https://github.com/julthep/tulip> and <https://github.com/julthep/tulip.js> respectively.

Furthermore, we experimented with TULIP vocabulary using Wikipedia as the data source by converting a number of articles into TULIP data format. The result datasets can be accessed via SPARQL endpoint at <http://tlpedia.org/sparql/> or downloaded at <http://tlpedia.org/datasets/>. These datasets will be updated periodically, and the number of imported articles will be increased on a regular basis.

## 6. Conclusion

Our proposal is different from existing research, which mainly efforts on transforming data from tables and lists into facts in various formats. TULIP focuses on extracting data from tables and lists into the dataset in the form of five-star Linked Data. Also, the RDF triples result can be used to recreate tables and lists in the same format as the source data because the designed schema focuses on the ability to preserve the structure of the original table and list. Another essential feature is that the acquired RDF triples can also be embedded in a package file such as XML or HTML with RDFa to be used to create tables and lists on a Webpage as an integrated dataset.

TULIP can also be applied to many applications since it is designed to be very flexible. Implementers can choose to use any of its schema models, depending on their usage. It also supports many types of data and is extensible. Actually, TULIP also supports the Document Object Model (DOM) which can transform paragraphs or text blocks into the same structure. It can be used to transforms Wikipedia articles into TULIP format as a five-star open dataset so that the Semantic Web application can consume Linked Data more conveniently. All of this is to create a data structure that is not just machine-readable but will be machine-understandable.

## **Author details**

Julthep Nandakwang\* and Prabhas Chongstitvatana  
Department of Computer Engineering, Chulalongkorn University, Bangkok,  
Thailand

\*Address all correspondence to: [julthep@nandakwang.com](mailto:julthep@nandakwang.com)

## **IntechOpen**

---

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Spivack N. Web 3.0–The Best Official Definition Imaginable [Internet]. 2007. Available from: <http://www.novaspivack.com/technology/web-3-0-the-best-official-definition-imaginable> [Accessed: 20 December 2019]
- [2] Berners-Lee T, Cailliau R, Luotonen A, Nielsen HF, Secret A. The world-wide web. *Communications of the ACM*. 1994;37(8):76-82
- [3] Berners-Lee T, Hendler J, Lassila O. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*. 1 May 2001;284(5):34-43
- [4] Bratt S. Semantic Web and Other W3C Technologies to Watch [Internet]. W3C; 2006. Available from: <https://www.w3.org/2006/Talks/1023-sb-W3CTechSemWeb> [Accessed: 20 December 2019]
- [5] Shadbolt N, Hall W, Berners-Lee T. The semantic web revisited. *Intelligent Systems, IEEE*. 2006;21(3):96-101
- [6] Bizer C, Heath T, Berners-Lee T. Linked Data - the Story So Far. *Semantic Services. Interoperability and Web Applications: Emerging Concepts*. 2009: 205-227
- [7] Bizer C. The emerging web of linked data. *IEEE Intelligent Systems*. 2009; 24(5)
- [8] Färber M, Ell B, Menne C, Rettinger A. A comparative survey of DBpedia, freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web Journal*. July 2015;1(1):1-5
- [9] Ringler D, Paulheim H. One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & co. In: *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. Cham: Springer; 25 September 2017:366-372
- [10] Färber M, Bartscherer F, Menne C, Rettinger A. Linked data quality of DBpedia, freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*. 2018;9(1): 77-129
- [11] Pillai SG, Soon L-K, Haw S-C. Comparing DBpedia, Wikidata, and YAGO for web information retrieval. In: *Intelligent and Interactive Computing*. Singapore: Springer; 2019. pp. 525-535
- [12] Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, et al. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*. 1 January 2015;6(2):167-195
- [13] Auer S, Lehmann J. What have Innsbruck and Leipzig in common? Extracting semantics from wiki content. In: *Franconi E, Kifer M, May W, editors. 4th European Semantic Web Conference, ESWC 2007, June 3–7, 2007 Proceedings; 2007/01/01. Innsbruck, Austria: Springer Berlin Heidelberg; 2007. pp. 503-517*
- [14] Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, et al. DBpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*. 2009;7(3):154-165
- [15] Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. DBpedia: A nucleus for a web of open data. In: *Aberer K, Choi K-S, Noy N, Allemang D, Lee K-I, Nixon L, et al., editors. 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, November 11-15, 2007 Proceedings; 2007/01/01. Busan, Korea: Springer Berlin Heidelberg; 2007. pp. 722-735*

- [16] Isbell J, Butler MH. Extracting and re-using structured data from wikis. Bristol: Digital Media Systems Laboratory of Hewlett-Packard Development Company; November 14, 2007. Report No.: HPL-2007-182
- [17] Suchanek FM, Kasneci G, Weikum G. A core of semantic knowledge unifying WordNet and Wikipedia. In: YAGO, editor. Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta, Canada: ACM; 2007. p. 1242667
- [18] Navigli R, Ponzetto SP. BabelNet: Building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden. 1858704: Association for Computational Linguistics; 2010. pp. 216-225
- [19] Gd M, Weikum G. Inducing multilingual taxonomies from Wikipedia. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, ON, Canada. 1871577: ACM; 2010. pp. 1099-1108
- [20] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 9 June 2008. pp. 1247-1250
- [21] Bollacker K, Tufts P, Pierce T, Cook R, editors. A platform for scalable, collaborative, structured information integration. In: International Workshop on Information Integration on the Web (IIWeb'07); 2007 July
- [22] Dong X, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, New York, USA. 2623623: ACM; 2014. pp. 601-610
- [23] Bollacker K, Cook R, Tufts P, editors. Freebase: A shared database of structured general human knowledge. In: Proceedings of the 22nd national conference on Artificial intelligence-2. Vol 7. 22 July 2007. pp. 1962-1963
- [24] Vrandečić D, Krötzsch M. Wikidata: A free collaborative knowledgebase. Communications of the ACM. 2014; 57(10):78-85
- [25] Vrandečić D. The rise of Wikidata. IEEE Intelligent Systems. 2013;28(4): 90-95
- [26] Abián D, Guerra F, Martínez-Romanos J, Trillo-Lado R. Wikidata and DBpedia: A comparative study. In: Semantic Keyword-Based Search on Structured Data Sources. Cham: Springer; 11 September 2017. pp. 142-154
- [27] Ismayilov A, Kontokostas D, Auer S, Lehmann J, Hellmann S. Wikidata through the eyes of DBpedia. Semantic Web. 2018;9(4):493-503
- [28] Frey J, Hofer M, Obraczka D, Lehmann J, Hellmann S. DBpedia FlexiFusion the best of Wikipedia> Wikidata> your data. In: International Semantic Web Conference. Cham: Springer; 26 October 2019. pp. 96-112
- [29] Lenat DB. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM. 1995; 38(11):33-38
- [30] Lenat DB. The voice of the turtle: Whatever happened to AI? AI Magazine. 2008;29(2):11
- [31] Lenat DB. Building a machine smart enough to pass the Turing test. In: Epstein R, Roberts G, Beber G, editors.



- Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer. Dordrecht: Springer Netherlands; 2009. pp. 261-282
- [32] Han L, Finin T, Parr C, Sachs J, Joshi A. RDF123: From Spreadsheets to RDF. In: Sheth A, Staab S, Dean M, Paolucci M, Maynard D, Finin T, et al., editors. *The Semantic Web - ISWC 2008. Lecture Notes in Computer Science*. 5318. Berlin Heidelberg: Springer; 2008. pp. 451-466
- [33] Sahoo SS, Halb W, Hellmann S, Idehen K, Thibodeau T Jr, Auer S, et al. A survey of current approaches for mapping of relational databases to RDF. W3C RDB2RDF Incubator Group Report; 2009
- [34] Yang Y. *Web Table Mining and Database Discovery*. Doctoral Dissertation, Simon Fraser University; 2002
- [35] Yang Y, Luk W-S. A framework for web table mining. In: *Proceedings of the 4th International Workshop on Web Information and Data Management*. McLean, Virginia, USA. 584940: ACM; 2002. pp. 36-42
- [36] Pivk A, Cimiano P, Sure Y. From Tables to Frames. In: McIlraith SA, Plexousakis D, van Harmelen F, editors. *The Semantic Web – ISWC 2004: Third International Semantic Web Conference, Hiroshima, Japan, November 7-11 2004, Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 166–181.
- [37] Embley DW, Tao C, Liddle SW. Automatically extracting ontologically specified data from HTML tables of unknown structure. In: Spaccapietra S, March ST, Kambayashi Y, editors. *Conceptual Modeling — ER 2002: 21st International Conference on Conceptual Modeling Tampere, Finland, October 7–11, 2002 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. pp. 322-337
- [38] Tijerino YA, Embley DW, Lonsdale DW, Nagy G, editors. *Ontology generation from tables*. In: *Proceedings of the fourth international conference on web information systems engineering, WISE; 2003*. pp. 10-12
- [39] Tijerino YA, Embley DW, Lonsdale DW, Ding Y, Nagy G. *Towards ontology generation from tables*. *World Wide Web*. 2005;8(3):261-285
- [40] Chu X, He Y, Chakrabarti K, Ganjam K. *TEGRA: Table extraction by global record alignment*. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. Melbourne, Victoria, Australia. 2723725: ACM; 2015. pp. 1713-1728
- [41] Eberius J, Werner C, Thiele M, Braunschweig K, Dannecker L, Lehner W. *DeExcelerator: A framework for extracting relational data from partially structured documents*. In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. San Francisco, California, USA. 2508210: ACM; 2013. pp. 2477-2480
- [42] Venetis P, Halevy A, Madhavan J, Paşca M, Shen W, Wu F, et al. *Recovering semantics of tables on the web*. *Proceedings of the VLDB Endowment*. 2011;4(9):528-538
- [43] Cafarella MJ, Halevy A, Wang DZ, Wu E, Zhang Y. *WebTables: Exploring the power of tables on the web*. *Proc VLDB Endow*. 2008;1(1):538-549
- [44] Balakrishnan S, Halevy AY, Harb B, Lee H, Madhavan J, Rostamizadeh A, et al., editors. *Applying WebTables in practice*. In: *CIDR*. 2015
- [45] Cafarella MJ, Halevy AY, Zhang Y, Wang DZ, Wu E, editors. *Uncovering the Relational Web*. *WebDB*; June 13, 2008

- [46] Wang Y, Hu J, editors. A machine learning based approach for table detection on the web. In: Proceedings of the 11th International Conference on World Wide Web; 2002 May. Honolulu, Hawaii, USA: ACM; 2002. p. 511478
- [47] Cafarella MJ, Halevy A, Khoussainova N. Data integration for the relational web. Proc VLDB Endow. 2009;2(1):1090-1101
- [48] Elmeleegy H, Madhavan J, Halevy A. Harvesting relational tables from lists on the web. Proc VLDB Endow. 2009;2(1):1078-1089
- [49] Wong YW, Widdows D, Lokovic T, Nigam K. Scalable attribute-value extraction from semi-structured text. In: IEEE International Conference on Data mining workshops, 2009 ICDMW'09. IEEE; 6 December 2009. pp. 302-307
- [50] Gonzalez H, Halevy A, Jensen CS, Langen A, Madhavan J, Shapley R, et al. Google fusion tables: Data management, integration and collaboration in the cloud. In: Proceedings of the 1st ACM Symposium on Cloud Computing. Indianapolis, Indiana, USA. 1807158: ACM; 2010. pp. 175-180
- [51] Lehmborg O, Ritze D, Meusel R, Bizer C. A large public corpus of web tables containing time and context metadata. In: Proceedings of the 25th International Conference Companion on World Wide Web. Montréal, Québec, Canada. 2889386: International World Wide Web Conferences Steering Committee; 2016. pp. 75-76
- [52] Ritze D, Bizer C. Matching web tables to DBpedia—a feature utility study. Context. 2017;42(41):19-31
- [53] Bhagavatula CS, Noraset T, Downey D. Methods for exploring and mining tables on Wikipedia. In: Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics. Chicago, Illinois. 2501516: ACM; 2013. pp. 18-26
- [54] Shafranovich Y. Common Format and MIME Type for Comma-Separated Values (CSV) Files [Internet]. IETF; 2005. Available from: <https://tools.ietf.org/html/rfc4180> [Accessed: 20 December 2019]
- [55] Raggett D. HTML Tables [Internet]. IETF; 1996. Available from: <https://tools.ietf.org/html/rfc1942> [Accessed: 20 December 2019]
- [56] Berners-Lee T, Connolly D. Hypertext Markup Language - 2.0 [Internet]. IETF; 1995. Available from: <https://tools.ietf.org/html/rfc1866> [Accessed: 20 December 2019]
- [57] Abdallah SA, Ferris B. The Ordered List Ontology Specification [Internet]. 2010. Available from: <http://purl.org/ontology/olo/core#> [Accessed: 20 December 2019]
- [58] Ciccarese P, Peroni S. The collections ontology: Creating and handling collections in OWL 2 DL frameworks. Semantic Web. 2014;5(6): 515-529
- [59] Leigh J, Wood D. An ordered RDF list. W3C workshop — RDF next steps; June 26–27, 2010; National Center for Biomedical Ontology (NCBO), Stanford, Palo Alto, CA, USA: W3C; 2010



# Linked Open Data: State-of-the-Art Mechanisms and Conceptual Framework

*Kingsley Okoye*

## Abstract

Today, one of the state-of-the-art technologies that have shown its importance towards data integration and analysis is the linked open data (LOD) systems or applications. LOD constitute of machine-readable resources or mechanisms that are useful in describing data properties. However, one of the issues with the existing systems or data models is the need for not just representing the derived information (data) in formats that can be easily understood by humans, but also creating systems that are able to process the information that they contain or support. Technically, the main mechanisms for developing the data or information processing systems are the aspects of aggregating or computing the metadata descriptions for the various process elements. This is due to the fact that there has been more than ever an increasing need for a more generalized and standard definition of data (or information) to create systems capable of providing understandable formats for the different data types and sources. To this effect, this chapter proposes a semantic-based linked open data framework (SBLODF) that integrates the different elements (entities) within information systems or models with semantics (metadata descriptions) to produce explicit and implicit information based on users' search or queries. In essence, this work introduces a machine-readable and machine-understandable system that proves to be useful for encoding knowledge about different process domains, as well as provides the discovered information (knowledge) at a more conceptual level.

**Keywords:** LOD, semantics, ontologies, metadata creation, data integration, process description, information retrieval, information extraction, information systems

## 1. Introduction

Linked Open Data (LOD) is a term used to refer to tools or platforms that support freely-connected (interlinked) resources or frameworks to allow for collection and integration of data (usually derived from various sources or formats) and provide useful information that can be accessed by machines or humans. Typically, LOD supported tools or platforms is expected to allow for both simple or complex oriented lookup for information access through some form of predefined language or mechanisms (e.g. using scripts or query-based languages such SQL, HTML, SPARQL, Description Logics, RDF graphs of the Triples form, XML, etc.) [1–3]. Technically, LOD is semantically defined as a knowledge graph [3] that vents in the form of

semantical web or schema (e.g. using ontologies) [4–6] of interconnected data [7]. According to Snyder et al. [7], LOD has since been epitomized as a way of improving the process of discovering useful information or resources by creating a series of robust links between related concepts or items.

The work done in this chapter notes that one of the main challenges with LOD has been on how to create systems or methods that are capable of providing an understandable format (both machine-readable and machine-understandable) for the various datasets that may come from different sources, as well as, making the derived formats or standards explicable across the several platforms. To this end, the work proposes a semantic-based LOD framework (SBLODF) that provides an additional function to LOD that allows for formal integration of the process elements or concepts through metadata creation (process description) using the semantic technologies or schema. This is called Semantic-based Linked Open Data.

## **2. Preliminaries**

### **2.1 Semantics? the missing link in LOD systems**

Research on why “domain knowledge” is useful in bridging the semantic gap in existing systems or applications that aims to store and/or process data has long been discussed in existing works of literature [1, 6, 8–11]. Whereas, Declerck et al. [1] note that one of the main aims of LOD supported systems is to develop new ways or methods for construing data values (interlinks) that are applicable to a broad range of applications or platforms (based on language technologies or resource descriptions) through semantic technologies. Wang [12] notes that contemporary studies on LOD methods and tools are mainly directed towards ascertaining different levels or types of process instances (entities), thereby resulting in the central task of finding relationships (schema-level) or links that exist amongst the LOD datasets or models in question being ignored.

According to Wang [12], ontological representations (mappings) are a very crucial way of solving the data heterogeneity or missing link. Moreover, ontologies can be described as an essential tool that proves useful towards establishing the semantic-level links in LOD [6, 13–18]. For example, Selvan et al. [19] proposed an ontology-based recommender system that is built on cloud services to store and retrieve data for further analysis using Type-2 fuzzy logic.

Studies have shown that there exists a (semantic) gap between different datasets and the various tools/algorithms that are applied to analyze or understand the data including results of the analysis in all stages of the data processing; ranging from the data pre-processing to implementation of the algorithms, and the interpretation of the results [6, 8, 11, 16]. For instance, data pre-processing usually involves the process of filtering and cleaning of data, standardization by defining formats for its integration, transformation and properties extraction and retrieval of the defined formats/structures, and then selected for the purpose of analysis. Nevertheless, in many settings, there exist the issue of semantic gaps in the several phases of the data pre-processing. For example, we note that in the absence of considering the formal structure (semantics) of the data models, most of the resulting systems have resort to empirical or ad-hoc methods to determine the quality of the underlying datasets or concepts. Whereas, it is certain that data semantics is necessary for understanding the relations that exist amongst the different process elements in the models, especially during the standardization and transformation step. Thus far, it is important to determine the correlation between the different data elements by taking into account the underlying properties/attributes of the data when performing

data standardization or processing at large. Apparently, tightly (closely) correlated attributes can be generalized into one combined attribute or classification for the purpose of tractability and conceptualized analysis.

Typically, in terms of the different application domains and rule-based information extraction systems, Yankova [20] conducted a semantic-based identity resolution and experiment that aims to identify conceptual information expressed within a domain ontology. The experiment was based on a generic and adaptable human language technology. In the experimentation, they extracted company information from several sources and update the existing ontologies with the resolved entities. The method for information extraction is a rule-based system they referred to as Identity Resolution Framework (IdRF) built using Proton [20] that provides a general solution to identifying known and new facts in a certain domain, and can also be applied to other domains regardless of the type of entities that may need to be resolved. Moreover, input to the IdRF includes different entities together with their associated properties and values, and the expected output is an integrated representation of the entities that are consequently resolved to have new properties or values within the ontology.

On the one hand, ontologies have shown to be beneficial in such data processing or conceptualization scenarios [21–23]. Ontologies are formal structures that are used to capture knowledge about some specific domain processes of interest [24–25]. Technically, the “ontologies” or formal expressions (taxonomies) per se are used to describe concepts within process domains as well as the relationships that hold between those concepts. Ontologies range from the tools or mechanisms used to create the taxonomies, to the population of the classified elements or database schemas to fully axiomatized theories [11]. Practically, ontologies are used by the domain experts to (manually, semi-automatic, or automatically) fill the semantic gaps that are allied to the data analysis procedures and models.

On the other hand, it is also noteworthy to mention that ontologies are now central to many applications; such as scientific knowledge portals, information management and integration systems, electronic commerce and web services, etc. which are all grounded or built on the LOD scheme.

## **2.2 State-of-the-art: semantic schema for data integration and processing**

Indeed, several areas of application and definition of ontologies (schema) have been noted in the current works of literature especially as it concerns the varied domains of interest. For example, Hashim [26] notes that the term “ontology” is borrowed from the philosophy field that is concerned with being or existence, and further mention that in context of computer and information science, it symbolizes as an “artefact that is designed to model any domain knowledge of interest”. Ontology has also been broadly used in many sub-fields of the computer science and AI, particularly in data pre-processing, management, and LOD related areas such as intelligent information integration and analysis [27], cooperative information management systems [28], knowledge engineering and representation [29], information retrieval [30], information extraction [31], ontology-based information extraction systems [13, 15, 32–34], database management systems [35–37], and semantic-based process mining and analysis [10, 16, 18, 38–40].

Gruber [25] describes the ontological concept or notion as “a formal explicit specification of a conceptualization”. To date, the aforementioned breadth has been the most widely applied and cited definition of ontologies within the computer science field. The description means that ontologies are able to explicitly define (i.e. specify) concepts and relationships that are paramount for modeling any given process or domain of interest. Moreover, with such expressive application

or schema, it means that the processes can be represented in the form of classes, relations, individuals, and axioms (C,R,I,A). Thus, we note that the structural layer of ontologies can be defined as a *quadruple* which are construed on connected sets of taxonomies (RDF + Axioms) or yet formal structure (Triple + Facts). Whereby the *subjects* include the represented class(es), *C*, the *objects* include the individual process elements or instances, *I*, the *predicates* are used to express the relationships, *R*, that exist amongst the subjects and objects, and then sets of axioms that state facts, *A*, [11]. Thus;

$$\text{Ont} = (C,R,I,A) \quad (1)$$

Following the aforementioned definition of the ontological concept or schema, this work note that ontologies serve and are built to perform the main functional mechanisms for the integration of data models for the various systems (e.g. LOD) as follows:

- *Conceptualisation*: method used to represent abstract models of a phenomenon in real-world settings. This is done by identifying suitable domain (semantic) relationships that exist amidst the process elements (concepts) through formal definitions in what can be called declarative axioms that allow for the resultant models to be represented (conceptualisation) declaratively.
- *Explicitness*: procedures that allow or support the different types of concepts and restrictions on their use (properties assertions) to be defined explicitly.
- *Formality*: expressions which are defined to prevent unexpected interpretation of the C,R,I,A as quadruple (e.g. concepts and notations, relationships, properties restrictions, etc.). Thus, it enables the resultant systems or models to be machine-readable and machine-understandable, respectively.

### 3. Proposed semantic-based LOD framework (SBLODF)

The representation (modeling) of knowledge using ontologies (e.g. taxonomies) helps in organizing *metadata* for complex information or data structures. According to Sheth et al. [41], description of real-time processes through metadata creation provides a syntactic as well as semantic way of representing information about the resources that are encoded as instances (entities) in ontological form. Besides, the formal representation of ontologies and the underlying metadata created as a result of the representations allows for automatic reasoning of the processes by making references (inference) to the defined concepts [42]. Indeed, with such reasoning aptitude, the process analysts or owners are able to ensure specification of the process domains (knowledge) in view in an ontological form that can logically be interpreted in an apt way. Consequently, this permits for automatic reasoning of the different concepts to derive an explicit/implicit knowledge about the process domains in question [43].

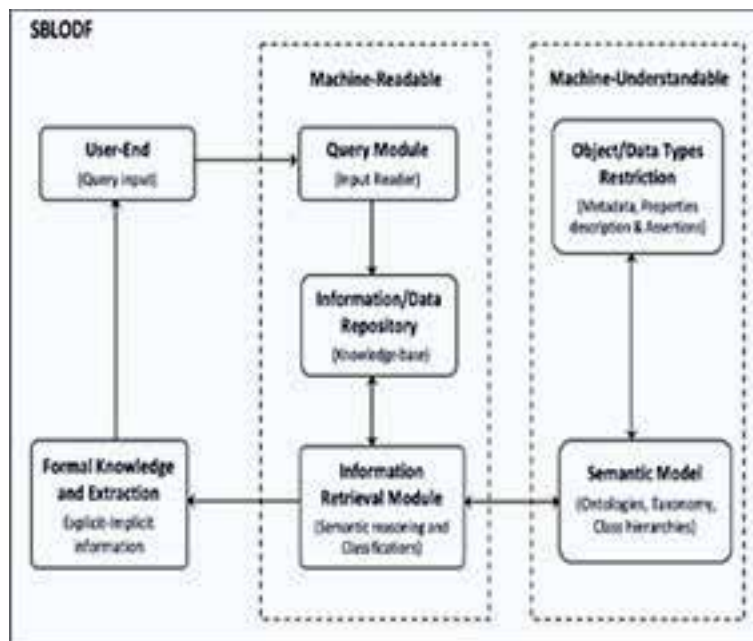
Therefore, the main benefits of ontologies for formal integration of datasets and models in any shape or platform can be summarized in two forms: (i) encoding knowledge about the specific process domains, and (ii) conceptual analysis and reasoning of the processes at more abstraction levels as described in detail in the following section.

### 3.1 Architecture of the SBLODF framework

Information retrieval and structuring of the different sets of data that are stored in several databases or knowledge-base are usually performed in alignment with the users' query [38]. As gathered in **Figure 1**, the supported formats may be a list of document files or keywords issued to the system through the query module (functional operators). In turn, the retrieval module references the properties descriptions (conceptual assertions) that underlie the (semantic) models to produce information that is relevant to the users' query. For example, using the superClass-subClass hierarchies that are usually defined in a taxonomical form in ontologies. This is done through the classification process (e.g. classifying by using a reasoner) to compute the relevant information (e.g. individual entities or process instances) that fulfills the properties restriction by definition [44]. Technically, the most fitting (related) concepts are then presented to the user in a formal way, e.g. explicitly and implicitly.

Furthermore, we note that information retrieval and extraction systems such as the SBLODF framework (**Figure 1**) typically do not only support unstructured data or documents (e.g. textual data), but it also deals with semi-structured and structured data. This is where the semantic technologies and such type of systems (which combines the information retrieval (IR) with information extraction (IE) features) [38] becomes greatly beneficial. Functionally, the resulting system allows for merging and manipulation of structured, semi-structured, and unstructured data through the search (query) modules by enabling a conceptual intersection or reasoning between the different elements as contained in the system. Thus, the SBLODF is referred to as a conceptualization method or information processing system that combines the features of the machine-readable and machine-understandable systems or mechanisms.

For example, enterprise vendors such as FAST (a Microsoft subsidiary) incorporated analytical search functions to support data visualization and reporting into



**Figure 1.** Semantic-based linked open data framework (SBLODF).



their products [38, 45]. Moreover, Ingvaldsen [38] notes that the business process intelligence (BPI) solutions and offerings can also benefit from such a combination (IR and IE supported systems) by giving the users a search facilitated (data analysis) environment to harvest/harness data from both structured and unstructured data sources. Thus far, giving the users a more flexible environment for accessing relevant data items.

Interestingly, semantic-based information retrieval and extraction systems as illustrated in **Figure 1**, represents to be a step further in supporting the BPI's by providing additional modules or components that allows for integrating metadata description (e.g. ontologies) to the system design or framework. The semantic-based components (see: **Figure 1**) aims to add a machine tractable and/or re-purposeable layer of annotations that are relative to ontologies in order to complement the existing web of information and data analysis procedures, or yet, the omnipresence of natural language hypertext [4, 46, 47]. Perhaps, this is fundamentally done through the creation of semantic annotations [11, 23] and linking of the different concepts or modules to ontologies. In turn, the semantically motivated process or models turns out to become automatic or semi-automatic in nature and allows for ample integration of the LOD frameworks due to creation, interrelation, or application of the ontologies (semantic schema). Besides, this has led to the advancement of hybrid intelligent systems such as the ontology-based information extraction systems (OBIE) [9, 13, 15]. Explaining why IE and semantic technologies can be used to bring together a common language or syntax upon which the LOD systems or web search are built specially given the ever-needed formal knowledge or tools for information (data) access and utilization.

Some examples of state-of-the-art tools or systems that trails to support the semantic-based LOD framework or search include; KIM (knowledge and information management system) [31, 48] an extendable platform for information management that seemingly offers IE-based functions for metadata creation and search. Technically, KIM consists of a set of front-end (user-interface) for online information search by offering semantically-enhanced browsing features.

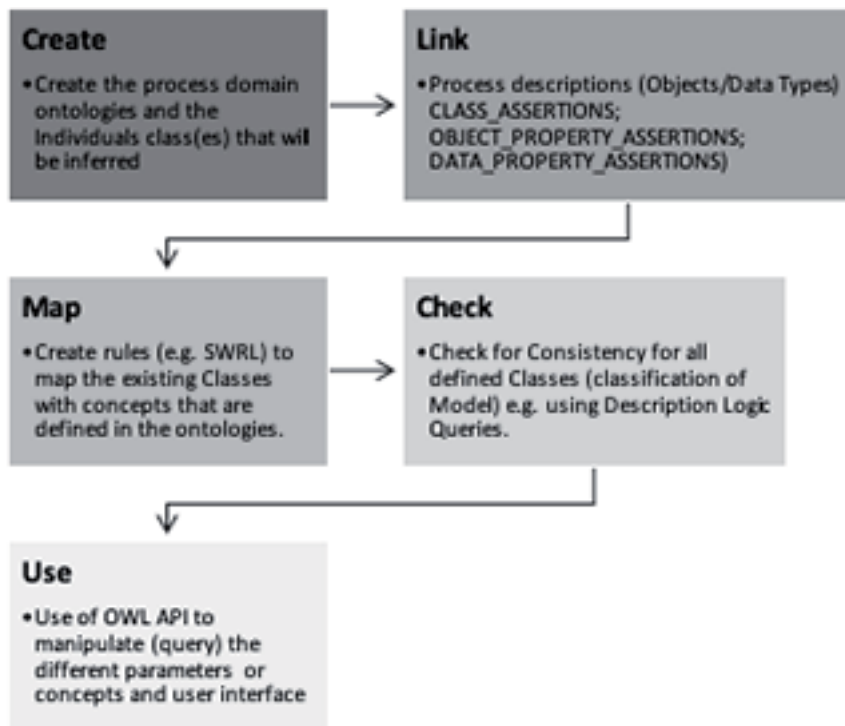
Another tool that tends to support the semantic-based LOD, such as the SBLODF framework described in this chapter, is Magpie [49]. Magpie is developed and implemented as an add-on to web browsers by using IE mechanisms to support collaborative information interpretation and modeling of the extracted knowledge from the web. As illustrated in **Figure 1**, it annotates the different web pages with metadata descriptions in an automated manner by automatically populating ontologies from the relevant (web) sources. Thus, the application (Magpie) is interoperable with ontologies or semantic schema. Moreover, it is important to mention that one of the fundamental elements of the tool (Magpie) that is pertinent to this work is the fact that it makes use of ontologies to provide specific (tailored content) information to the users.

There are several other platforms that can be referred to also support the SBLODF framework. This includes the SemTag [50] which utilizes IE facilities or function to support large scale semantic annotations and process descriptions using TAP ontology. As described in **Figure 1**, SemTag functions by performing annotation of all defined mentions (references) of any given process instance or entity in the ontology (TAP) through a lookup phase. This lookup process is then followed by the disambiguation phase during which it assigns the right classes (or establishes instances that do not correspond to a class in the TAP) using a vector-space model [50].

#### 4. Implementation components of the semantic-based linked open data framework (SBLODF)

The work describes in this section (Figure 2) how the semantic schema is used to support the development of the LOD framework. Ontology-based information retrieval and extraction systems such as the SBLODF (Figure 1) are construed on the main building blocks [31]:

- *Named Entity recognition* (NE) which trails to find and classifies the different concepts that can be found within the model or knowledge-base.
- *Co-reference resolution* (CO) which identifies the relations or association that co-exist amongst the concepts or entities.
- *Template Element construction* (TE) that adds descriptive information (meta-data) to the classified NE through the CO component.
- *Template Relation construction* (TR) that locates the links or references between the TE (entities), and
- *Scenario Template production* (ST) that matches (fits) the TE and TR components into a specified scenario or process instance.



**Figure 2.** Implementing the semantics components in SBLODF using create-link-map-check-use (CLMCU) procedure [11].

Interestingly, Dou et al. [8] note that a well-designed information retrieval or data/process mining system should present the outcomes or discovered information in a formal and structured format qua being interpreted as domain knowledge, or yet, utilized to further augment the existing system. Besides, the work [8] states that ontological schema is one of the most effective ways to formally represent any given type of data or process models. This is due to the fact that concepts defined within ontologies can be expressed or represented as set(s) of annotated terms and/or relations that aims to support information extraction and association rule mining systems especially with those allied to the ontology-based information and extraction (OBIE) [9].

To this effect, this current study note that to implement the aforementioned functionalities of the ontology-based systems in the SBLODF framework, the extracted information or models from the standard process mining (management) or analysis tools/sources needs to be represented as sets of annotated terms (that links or connects the defined terms) in an ontological form using the create-link-map-check-use (CLMCU) incremental or semantic modeling procedure [6, 11].

As illustrated in **Figure 2**, the resultant class hierarchies or taxonomy (ontologies) tends to provide a way of formally representing the defined (annotated) terms or concepts in a structured format by ascertaining the relationships (association) that co-exist amongst the several entities within the process model. Henceforth, the process descriptions and assertions are realized by encoding the process model in the formal structure or taxonomy, thus far ontologies, for the information/knowledge extraction to follow. In the end, the system is integrated or manipulated with an inference engine (e.g. reasoner or classifier) that performs semantic reasoning by uncovering the different levels of the ontological classification and process elements to produce the (inferred) information (knowledge) based on the input queries or users search that displays to be closer to human understanding (machine-understandable).

## **5. Data analysis and implementation results**

For the data analysis and implementation in this section of the chapter; the work uses dataset about a real-time business process provided by the IEEE CIS Task Force on Process Mining [51] to illustrate how the proposed method is capable of performing the information retrieval and extraction process by integrating the different components of the SBLODF framework, as described in **Figure 1**. Typically, this is done by enabling a conceptual intersection or reasoning between the different elements/components which are supported by the system. These functions ranges from the user input query or search module to the information retrieval module or input reader (machine-readable component), and then, the metadata descriptions/assertions, ontological modeling and class hierarchies (taxonomy) to the provision of formal knowledge (explicit and implicit information) that can be easily understood by humans in real-world settings. Fundamentally, the work note the key function of the SBLODF framework to be in its capability to utilize the semantic concepts to perform automatic (semantic) reasoning/inferences capable of discovering useful models and conceptual information from the dataset. Henceforth, the SBLODF implementation allows the meaning of the process elements to be enhanced through the use of property description languages and classification of the discoverable entities, for example, using the Web Ontology Language (OWL) [4], Semantic Web Rule Language (SWRL) [52], and Description Logic (DL) [2].

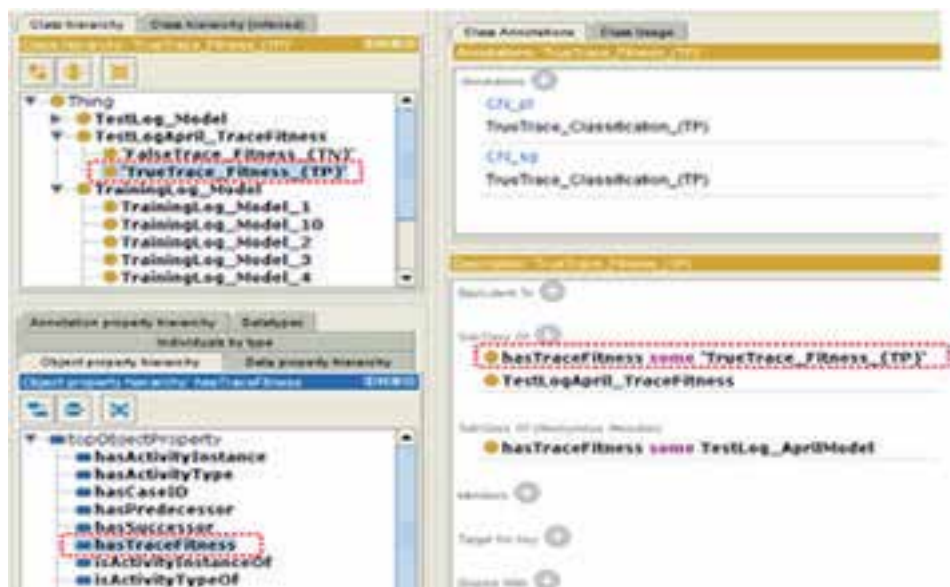
Practically, as shown earlier in **Figure 2**, the ontological schema or framework trails to connect the different sets of discoverable entities in the model with their

class membership, or yet with a fixed literal, and can also describe the sub assumption hierarchies (taxonomies) that exists between the various classes including the relationships that they share within the underlying model. Moreover, the different class(es) are consequently instantiated with the set of individuals,  $I$ , and can also contain the various set of axioms,  $A$ , which states facts. For instance, the true positive elements, i.e., what is true and fitting within the model, and true negatives, i.e., what is true and not fitting in the model.

To illustrate this, the work analyzes the data provided in Ref. [51], by making use of the object properties (see: **Figure 2**) to describe the different classes that can be found within the semantic model developed with Protégé Editor for the purpose of this work. As shown in **Figure 3**, it used the “hasTraceFitness” object property to describe the classes or entities in the test data log that has a “TrueTrace\_Classification\_(TP)” or “FalseTrace\_Classification\_(TN)”

Moreover, as defined in Section 2.2 and Section 4 (**Figure 2**), if we Let  $A$ , be the set of all process executions or actions that can be performed within the semantic model. A process action  $a \in A$  is characterized by a set of input parameters  $In_a \in P$  which is required for the execution of  $a$ , and a set of output parameters  $Out_a \subseteq P$  which is produced by  $a$  after the execution or search query. Thus, with such function, the extraction and automatic reasoning (e.g. classification) of the process parameters is enabled and/or supported by the model. Perhaps, the key purpose of implementing the framework is to match the questions one would like to answer about attributes/relationships the process instances share amongst themselves within the knowledge-base by linking to the concepts (inferred classes) described in the model.

As shown in **Table 1**, based on the features of the provided datasets [51], the work applies the cross-validation technique to analyze the training and test sets. The traces were computed and recorded according to the *reasoner* response, and the classifier (reasoner) was tested on the resulting individuals by assessing its performance with respect to the correctly classified traces. As an example, the following DL queries/syntax [2] represents as set of input parameters (search query) the work executed in order to output the set of traces that can be found within the



**Figure 3.**  
Example of object property description and assertion for the true trace classification.

	Model1	Model2	Model3	Model4	Model5	Model6	Model7	Model8	Model9	Model10
Trace_1	TP*	TN*	TP*	TN*	TN*	TN*	TP*	TP*	TP*	TP*
Trace_2	TN*	TN*	TP*	TP*	TP*	TP*	TP*	TN*	TP*	TP*
Trace_3	TP*	TP*	TP*	TN*	TN*	TN*	TN*	TP*	TP*	TN*
Trace_4	TP*	TP*	TN*	TP*	TN*	TP*	TN*	TP*	TP*	TN*
Trace_5	TN*	TN*	TN*	TP*	TN*	TP*	TN*	TP*	TP*	TN*
Trace_6	TP*	TN*	TN*	TP*	TN*	TP*	TP*	TN*	TN*	TP*
Trace_7	TN*	TP*	TP*	TN*	TN*	TP*	TN*	TP*	TN*	TN*
Trace_8	TN*	TP*	TP*	TP*	TN*	TN*	TP*	TP*	TP*	TP*
Trace_9	TP*	TN*	TP*	TN*	TP*	TN*	TP*	TP*	TN*	TP*
Trace_10	TP*	TN*	TP*	TN*	TN*	TN*	TP*	TP*	TP*	TP*
Trace_11	TN*	TP*	TP*	TP*	TP*	TN*	TN*	TN*	TN*	TP*
Trace_12	TP*	TN*	TN*	TP*	TP*	TP*	TP*	TN*	TP*	TN*
Trace_13	TP*	TP*	TN*	TN*	TP*	TN*	TN*	TN*	TN*	TP*
Trace_14	TN*	TP*	TN*	TN*	TN*	TN*	TN*	TP*	TN*	TP*
Trace_15	TP*	TN*	TN*	TN*	TP*	TP*	TN*	TN*	TN*	TN*
Trace_16	TN*	TN*	TN*	TP*	TP*	TN*	TN*	TN*	TP*	TN*
Trace_17	TP*	TP*	TP*	TP*	TP*	TP*	TP*	TN*	TN*	TP*
Trace_18	TN*	TP*	TN*	TN*	TP*	TP*	TP*	TN*	TN*	TN*
Trace_19	TN*	TP*	TP*	TP*	TN*	TP*	TP*	TP*	TN*	TN*
Trace_20	TN*	TN*	TN*	TN*	TP*	TN*	TN*	TN*	TP*	TN*
True positives (TP):	10	10	10	10	10	10	10	10	10	10
False positives (FP):	0	0	0	0	0	0	0	0	0	0

	Model1	Model2	Model3	Model4	Model5	Model6	Model7	Model8	Model9	Model10
True negatives (TN):	10	10	10	10	10	10	10	10	10	10
False negatives (FN)	0	0	0	0	0	0	0	0	0	0
No. of traces correctly classified	20	20	20	20	20	20	20	20	20	20

*Note: cells with gold sign (\*) indicates traces that were correctly classified by the reasoner which equals to 200 traces out of 200.*

**Table 1.**  
 Classification results and performance of the discovered models.

defined model that has “TrueTrace\_Fitness\_(TP)” and “FalseTrace\_Fitness\_(TN)” respectively.

“TestLog\_(forSpecifiedClass) and hasTraceFitness some  
"TrueTrace\_Fitness\_(TP)”.

“TestLog\_(forSpecifiedClass) and hasTraceFitness some  
'FalseTrace\_Fitness\_(TN)”.

Thus, as reported in **Table 1**, each results of the classification process for the discovered models, i.e., the true positives and true negatives traces, were determined.

From the results of the classification method (**Table 1**), we note for each run set of parameters retrieved from the model that the commission error, otherwise referred to as error-rate (false positives (FP) and false negatives (FN)) was null, thus, equal to 0. This means that the reasoner (classifier) did not make critical mistakes. For instance, a case whereby a trace could be considered to be an instance of a class while it is categorically an instance of another class. In the same vein, the work notes that the accuracy rate (i.e., true positives (TP) and true negatives (TN)) when determining the different traces and classifications was very high, thus, correct, and were consistently observed for all the test sets.

## **6. Discussion and conclusion**

LOD systems or frameworks and algorithms are fundamentally aimed to provide a standard platform for integrating/analyzing different datasets or models to extract snippets of information that are relevant to the users, independent of the various formats or syntax. In other words, LOD stands as the bridge between the different data formats/sources and knowledge acquisition or information retrieval. For example, Cunningham [31] notes that the process of extracting information from the several sources may simply imply taking text documents, speech, graphics, etc., as input and produces fixed-format unambiguous data (or information) as output. In turn, the discovered information or data may be directly displayed to the users, stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in IR-supported applications such as the web search technologies, internet, or search engines like Google, Bing, etc.

Studies have shown that IE technologies may be distinctive from the IR systems or functions. Whereas, Cunningham [31] notes that the IR systems aims to find relevant information (e.g. texts) and presents them to the users, an IE application analyses the texts and presents only the specific information from the text that the user is interested in. Apparently, this kind of tailored information analysis is where ontology-based information extraction systems such as the SBLODF framework described in this chapter construes its incentives.

For example, a user of an IR-supported system wanting information on higher educational institutions that offers a particular course would enter a list of relevant words or keywords in the search module and receive in return a set of documents (e.g. various university prospectus, course guidelines, etc.) that contain likely matches based on the keywords. In turn, the user would read through the matches or documents and extract the requisite information they need themselves, or yet store them on their computer storage for future reference. Nonetheless, unlike IR, an IE system would automatically populate a list of tables or spreadsheets directly with the names of relevant universities and their course offerings making it easier for the users to extract or learn the specific information they need or seek to acquire.

However, there may also exist some limitations with IE supported LOD frameworks or systems when compared to IR only. One of the limitations is that IE systems are more difficult and knowledge-intensive to build and are to a certain extent tied to particular domains or case scenarios. Also, IEs are more computationally intensive than IRs. Although, on the other hand, when compared to applications where there are large text or document volumes, IEs are potentially much more efficient than IRs due to the capacity of dramatically reducing the amount of time people may spend reading through text documents to find the relevant information. Perhaps, the aforementioned benefit of the IEs is only possible as a result of applying the ontological (semantics) schema to represent and manipulate the underlying information as described in Section 3 of this chapter.

Moreover, in settings where the results need to be presented, for example, in several languages; the fixed-format and unambiguous nature of IEs outputs make the information retrieval process relatively direct when compared to the full translation facilities that are consequently needed for interpretation of the multilingual texts found by IRs. Indeed, this means that IEs only present the specific information in a form that the user is interested in, and this feature is where the ontology-based IE systems are more powerful given that ontology is one of such tools that have the capability of providing information in a structured format. For instance, the automatic population of the different class hierarchies in ontologies within OBIE [9] applications is capable of formally identifying process instances or element within a text file that belongs to or references certain concepts in the pre-defined ontologies, and then trails to add those instances to the model in the right locations.

Having said that, we note that OBIE systems such as the SBLODF attempts to classify the several entities in a more scalar way; as there may be different categories to which an entity can belong to and cataloging the discrepancies between those classifications is more or less straightforward when using the OBIE framework [17].

Furthermore, to explain the application of the OBIE concept in the context of information retrieval and extraction or semantic-based knowledge representation, Yankova [20] refers to an identity resolution method of deciding whether an instance extracted from a text by an IE application refers to a known entity within a target domain ontology. Technically, the authors [20] developed a customizable rule-based framework for identity resolution and merging that uses ontologies for knowledge representation by using customizable identity criteria put in place to decide on the similarity between two process instances or entities. The criteria utilizes ontological operations and similarity computation between extracted and stored values that are weighted. Besides, the weighting criteria are routinely specified according to the type of entities and the application domain.

Accordingly, studies have also shown that aggregation of the extracted information from the different data sources has greater advantages (e.g. complementing partial information from one source to another, increasing the confidence of the extracted information, and storage of updated information within the knowledge bases) [11, 14, 15, 17, 20, 23, 53]. Truly, the resultant methods prove to provide standard structures for resolving the identities or properties description of the different class(es) of entities (process instances) by using ontologies as the core (fundamental) knowledge representation tools that help to provide the formal descriptions that are complemented with semantics.

Interestingly, Yankova [20] reveals that one fundamental problem to be addressed when providing a structure for distribution of the conceptual knowledge such as with OBIE systems; is that of identifying and merging the instances extracted from the multiple sources. Basically, the process should aim at identifying newly extracted facts, e.g. from the derived models, and linking them to



their previous references or mentions. To this effect, we note that ontology-based systems, in general, poses two main challenges that are directed towards [31]:

- identification of the concepts (e.g., process instances or entities) within the ontologies, and
- automatic population of the ontologies with newly (inferred or classified) instances.

Perhaps, it is also important to mention that when the ontologies are populated with the process instances or concepts assertions; the ultimate function of the resultant (OBIE) systems would simply be to manipulate the process elements, for example, by uncovering the relationships that exist amongst the process instances and revealing those to the users or search initiators based on the query modules [6, 9, 16, 31, 44, 54]. Moreover, for rule-based systems like OBIE, such procedures are relatively unswerving. But for learning-based IE systems, it appears to be more problematic due to the fact that training data are most often required to train the models, and collecting the necessary training data is, on the other hand, likely to be cumbersome/bottleneck [31]. Although to resolve such issues, new training datasets may need to be created either manually or semi-automatically; which are a lot of the time is time-consuming and/or burdensome task.

However, new and emerging systems/methods are being developed with the aim to help address such *metadata creation* problems for knowledge management or data analysis to support the IE and LOD at large [1, 11–15, 23, 33, 55, 56]. Moreover, unlike the traditional IE systems where the extracted facts (or information) are only classified as belonging to pre-defined types, an ontology-based (semantic) IE system (such as the SBLODF) seeks to identify, analyze and represent information at the conceptual (abstraction) levels by establishing a link (references) between the entities residing in the underlying systems' knowledge-bases and their mentions within the contextual domain. Henceforth, semantically-based LOD systems should not only support the formal representation of the different domains. But should also, on the other hand, provide information about the several known entities and their properties descriptions. Thus, ontology-based LOD systems such as the SBLODF introduced in this chapter must integrate well-defined entities with their semantic descriptions for an efficient explicit and implicit information extraction and/or analysis, i.e., machine-readable and machine-understandable system.

## **Acknowledgements**

The author would like to acknowledge the technical support of Writing Lab, TecLabs, Tecnológico de Monterrey, in the publication of this work.

## Author details

Kingsley Okoye

Writing Lab, TecLabs, Office of the Vice President for Research and Technology Transfer, Tecnológico de Monterrey, Monterrey, CP 64849, Nuevo Leon, Mexico

\*Address all correspondence to: [kingsley.okoye@tec.mx](mailto:kingsley.okoye@tec.mx)

## IntechOpen

---

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] T. Declerck et al., “Recent Developments for the Linguistic Linked Open Data Infrastructure,” in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 5660-5667.
- [2] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF. *The Description Logic Handbook: Theory, Implementation and Applications*, 2nd Ed. Cambridge University Press; 2007
- [3] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, and K. Srinivas, “SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems,” Springer, Cham, 2020, pp. 514-530.
- [4] S. Bechhofer et al., “OWL Web Ontology Language Reference,” Technical report W3C Proposed Recommendation, Manchester, UK, 2004.
- [5] C. D’Amato, N. Fanizzi, and F. Esposito, “Query answering and ontology population: An inductive approach,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 5021 LNCS, pp. 288-302, doi: 10.1007/978-3-540-68234-9\_23.
- [6] K. Okoye, S. Islam, and U. Naeem, “Ontology: Core Process Mining and Querying Enabling Tool,” in *Ontology in Information Science*, C. Thomas, Ed. IntechOpen, 2018, pp. 145-168.
- [7] Snyder E, Lorenzo L, Mak L. *Linked open data for subject discovery: Assessing the alignment between Library of Congress vocabularies and Wikidata*. In: *International Conference on Dublin Core and Metadata Applications*. 2019
- [8] D. Dou, H. Wang, and H. Liu, “Semantic data mining: A survey of ontology-based approaches,” in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, IEEE ICSC 2015*, 2015, pp. 244-251, doi: 10.1109/ICOSC.2015.7050814.
- [9] Wimalasuriya DC, Dou D. *Ontology-based information extraction: An introduction and a survey of current approaches*. *Journal of Information Science*. Jun. 2010;**36**(3):306-323. DOI: 10.1177/0165551509360123
- [10] A. K. A. De Medeiros and W. M. P. Van Der Aalst, “Process mining towards semantics,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4891 LNCS, T. Dillon, E. Chang, R. Meersman, and K. Sycara, Eds. Springer, Berlin, Heidelberg, 2009, pp. 35-80.
- [11] Okoye K, Islam S, Naeem U, Sharif MS, Sharif MhD S. *Semantic-based process mining technique for annotation and modelling of domain processes*. *Int. J. Innovative Computing & Information Control*. 2020;**16**(3):899-921
- [12] Wang T. *Aligning the large-scale ontologies on schema-level for weaving Chinese linked open data*. *Cluster Comput*. Mar. 2019;**22**(2):5099-5114. DOI: 10.1007/s10586-018-1732-z
- [13] D. Calvanese, M. Montali, A. Syamsiyah, and W. M. P. van der Aalst, “Ontology-driven extraction of event logs from relational databases,” in *Lecture Notes in Business Information Processing*, 2016, vol. 256, pp. 140-153, doi: 10.1007/978-3-319-42887-1\_12.
- [14] De Giacomo G, Lembo D, Lenzerini M, Poggi A, Rosati R. *Using*

- ontologies for semantic data integration. In: Flesca S, Greco S, Masciari E, Saccà D, editors. *A Comprehensive Guide through the Italian Database Research over the Last 25 Years*. Springer: Cham; 2018. pp. 187-202
- [15] D. Calvanese, T. E. Kalayci, M. Montali, and S. Tinella, "Ontology-based data access for extracting event logs from legacy data: The onprom tool and methodology," in *Lecture Notes in Business Information Processing*, vol. 288, W. Abramowicz, Ed. Springer Verlag, 2017, pp. 220-236.
- [16] A. K. A. de Medeiros, W. van der Aalst, and C. Pedrinaci, "Semantic process mining tools: core building blocks," in *ECIS, Ireland, June 2008*, 2008, pp. 1953-1964.
- [17] Maynard D, Peters W, Li Y. Evaluating evaluation metrics for ontology-based applications: Infinite reflection. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. 2008
- [18] A. H. Cairns, J. A. Ondo, B. Gueni, M. Fhima, M. Schwarcfeld, C. Joubert and N. Khelifa, "Using semantic lifting for improving educational process models discovery and analysis," in *CEUR Workshop Proceedings*, 2014, pp. 150-161.
- [19] Selvan NS, Vairavasundaram S, Ravi L. Fuzzy ontology-based personalized recommendation for internet of medical things with linked open data. *Journal of Intelligent Fuzzy Systems*. Jan. 2019;**36**(5):4065-4075. DOI: 10.3233/JIFS-169967
- [20] Yankova M, Saggion H, Cunningham H. *Semantic-Based Identity Resolution and Merging for Business Intelligence*. UK: Sheffield; 2008
- [21] N. Khasawneh and C. C. Chan, "Active user-based and ontology-based Web log data preprocessing for Web usage mining," in *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)*, WI'06, 2006, pp. 325-328, doi: 10.1109/WI.2006.32.
- [22] D. Perez-Rey, A. Anguita, and J. Crespo, "OntoDataClean: Ontology-based integration and preprocessing of distributed data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006, vol. 4345 LNBI, pp. 262-272, doi: 10.1007/11946465\_24.
- [23] K. Okoye, "Technique for annotation of fuzzy models: A semantic fuzzy mining approach," in *Frontiers in Artificial Intelligence and Applications*, 2019, vol. 320, pp. 65-75, doi: 10.3233/FAIA190166.
- [24] Gruber TR. A translation approach to portable ontology specifications. *Knowledge Acquisition*. Jun. 1993;**5**(2):199-220. DOI: 10.1006/knac.1993.1008
- [25] Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum. - Comput. Stud.* Nov. 1995;**43**(5-6):907-928. DOI: 10.1006/ijhc.1995.1081
- [26] Hashim H. Ontological structure representation in reusing ODL learning resources. *Asian Assoc. Open Univ. J.* Aug. 2016;**11**(1):2-12. DOI: 10.1108/aaouj-06-2016-0008
- [27] Seng JL, Kong IL. A schema and ontology-aided intelligent information integration. *Expert Systems with Applications*. Sep. 2009;**36**(7):10538-10550. DOI: 10.1016/j.eswa.2009.02.067
- [28] Ouksel AM, Sheth A. Semantic interoperability in global information systems: A brief introduction to the research area and the special section.

- SIGMOD Rec. Dec. 1999;**28**(1):5-12.  
DOI: 10.1145/309844.309849
- [29] Brewster C, O'Hara K. Knowledge representation with ontologies: Present challenges-future possibilities. *International Journal of Human Computer Studies*. Jul. 2007;**65**(7):563-568. DOI: 10.1016/j.ijhcs.2007.04.003
- [30] Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge University Press; 2008
- [31] Cunningham H. *Information Extraction, Automatic*. UK: Sheffield; 2005
- [32] H. M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature," *PLoS Biol.*, vol. 2, no. 11, Nov. 2004, doi: 10.1371/journal.pbio.0020309.
- [33] H. M. Müller, K. M. Van Auken, Y. Li, and P. W. Sternberg, "Textpresso Central: A customizable platform for searching, text mining, viewing, and curating biomedical literature," *BMC Bioinformatics*, vol. 19, no. 1, Mar. 2018, doi: 10.1186/s12859-018-2103-8.
- [34] S. A. Hosseini, A.-R. H. Tawil, H. Jahankhani, and M. Arandi, "Towards an Ontological Learners' Modelling Approach for Personalised E-Learning," *Int. J. Emerg. Technol. Learn.*, vol. 8, no. 2, p. 4, 2013.
- [35] Alkharouf NW, Jamison DC, Matthews BF. Online analytical processing (OLAP): A fast and effective data mining tool for gene expression databases. *Journal of Biomedicine & Biotechnology*. Jun. 2005;**2005**(2):181-188. DOI: 10.1155/JBB.2005.181
- [36] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati, "Linking data to ontologies," in *Journal on Data Semantics X.*, vol. 4900 LNCS, S. Spaccapietra, Ed. Springer Verlag, 2008, pp. 133-173.
- [37] C. Snae and M. Brückner, "Ontology-Driven E-Learning System Based on Roles and Activities for Thai Learning Environment," *Interdiscip. J. e-Skills Lifelong Learn.*, vol. 3, pp. 001-017, 2007, doi: 10.28945/382.
- [38] Ingvaldsen JE. *Semantic Process Mining of Enterprise Transaction Data*. Norway; 2011
- [39] K. Okoye, A. R. H. Tawil, U. Naeem, S. Islam, and E. Lamine, "Using semantic-based approach to manage perspectives of process mining: Application on improving learning process domain data," in 2016 IEEE International Conference on Big Data, BigData2016, 2016, Washington DC, USA, pp. 3529-3538, doi: 10.1109/BigData.2016.7841016.
- [40] Okoye K, Naeem U, Islam S. Semantic fuzzy mining: Enhancement of process models and event logs analysis from syntactic to conceptual level. *Int. J. Hybrid Intell. Syst.* Nov. 2017;**14**(1-2):67-98. DOI: 10.3233/his-170243
- [41] Sheth A, Bertram C, Avant D, Hammond B, Kochut K, Warke Y. Managing semantic content for the web. *IEEE Internet Computing*. Jul. 2002;**6**(4):80-87. DOI: 10.1109/MIC.2002.1020330
- [42] P. Dolog and W. Nejdl, "Semantic web technologies for the adaptive web," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2007, vol. 4321 LNCS, pp. 697-719, doi: 10.1007/978-3-540-72079-9\_23.
- [43] Yarandi M. *Semantic Rule-Based Approach for Supporting Personalised*

Adaptive E-Learning. United Kingdom: University of East London; 2013

[44] K. Okoye, A. R. H. A.-R. H. Tawil, U. Naeem, and E. Lamine, "Discovery and enhancement of learning model analysis through semantic process mining," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, 2016, vol. 8(2016), pp. 093-114

[45] De Leoni M, Adams M, Van Der Aalst WMP, Ter Hofstede AHM. Visual support for work assignment in process-aware information systems: Framework formalisation and implementation. *Decision Support Systems*. Dec. 2012;**54**(1):345-361. DOI: 10.1016/j.dss.2012.05.042

[46] Fensel D, Hendler JA, Lieberman H, Wahlster W, Berners-Lee T, Lieberman H. *Spinning the Semantic Web : Bringing the World Wide Web to its Full Potential*. MIT Press; 2003

[47] J. Davies, D. Fensel, and F. Van Harmelen, *Towards the semantic web : ontology-driven knowledge management*. J. Wiley, 2003.

[48] Popov B, Kiryakov A, Ognyanoff D, Manov D, Kirilov A. KIM - a semantic platform for information extraction and retrieval. *Natural Language Engineering*. Sep. 2004;**10**(3-4):375-392. DOI: 10.1017/S135132490400347X

[49] J. Domingue, M. Dzbor, and E. Motta, "Magpie: supporting browsing and navigation on the semantic web," in *Proceedings of the 9th international conference on Intelligent user interface - IUI '04*, 2004, pp. 191-197, doi: 10.1145/964442.964479.

[50] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. V. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, J. Y. Zien, "SemTag and seeker: Bootstrapping the semantic web via automated semantic annotation," in *Proceedings of the 12th International*

*Conference on World Wide Web, WWW 2003*, 2003, pp. 178-186, doi: 10.1145/775152.775178.

[51] J. Carmona, M. de Leoni, B. Depair, and T. Jouck, "IEEE CIS Task Force on Process Mining - Process Discovery Contest", 1st Edition, 2016 [https://www.win.tue.nl/ieeetfpm/doku.php?id=shared:edition\\_2016](https://www.win.tue.nl/ieeetfpm/doku.php?id=shared:edition_2016)

[52] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Groszof, and M. Dean, "SWRL: A Semantic Web Rule Language Combining OWL and RuleML", W3C Member Submission. <https://www.w3.org/Submission/SWRL/>

[53] K. Okoye, S. Islam, U. Naeem, M. S. M. S. Sharif, M. A. M. A. Azam, and A. Karami, "The application of a semantic-based process mining framework on a learning process domain," in *Advances in Intelligent Systems & Computing*, 2019, vol. 868, pp. 1381-1403, doi: 10.1007/978-3-030-01054-6\_96.

[54] Okoye K, Tawil ARH, Naeem U, Lamine E. A semantic reasoning method towards ontological model for automated learning analysis. *Advances in Intelligent Systems & Computing*. 2016;**419**:49-60

[55] Okoye K, *Applications and Developments in Semantic Process Mining*. IGI Global Publishers. Hershey, USA. 2020

[56] Polyvyanyy A, Ouyang C, Barros A, van der Aalst WMP. Process querying: Enabling business intelligence through query-based process analytics. *Decision Support Systems*. Aug. 2017;**100**:41-56. DOI: 10.1016/j.dss.2017.04.011



# Analysis of Effective Load Balancing Techniques in Distributed Environment

*Anju Shukla, Shishir Kumar and Harikesh Singh*

## Abstract

Computational approaches contribute a significance role in various fields such as medical applications, astronomy, and weather science, to perform complex calculations in speedy manner. Today, personal computers are very powerful but underutilized. Most of the computer resources are idle; 75% of the time and server are often unproductive. This brings the sense of distributed computing, in which the idea is to use the geographically distributed resources to meet the demand of high-performance computing. The Internet facilitates users to access heterogeneous services and run applications over a distributed environment. Due to openness and heterogeneous nature of distributed computing, the developer must deal with several issues like load balancing, interoperability, fault occurrence, resource selection, and task scheduling. Load balancing is the mechanism to distribute the load among resources optimally. The objective of this chapter is to discuss need and issues of load balancing that evolves the research scope. Various load balancing algorithms and scheduling methods are analyzed that are used for performance optimization of web resources. A systematic literature with their solutions and limitations has been presented. The chapter provides a concise narrative of the problems encountered and dimensions for future extension.

**Keywords:** load balancing, resource management, resource scheduling, load measurement, fault tolerance

## 1. Introduction

The performance of any web server has been affected by the web traffic usually, and the web server makes a slow response because it gets overloaded. Due to the increased traffic over the Internet, a web server faces challenges to serve the large number of users with high-speed availability. Therefore, the concept of resource confederation comes in existence. The popular Google web server works on the same concept. It distributes the user's query in different web servers which are geographically distributed at various locations. The web server requires several mechanisms to deal with linked open data (LOD) to fulfill the user's request [1, 2]. Load balancing plays a vital role in the operation of distributed and parallel computing. It partitioned the incoming workload into smaller tasks that are assigned to computational resources for concurrent execution. The load may be CPU capacity, memory size, network load, delay, etc. The reason behind load balancing is to handle requests of multiple users without degrading the performance of web server. Load balancer receives requests from user, determines the load on available



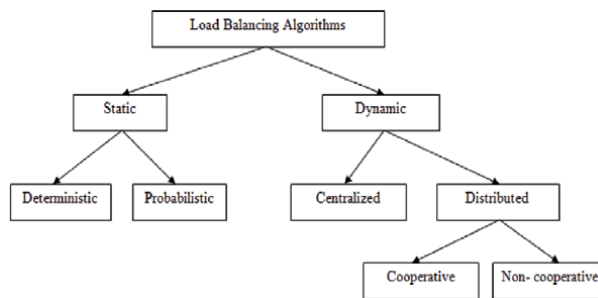
resources, and sends request to the server which is lightly loaded. The major functions of load balancer are as follows:

- Distributes incoming traffic across multiple computational resources
- Determines resource availability and reliability for task execution
- Improves resource utilization
- Increases client satisfaction
- Provides fault tolerance and flexible framework by adding or subtracting resources as demand occurs

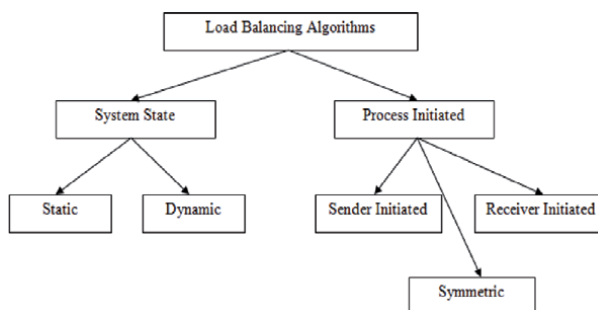
The load balancing algorithm (LBA) exists based on current state of the system as shown in **Figure 1**. Static algorithms require prior information about the system characteristics such as processing capability, memory, number of active connections etc., while DLB algorithms use current status of the system to make scheduling decisions.

Load balancing significantly improves global system performance in terms of throughput and resource utilization. The several reasons to use LBAs are as follows: cost optimization, fault tolerance ability, system adaptability and extensibility, decreased response time, idle time of resources, increased throughput, reliability, and prevents starvation [3, 4].

To incorporate these benefits, it is important to select the suitable load balancing algorithm (LBA) for web resources in distributed environment [5, 6]. Based on a process characteristic, LBAs can be categorized in three categories as shown in **Figure 2**. Both sender-initiated and receiver-initiated algorithms use different



**Figure 1.**  
Classification based on System state.



**Figure 2.**  
Classification based on Process Origination.

transfer and location policies for implementing load balancing. Symmetric-initiated algorithms eliminate the preemption condition of receiver-initiated algorithm and offer two algorithms: above-average algorithm and adaptive algorithm. Above-average algorithm uses an acceptable range for deciding whether a node is sender-initiated or receiver-initiated.

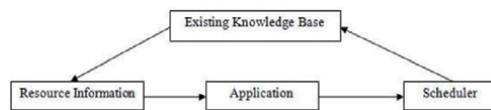
### 1.1 Static load balancing

Static load balancing approaches use the prior information of tasks, computing resources or processing element, and network detail as shown in **Figure 3**. The task can be submitted to any processing element using two methods:

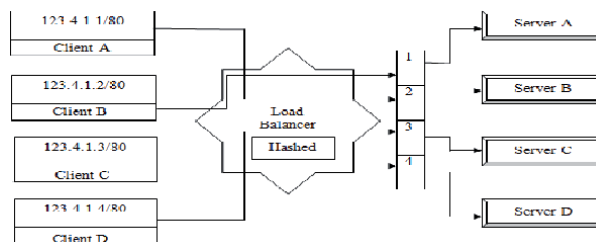
- Stateless method
- State-based method

In stateless method, selection of a processing element (PE) is done without having any awareness of the system environment, while in a state-based method, selecting a PE requires information of the system condition [7]. Stateless methods are simple to implement, but it provides one-to-one interaction between the client and server at a time as shown in **Figure 4**.

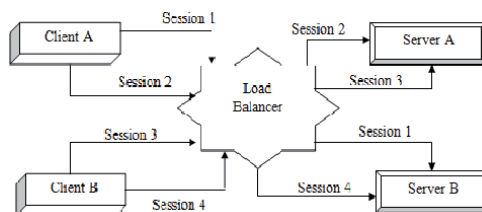
**Figure 5** represents the stateful load balancing method; the load balancer keeps track for all the sessions, and decisions are taken based on server load. Various stateless techniques exist for selecting the processing element such as RR-LBA, weighted



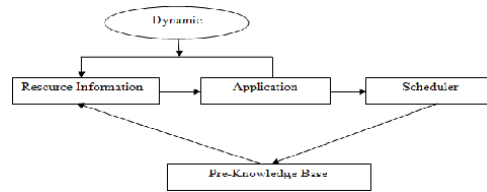
**Figure 3.**  
 Static load balancing.



**Figure 4.**  
 Stateless static load balancing.



**Figure 5.**  
 Stateful static load balancing.



**Figure 6.**  
*Dynamic load balancing.*

round robin (WRR)-LBA, and random allocation algorithm [7]. However, these algorithms have limited scope due to the dynamic nature of distributed environment.

## 1.2 Dynamic load balancing

It varies from the SLB algorithms in which clients' requests are distributed among available resources at run time. The LB assigns the request based on the dynamic information collected from all the resources as shown in **Figure 6**.

DLB algorithms can be classified in two categories—distributed and non-distributed. In distributed DLB, all computing resources are equally responsible for balancing the load. The responsibility of load balancing is shared among all the resources. But in non-distributed algorithms, each resource performs independently to accomplish the common goal. Generally, distributed DLB algorithms generated more message overhead than non-distributed DLB due to its interaction with all the resources.

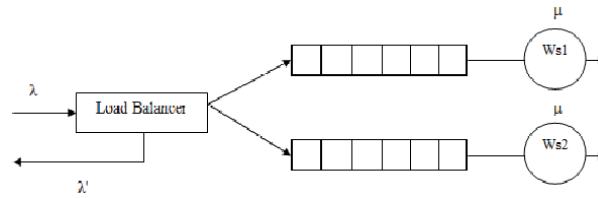
Distributed algorithms perform better in fault conditions as it degrades only the sectional of the system instead of global system performance. Non-distributed algorithms are further classified into two categories—centralized and semi-centralized. In centralized algorithm, a central server is responsible for executing load balancing algorithm. In semi-centralized, servers are arranged in clusters, and load balancing within the cluster is managed centrally.

## 2. Challenges of load balancing

The load balancer implements several load balancing algorithms to determine the suitable resource. However, it faces several issues while distributing the load across available resources. Several major issues with their respective solutions are presented in the next section.

### 2.1 Increased web traffic

Over the last few years, the web traffic is increased very rapidly due to numerous registered websites and online transactions. As the numbers of requests are increased, the response of server becomes slow due to the limited number of open connections. The requests are added to the overall processing capability of resources. When incoming requests go beyond the capability of the resource, a resource crashes or fault occurs. Several authors analyzed and suggested the solution to resolve the issue. The first solution is the server upgradation in which requests are handled by a more powerful server for a while. But, scalability, interruption, and maintenance issues are associated with this solution. Another solution is the outsourcing in which requests are sent to another suitable server for speedy response. But this approach is costly and has limited control over the QoS issues. Chen et al. [8] suggested that the web page size and number of users both affect the system response time.



**Figure 7.**  
*Centralized queueing model.*

The most favorable solution is to use the multiple servers with an efficient load balancer which balances the load among servers. The performances of these servers are analyzed through queueing models or waiting line models. Broadly, two types of load balancing models are used to analyze the web server performance. Each approach has its benefits, applications, and limitations.

### 2.1.1 Centralized queueing model for load balancing

In this mechanism, homogeneous servers with finite buffer sizes are used as shown in **Figure 7**. The load balancer receives request from the user and redirects the request among servers using one of these routing policy:

- Random policy
- RR policy
- Shortest queue policy

Zhang and Fan [9] compared these policies in terms of rejection rate and system response time. They analyzed that these algorithms perform well when traffic is light. But when web traffic becomes high, shortest queue policy performs better than random and RR policy. The number of rejections in RR and random policy is increased as the traffic increases. Singh and Kumar [10] presented a queueing algorithm for measuring the overloading and serving capacity of server in distributed load balancing environment. The algorithm performs better in both homogeneous and heterogeneous environment than the remaining capacity (RC) and server content based queue (QSC) algorithms.

### 2.1.2 Distributed queueing model for load balancing

These mechanisms address the network latency issue also, which avoids network congestion. The queueing models follow certain arrival and distribution rules to distribute the requests. Zhang and Fan [9] suggested that distributed queueing models perform well in heavy traffic conditions. Routing decisions are taken on the basis of queue length differences of web servers. The collected information is used in traffic distribution for improving the performance of web servers. Singh and Kumar [11] suggested that task completion time directly affects the queue length of the web server. They presented a model based on the ratio factor of the task's average completion time. The model is compared with the model presented by Birdwell et al. [12], and it performs better for two performance metrics: average queue length and average waiting time of web servers.

Li et al. [13] analyzed network delay and presented a delay controlled load balancing approach for improving network performance. However, the approach has limited applicability and is suitable for stable path states.

Kamali et al. [14] used queueing theory to monitor the network traffic and its simulation is performed in a homogeneous as well as heterogeneous network environment. The traffic monitoring is required for the calculation of confidence and efficiency parameters from steady operations of the network. Based on queueing theory and little’s law network congestion rate is balanced.

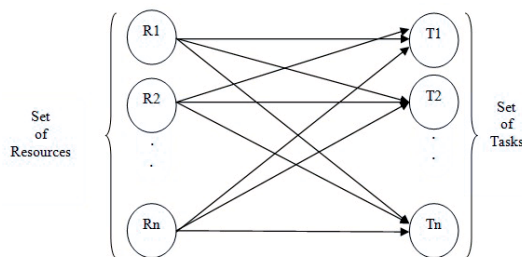
## 2.2 Resource selection and task allocation

Many researchers have addressed the problem of resource selection and task allocation for the fair perspective of load balancing. It is the responsibility of load balancer to map resource and task before actual execution as shown in **Figure 8**.

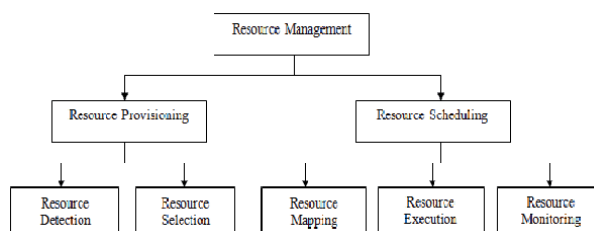
The resource management consists of two major functions: resource provisioning and resource scheduling. In resource provisioning, the user submits task to the broker with various predefined QoS constraints. The broker is responsible to find the suitable resource for task execution. The resource scheduling is all about mapping and execution of task on the appropriate resource . It comprises of three major functions: resource mapping, resource execution and resource monitoring as shown in **Figure 9**. Various types of resources that need to be managed are shown in **Figure 10**.

Hao et al. [15] categorized the resource in three categories—underloaded, normal loaded, and overloaded. The scheduler assigns the task to underloaded or normal-loaded resources only. Chang et al. [16] categorized the resources into L discrete levels for selecting the fittest resource for task execution. Arabnejad and Barbosa [17] presented a budget-based task scheduling and calculated the worthiness of all the resources for resource selection.

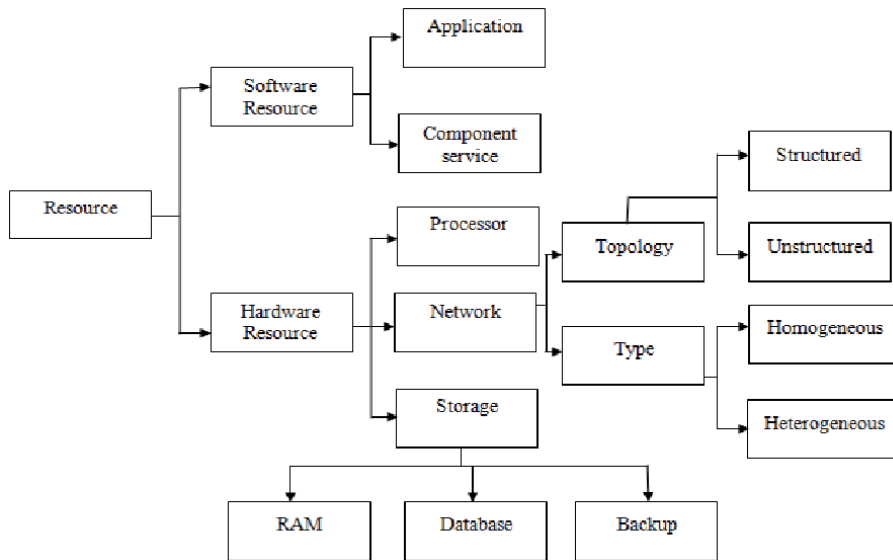
Naik et al. [18] presented a value function to select a resource for task execution. A value function is calculated using completion ratio and historic information of a resource. For minimizing the data transfer between the resources, Cheng et al. [19] used a hypergraph which identifies task and data dependency. Tasks that use similar data are assigned to the same resource to decrease the cost indirectly. Abdelrouf et al. [20] used a genetic algorithm for producing chromosomes. A fitness function



**Figure 8.**  
Task and resource allocation model.



**Figure 9.**  
Resource management classifications.



**Figure 10.**  
*Types of resources for Management.*

is used for generating chromosomes. Individuals who have higher fitness value will only proceed for further chromosome reproduction.

Murugesan and Chellappan [21] suggested deadline and budget-based resource selection method for divisible workloads. The method assigns the appropriate resource in terms of cost from the list of available resources. Shah et al. [22] also claimed a linear programming-based resource allocation method for divisible workloads. The job is categorized in appropriate sizes to allocate on available resources. Singh and Kumar [23] improved the resource selection method presented by Singhal et al. [24] by determining the task workload and resource availability, respectively. Ang et al. [25] introduced a resource allocation mechanism by considering the requirement of user as well as service provider.

Various researchers suggested numerous techniques for heterogeneous task allocation. Raman et al. [26] provide improvements of traditional round robin (RR) task scheduling which performs well when all the resources have equal serving capacity. In heterogeneous environment, it does not provide prominent results. Therefore, a weight is assigned to each server which represents the priority of selecting a server. The algorithm performs well in distributing the load more efficiently than RR scheduling algorithm.

Pham and Huh [27] analyzed the cloud-fog environment and presented a task scheduling algorithm. The suggestion behind the presented algorithm is the association between fog nodes and cloud nodes to decrease the makespan and price of cloud resources. If the computation is not feasible on fog node, then tasks are executed on cloud node. Several constraints like deadline and budget can enhance the algorithm efficiency and applicability.

Wu [28] presented a task scheduling for embedded systems to enhance the performance of real applications in CloudIoT paradigm. These approaches are used in real-time networks where time constraints are strictly followed. The algorithm increases the scheduling success rate of real-time task on heterogeneous web servers. Moschakis and Karatza [29] analyzed the workload generated by IoT devices and scheduled them on multi cloud-based system. The least loaded server is selected by global dispatcher for scheduling IoT jobs.

Grandinetti et al. [30] presented an offline mathematical formula to improve task scheduling and average waiting time. Xu et al. [31] presented a task scheduling algorithm which determines crossover and mutation operations for mapping between tasks and resources. Kamalinia and Ghaffari [32] addressed the task scheduling as an NP complete problem. They also used a genetic algorithm to design task scheduling problem to improve makespan and resource efficiency. The presented scheduling algorithm reduces the communication cost among processors by using meta-heuristic methods.

### **2.3 Load measurement**

The load measurement is very important and crucial activity in distributed environment. Various load balancing algorithms determine resource load condition before real implementation of task. Various performance metrics like fault tolerance, waiting time, response time, etc. can effectively be optimized by measuring the current load of a resource. Many authors addressed this issue and presented various resource provisioning techniques for effective distribution of incoming load.

Patel and Tripathy [33] categorized the resources in three categories: under-loaded, normal-loaded, and overloaded to manage the load of available resources. Before assigning a task to a resource, the scheduler checks the current load of each resource and selects the underloaded or normal-loaded resource for task execution. Task length, processing element capacity, and deadline constraints are the factors that are considered to determine the current load of each resource. If a resource becomes overloaded, the unfinished tasks are shifted to another suitable resource for completing their execution. Checkpoint mechanism is used to save and resume the task state which greatly decreases the average response time and task resubmission time and improves the system throughput.

Liu et al. [34] advised that resource provisioning techniques may balance the resource load effectively. They presented peer load balance provision compares the demand and resource capacity by considering requirement of both customer and service provider. The presented mechanism reduced the cost and average response time than other existing methods.

Rathore and Chana [35] determined a dynamic threshold value based on standard deviation for load balancing and job migration. For job migration, the resources are categorized and the average load of each cluster is compared with processing element's threshold value. For load balancing, tasks are selected randomly either from underloaded or overloaded resource collection.

Kaushik and Vidyarthi [36] consider various parameters for effective job scheduling and resource allocation. The presented model selects the best cluster in terms of increased system reliability and reduced energy consumption and balances the system load efficiently. The customer can prioritize their choices to select the suitable cluster for task execution. An effective approach for determining job migration overhead can increase the model adaptability in real scenarios.

### **2.4 Cost optimization**

Load balancing algorithm maps task to various heterogeneous resources based on predefined objectives. The major objective of load balancer is to optimize task completion time, resource cost, and its utilization. Several authors addressed the cost issue and provide possible solutions for its optimization.

Garg and Singh [37] suggested an adaptive workflow scheduling (AWS) by considering resource cost and communication cost between task and resources. Due to heterogeneous nature of resources, final cost is calculated periodically. Arabnejad and Barbosa [17] presented a task scheduling algorithm which works in two

phases - task selection phase and processor selection phase. For selecting the task, priority is assigned by computing the rank. For processor selection, worthiness of all processors is calculated and selects the processor with highest worthiness value.

Chaisiri et al. [38] analyzed the resource provisioning phases and suggested that reservation method provides reduced cost than on demand methods. Broadly, there are three stages in resource provisioning:

- Resource reservation
- Resource expanding
- Resource on demand

In the first stage, the cloud broker arranged the resources in advance without experiencing the customer requirement. In the second phase, the customer requirement and resource cost are comprehended, and the resource overutilization or underutilization is identified. If customer requirement is greater than reserved resources, the broker could request for additional resources on pay-per-use basis. Here, the on demand phase started. In on demand phase, the customer must know the appropriate future requirement which is difficult to estimate in cloud environment.

Singh and Kumar [39] presented a cost optimization method based on process activity. Processing cost and waiting time are determined by using activity time, resource utilization, and variability factor to check the method efficiency. Bittencourt and Madeira [40] presented a cost optimization method for hybrid cloud. The clouds can be categorized in three categories based on resource availability: public cloud, private cloud, and hybrid cloud. A user can use the services of public cloud by using pay-per-use method. Private clouds belong to individuals and offer free variety of services. In hybrid cloud, resources from public cloud are aggregated as per requirement. Bittencourt and Madeira [40] identified the method for appropriate resource.

Cao et al. [41] analyzed that each task is different from each other in cloud environment. They suggested an activity-based task scheduling approach for task reduction. The presented algorithm performs well than traditional task assignment approaches in terms of cost reduction.

Efficient resource provisioning plays a vital role in reducing the cost of task execution. Suresh and Varatharajan [42] presented a particle swarm optimization (PSO)-based resource provisioning algorithm. PSO is adopted to select the appropriate resource for cost optimization. Three performance metrics task execution time, memory usage, and cost are evaluated and compared with other existing methods. The simulation result shows that the presented PSO-based algorithm provides minimum execution time and memory usage with least cost than other state-of-the-art methods.

Salehan et al. [43] suggested auction-based resource allocation to meet the requirement of the customer and service provider. At the time of scheduling, resources are assigned to users that have highest bids. The algorithm provides highest profit and satisfies both the customers and service providers for multiple criteria than other existing methods. Nezarat and Dastghaibyfarid [44] map the resource allocation mechanism to economic-based supply and demand problem which provides better functionality with 17% profit with other existing methods.

Netjinda et al. [45] suggested a task scheduling for workflow applications. These workflow applications consist of dependent task with deadline constraints. The aim is to select the least cost cloud resource through PSO for workflow-based task execution. The effective task scheduling decreases the execution time which directly affects the final cost. By considering communication overhead, the model effectiveness and applicability can be increased in real cloud environment.



Chunlin and Layuan [46] presented a resource provisioning method for mobile clients. The mobile devices greatly depend on cloud resources for accessing data and performing operations. The aim is to select the optimal resource at least cost. The service provider executes the tasks on appropriate resources to get the maximum profit.

## **2.5 Fault tolerance**

Fault tolerance is a mechanism that provides the estimated quality results even in the presence of faults. A system with its components and services can consider reliable only if it has fault tolerance capability. Therefore, fault tolerance issue has got a noticeable attention by the research community over the last decades [47].

Fault tolerance techniques can be categorized into two: proactive and reactive. Proactive techniques are prevention techniques that determine the controlled state for fault tolerance before they occur. The systems are continuously monitored for fault estimation. Proactive fault tolerance can be implemented in three ways: self-healing, preemption migration, and system rejuvenation. In self-healing, fault recovery procedures are periodically applied for autonomous recovery. In preemptive migration, the tasks are shifted from fault probable resource to another resource. System rejuvenation is the mechanism in which periodic backups are taken for cleaning and removing errors from the system.

Another category is the reactive approaches that deal with faults after their occurrence. Reactive fault tolerance can also be implemented in three ways: job replication, job migration, and checkpoint. In job replication, several instances or copies of the same task make available on different resources. If one instance fails, task is executed on another instance. In job migration, tasks are migrated to another suitable resource for completing its execution. In checkpoint, task states are periodically saved and restarted from the last saved state instead of from the very beginning [47]. Several authors suggested fault tolerance mechanism and recovery solutions to resolve the issue.

Patel et al. [48] addressed resource failure issues and presented a checkpoint based recovery mechanism for task execution. If task does not complete its execution within deadline, then another suitable resource is selected for completing its execution. Before transferring it to another suitable resource, task state is saved and resumed for further execution through checkpoint. This results in reduced execution time, response time, and improved throughput than other existing methods.

Generally checkpoint increases the execution time that directly affects the execution cost. Egwutuoha et al. [49] use the process of redundant technique to reduce the task execution time. The presented technique is pretty good and reduces up to 40% checkpoint overhead. Choi et al. [50] identify the malicious users to provide fault tolerance scheduling in cloud environment. Any user which only use cloud services and reject other requests is treated as malicious user. The reputation is calculated to determine the malicious users. The work can be implemented to improve network reliability and task execution time in cloud paradigm.

Mei et al. [51] suggested that replication-based fault tolerance approaches waste lots of resources and also compromise with makespan. To resolve the issue, Mei et al. [51] presented fault tolerance scheduling mechanism that ensures successful completion of task execution. The limitation of replication is avoided by rescheduling the task for further execution. If scheduler identifies the failure, it reassigns task to another suitable resource and saves the wastage of resources. This mechanism reduces resource consumption and task execution time. However, costs are presumed for implementing scheduling, which limits the model applicability in real scenario.

Nazir et al. [52] use fault index for maintaining the history of resources. Fault index is determined based on successful and unsuccessful task completion on particular resource. Based on fault index value, grid broker replicates the task that

can be used when fault occurs. Budget and time constraints are also considered at the time of task scheduling. The presented mechanism satisfies various QoS requirement, increases the reliability, and performs consistent in the existence of fault also.

Qureshi et al. [53] combined two fault tolerance techniques to inherit the favorable aspects. They perform hybridization of alternate task with retry and checkpoint mechanism and evaluate various performance metrics. The simulation result shows that alternate task with checkpoint mechanism performs better and improves system throughput than other existing methods.

Cloud facilitates the storage and access heterogeneous data in a distributed remote network. Due to dynamicity, network congestion and system faults are key factors for fault occurrence. Preventing the network from congestion and selection of suitable servers can avoid the fault conditions. Tamilvizhi and Parvathavarthini [54] suggested the concept of square matrix multiplication to manage the network traffic and avoid network congestion. The resource monitor predicts the fault conditions and uses migration policies to avoid system failure. The presented fault tolerance mechanism provides reduced cost with less energy consumption.

Garg and Singh [55] observed various fault conditions and suggested a fault tolerance-based task scheduling algorithm in grid environment. A genetic algorithm is used to determine the resource capacity for task scheduling. The presented approach increased system reliability and reduced task execution time in grid environment.

## 2.6 Interoperability issue

Interoperability refers efficient migration and integration of heterogeneous applications and data to get the seamless services across domains. Various distributed applications exist to provide millions of services that differ in the services they offered:

- Distributed computing is a collection of various heterogeneous components that are located at remote locations, which coordinate with each other by message passing. Each component or processor has its own memory. It is a kind of parallel computing in which a task is split into subtasks to run on multiple components simultaneously.
- Grid computing is a network of computer resources that are connected to solve a complex problem. Each resource is loosely linked and runs independent task to achieve a common goal. Grid computing may be classified on the basis of scale and functionality. On the basis of scale, grid computing may be classified into two categories (**Table 1**), i.e., cluster grid and enterprise grid. Cluster means a group of similar kind of entities. So cluster grid provides services to

S. no	Classification criteria	Types of grid	Characteristics
1	Scale	Cluster	Computational services are limited to a group or a department
		Enterprise	Provides services within an enterprise
2	Functionality	Global	Comprises of collection of cluster grid
		Computational	Acts as an integrated processing resource
		Data	Coordinate and manage database information which is located at remote locations

**Table 1.**  
*Classification of grid.*

the group or departmental level. Another type is enterprise grid that provides to share resources within the enterprise.

- Cloud computing provides on demand computer resources such as storage or computational resources without direct involvement of users. It has effective data management and computing framework for executing task in parallel to improve various QoS metrics.
- Fog computing is the extension of cloud computing which consists of multiple fog nodes that are directly connected to the physical devices. The difference between both technologies is that cloud is a centralized system while fog is distributed but decentralized system.
- CloudIoT is an innovative trend which connects and manages millions of devices in very cost-effective manners that are dispersed globally. Cloud can profit by IoT to deal with real-world things by sharing the pool of highly computational resources rather than having local servers or personal devices to handle applications [56, 57].

Various authors analyzed the interoperability issues that are briefly presented with their respective solutions. Aazam et al. [58] focused on analyzed two complementary technologies: cloud computing and IoT. Various challenges and integration issues of CloudIoT framework are discussed. Data analysis, service provisioning, and storage are the future dimensions to improve the performance of CloudIoT model.

Botta et al. [59] also analyzed the integration issues of cloud and IoT. Both the technologies are analyzed separately based on applications, technology, issues, and challenges. The details of existing platforms and projects are presented that are currently implementing CloudIoT. Standardization, address resolution, multi-networking, and developments of APIs are some future directions to provide full potential to CloudIoT framework. Khodkari et al. [60] present the significance and requirement of CloudIoT paradigm. They presented complementary aspects of cloud computing and IoT and assure the QoS by evaluating the integrity requirement of both the technologies.

Bonomi et al. [61] analyzed characteristics, services, and applications of fog computing. They determined the importance of collaboration of fog and cloud and address that some applications need both cloud globalization and fog localization like big data and analytics.

The linked open data (LOD) provides a new dimension for various heterogeneous interoperability issues based on Web server architectures. These issues require attention to support heterogeneous description principles that is necessary to deal with different data from web resources. The LOD interoperability follows bottom up approach to establish the strong relationships among datasets. Various researchers addressed LOD interoperability issues and presents respective solutions to meet with the users demand [1, 2].

## **2.7 QoS issues**

The user submits the tasks with various QoS constraints (cost execution time, energy consumption, delay, etc.) to improve the performance in distributed environment. Researchers addressed several QoS issues and provide the solutions for meeting the objective. Aron and Chana [62] observed various QoS issues and

identified four issues, i.e., cost, reliability, security, and time, for resource provisioning in grid environment. Service-level agreement (SLA) reduced the complexity of resource provisioning by maintaining up-to-date information of all the resources. The presented approach performs better in terms of resource utilization, cost, and customer satisfaction.

Popularity of cloud computing increased burden on distributed data centers. These data centers consumes excessive amount of energy to provide services and fulfill consumer satisfaction. Horri et al. [63] identified overloaded and underloaded servers and shift load from overloaded to under loaded resources. This makes a trade-off between energy consumption and SLA. Hoseiny Farahabady et al. [64] suggested an objective function to reduce cost and performance improvement for resource allocation mechanism. Two test cases are considered: tasks with known running time and tasks with unknown running time. They listed Monte Carlo method to determine the task's unknown values (**Table 2**).

Author	Technique	Strength	Limitations & Future Scope	Technology
Shah et al. [22]	Cost based resource allocation	Cost	Performance optimization	Grid
Murugesan & Chellappan [21]	Resource allocation	Cost	Budget and time constraint	Grid
Singhal et al. [24]	Parallel execution-based resource allocation	Cost	Simulation on real cloud	Grid
Singh and Kumar [39]	Cost based resource allocation	Cost	Time constraint	Grid
Chang et al. [16]	Execution time prediction	Execution time	Fault tolerance	Grid
Hao et al. [15]	Resource selection mechanism	Makespan, finished rate, resubmitted time	Task deadline consideration	Grid
Arabnejad and Barbosa [17]	Resource selection mechanism	Execution time, cost	Fault tolerance	Grid
Garg and Singh [55]	Task scheduling and resource selection mechanism	Execution time	Cost optimization	Grid
Singh and Kumar [10]	Queueing model	Response time, drop rate, server utilization	Fault tolerance	Grid
Singh and Kumar [23]	Resource selection model	Response time, drop rate, server utilization	Other optimization metrics consideration	Grid
Patel et al. [48]	Resource selection and task scheduling	Execution time, response time, throughput, resubmitted time	Fault tolerance mechanism and associated overhead	Grid

**Table 2.** Existing load balancing techniques and future scope.

### **3. Conclusion**

In summary, a web server system uses several load balancing techniques for distributing its load among available web resources. These resources are getting costlier day-by-day; therefore, efficient cost optimization mechanisms are required to support by small organization or industry. In this chapter, several load balancing issues have been identified for efficient use of web resources in distributed environment. Detailed description of existing approaches, strength, limitation, and future scope has been analyzed and an adequate radiance has been thrown to these techniques. On the basis of abovementioned issues, several future dimensions have been identified that will be beneficial for the research community to achieve various objectives:

- Development of a resource allocation model which considers resource as well as task characteristics to optimize various QoS metrics
- Development of a fault tolerance load balancing model for partial executed tasks due to resource failure and construction of a resource selection policy for task execution
- Analysis of contextual relationship among CloudIoT issues and optimization through effective scheduling
- Development of an execution time prediction model for efficient resource provisioning, selection, and scheduling.


### **Author details**

Anju Shukla, Shishir Kumar\* and Harikesh Singh  
Department of Computer Science and Engineering, Jaypee University of  
Engineering and Technology, Guna, MP, India

\*Address all correspondence to: [dr.shishir@yahoo.com](mailto:dr.shishir@yahoo.com)

### **IntechOpen**

---

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Bermes E. Convergence and interoperability: A linked data perspective. In: IFLA World Library and Information Congress. Vol. 77. 2011. pp. 1-12
- [2] Hidalgo-Delgado Y, Xu B, Marino-Molerio AJ, Febles-Rodriguez JP, Leiva-Mederos AA. A linked data-based semantic interoperability framework for digital libraries. *Revista Cubana de Ciencias Informáticas*. 2019; **13**(1):14-30
- [3] Alakeel AM. A guide to dynamic load balancing in distributed computer systems. *International Journal of Computer Science and Information Security*. 2010; **10**(6):153-160
- [4] Khan RZ, Ali MF. An efficient diffusion load balancing algorithm in distributed system. *International Journal of Information Technology and Computer Science*. 2014; **6**(8):65-71
- [5] Khatchadourian S, Consens MP. ExpLOD: Summary-based exploration of interlinking and RDF usage in the linked open data cloud. In: *Extended Semantic Web Conference*; May 2010. pp. 272-287
- [6] Schwarte A, Haase P, Hose K, Schenkel R, Schmidt M. Fedx: Optimization techniques for federated query processing on linked data. In: *International Semantic Web Conference*. Berlin, Heidelberg: Springer; October 2011. pp. 601-616
- [7] Kumar B, Richhariya V. Load Balancing of Web Server System Using Service Queue Length. *M.tech Scholar (CSE) Bhopal*. Vol. 5(5). 2014. Available from: [http://www.ijetae.com/files/Volume4Issue5/IJETA\\_0514\\_14.pdf](http://www.ijetae.com/files/Volume4Issue5/IJETA_0514_14.pdf)
- [8] Chen C, Bai Y, Chung C, Peng H. Performance measurement and queueing analysis of web servers with a variation of webpage size. In: *Proceedings of the International Conference on Computer Applications and Network Security*. 2011. pp. 170-174
- [9] Zhang Z, Fan W. Web server load balancing: A queueing analysis. *European Journal of Operational Research*. 2008; **186**(2):681-693
- [10] Singh H, Kumar S. WSQ: Web server queueing algorithm for dynamic load balancing. *Wireless Personal Communications*. 2015a; **80**(1):229-245
- [11] Singh H, Kumar S. Analysis & minimization of the effect of delay on load balancing for efficient web server queueing model. *International Journal of System Dynamics Applications*. 2014; **3**(4):1-16
- [12] Birdwell JD, Chiasson J, Tang Z, Abdallah C, Hayat MM, Wang T. Dynamic time delay models for load balancing. Part I: Deterministic models. In: *Advances in Time-Delay Systems*. Berlin, Heidelberg: Springer; 2004. pp. 355-370
- [13] Li M, Nishiyama H, Kato N, Mizutani K, Akashi O, Takahara A. On the fast-convergence of delay-based load balancing over multipaths for dynamic traffic environments. In: *International Conference on Wireless Communications and Signal Processing*. October 2013. pp. 1-6
- [14] Kamali SH, Hedayati M, Izadi AS, Hoseiny HR. The monitoring of the network traffic based on queuing theory and simulation in heterogeneous network environment. In: *International Conference on Computer Technology and Development*; November 2009. pp. 322-326
- [15] Hao Y, Liu G, Wen N. An enhanced load balancing mechanism based on deadline control on GridSim. *Future Generation Computer Systems*. 2012; **28**(4):657-665

- [16] Chang RS, Lin CF, Chen JJ. Selecting the most fitting resource for task execution. *Future Generation Computer Systems*. 2011;27(2):227-231
- [17] Arabnejad H, Barbosa JG. A budget constrained scheduling algorithm for workflow applications. *Journal of Grid Computing*. 2014;12(4):665-679
- [18] Naik KJ, Jagan A, Narayana NS. A novel algorithm for fault tolerant job scheduling and load balancing in grid computing environment. In: *International Conference on Green Computing and Internet of Things*. 2015. pp. 1113-1118
- [19] Cheng B, Guan X, Wu H. A hypergraph based task scheduling strategy for massive parallel spatial data processing on master-slave platforms. In: *23rd International Conference on Geoinformatics*. 2015. pp. 1-5
- [20] AbdElrouf W, Yousif A, Bashir MB. High exploitation genetic algorithm for job scheduling on grid computing. *International Journal of Grid and Distributed Computing*. 2016;9(3):221-228
- [21] Murugesan G, Chellappan C. An economic allocation of resources for divisible workloads in grid computing paradigm. *European Journal of Scientific Research*. 2011;65(3):434-443
- [22] Shah SNM, Mahmood AKB, Oxley A. Modified least cost method for grid resource allocation. In: *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*. 2010. pp. 218-225
- [23] Singh H, Kumar S. Optimized resource allocation mechanism for web server grid. In: *IEEE UP Section Conference on Electrical Computer and Electronics*. 2015b. pp. 1-6
- [24] Singhal S, Kumar M, Kant K. An economic allocation of resources in grid environment. In: *International Conference on Information Systems and Computer Networks*. March 2013. pp. 185-190
- [25] Ang TF, Por LY, Liew CS. Dynamic pricing scheme for resource allocation in multi-cloud environment. *Malaysian Journal of Computer Science*. 2017;30(1):1-17
- [26] Raman K, Subramanyam A, Rao AA. Comparative analysis of distributed web server system load balancing algorithms using qualitative parameters. *VSRD International Journal of Computer Science and Information Technology*. 2011;8:592-600
- [27] Pham XQ, Huh EN. Towards task scheduling in a cloud-fog computing system. In: *18th Asia-Pacific Network Operations and Management Symposium*. October 2016. pp. 1-4
- [28] Wu DH. Task optimization scheduling algorithm in embedded system based on internet of things. *Applied Mechanics and Materials*. 2014;513:2398-2402
- [29] Moschakis IA, Karatza HD. Towards scheduling for Internet of Things applications on clouds: A simulated annealing approach. *Concurrency and Computation: Practice and Experience*. 2015;27(8):1886-1899
- [30] Grandinetti L, Pisacane O, Sheikhalishahi M. An approximate—Constraint method for a multi-objective job scheduling in the cloud. *Future Generation Computer Systems*. 2013;29(8):1901-1908
- [31] Xu Y, Li K, Hu J, Li K. A genetic algorithm for task scheduling on heterogeneous computing systems using multiple priority queues. *Information Sciences*. 2014;270:255-287
- [32] Kamalinia A, Ghaffari A. Hybrid task scheduling method for cloud

- computing by genetic and DE algorithms. *Wireless Personal Communications*. 2017;**97**(4):6301-6323
- [33] Patel DK, Tripathy C. An improved approach for load balancing among heterogeneous resources in computational grids. *Engineering with Computers*. 2015;**31**(4):825-839
- [34] Liu L, Mei H, Xie B. Towards a multi-QoS human-centric cloud computing load balance resource allocation method. *The Journal of Supercomputing*. 2016;**72**(7):2488-2501
- [35] Rathore N, Chana I. Variable threshold-based hierarchical load balancing technique in grid. *Engineering with Computers*. 2015;**31**(3):597-615
- [36] Kaushik A, Vidyarthi DP. An energy-efficient reliable grid scheduling model using NSGA-II. *Engineering with Computers*. 2016;**32**(3):355-376
- [37] Garg R, Singh AK. Adaptive workflow scheduling in grid computing based on dynamic resource availability. *Engineering Science and Technology an International Journal*. 2015;**18**(2):256-269
- [38] Chaisiri S, Lee BS, Niyato D. Optimization of resource provisioning cost in cloud computing. *IEEE Transactions on Services Computing*. 2012;**5**(2):164-177
- [39] Singh H, Kumar S. Resource cost optimization for dynamic load balancing on web server system. *International Journal of Distributed and Cloud Computing*. 2014;**2**(1):7-18
- [40] Bittencourt LF, Madeira ERM. HCOG: A cost optimization algorithm for workflow scheduling in hybrid clouds. *Journal of Internet Services and Applications*. 2011;**2**(3):207-227
- [41] Cao Q, Wei ZB, Gong WM. An optimized algorithm for task scheduling based on activity based costing in cloud computing. In: 3rd International Conference on Bioinformatics and Biomedical Engineering. June 2009. pp. 1-3
- [42] Suresh A, Varatharajan R. Competent resource provisioning and distribution techniques for cloud computing environment. *Cluster Computing*. 2017:1-8
- [43] Salehan A, Deldari H, Abrishami S. An online valuation-based sealed winner-bid auction game for resource allocation and pricing in clouds. *The Journal of Supercomputing*. 2017;**73**(11):4868-4905
- [44] Nezarat A, Dastghaibyfarid G. A game theoretical model for profit maximization resource allocation in cloud environment with budget and deadline constraints. *The Journal of Supercomputing*. 2016;**72**(12):4737-4770
- [45] Netjinda N, Sirinaovakul B, Achalakul T. Cost optimal scheduling in IaaS for dependent workload with particle swarm optimization. *The Journal of Supercomputing*. 2014;**68**(3):1579-1603
- [46] Chunlin L, Layuan L. Cost and energy aware service provisioning for mobile client in cloud computing environment. *The Journal of Supercomputing*. 2015;**71**(4):1196-1223
- [47] Hasan M, Goraya MS. Fault tolerance in cloud computing environment: A systematic survey. *Computers in Industry*. 2018;**99**:156-172
- [48] Patel DK, Tripathy D, Tripathy C. An improved load-balancing mechanism based on deadline failure recovery on GridSim. *Engineering with Computers*. 2016;**32**(2):173-188
- [49] Egwutuoha IP, Chen S, Levy D, Selic B. A fault tolerance framework



for high performance computing in cloud. In: 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. May 2012. pp. 709-710

[50] Choi S, Chung K, Yu H. Fault tolerance and QoS scheduling using CAN in mobile social cloud computing. *Cluster Computing*. 2014;**17**(3):911-926

[51] Mei J, Li K, Zhou X, Li K. Fault-tolerant dynamic rescheduling for heterogeneous computing systems. *Journal of Grid Computing*. 2015;**13**(4):507-525

[52] Nazir B, Qureshi K, Manuel P. Replication based fault tolerant job scheduling strategy for economy driven grid. *The Journal of Supercomputing*. 2015;**62**(2):855-873

[53] Qureshi K, Khan FG, Manuel P, Nazir B. A hybrid fault tolerance technique in grid computing system. *The Journal of Supercomputing*. 2011;**56**(1):106-128

[54] Tamilvizhi T, Parvathavarthini B. A novel method for adaptive fault tolerance during load balancing in cloud computing. *Cluster Computing*. 2017:1-14

[55] Garg R, Singh AK. Fault tolerant task scheduling on computational grid using checkpointing under transient faults. *Arabian Journal for Science and Engineering*. 2014;**39**(12):8775-8791

[56] Singh S, Chana I. A survey on resource scheduling in cloud computing: Issues and challenges. *Journal of Grid Computing*. 2016;**14**(2):217-264

[57] Yassa S, Chelouah R, Kadima H, Granado B. Multi-objective approach for energy-aware workflow scheduling in cloud computing environments. *The Scientific World Journal*. 2013:1-13

[58] Aazam M, Khan I, Alsaffar AA, Huh EN. Cloud of things: Integrating

Internet of Things and cloud computing and the issues involved. In: Proceedings of 2014 11th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan. 2014. pp. 414-419

[59] Botta A, De Donato W, Persico V, Pescapé A. Integration of cloud computing and internet of things: A survey. *Future Generation Computer Systems*. 2016;**56**:684-700

[60] Khodkari H, Maghrebi SG, Branch R. Necessity of the integration Internet of Things and cloud services with quality of service assurance approach. *Bulletin de la Société Royale des Sciences de Liège*. 2016;**85**(1):434-445

[61] Bonomi F, Milito R, Zhu J, Addepalli S. Fog computing and its role in the internet of things. In: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing. August 2012. pp. 13-16

[62] Aron R, Chana I. Formal QoS policy based grid resource provisioning framework. *Journal of Grid Computing*. 2012;**10**(2):249-264

[63] Horri A, Mozafari MS, Dastghaibifard G. Novel resource allocation algorithms to performance and energy efficiency in cloud computing. *The Journal of Supercomputing*. 2014;**69**(3):1445-1461

[64] HoseinyFarahabady M, Lee YC, Zomaya AY. Randomized approximation scheme for resource allocation in hybrid-cloud environment. *The Journal of Supercomputing*. 2014;**69**(2):576-592

# Study on IoT and Big Data Analysis of 12" 7 nm Advanced Furnace Process Exhaust Gas Leakage

*Kuo-Chi Chang, Kai-Chun Chu, Hsiao-Chuan Wang, Yuh-Chung Lin, Tsui-Lien Hsu and Yu-Wen Zhou*

## Abstract

Modern FAB uses a large number of high-energy processes, including plasma, CVD, and ion implantation. Furnaces are one of the important tools for semiconductor manufacturing. According to the requirements of conversion production management, FAB installed a set of IoT-based research based on 12" 7 nm-level furnaces chip process. Two furnace processing tool measurement points were set up in a 12-inch 7 nm-level factory in Hsinchu Science Park, Taiwan, this is a 24-hour continuous monitoring system, the data obtained every second is sequentially send and stored in the cloud system. This study will be set in the cloud database for big data analysis and decision-making. The lower limit of TEOS, C<sub>2</sub>H<sub>4</sub>, CO is 0.4, 1.5, 1 ppm. Semiconductor process, so that IoT integration and big data operations can be performed in all processes, this is an important step to promote FAB intelligent production, and also an important contribution to this research.

**Keywords:** IoT, big data, furnace, exhaust gas, gas leakage, 7 nm chip process

## 1. FAB advanced furnace process of 12" 7 nm

The semiconductor plant investment exceeds 3 billion US dollars, and the basic operating cost per day exceeds 6 million US dollars, although this process is very important, especially in commercial key sizes below 12 nm (**Figure 1**) [1, 2]. However, modern semiconductor manufacturing has five major difficulties, which are summarized in **Table 1**. Semiconductor refers to the material whose conductivity is between conductor and insulator at normal temperature. Semiconductors are used in integrated circuits, consumer electronics, communication systems, photovoltaic power generation, lighting applications, high-power power conversion and other fields. Such as diode is a device made of semiconductor. Whether from the perspective of technology or economic development, the importance of semiconductors is very huge [3, 4]. 5G and artificial intelligence technologies are the main application areas of advanced technology. For example, in 2019, many mobile phone manufacturers launched 5G models, and most of these models used

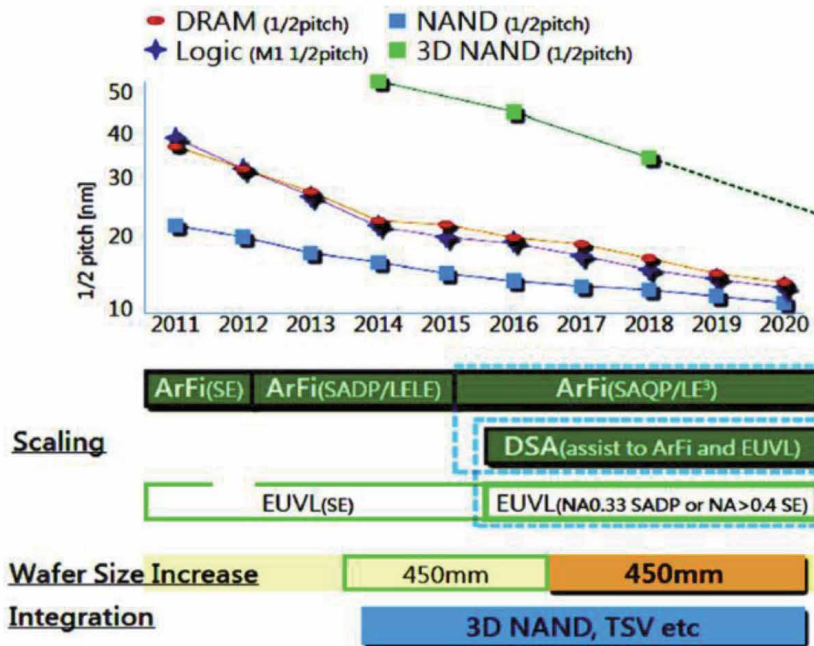


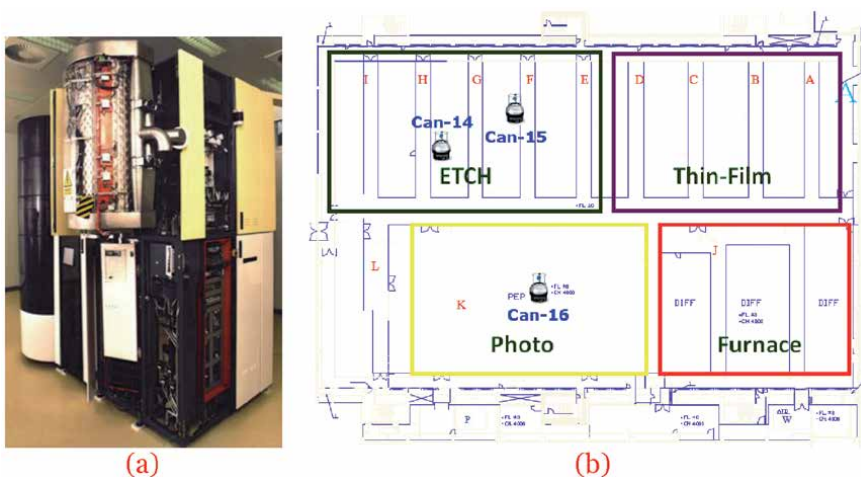
Figure 1. Development trend of key dimensions of wafer manufacturing in 2020.

12.01 ~ 7.01 advanced technology baseband chips, such as Qualcomm Snapdragon X50, MediaTek Helio M70, Intel XMM8000 series, Samsung Exynos 5000 series, Hisilicon Balong 5000 series, etc. This creates a potential demand for semiconductor equipment. 5G and artificial intelligence will not only bring the semiconductor equipment market back to a short term in 2020, but also support the development of the semiconductor equipment industry in the long term. The research and development organization predicts the technology trend in 2020, and the research and development of Toyo, a subsidiary of Jibang, predicts that the semiconductor industry will gradually come out of the bottom in 2020 with the continuous increase in demand for 5G, AI, automotive and other emerging applications. Among them, 5G transformative technologies are the most critical. The Industrial Intelligence Institute (MIC) of the China Resources Planning Association pointed out that this year, about 56 telecom operators in 32 countries have announced the deployment of 5G networks, of which 39 telecom operators have officially opened 5G services. It is estimated that by 2020, there will be 170 telecom providers worldwide providing 5G commercial services [5, 6].

Combining the foregoing, in order to achieve high accuracy and high throughput, modern FAB uses a large number of high-energy processes, such as plasma, CVD, and ion implantation. This furnace is one of the important tools for semiconductor manufacturing (Figure 2). Due to the high energy, the physicochemical changes in each reactor are very complex, and it is often impossible to determine the type and concentration of the by-products produced, and the type and concentration of these by-products often change randomly. These by-products usually cause the following effects on FAB, including: (1) Incompatibility between by-products may increase the toxicity or explosiveness of the gas in the pipeline; (2) By-products may corrode the exhaust pipe or make it brittle (3) If the type and concentration of by-products cannot be determined, it is impossible to select the appropriate exhaust gas treatment equipment; (4) The currently used processing

Topic	Explanation
Integration is getting higher and higher	The number of transistors integrated on a chip is increasing, from the 1960s to the present, from one transistor to more than 10 billion.
The higher the accuracy requirements, the more difficult the process.	The key size was reduced from 1um in 1988 to 5 nm in 2020, a reduction of 99.5%. From this perspective, the difficulty of integrated circuit manufacturing is gradually increasing, and the acceleration of the difficulty is also increasing.
Difficult to break through some specific single-point technology	The process technology that constitutes the smallest unit of the semiconductor manufacturing process is the single-point technology, especially the photolithography process. The manufacturing process of complex circuits exceeds 500 processes. These processes are carried out under precision instruments, which are not clear to human eyes.
Need to integrate multiple technologies	The difficulty of integrating technology lies in how to complete a technological process with low cost, satisfactory specifications and complete operation from an unlimited combination of component technologies in a short time. This includes operators, process equipment, raw materials, process parameters and methods, FAB environmental conditions and so on.
Mass production technology	The process flow constructed by the R & D center through integrated technology is transferred to the mass production plant. Strict exact copying is basically impossible. Even if the equipment in the development center and the mass production plant are the same, the same result may not be obtained under the same process conditions. This is because even with the same equipment, there will be a slight performance difference between the two machines. Therefore, with the continuous improvement of the degree of semiconductor precision, the problem of machine difference is becoming more and more obvious.

**Table 1.**  
 Summary five major difficulties of modern semiconductor manufacturing.



**Figure 2.**  
 The layout and actual state of the FAB 12-inch 7 nm furnace equipment selected for this study. (a) 12 advanced furnace process tool, (b) FAB layout.

equipment may be damaged, which may affect the processing efficiency [7–9]. As of 2020, there are already 37 semiconductor wafers FABs in Taiwan. Taking the 12-inch FAB as an example, there are about 420 various types of main process machines in a manufacturing plant with a monthly capacity of 50,000 wafers, of which about 63 are thin film process machines, photolithography process machine also has about 55, summary of Taiwan IC manufacturing FABs show in **Table 2** [10].

Treat 100,000 pieces of factory space as a large factory, that is, all machines with similar functions are only divided into a group. The layout of the machine group is planned using a typical “non-shaped” pattern (please refer **Figure 2(b)**). A FAB is divided into several rectangular blocks, and each block is called a processing area (bay). Machines in the same machine group or machines with similar functions are placed on the same bay as the principle. The bay is located on the two sides of the FAB. In the center is the among-bay goods material handling system, which is responsible for the transportation between bay and bay. Within each bay, there is also a within-bay goods material handling system, which is responsible for the machine and transport with the machine [11].

However, due to the FAB12" 7 nm stove based on the above production management requirements, the FTIR system was installed in this study, which includes (1) confirming the characteristics of harmful process exhaust gas; (2) evaluating the processing efficiency of various process equipment process exhaust; (3) Conduct a hazard exposure assessment survey during machine maintenance and repair; (4) Confirm the concentration and source of harmful gases and particles in the clean room operating environment; (5) Identify harmful substances in the pipeline. It is intelligent with the hardware system, and the IoT module is added to the original module. In this study, various process parameters and information required by FAB are continuously obtained in the 12" stove [12, 13].

Trade names	5" FAB	6" FAB	8" FAB	12" FAB
AMPI	—	1	—	—
Liteon	—	1	—	—
EPISIL	1	2	—	—
Micron	—	—	—	3 (Taichung * 1 + Hua Ya * 2)
MOSEL	—	1	—	—
MXIC 2	—	1	1	1 (Hsinchu Science Park)
NanYa	—	—	1	1 (Taishan Nanlin Park)
PSC	—	—	—	3 (Hsinchu Science Park)
Maxchip	—	—	1	—
TSMC	—	1	7	5 (Hsinchu Science Park *3 + Tainan Science Park*2)
UMC	—	1	6	3 (Hsinchu Science Park)
VISC	—	—	3	—
Winbond	—	—	—	2 (Taichung Science Park +Tainan Science Park)
Win Semiconductor	—	1	—	—
Total	1	9	19	17

Explanation: — means “none.”

**Table 2.**  
Summary of Taiwan IC manufacturing FABs.

Under this premise, in order to make the above software and hardware system intelligent, add the IoT module to the original module, so that we can continuously obtain various process parameters and information required by FAB in the 12" furnace process tool. Through 24 hr of continuous Processing and thousands of processing machines, we will obtain a large amount of data to confirm the above production requirements, so that we can effectively master the FAB characteristics, improve production efficiency, improve product yield and establish a safe and healthy product line and employee working environment It is an important contribution of this research [14–16].

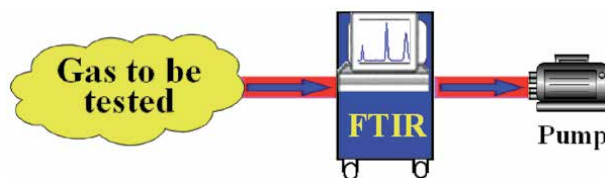
## 2. Methodology and study procedure

### 2.1 RFID and IoT technology

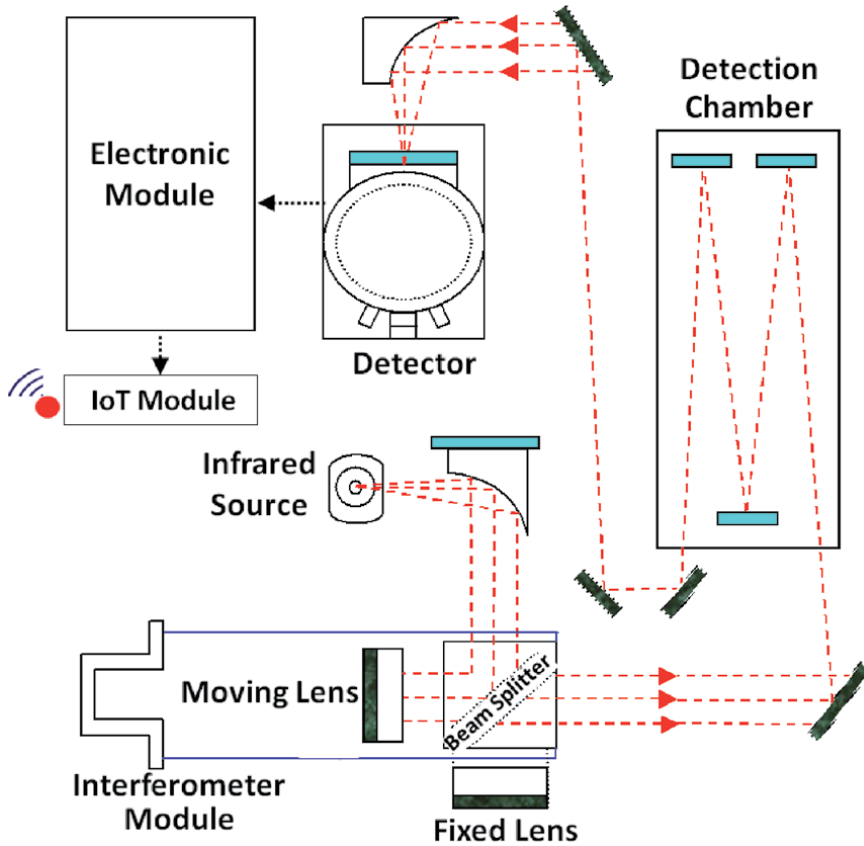
The core of the FTIR is the Michelson interferometer. Its principle is that the two infrared beams after the infrared light source is split by the beam splitter are respectively directed to the fixed mirror and the moving mirror, and then combined into A single infrared ray, due to the difference in optical path formed by the moving mirror, makes the final combined infrared ray form an infrared beam of different energy due to destructive and constructive interference. It has a fast analysis speed, is not destructive to the sample, and can analyze solid liquid and gas samples, making it gradually become an indispensable qualitative tool for material analysis. In certain circumstances, it can even achieve the ability to quickly screen quantitative [17].

The instrument used in this study is aspirated FTIR. The pumped FTIR uses a pump to introduce the gas to be tested into the FTIR detection chamber for immediate analysis. The measurement method is shown in **Figure 3**. The main components of the pumped FTIR include infrared sources, interferometers, beam splitters, fixed mirrors, moving mirrors and gas chambers, detectors and electronic modules, etc. In addition, there must be a sampling tube and pump and other gas samples into the closed cavity In addition to the computer used for data collection and data analysis and appropriate software, this study also added an IoT module to the existing FTIR, allowing the FTIR to transmit and calculate. With the cloud, pumped FTIR's the instrument configuration is shown in **Figure 4** [18, 19]. For IoT part, this study starts with the sensor, imports the sensing signal into the electronic module, and exports the signal to the cloud system through the WiFi module, in this way, this study obtain FTIR sensing data 24 hr, and this system is set to obtain data once per second.

The basic design of the infrared spectrometer is to emit a beam of light to the measurement area and measure the amount of intensity change after the beam passes through the gas to be tested. Since each gas molecule has its specific infrared light absorption coefficient, when a light beam passes through the measurement region, a specific gas molecule absorbs light of a specific wavelength, so that the



**Figure 3.**  
*Schematic diagram of instrument configuration of gas-type FTIR spectrometer.*



**Figure 4.**  
The instrument configuration of the pumped FTIR.

intensity of the light beam in this wavelength band is weakened, and the ratio of light intensity before and after absorption is The concentration of the gas is directly related. The absorption band and intensity of the gas sample can be measured to know the composition and concentration contained in the gas. For a maximum path difference  $d$  adjacent wavelengths  $\lambda_1$  and  $\lambda_2$  will have  $n$  and  $(n + 1)$  cycles respectively in the interferogram. The corresponding frequencies are  $\nu_1$  and  $\nu_2$ , and the membership function in the following Eqs. (1)~(5) [20, 21]:

$$d = n\lambda_1 \text{ and } d = (n + 1)\lambda_2 \quad (1)$$

$$\lambda_1 = d/n \text{ and } \lambda_2 = d/(n + 1) \quad (2)$$

$$\nu_1 = 1/\lambda_1 \text{ and } \nu_2 = 1/\lambda_2 \quad (3)$$

$$\nu_1 = n/d \text{ and } \nu_2 = (n + 1)/d \quad (4)$$

$$\nu_2 - \nu_1 = 1/d \quad (5)$$

FTIR mainly emits a beam of light to the measurement area and measures the intensity change of the beam after passing the gas to be measured. Since each gas molecule has its specific infrared light absorption coefficient, when the light beam passes through the measurement area, the specific gas molecule will absorb light of a specific wavelength, so that the intensity of the light beam in this band is reduced, and the ratio of the light intensity before and after absorption The concentration of the gas is directly related, and the absorption band and intensity of the gas sample



can be calculated to know the composition and concentration of the gas. **Figure 5** shows the absorption spectrum of several gas molecules in the infrared range.

## 2.2 Cloud system and big data analysis technology

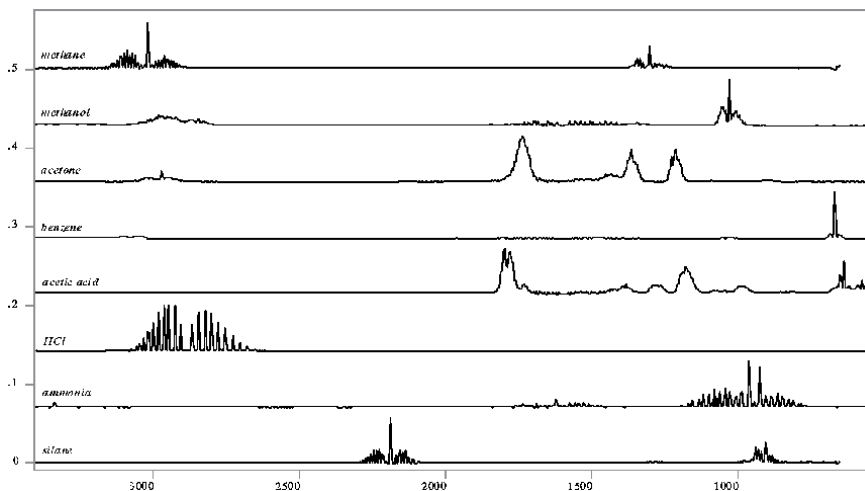
This research is based on the fact that FTIR operates continuously for 24 hr and captures signals every second to obtain data throughout the year as a basis for big data. The main research significance is not to grasp the huge data information, but to professionally process these meaningful data, so that FAB factory managers can know whether the exhaust emissions meet the alert or warning potential under the sensing trend. Most of the research uses the big data platform for deployment, debugging and maintenance. This study is connected to the Chief Cloud eXchange (CCX) cloud database platform and the Spark big data platform. At the same time, these two platforms are also more suitable for primary systems to implement systems. **Figure 6** is the research cloud computing system and database architecture.

For the big data of the chip process exhaust obtained by RFID, full consideration should be given to (1) big data life cycle, (2) big data technology ecology, (3) big data acquisition and preprocessing, (4) big data storage and management, (5) Big data computing model and system.

The big data analysis method used in this study is described as follows.  $U$  is defined as the non-empty initial universe of the object. Then define  $E$  as a set of parameters related to the object in  $U$ . Let  $P(U)$  be the power set of  $U$ , and  $A \subset E$ . A pair  $(F, A)$  is called a soft set on  $U$ , where  $F$  is the mapping given by a  $F: A \rightarrow (U)$ . In other words, the soft set on  $U$  is a parameterized family of Universe  $U$  subsets.

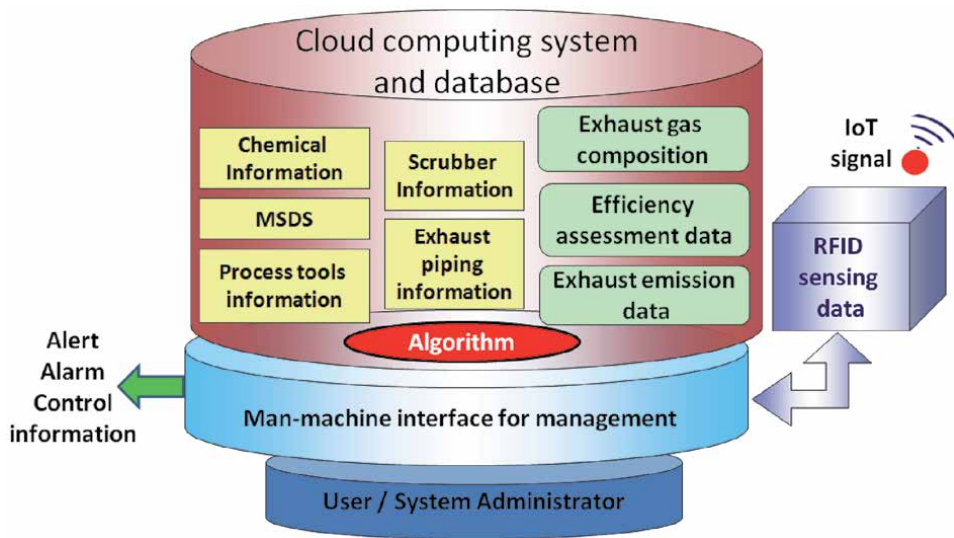
In addition, if the universe set  $U$  is a non-empty finite set, and  $\sigma$  is the equivalent relationship on  $U$ . Then  $(U, \sigma)$  is called approximate space. If  $X$  is a subset of  $U$ , then  $X$  can be written as a union of equivalent classes of  $U$  or not. If  $X$  can be written as a union of equivalent classes of  $U$ , then  $X$  is definable, otherwise it is undefinable. If  $X$  is undefinable, it can be approximated as two definable subsets, called the upper and lower approximation of  $X$ , as shown below Eq. (6) [22, 23].

$$\begin{aligned} \underline{app}(X) &= \cup\{[x]_{\sigma} : [x]_{\sigma} \subseteq X\}, \\ \overline{app}(X) &= \cup\{[x]_{\sigma} : [x]_{\sigma} \cap X \neq \emptyset\}. \end{aligned} \quad (6)$$



**Figure 5.**  
 The absorption spectrum of several gas molecules in the infrared range.





**Figure 6.**  
The research cloud computing system and database architecture.

The process of data decomposition is defined as follows: Let  $X$  be defined as the number of groups and  $Y$  as several data, as shown below Eqs. (7)~(8).

$$X = (Y/10,000) \tag{7}$$

If  $X$  contains remainder, then

$$X = X + 1 \tag{8}$$

Where the number of groups will be added to 1.

Algorithm (1): The most optimized attribute set searching algorithm.

Input: Optimized reduct sets,  $R_1$  until  $R_n$  Output: The most optimal reduct set.

if Reduct set  $R$  has more than one value then.

Select the highest number of attribute values,  $HR$  if  $HR$  does not have the same number with attribute value AND  $HR$  has more than one value then.

Select the first reduction set,  $FR$  of attribute values.

else

Proceed to the next process

else

Proceed to the next process

Algorithm (2): Soft set parameter reduction algorithm.

In tabular representation, let  $(F, P)$  represent the soft set. If  $Q$  is the reduction of  $P$ , the soft set reduction set is defined as  $(F, Q)$  of the soft set  $(F, P)$  where  $P \subset E$ .

Input: A soft set  $(F, E)$ , set  $P$ .

Output: Optimal decision.

Input the set  $P$  of choice parameters.

Find all reducts of  $(F, P)$ .

Select one reduct set  $(F, Q)$  of  $(F, P)$ .

Find weighted table of soft set  $(F, Q)$  according to the decided weights.

Find  $k$ , for which  $c_k = \max c_i$ .

$h_k$  is the optimal choice of value for the selected object. If  $k$  has more than one value, any one of the benefits could be chosen.

$c_i$  is the choice of value of an object  $h_i$  where  $c_i = \sum_j h_{ij}$  and  $h_{ij}$  is the entries in the table of the reduct soft set.

Algorithm (3): Rough set parameter reduction algorithm.

Input: An information system  $S = (U, A, V, f)$ .

$U$  is a finite nonempty set object.

$A$  is a finite nonempty set of attributes.

$V$  is a nonempty set of values.

$f$  is an information function that maps an object in  $U$  to exactly one value in  $V$ .

Output: Simplified reduct sets.

Input the information Table  $S$ .

Discretization of data.

Forming up the  $n \times n$  discernibility matrix. The elements of  $S$  table is defined as  $d(x, y) = a \in A \mid f(x, a) \neq f(y, a)$ ,  $d(x, y)$  is an attribute.

set distinguishing  $x$  and  $y$ . For each attribute  $a \in A$ , if  $d(x, y) = a_1, a_2, \dots, a_k \neq \emptyset$ .

Formulate the Boolean function  $a_1 \vee a_2 \dots \vee a_k$  or discernibility function which represented by  $\sum d(x, y)$  as indicated:  $F(A) = \prod_{(x,y) \in U \times U} \sum d(x, y)$ .

If  $d(x, y) = \emptyset$ , constant 1 will be assigned to the Boolean function.

Execute the attribute reduction process based on the simplified Boolean function.

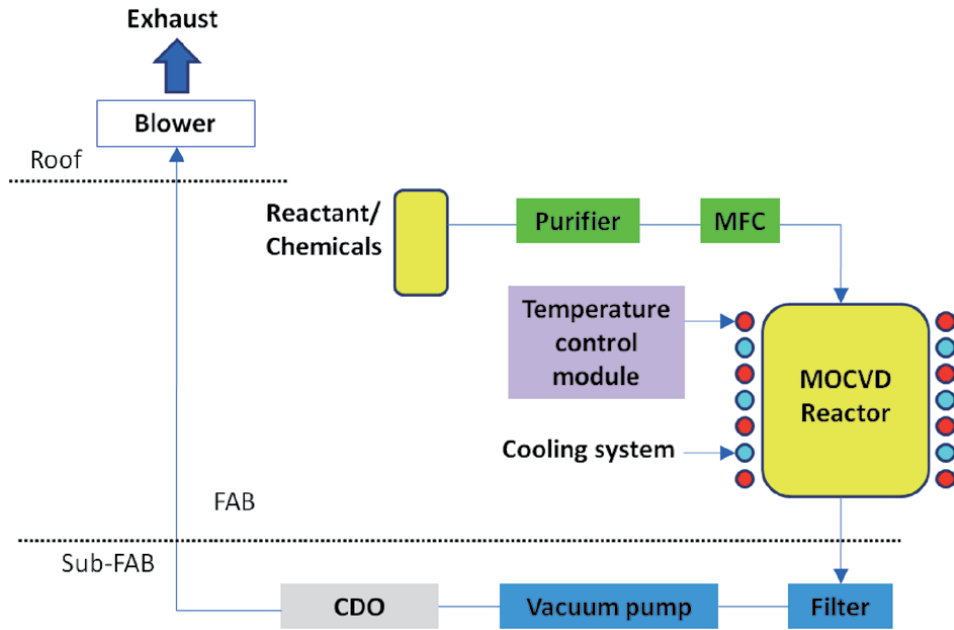
New optimized reduct sets are generated.

### 2.3 Process design and discussion

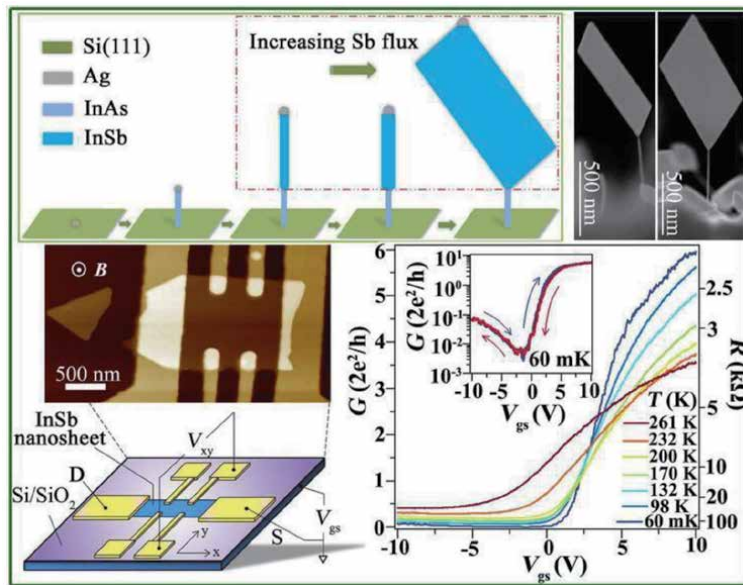
In the reactor, chemical reaction is used to form the reactant (usually a gas) into a solid product, and a thin film is deposited on the surface of the wafer. This process is called CVD (Chemical Vapor Deposition). This process has (1) good step coverage, (2) energy with high aspect ratio gap filling, (3) good thickness uniformity, (4) high pure and dense, (5) when ratio can be controlled, (6) low stress for high film quality, (7) good electrical properties, (8) base plate and excellent film adhesion characteristics. **Figure 7** shows the system architecture of CVD in FAB.

The precipitation of products during the CVD process can be divided into the following steps: (1) the source gas diffuses to the substrate surface, (2) the substrate adsorbs the source gas, (3) the substances adsorbed on the substrate react chemically on its surface, (4) The precipitated material diffuses on the surface of the substrate, (5) The reaction product is separated from the gas-phase reactant, (6) The precipitated non-volatile material is deposited on the substrate surface by diffusion and the like. The composition, structure and performance of the products obtained in this chemical reaction can be controlled by changing the parameters of the reaction. The reaction parameters mainly include the type of gas, the gas reaction concentration, the delivery method of the reactant, the gas flow rate, the total gas pressure and the Area pressure, heating method, substrate material, substrate surface state, substrate reaction temperature, temperature distribution and gradient, etc.

When entering the 7-nanometer process, the channel material of the semiconductor PN junction must also be changed. Since the electron mobility of silicon is  $1500 \text{ cm}^2 / \text{Vs}$ , and germanium can reach  $3900 \text{ cm}^2 / \text{Vs}$ , and the implementation



**Figure 7.**  
The system architecture of CVD in FAB.



**Figure 8.**  
7 nm process device structure and characteristics.

voltage of silicon devices is 0.75 ~ 0.8 V, while the germanium devices are only 0.5 V, so germanium was considered to be the preferred material for MOSFET transistors, the first 7-nanometer wafer in IBM Lab used Ge-Si material. IMEC researched new germanium-doped materials and screened two channel materials that can be used for 7 nm: one is PFET composed of 80% germanium and the other is 25 ~ 50% mixed germanium FET Or 0 ~ 25% NFET mixed with germanium (Figure 8) [24, 25].

characteristic	LP-TEOS	PE-TEOS	AP-TEOS/O <sub>3</sub>
Deposition temperature (°C)	650 ~ 750	300 ~ 400	350 ~ 450
Operating pressure (Torr)	1 ~ 10	0.1 ~ 5	500 ~ 700
Sedimentary composition	SiO <sub>2</sub>	SiO <sub>2</sub> :H	SiO <sub>2</sub>
Density(g/cm <sup>3</sup> )	2.2	2.3	2.15
Refractive index	1.43 ~ 1.46	1.47 ~ 1.5	1.45
Dielectric constant	4.0	4.1 ~ 4.9	4.4
BOE(100:1) Etching rate ( )	30	400	1200
Stress value (dyne/cm <sup>2</sup> )	1 ~ 3 × 10 <sup>9</sup>	-(1 ~ 5 × 10 <sup>9</sup> )	10 <sup>8</sup> ~ 3 × 10 <sup>9</sup>

**Table 3.**  
 Summary of process parameters and characteristics of TEOS.

Air pollutants	Equipment efficiency standard	Total control standards
Volatile Organic Compounds	>90%	<0.6 kg/hr (calculated based on methane)
Trichloroethylene	>90%	<0.02 kg/hr
Nitric acid, hydrochloric acid, phosphoric acid and hydrofluoric acid	>95%	<0.6 kg/hr
Sulfuric acid droplets	>95%	<0.1 kg/hr

**Table 4.**  
 Standards of FAB total emissions.

The silicon dioxide in the device, in this study, uses TEOS as the raw material, Tetraethyl Orthosilicate, the chemical formula is Si (OC<sub>2</sub>H<sub>5</sub>)<sub>4</sub>. High boiling point (about 169° C under normal pressure), store and use in liquid form. TEOS is liquid at room temperature and normal pressure. In order to increase the use of CVD process and the stability of the process, the TEOS container (about 40 ~ 70°C) is heated during use to increase its saturated vapor pressure. In the gaseous use of TEOS in the deposition reaction of CVD. The process parameters and characteristics of TEOS are summarized in **Table 3**.

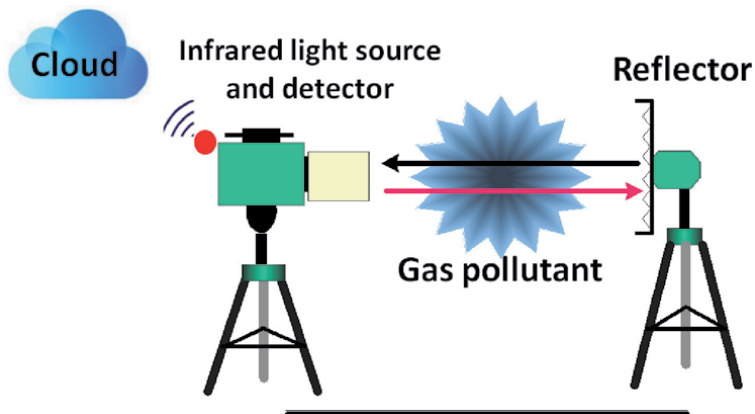
According to the “Semiconductor Manufacturing Air Pollution Control and Emission Standards” announced by the Environmental Protection Agency of the Taiwan Government, air pollutants produced in the process should be discharged after being purified by the appropriate system, where the efficiency of the system or the total emissions of the entire factory should be Meet the standards listed in the **Table 4**.

### 3. FTIR sensing system of IoT and experiment settings

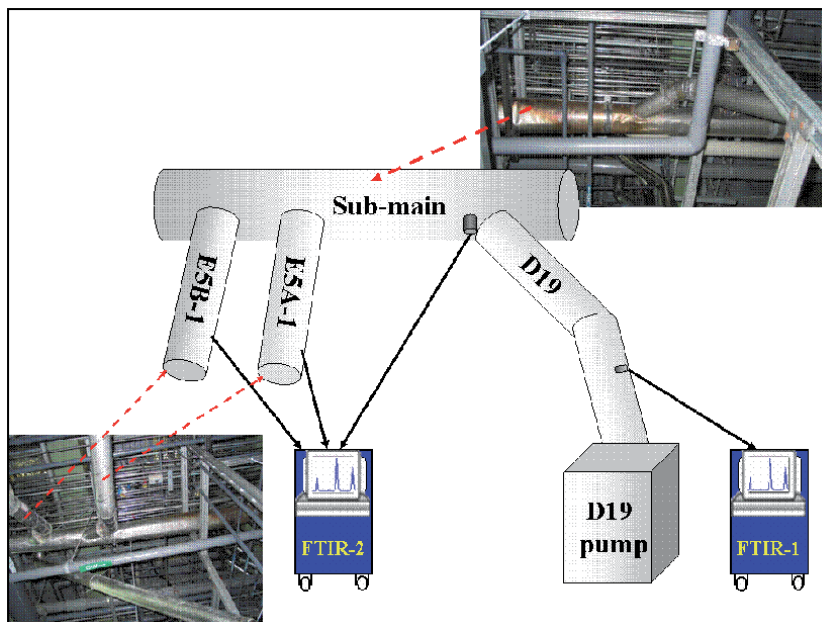
In order to achieve high precision and high output, modern high-energy processes such as plasma are important tools for semiconductor manufacturing. The physicochemical changes that occur in each reactor due to high energy are quite complicated, and the type and concentration of by-products cannot often be determined. These by-products usually have the following effects on the plant. (1) Incompatibility between by-products may increase the toxicity or explosiveness of

the gas in the pipeline; (2) By-products may cause corrosion or embrittlement of exhaust pipe materials; (3) If the type and concentration of by-products cannot be determined, appropriate exhaust gas treatment equipment may be selected; (4) damage may be caused to the currently used treatment equipment, which may affect treatment efficiency. Based on the aforementioned production management requirements, the 12" furnace FTIR system installed by FAB includes (1) confirmation of the characteristics of hazardous process exhaust gas; (2) evaluation of the processing efficiency of various process exhaust gas treatment equipment; (3) investigation of hazardous sources. Condition assessment during machine maintenance and repair; (4) Confirm the concentration and source of harmful gases and particulate matter in the clean room operating environment [26–28].

In this study, open-path FTIR was used to monitor the air quality of clean room to ensure the air quality of the working environment and the health of employees.

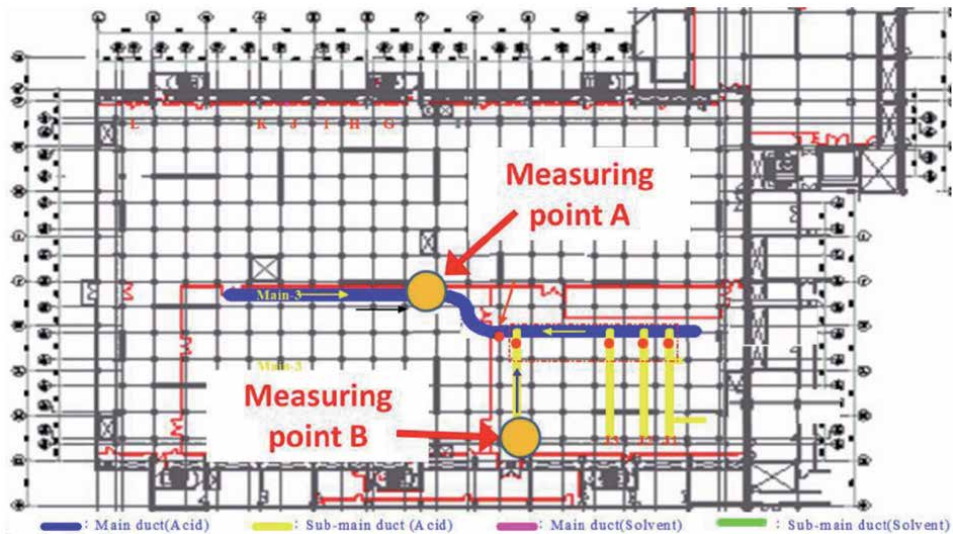


**Figure 9.**  
The FTIR field setting architecture.



**Figure 10.**  
This study was set on the site to set the exhaust line of the furnace control process.

The principle of measurement is the same as the principle of Extractive FTIR, but the closed cavity is changed to open type and integrated IoT mechanism to connect to the cloud, which is suitable for a variety of gaseous pollutants (including organic gases and inorganic gaseous pollutants) in the atmosphere are monitored (Figure 9). Figure 10 shows the situation where this study was set on the site to set the exhaust line of the furnace control process.



**Figure 11.**  
 The furnace process tools area measurement point distribution in this study.

Process tools	BPSG of A point	BPSG of b point
TMB flow	30 sccm	30 sccm
TEOS flow	300 sccm	300 sccm
PH <sub>3</sub> flow	0.77 slm	0.8 slm
Chamber pressure	1.1 torr	0.8 torr

**Table 5.**  
 The process parameters of the on-site process tools during our experiment.

concentration compound	Inlet		Outlet		Efficiency (%)
	Max (ppm)	Average (ppm)	Max (ppm)	Average (ppm)	
TEOS	937	850	47	42	86%
TMB <sup>*</sup>	1430X	870X	11X	0.3	>99%
C <sub>2</sub> H <sub>4</sub>	2108	2068	225	220	69%
CH <sub>3</sub> OH	1335	1194	N.D.	N.D.	>99%

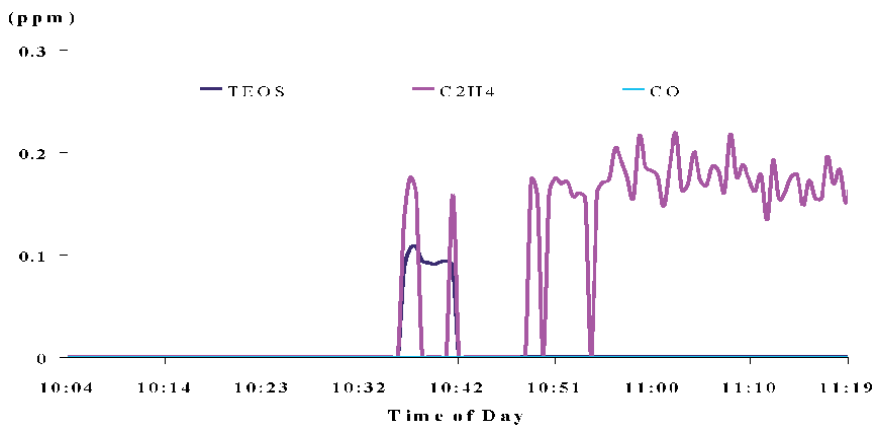
Front and rear gas flow details of exhaust gas treatment equipment: N<sub>2</sub>-pump = 98 L/min; CDO-air = 89 L/min; CDO-N<sub>2</sub> = 47 L/min; CDO outlet air = 57 L/min.  
<sup>\*</sup>TMB does not have FTIR standard spectrum.

**Table 6.**  
 The processing parameters of the on-site machine exhaust gas treatment equipment.

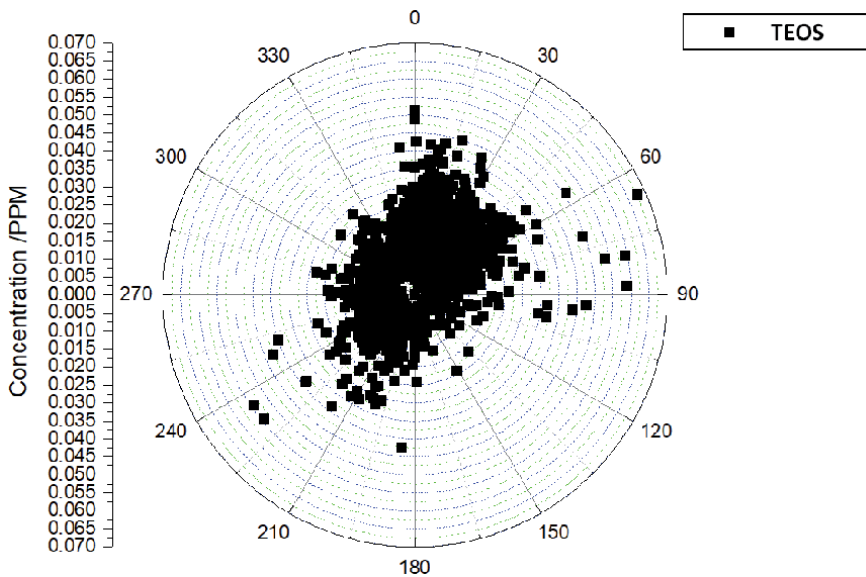
This study set up two measuring points in the 12" factory of Hsinchu Science Park in Taiwan, as shown in **Figure 11**. **Table 5** shows the process parameters of the on-site process tools during our experiment, and **Table 6** shows the processing parameters of the on-site machine exhaust gas treatment equipment. Among them, Inlet flow rate ( $Q_i$ ) estimated from the TEOS injection = 89 LPM, Initial outlet flow rate ( $Q_o$ ) estimated from the TEOS injection = 292 LPM. Therefore, dilution ratio =  $Q_o/Q_i = 292/89 = 3.3$ .

#### 4. FTIR IoT experiment result and big data analysis

From the measurement data of this study, it can be found (**Figure 12**) that the main reactant of the thin film process is TEOS for BPSG, so almost all reactions are carried out in the reaction chamber, or become a composite, which does not exist in



**Figure 12.** Main exhaust pipe concentration trend (A point).

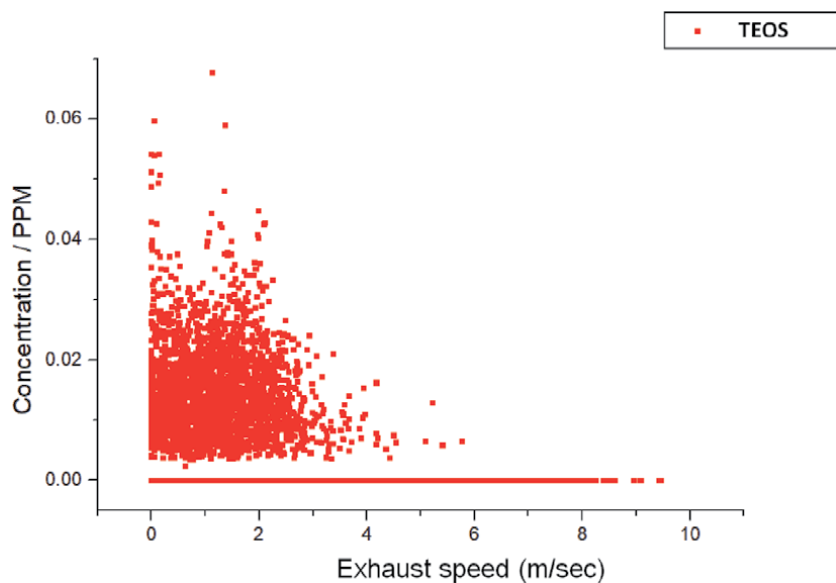


**Figure 13.** Results of scatter diagram for calculation of concentration distribution for one consecutive year (A point).

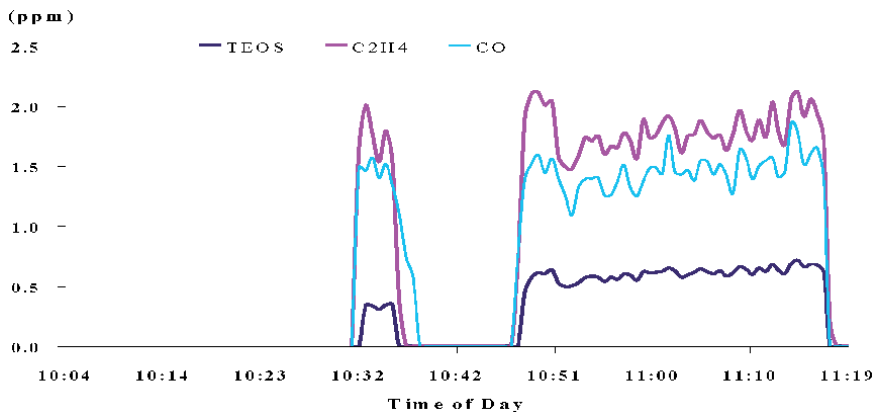


the main exhaust gas pipeline, the beginning of the reaction concentration between 0.08 ~ 0.1 ppm only. C<sub>2</sub>H<sub>4</sub> is mainly used for cleaning the reaction chamber, concentration between 0.15 ~ 0.25 ppm. Therefore, when the main process is carried out, the high concentration of the input will be cleaned, so the high concentration state can be seen in the main exhaust pipe. On the other hand, the CO is active because it is in the process of production, so it is difficult to find the concentration of the main exhaust pipe. **Figures 13** and **14** are the results of the study after optimizing the concentration values measured at the day of point A.

**Figure 15** shows secondary main exhaust pipe concentration trend. Due to the proximity of the process chamber, the concentrations are clearly detected, especially CO is more obvious, concentration between 1 ~ 1.7 ppm, and TEOS is liquid because it is normal and concentration between 0.4 ~ 0.6 ppm, so although the concentration near the reaction chamber is high, condensation occurs when entering the low temperature zone, so only this The section pipeline is measured, and the main exhaust pipe is not obvious. The C<sub>2</sub>H<sub>4</sub> concentration is between 1.5 ~ 2.0 ppm.



**Figure 14.** Achievement of the calculation of the concentration value distribution for one consecutive year (A point).



**Figure 15.** Secondary main exhaust pipe concentration trend (B point).



According to the OSHA regulations, this study is set in the cloud database for big data analysis and decision making, when the upper limit of TEOS, C<sub>2</sub>H<sub>4</sub>, CO are 0.6, 2.0, 1.7 ppm; the lower limit of TEOS, C<sub>2</sub>H<sub>4</sub>, CO is 0.4, 1.5, 1 ppm. The application architecture of this study can be extended to other semiconductor processes, so that IoT integration and big data operations can be performed for all processes, this is an important step in promoting FAB intelligent production and an important contribution of this study.

## **5. Conclusion and suggestion**

In order to achieve high precision and yield, modern FABs use a large number of high-energy processes such as plasma, CVD and ion implantation, the furnace is one of the important tools of semiconductor manufacturing. The FAB installed FTIR system due to the 12" furnace tools based on the aforementioned production management requirements. The principle of measurement is the same as the principle of Extractive FTIR, but the closed cavity is changed to open type and integrated IoT mechanism to connect to the cloud, which is suitable for a variety of gaseous pollutants (including organic gases and inorganic gaseous pollutants) in the atmosphere are monitored in this study. This study set up two measuring points of furnace process tools in the 12" factory of Hsinchu Science Park in Taiwan. This study obtained FTIR measurements, and according to the OSHA regulations, this study is set in the cloud database for big data analysis and decision making, when the upper limit of TEOS, C<sub>2</sub>H<sub>4</sub>, CO are 0.6, 2.0, 1.7 ppm; the lower limit of TEOS, C<sub>2</sub>H<sub>4</sub>, CO is 0.4, 1.5, 1 ppm. The application architecture of this study can be extended to other semiconductor processes, so that IoT integration and big data operations can be performed for all processes, this is an important step in promoting FAB intelligent production and an important contribution of this study.

## Author details

Kuo-Chi Chang<sup>1,2\*</sup>, Kai-Chun Chu<sup>1</sup>, Hsiao-Chuan Wang<sup>3</sup>, Yuh-Chung Lin<sup>1</sup>,  
Tsui-Lien Hsu<sup>4</sup> and Yu-Wen Zhou<sup>1</sup>

1 Fujian University of Technology, China

2 College of Mechanical and Electrical Engineering, National Taipei University of  
Technology, Taiwan

3 Institute of Environmental Engineering, National Taiwan University, Taiwan

4 Institute of Construction Engineering and Management, National Central  
University, Taiwan

\*Address all correspondence to: [albertchangxuite@gmail.com](mailto:albertchangxuite@gmail.com)

## IntechOpen

---

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Lu CC, Chang KC, Chen CY. Study of high-tech process furnace using inherently safer design strategies (IV). The advanced thin film manufacturing process design and adjustment. *Journal of Loss Prevention in the Process Industries*. 2016;**40**:378-395
- [2] Chen C-Y, Chang K-C, Wang G-B. Study of high-tech process furnace using inherently safer design strategies (I) temperature distribution model and process effect. *Journal of Loss Prevention in the Process Industries*. 2013;**26**(6):1198-1211, ISSN 0950-4230. DOI: 10.1016/j.jlp.2013.05.006
- [3] Lu CC, Chang KC, Chen CY. Study of high-tech process furnace using inherently safer design strategies (III) advanced thin film process and reduction of power consumption control. *Journal of Loss Prevention in the Process Industries*. 2016;**43**:280-291
- [4] Chen C-Y, Chang K-C, Lu C-C, Wang G-B. Study of high-tech process furnace using inherently safer design strategies (II). Deposited film thickness model. *Journal of Loss Prevention in the Process Industries*. 2013;**26**(1):225-235, ISSN 0950-4230. DOI: 10.1016/j.jlp.2012.11.004
- [5] Chang KC, Chu KC, Wang HC, Lin YC, Pan JS. Energy saving technology of 5G base station based on internet of things collaborative control. *IEEE Access*. 2020;**8**:32935-32946
- [6] Li S, Da Xu L, Zhao S. 5G internet of things: A survey. *Journal of Industrial Information Integration*. 2018;**10**:1-9, ISSN 2452-414X. DOI: 10.1016/j.jii.2018.01.005
- [7] Sze SM, Ng KK. *Physics of Semiconductor Devices*. 3rd ed. Canada: Wiley; 2006. ISBN: 978-0-471-14323-9
- [8] Chang K-C, Lin Y-C, Chu K-C. Mobile edge computing technology and local shunt design. *The Frontiers of Society, Science and Technology*. 2019; **1**(10):135-140. DOI: 10.25236/FSST.2019.011017
- [9] Chen C-Y, Chang K-C, Huang C-H, Lu C-C. Study of chemical supply system of high-tech process using inherently safer design strategies in Taiwan. *Journal of Loss Prevention in the Process Industries*. 2014;**29**:72-84, ISSN: 0950-4230. DOI: 10.1016/j.jlp.2014.01.004
- [10] Zhou YW et al. Study on IoT and big data analysis of furnace process exhaust gas leakage. In: Pan JS, Li J, Tsai PW, Jain L, editors. *Advances in Intelligent Information Hiding and Multimedia Signal Processing*. Smart Innovation, Systems and Technologies. Vol 156. Singapore: Springer; 2020
- [11] Chang KC, Chu KC, Chen T, Lee YW, Lin YC, Nguyen T. Study of the high-tech process mechanical integrity and electrical safety. In: 14th International Microsystems, Packaging, Assembly and Circuits Technology Conference (IMPACT). Taipei, Taiwan: IEEE; 2019. pp. 162-165
- [12] Chang KC, Chu KC, Horng D, Lin JC, Yi-Chun Chen V. Study of wafer cleaning process safety using inherently safer design strategies. In: 2018 13th International Microsystems, Packaging, Assembly and Circuits Technology Conference (IMPACT). Taipei, Taiwan: IEEE; 2018. pp. 218-221
- [13] Loke ALS et al. Analog/mixed-signal design challenges in 7-nm CMOS and beyond. In: 2018 IEEE Custom Integrated Circuits Conference (CICC). San Diego, CA: IEEE; 2018. pp. 1-8
- [14] Khan MA, Salah K. IoT security: Review, blockchain solutions, and open challenges. *Future Generation Computer Systems*. 2018;**82**:395-411. ISSN: 0167-739X. DOI: 10.1016/j.future.2017.11.022

- [15] Chang K-C, Chu K-C, Wang H-C, Lin Y-C, Pan J-S. Agent-based middleware framework using distributed CPS for improving resource utilization in smart city. *Future Generation Computer Systems*. 2020; **108**:445-453. ISSN 0167-739X. DOI: 10.1016/j.future.2020.03.006
- [16] Novo O. Blockchain meets IoT: An architecture for scalable access management in IoT. *IEEE Internet of Things Journal*. 2018;5(2):1184-1195
- [17] Müsellim E, Tahir MH, Ahmad MS, Ceylan S. Thermokinetic and TG/DSC-FTIR study of pea waste biomass pyrolysis. *Applied Thermal Engineering*. 2018;137:54-61. ISSN: 1359-4311. DOI: 10.1016/j.applthermaleng.2018.03.050
- [18] Huang M, Li Z, Huang B, Luo N, Zhang Q, Zhai X, et al. Investigating binding characteristics of cadmium and copper to DOM derived from compost and rice straw using EEM-PARAFAC combined with two-dimensional FTIR correlation analyses. *Journal of Hazardous Materials*. 2018;344:539-548. ISSN: 0304-3894. DOI: 10.1016/j.jhazmat.2017.10.022
- [19] Petit T, Puskar L. FTIR spectroscopy of nanodiamonds: Methods and interpretation. *Diamond and Related Materials*. 2018;89:52-66. ISSN: 0925-9635. DOI: 10.1016/j.diamond.2018.08.005
- [20] Amesimenu DK et al. Home appliances control using android and arduino via bluetooth and GSM control. In: Hassanien AE, Azar A, Gaber T, Oliva D, Tolba F, editors. *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*. AICV 2020. *Advances in Intelligent Systems and Computing*. Vol 1153. Cham: Springer; 2020
- [21] Primpke S, Wirth M, Lorenz C, et al. Reference database design for the automated analysis of microplastic samples based on Fourier transform infrared (FTIR) spectroscopy. *Analytical and Bioanalytical Chemistry*. 2018;410:5131-5141. DOI: 10.1007/s00216-018-1156-x
- [22] Mohamad M, Selamat A, Krejcar O, Fujita H, Wu T. An analysis on new hybrid parameter selection model performance over big data set. *Knowledge-Based Systems*. 2020;192: 105441. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2019.105441
- [23] Tao F, Cheng J, Qi Q, et al. Digital twin-driven product design, manufacturing and service with big data. *International Journal of Advanced Manufacturing Technology*. 2018;94: 3563-3576. DOI: 10.1007/s00170-017-0233-1
- [24] Jang D et al. Self-heating on bulk FinFET from 14nm down to 7 nm node. In: 2015 IEEE International Electron Devices Meeting (IEDM). Washington, DC: IEEE; 2015. pp. 11.6.1-11.6.4
- [25] Tian H, Chang K-C, Chen JS. Application of hyperbolic partial differential equations in global optimal scheduling of UAV. *Alexandria Engineering Journal*. 2020. DOI: 10.1016/j.aej.2020.02.013 [Available online 21 February 2020, in press]
- [26] Chang KC et al. Study on hazardous scenario analysis of high-tech facilities and emergency response mechanism of science and technology parks based on IoT. In: Pan JS, Lin JW, Liang Y, Chu SC, editors. *Genetic and Evolutionary Computing. ICGEC 2019. Advances in Intelligent Systems and Computing*. Vol. 1107. Singapore: Springer; 2020
- [27] Chang KC, Chu KC, Lin YC, Nguyen T, Pan J. Study of inherently safer design strategy application for IC process power supply system. In: 14th International Microsystems, Packaging, Assembly and Circuits Technology

Conference (IMPACT). Vol. 2019.  
Taipei, Taiwan: IEEE; 2019. pp. 158-161

[28] Chu KC, Horng DC, Chang KC.  
Numerical optimization of the energy  
consumption for wireless sensor  
networks based on an improved ant  
colony algorithm. *IEEE Access*. 2019;7:  
105562-105571

# Financial Time Series Analysis via Backtesting Approach

*Monday Osagie Adenomon*

## Abstract

This book chapter investigated the place of backtesting approach in financial time series analysis in choosing a reliable Generalized Auto-Regressive Conditional Heteroscedastic (GARCH) Model to analyze stock returns in Nigeria. To achieve this, The chapter used a secondary data that was collected from [www.cashcraft.com](http://www.cashcraft.com) under stock trend and analysis. Daily stock price was collected on Zenith bank stock price from October 21st 2004 to May 8th 2017. The chapter used nine different GARCH models (standard GARCH (sGARCH), Glosten-Jagannathan-Runkle GARCH (gjRARCH), Exponential GARCH (Egarch), Integrated GARCH (iGARCH), Asymmetric Power Autoregressive Conditional Heteroskedasticity (ARCH) (apARCH), Threshold GARCH (TGARCH), Non-linear GARCH (NGARCH), Nonlinear (Asymmetric) GARCH (NAGARCH) and The Absolute Value GARCH (AVGARCH) with maximum lag of 2. Most the information criteria for the sGARCH model were not available due to lack of convergence. The lowest information criteria were associated with apARCH (2,2) with Student t-distribution followed by NGARCH(2,1) with skewed student t-distribution. The backtesting result of the apARCH (2,2) was not available while eGARCH(1,1) with Skewed student t-distribution, NGARCH(1,1), NGARCH(2,1), and TGARCH (2,1) failed the backtesting but eGARCH (1,1) with student t-distribution passed the backtesting approach. Therefore with the backtesting approach, eGARCH(1,1) with student distribution emerged the superior model for modeling Zenith Bank stock returns in Nigeria. This chapter recommended the backtesting approach to selecting reliable GARCH model.

**Keywords:** financial, time series, backtesting, GARCH, ARCH-LM

## 1. Introduction

Time series is a series of observation collected with respect to time. The time could be in minutes, hours, daily, weekly, monthly, yearly etc. Time series data can be seen and applied in all fields of endeavors such as engineering, geophysics, business, economics, finance, agriculture, medical sciences, social sciences, meteorology, quality control etc. [1] but this chapter focused on financial time series analysis.

In the field of time series analysis, Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) models are popular and excellent for modeling and forecasting univariate time series data as proposed by Box and Jenkins in 1970 but many times these models failed in analyzing and

forecasting financial time series [2], this is because the ARMA and ARIMA models are used to model conditional expectation of a process but in ARMA model, the conditional variance is constant. This means that ARMA model cannot capture process with time-varying conditional variance (volatility) which is mostly common with economic and financial time series data [3].

In economic and financial time series literatures, time-varying is more common than constant volatility, and accurate modeling of time volatility is of great importance in financial time series analysis by financial econometricians [4]. In practice, financial time series contains uncertainty, volatility, excess kurtosis, high standard deviation, high skewness and sometimes non normality [3]. Therefore, to model and capture properly the characteristics of financial time series models such as Auto-Regressive Conditional Heteroscedastic (ARCH), Generalized Auto-Regressive Conditional Heteroscedastic (GARCH), multivariate GARCH, Stochastic volatility (SV) and various variants of the models have been proposed to handle these characteristics of financial time series [5]. This chapter would focus on univariate GARCH models. In practice, the backtesting approach compliment the estimated GARCH model, in order to select a reliable GARCH model useful for real life application.

This book chapter aimed at obtaining reliable GARCH model via backtesting approach using daily Zenith bank Nigeria plc stock returns.

## **2. Empirical literature reviews of previous studies**

Emenogu, et al. [3] modeled and forecasted the Guaranty Trust (GT) Bank daily stock returns from January 22,001 to May 82,017 data set collected from a secondary source. The ARMA-GARCH models, persistence, half-life and backtesting were used to analyzed the collected data using student t and skewed student t-distributions, and the analyses are carried out R environment using rugarch and performanceAnalytics Packages. The study revealed that using the lowest information criteria values only could be misleading rather we added the use of backtesting. The ARMA(1,1)-GARCH(1,1) models fitted exhibited high persistency in the daily stock returns while the days it takes for mean-reverting of the models is about 5 days, but unfortunately the models failed backtesting. The results further revealed ARMA(1,1)-eGARCH (2,2) model with student t distribution provides a suitable model for evaluating the GT bank stock returns among the competing models while it takes less than 30 days for the persistence volatility to return back to its average value of the stock returns. They recommended that researchers should adopt backtesting approach while fitting GARCH models while GT bank stocks investor should be assured that no matter the fluctuations in the stock market, the GT bank stock returns has the ability to returns to its mean price return.

Asemota and Ekejiuba [6] examined the volatility of banks equity weekly returns for six banks (coded B1 to B6) using GARCH models. Results reveal the presence of ARCH effect in B2 and B3 equity returns. In addition, the estimated models could not find evidence of leverage effect. On evaluating the estimated models using standard criteria, EGARCH (1, 1) and CGARCH (1, 1) model in Student's t-distribution are adjudged the best volatility models for B2 and B3 respectively. The study recommends that in modeling stock market volatility, variants of GARCH models and alternative error distribution should be considered for robustness of results. The study also recommended for adequate regulatory effort by the CBN over commercial banks operations that will enhance efficiency of their stocks performance and reduce volatility aimed at boosting investors' confidence in the banking sector.

Adigwe, et al. [7], examined the effect of stock market development on Nigeria's economic growth. The objective of the study was to determine if stock market development significantly impact on the country's economic growth. Secondary data were employed for the study covering 1985 to 2014. Ordinary Least Square (OLS) econometric technique was used for the time series analysis in which variations in economic growth was regressed on market capitalisation ratio to GDP, value of stock traded ratio to GDP, trade openness and inflation rate. The analysis revealed that stock market has the potentials of growth inducing, but has not contributed meaningfully to Nigerian economic growth, since only 26.5% of variations in economic growth were explained by the stock market development variables. Based on this, they suggested for an encouragement of more investors in the market, improvement in the settlement system and ensuring investors' confidence in the market.

Yaya, et al. [8] examined the application of nonlinear Smooth Transition- Generalized Autoregressive Conditional Heteroscedasticity (ST-GARCH) model of Hagerud on prices of banks' shares in Nigeria. The methodology was informed by the failure of the conventional GARCH model to capture the asymmetric properties of the banks' daily share prices. The asymmetry and non-linearity in the model dynamics make it useful for generating nonlinear conditional variance series. From the empirical analysis, we obtained the conditional volatility of each bank's share price return. The highest volatility persistence was observed in Bank 6, while Bank 12 had the least volatility. Evidently, about 25% of the investigated banks exhibited linear volatility behavior, while the remaining banks showed nonlinear volatility specifications. Given the level of risk associated with investment in stocks, investors and financial analysts could consider volatility modeling of bank share prices with variants of the ST-GARCH models. The impact of news is an important feature that relevant agencies could study so as to be guided while addressing underlying issues in the banking system.

Emenike and Aleke [9] studied the daily closing prices of the Nigerian stocks from January 1996 to December 2011 used the asymmetric GARCH variants. Their result showed strong evidence of asymmetric effects in the stock returns and therefore proposed EGARCH as performing better than other asymmetric rivals.

Arowolo [10] examined the forecasting properties of linear GARCH model for daily closing stocks prices of Zenith bank Plc in the Nigerian Stock Exchange. The Akaike and Bayesian Information Criteria (AIC and BIC) techniques were used to obtain the order of the GARCH (p,q) that best fit the Zenith Bank return series. The information criteria identified GARCH (1,2) as the appropriate model. His result further supported the claim that financial data are leptokurtic.

Emenike and Ani [11], examined the nature of volatility of stock returns in the Nigerian banking sector using GARCH models. Individual bank indices and the All-share Index of the Nigerian Stock Exchange were evaluated for evidence of volatility persistence, volatility asymmetry and fat tails using data from 3 January 2006 to 31 December 2012. Results obtained from GARCH models suggest that stock returns volatility of the Nigerian banking sector move in cluster and that volatility persistence is high for the sample period. The results also indicate that stock returns distribution of the banking sector is leptokurtic and that sign of the innovations have insignificant influence on the volatility of stock returns of the banks. Finally, the findings of this study show that the degree of volatility persistence is higher for the All Share Index than for most of the banks.

Abubakar and Gani [12] re-examined the long run relationship between financial development indicators and economic growth in Nigeria over the period 1970–2010. The study employed the Johansen and Juselius (1990) approach to cointegration and Vector Error Correction Modeling (VECM). Their findings



revealed that in the long-run, liquid liabilities of commercial banks and trade openness exert significant positive influence on economic growth, conversely, credit to the private sector, interest rate spread and government expenditure exert significant negative influence. The findings implied that, credit to the private sector is marred by the identified problems and government borrowing and high interest rate are crowding out investment and growth. The policy implications are financial reforms in Nigeria should focus more on deepening the sector in terms of financial instruments so that firms can have alternatives to banks' credit which proved to be inefficient and detrimental to growth, moreover, government should inculcate fiscal discipline.

### 3. Model specification

This study focuses on the GARCH models that are robust for forecasting the volatility of financial time series data; so GARCH model and some of its extensions are presented in this section.

#### 3.1 Autoregressive conditional heteroskedasticity (ARCH) family model

Every ARCH or GARCH family model requires two distinct specifications, namely: the mean and the variance Equations [13]. The mean equation for a conditional heteroskedasticity in a return series,  $y_t$  is given by

$$y_t = E_{t-1}(y_t) + \varepsilon_t \quad (1)$$

where  $\varepsilon_t = \varphi_t \sigma_t$ .

The mean equation in Eq. (1) also applies to other GARCH family models.  $E_{t-1}(\cdot)$  is the expected value conditional on information available at time  $t-1$ , while  $\varepsilon_t$  is the error generated from the mean equation at time  $t$  and  $\varphi_t$  is the sequence of independent and identically distributed random variables with zero mean and unit variance.

The variance equation for an ARCH(p) model is given by

$$\sigma_t^2 = \omega + \alpha_1 a_{t-1}^2 + \dots + \alpha_p a_{t-p}^2 \quad (2)$$

It can be seen in the equation that large values of the innovation of asset returns have bigger impact on the conditional variance because they are squared, which means that a large shock tends to follow another large shock and that is the same way the clusters of the volatility behave. So the ARCH(p) model becomes:

$$a_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = \omega + \alpha_1 a_{t-1}^2 + \dots + \alpha_p a_{t-p}^2 \quad (3)$$

Where  $\varepsilon_t \sim N(0,1)$  iid,  $\omega > 0$  and  $\alpha_i \geq 0$  for  $i > 0$ . In practice,  $\varepsilon_t$  is assumed to follow the standard normal or a standardized student- $t$  distribution or a generalized error distribution [14].

#### 3.2 Asymmetric power ARCH

According to Rossi [15], the asymmetric power ARCH model proposed by [16] given below forms the basis for deriving the GARCH family models.

Given that:

$$\begin{aligned}
 r &= \mu + a_t, \\
 \varepsilon_t &= \sigma_t \varepsilon_t, \\
 \varepsilon_t &\sim N(0, 1) \\
 \sigma_t^\delta &= \omega + \sum_{i=1}^p \alpha_i (|a_{t-i}| - \gamma_i a_{t-i})^\delta + \sum_{j=1}^q \beta_j \sigma_{t-j}^\delta,
 \end{aligned} \tag{4}$$

where

$$\begin{aligned}
 \omega &> 0, & \delta &\geq 0 \\
 \alpha_i &\geq 0 & i &= 1, 2, \dots, p \\
 -1 < \gamma_i < 1 & & i &= 1, 2, \dots, p \\
 \beta_j &> 0 & j &= 1, 2, \dots, q
 \end{aligned}$$

This model imposes a Box-Cox transformation of the conditional standard deviation process and the asymmetric absolute residuals. The leverage effect is the asymmetric response of volatility to positive and negative “shocks”.

### 3.3 Standard GARCH(p, q) model

The mathematical model for the sGARCH(p,q) model is obtained from Eq. (4) by letting  $\delta = 2$  and  $\gamma_i = 0, i = 1, \dots, p$  to be:

$$a_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i a_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \tag{5}$$

Where  $a_t = r_t - \mu_t$  ( $r_t$  is the continuously compounded log return series), and  $\varepsilon_t \sim N(0,1)$  iid, the parameter  $\alpha_i$  is the ARCH parameter and  $\beta_j$  is the GARCH parameter, and  $\omega > 0, \alpha_i \geq 0, \beta_j \geq 0$ , and  $\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1$ , [17].

The restriction on ARCH and GARCH parameters  $(\alpha_i, \beta_j)$  suggests that the volatility ( $a_i$ ) is finite and that the conditional standard deviation ( $\sigma_i$ ) increases. It can be observed that if  $q = 0$ , then the model GARCH parameter ( $\beta_j$ ) becomes extinct and what is left is an ARCH(p) model.

To expatiate on the properties of GARCH models, the following representation is necessary:

Let  $\eta_t = a_t^2 - \sigma_t^2$  so that  $\sigma_t^2 = a_t^2 - \eta_t$ . By substituting  $\sigma_{t-i}^2 = a_{t-i}^2 - \eta_{t-i}, (i = 0, \dots, q)$  into Eq. (3), the GARCH model can be rewritten as

$$a_t = \alpha_0 + \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) a_{t-i}^2 + \eta_t - \sum_{j=1}^q \beta_j \eta_{t-j}, \tag{6}$$

It can be seen that  $\{\eta_t\}$  is a martingale difference series (i.e.,  $E(\eta_t) = 0$  and  $\text{cov}(\eta_t, \eta_{t-j}) = 0, \text{ for } j \geq 1$ ). However,  $\{\eta_t\}$  in general is not an iid sequence.

A GARCH model can be regarded as an application of the ARMA idea to the squared series  $a_t^2$ . Using the unconditional mean of an ARMA model, results in this

$$E(a_t^2) = \frac{\alpha_0}{1 - \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i)}$$

provided that the denominator of the prior fraction is positive [14].  
When  $p = 1$  and  $q = 1$ , we have GARCH(1, 1) model given by:

$$\begin{aligned} a_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= \omega + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \end{aligned} \tag{7}$$

### 3.4 GJR-GARCH(p, q) model

The Glosten-Jagannathan-Runkle GARCH (GJR-GARCH) model, which is a model that attempts to address volatility clustering in an innovation process, is obtained by letting  $\delta = 2$ .

When  $\delta = 2$  and  $0 \leq \gamma_i < 1$ ,

$$\begin{aligned} \sigma_t^2 &= \omega + \sum_{i=1}^p \alpha_i (|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i})^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \\ &= \omega + \sum_{i=1}^p \alpha_i (|\varepsilon_{t-i}|^2 + \gamma_i^2 \varepsilon_{t-i}^2 - 2\gamma_i |\varepsilon_{t-i}| \varepsilon_{t-i}) + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \\ \sigma_t^2 &= \begin{cases} \omega + \sum_{i=1}^p \alpha_i^2 (1 + \gamma_i)^2 \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, & \varepsilon_{t-i} < 0 \\ \omega + \sum_{i=1}^p \alpha_i (1 - \gamma_i)^2 \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, & \varepsilon_{t-i} > 0 \end{cases} \end{aligned} \tag{8}$$

i.e.;

$$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i (1 - \gamma_i)^2 \varepsilon_{t-i}^2 + \sum_{i=1}^p \alpha_i \left\{ (1 + \gamma_i)^2 - (1 - \gamma_i)^2 \right\} S_i^- \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$$

$$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i (1 - \gamma_i)^2 \varepsilon_{t-1}^2 + \sum_{j=1}^q \beta_j \sigma_{t-1}^2 + \sum_{i=1}^p 4\alpha_i \gamma_i S_i^- \varepsilon_{t-1}^2$$

$$\text{where } S_i^- = \begin{cases} 1 & \text{if } \varepsilon_{t-i} < 0 \\ 0 & \text{if } \varepsilon_{t-i} \geq 0 \end{cases}$$

Now define

$$\alpha_i^* = \alpha_i (1 - \gamma_i)^2 \quad \text{and} \quad \gamma_i^* = 4\alpha_i \gamma_i,$$

then

$$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i (1 - \gamma_i)^2 \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-i}^2 + \sum_{i=1}^p \gamma_i^* S_i^- \varepsilon_{t-1}^2 \tag{9}$$

Which is the GJR-GARCH model [15].

But when  $-1 \leq \gamma_i < 0$ ,

Then recall Eq. (8)

$$\begin{aligned}\sigma_t^2 &= \omega + \sum_{i=1}^p \alpha_i (|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i})^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \\ &= \omega + \sum_{i=1}^p \alpha_i \left( |\varepsilon_{t-i}|^2 + \gamma_i^2 \varepsilon_{t-i}^2 - 2\gamma_i |\varepsilon_{t-i}| \varepsilon_{t-i} \right) + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \\ \sigma_t^2 &= \begin{cases} \omega + \sum_{i=1}^p \alpha_i^2 (1 - \gamma_i)^2 \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, & \varepsilon_{t-i} > 0 \\ \omega + \sum_{i=1}^p \alpha_i (1 + \gamma_i)^2 \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, & \varepsilon_{t-i} < 0 \end{cases}\end{aligned}$$

$$\begin{aligned}\sigma_t^2 &= \omega + \sum_{i=1}^p \alpha_i (1 + \gamma_i)^2 \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 + \sum_{i=1}^p \alpha_i \left\{ (1 - \gamma_i)^2 - (1 + \gamma_i)^2 \right\} S_i^+ \varepsilon_{t-i}^2 \\ &= \omega + \sum_{i=1}^p \alpha_i (1 + \gamma_i)^2 \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 + \sum_{i=1}^p \alpha_i \{ 1 + \gamma_i^2 - 2\gamma_i - 1 - \gamma_i^2 - 2\gamma_i \} S_i^+ \varepsilon_{t-i}^2\end{aligned}$$

Where

$$S_i^+ = \begin{cases} 1 & \text{if } \varepsilon_{t-i} > 0 \\ 0 & \text{if } \varepsilon_{t-i} \leq 0 \end{cases}$$

also define

$$\alpha_i^* = \alpha_i (1 + \gamma_i)^2 \text{ and } \gamma_i^* = -4\alpha_i \gamma_i,$$

then

$$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i^* \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 + \sum_{i=1}^p \gamma_i^* S_i^+ \varepsilon_{t-i}^2 \quad (10)$$

which allows positive shocks to have a stronger effect on volatility than negative shocks [15]. But when  $p = q = 1$ , the GJR-GARCH(1,1) model will be written as

$$\sigma_t^2 = \omega + \alpha \varepsilon_t^2 + \gamma S_t \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2. \quad (11)$$

### 3.5 IGARCH(1, 1) model

The integrated GARCH (IGARCH) models are unit- root GARCH models. The IGARCH (1, 1) model is specified in Grek [18] as

$$a_t = \sigma_t \varepsilon_t; \sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + (1 - \beta_1) a_{t-1}^2 \quad (12)$$

Where  $\varepsilon_t \sim N(0, 1)$  iid, and  $0 < \beta_1 < 1$ , Ali (2013) used  $\alpha_i$  to denote  $1 - \beta_i$ .

The model is also an exponential smoothing model for the  $\{a_t^2\}$  series. To see this, rewrite the model as.

$$\begin{aligned}\sigma_t^2 &= (1 - \beta_1) a_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \\ &= (1 - \beta_1) a_{t-1}^2 + \beta_1 [(1 - \beta_1) a_{t-2}^2 + \beta_1 \sigma_{t-2}^2] \\ &= (1 - \beta_1) a_{t-1}^2 + (1 - \beta_1) \beta_1 a_{t-2}^2 + \beta_1^2 \sigma_{t-2}^2.\end{aligned} \quad (13)$$

By repeated substitutions, we have

$$\sigma_t^2 = (1 - \beta_1)(a_{t-1}^2 + \beta_1 a_{t-2}^2 + \beta_1^2 a_{t-3}^2 + \dots), \quad (14)$$

which is the well-known exponential smoothing formation with  $\beta_1$  being the discounting factor [14].

### 3.6 TGARCH(p, q) model

The Threshold GARCH model is another model used to handle leverage effects, and a TGARCH(p, q) model is given by the following:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p (\alpha_i + \gamma_i N_{t-i}) a_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, \quad (15)$$

where  $N_{t-i}$  is an indicator for negative  $a_{t-i}$ , that is,

$$N_{t-i} = \begin{cases} 1 & \text{if } a_{t-i} < 0, \\ 0 & \text{if } a_{t-i} \geq 0, \end{cases}$$

and  $\alpha_i$ ,  $\gamma_i$ , and  $\beta_j$  are nonnegative parameters satisfying conditions similar to those of GARCH models, [14]. When  $p = 1, q = 1$ , the TGARCH(1, 1) model becomes:

$$\sigma_t^2 = \omega + (\alpha + \gamma N_{t-1}) a_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (16)$$

### 3.7 NGARCH(p, q) model

The Nonlinear Generalized Autoregressive Conditional Heteroskedasticity (NGARCH) Model has been presented variously in literature by the following scholars [19–21]. The following model can be shown to represent all the presentations:

$$h_t = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^q \gamma_i \varepsilon_{t-i} + \sum_{j=1}^p \beta_j h_{t-j} \quad (17)$$

Where  $h_t$  is the conditional variance, and  $\omega$ ,  $\beta$  and  $\alpha$  satisfy  $\omega > 0, \beta \geq 0$  and  $\alpha \geq 0$ . Which can also be written as

$$\sigma_t = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^q \gamma_i \varepsilon_{t-i} + \sum_{j=1}^p \beta_j \sigma_{t-j} \quad (18)$$

### 3.8 The exponential generalized autoregressive conditional heteroskedasticity (EGARCH) model

The EGARCH model was proposed by Nelson [22] to overcome some weaknesses of the GARCH model in handling financial time series pointed out by [23], In particular, to allow for asymmetric effects between positive and negative asset returns, he considered the weighted innovation

$$g(\varepsilon_t) = \theta \varepsilon_t + \gamma [|\varepsilon_t| - E(|\varepsilon_t|)], \quad (19)$$

where  $\theta$  and  $\gamma$  are real constants. Both  $\varepsilon_t$  and  $|\varepsilon_t| - E(|\varepsilon_t|)$  are zero-mean iid sequences with continuous distributions. Therefore,  $E[g(\varepsilon_t)] = 0$ . The asymmetry of  $g(\varepsilon_t)$  can easily be seen by rewriting it as

$$g(\varepsilon_t) = \begin{cases} (\theta + \gamma)\varepsilon_t - \gamma E(|\varepsilon_t|) & \text{if } \varepsilon_t \geq 0, \\ (\theta - \gamma)\varepsilon_t - \gamma E(|\varepsilon_t|) & \text{if } \varepsilon_t < 0. \end{cases} \quad (20)$$

An EGARCH( $m, s$ ) model, according to Dhamija and Bhalla [24] can be written as

$$a_t = \sigma_t \varepsilon_t, \ln(\sigma_t^2) = \omega + \sum_{i=1}^s \alpha_i \frac{|a_{t-i}| + \theta_i a_{t-i}}{\sigma_{t-i}} + \sum_{j=1}^m \beta_j \ln(\sigma_{t-j}^2), \quad (21)$$

Which specifically results in EGARCH (1, 1) being written as

$$a_t = \sigma_t \varepsilon_t \\ \ln(\sigma_t^2) = \omega + \alpha(|a_{t-1}| - E(|a_{t-1}|)) + \theta a_{t-1} + \beta \ln(\sigma_{t-1}^2) \quad (22)$$

where  $|a_{t-1}| - E(|a_{t-1}|)$  are iid and have mean zero. When the EGARCH model has a Gaussian distribution of error term, then  $E(|\varepsilon_t|) = \sqrt{2/\pi}$ , which gives:

$$\ln(\sigma_t^2) = \omega + \alpha\left(|a_{t-1}| - \sqrt{2/\pi}\right) + \theta a_{t-1} + \beta \ln(\sigma_{t-1}^2) \quad (23)$$

### 3.9 The absolute value GARCH (AVGARCH)

An asymmetric GARCH (AGARCH), according to Ali [25] is simply

$$a_t = \sigma_t \varepsilon_t; \sigma^2 = \omega + \sum_{i=1}^p \alpha_i |\varepsilon_{t-i} - b|^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, \quad (24)$$

While the absolute value generalized autoregressive conditional heteroskedasticity (AVGARCH) model is specified as:

$$a_t = \sigma_t \varepsilon_t; \sigma^2 = \omega + \sum_{i=1}^p \alpha_i (|\varepsilon_{t-i} + b| - c(\varepsilon_{t-i} + b))^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \quad (25)$$

### 3.10 Nonlinear (asymmetric) GARCH, or N(a)GARCH or NAGARCH

NAGARCH plays key role in option pricing with stochastic volatility because, as we shall see later on, NAGARCH allows you to derive closed-form expressions for European option prices in spite of the rich volatility dynamics. Because a NAGARCH may be written as

$$\sigma_{t+1}^2 = \omega + \alpha \sigma_t^2 (z_t - \delta)^2 + \beta \sigma_t^2 \quad (26)$$

And if  $z_t \sim IIDN(0, 1)$ ,  $z_t$  is independent of  $\sigma_t^2$  as  $\sigma_t^2$  is only a function of an infinite number of past squared returns, it is possible to easily derive the long run, unconditional variance under NGARCH and the assumption of stationarity:

$$\begin{aligned}
 E[\sigma_{t+1}^2] &= \bar{\sigma}^2 = \omega + \alpha E[\sigma_t^2(z_t - \delta)^2] + \beta E[\sigma_t^2] \\
 &= \omega + \alpha E[\sigma_t^2]E(z_t^2 + \delta^2 - 2\delta z_t) + \beta E[\sigma_t^2] \\
 &= \omega + \alpha \bar{\sigma}^2(1 + \delta^2) + \beta \bar{\sigma}^2
 \end{aligned} \tag{27}$$

Where  $\bar{\sigma}^2 = E[\sigma_t^2]$  and  $E[\sigma_t^2] = E[\sigma_{t+1}^2]$  because of stationary. Therefore

$$\bar{\sigma}^2[1 - \alpha(1 + \delta^2) + \beta] = \omega \Rightarrow \bar{\sigma}^2 = \frac{\omega}{1 - \alpha(1 + \delta^2) + \beta} \tag{28}$$

Which exists and positive if and only if  $\alpha(1 + \delta^2) + \beta < 1$ . This has two implications:

- i. The persistence index of a NAGARCH(1,1) is  $\alpha(1 + \delta^2) + \beta$  and not simply  $\alpha + \beta$ ;
- ii. a NAGARCH(1,1) model is stationary if and only if  $\alpha(1 + \delta^2) + \beta < 1$ .

See details in [22].

### 3.11 Persistence

The low or high persistency in volatility exhibited by financial time series can be determined by the GARCH coefficients of a stationary GARCH model. The persistence of a GARCH model can be calculated as the sum of GARCH ( $\beta_1$ ) and ARCH ( $\alpha_1$ ) coefficients that is  $\alpha + \beta_1$ . In most financial time series, it is very close to one (1) [26, 27]. Persistence could take the following conditions:

If  $\alpha + \beta_1 < 1$ : The model ensures positive conditional variance as well as stationary.

If  $\alpha + \beta_1 = 1$ : we have an exponential decay model, then the half-life becomes infinite. Meaning the model is strictly stationary.

If  $\alpha + \beta_1 > 1$ : The GARCH model is said to be non-stationary, meaning that the volatility ultimately detonates toward the infinitude [27]. In addition, the model shows that the conditional variance is unstable, unpredicted and the process is non-stationary [28].

### 3.12 Half-life volatility

Half-life volatility measures the mean reverting speed (average time) of a stock price or returns. The mathematical expression of half-life volatility is given as

$$Half - Life = \frac{\ln(0.5)}{\ln(\alpha_1 + \beta_2)}$$

It can be noted that the value of  $\alpha + \beta_1$  influences the mean reverting speed [27], which means that if the value of  $\alpha + \beta_1$  is closer to one (1), then the volatility shocks of the half-life will be longer.

### 3.13 Backtesting

Financial risk model evaluation or backtesting is an important part of the internal model's approach to market risk management as put out by Basle Committee on

Banking Supervision [29]. Backtesting is a statistical procedure where actual profits and losses are systematically compared to corresponding VaR estimates [30]. This book chapter adopted Backtesting techniques of [29]; The test was implemented in R using rugarch package and this test considered both the unconditional (Kupiec) and conditional (Christoffersen) coverage tests for the correct number of exceedances (see details in [31, 32]).

The unconditional (Kupiec) test also refer to as POF-test (Proportion of failure) with its null hypothesis given as

$$H_0 : p = \hat{p} = \frac{y}{T}$$

Here  $y$  is the number of exceptions and  $T$  is the number of observations and  $k$  is the confidence level. The test is given as

$$LR_{POF} = -2 \ln \left( \frac{(1-p)^{T-y} p^y}{\left[1 - \left(\frac{y}{T}\right)^{T-y} \left(\frac{y}{T}\right)^y\right]} \right).$$

Under the null hypothesis that the model is correct and  $LR_{POF}$  is asymptotically chi-squared ( $\chi^2$ ) distributed with degree of freedom as one (1). If the value of the  $LR_{POF}$  statistic is greater than the critical value (or p-value < 0.01 for 1% level of significant or p-value < 0.05 for 5% level of significant) the null hypothesis is rejected and the model then is inaccurate.

The Christoffersen's Interval Forecast Test combined the independence statistic with the Kupiec's POF test to obtained the joint test [30, 31]. This test examined the properties of a good VaR model, the correct failure rate and independence of exceptions, that is condition coverage (cc). the conditional coverage (cc) is given as

$$LR_{cc} = LR_{POF} + LR_{ind}$$

Where

$$LR_{ind} = \sum_{i=2}^n \left[ -2 \ln \left( \frac{p(1-p)^{u_i-1}}{\left(\frac{1}{u_i}\right) \left(1 - \frac{1}{u_i}\right)^{u_i-1}} \right) \right] - 2 \ln \left( \frac{p(1-p)^{u-1}}{\left(\frac{1}{u}\right) \left(1 - \frac{1}{u}\right)^{u-1}} \right)$$

Where  $u_i$  is the time between exceptions  $I$  and  $i-1$  while  $u$  is the sum of  $u_i$ .

If the value of the  $LR_{cc}$  statistic is greater than the critical value (or p-value < 0.01 for 1% level of significant or p-value < 0.05 for 5% level of significant) the null hypothesis is rejected and that leads to the rejection of the model.

### 3.14 Distributions of GARCH models

In this study we employed two innovations namely student t and skewed student t distributions they can account for excess kurtosis and non-normality in financial returns [28, 33].

The student t-distribution is given as

$$f(y) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{y^2}{v}\right)^{-\frac{(v+1)}{2}} ; -\infty < y < \infty$$



The Skewed student t-distribution is given as

$$f(y; \mu, \sigma, v, \lambda) = \begin{cases} bc \left( 1 + \frac{1}{v-2} \left( \frac{b \left( \frac{y-\mu}{\sigma} \right) + a}{1-\lambda} \right)^2 \right)^{-\frac{v+1}{2}}, & \text{if } y < -\frac{a}{b} \\ bc \left( 1 + \frac{1}{v-2} \left( \frac{b \left( \frac{y-\mu}{\sigma} \right) + a}{1+\lambda} \right)^2 \right)^{-\frac{v+1}{2}}, & \text{if } y \geq -\frac{a}{b} \end{cases}$$

Where  $v$  is the shape parameter with  $2 < v < \infty$  and  $\lambda$  is the skewness parameter with  $-1 < \lambda < 1$ . The constants  $a$ ,  $b$  and  $c$  are given as

$$a = 4\lambda c \left( \frac{v-2}{v-1} \right); b = 1 + 3(\lambda)^2 - a^2; c = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi(v-2)}\Gamma\left(\frac{v}{2}\right)}$$

Where  $\mu$  and  $\sigma$  are the mean and the standard deviation of the skewed student t distribution respectively.

#### 4. Method of data collection

The data used in this study is a secondary data that was collected from [www.ca.shcraft.com](http://www.ca.shcraft.com) under stock trend and analysis. Daily stock price was collected on Zenith bank stock price from October 21st 2004 to May 8th 2017.

The returns was calculated using the formula below

$$R_t = \ln P_t - \ln P_{t-1} \tag{29}$$

Where  $R_t$  is stock returns;  $P_t$  is the present stock price;  $P_{t-1}$  is the previous stock price and  $\ln$  is the natural logarithm transformation. Then total observation becomes 3070.

#### 5. Results and discussion

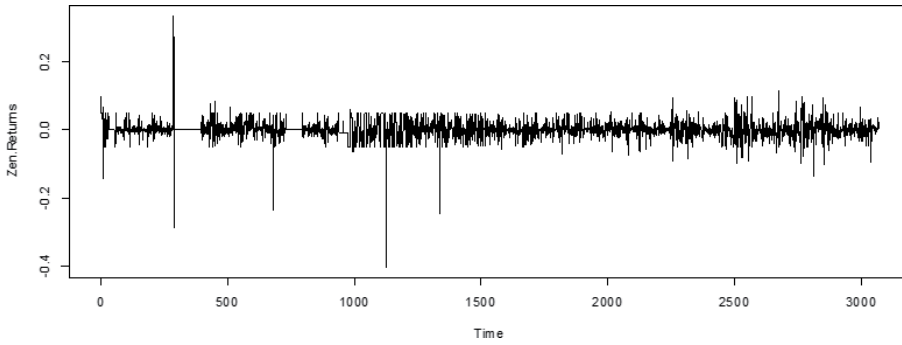
The section presented the results emanating from the analysis and discussions of results.

**Figure 1** below presented the plot of the log of Zenith Bank returns which is the first step in financial time series analysis. The plot revealed some spikes at the early part of the return series while later the series returns became stable.

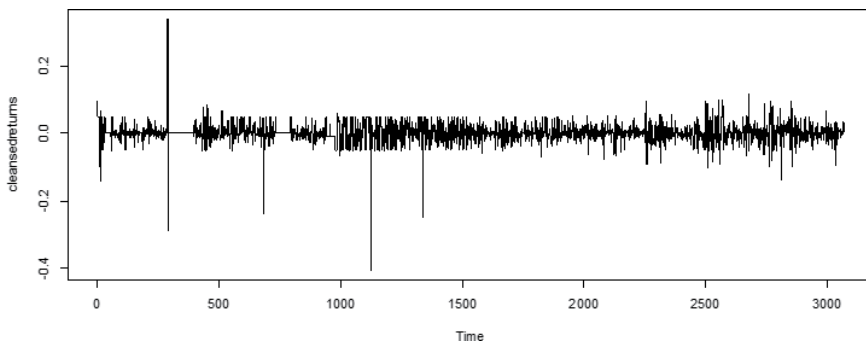
**Figure 2** below presented the plot of the cleansed log of Zenith Bank returns, this is necessary to remove any possible outlier that may be presents in the return series.

The **Table 1** below presented the descriptive statistics of the zenith bank return series. The **Table 1** revealed a maximum return as 0.338000 while minimum return as  $-0.405850$ . The average return as 0.000114 which signifies a gain in the stock for the period under study. The series is negatively skewed with high value of kurtosis. The return series is not normally distributed and the return series is stationary with presence of ARCH effects in the return series, these are typical characteristics of a financial return series [34, 35].

**Table 2** below presents the selection criteria values for daily zenith Bank stock returns based on the student and skewed student t-distributions. The log returns of



**Figure 1.**  
 The time plot of the log of zenith Bank returns.



**Figure 2.**  
 The time plot of the removal of possible outliers in the log of zenith Bank returns.

Statistic	Value
Mean	0.000114
Median	0.000000
Maximum	0.338000
Minimum	-0.405850
Std. Dev.	0.027600
Skewness	-1.267452
Kurtosis	33.76662
Jarque-Bera	121905.9 (p = 0.00000)
Number of Observation	3070
<b>Unit root testing</b>	
ADF	-47.11172 (p = 0.0000)
DF-GLS	-1.842682
PP	-46.52078 (p = 0.0000)
<b>ARCH test</b>	
Chi-squared = 123.05, df = 12, p-value <2.2e-16	

**Table 1.**  
 Descriptive statistics and unit root testing of zenith Bank stock returns.

Model	Information criteria	Std t innovation	Skewed stdt innovation
sGARCH (1,1)	Akaike	NA	NA
	Bayes		
	Shibata		
	Hannan-Quinn		
sGARCH (2,1)	Akaike	NA	NA
	Bayes		
	Shibata		
	Hannan-Quinn		
sGARCH(2,2)	Akaike	NA	-5.3667
	Bayes		-5.3530
	Shibata		-5.3667
	Hannan-Quinn		-5.3618
gjrGARCH(1,1)	Akaike	-5.5110	NA
	Bayes	-5.5012	
	Shibata	-5.5110	
	Hannan-Quinn	-5.5075	
gjrGARCH(2,1)	Akaike	-5.5493	NA
	Bayes	-5.5356	
	Shibata	-5.5493	
	Hannan-Quinn	-5.5444	
gjrGARCH(2,2)	Akaike	NA	NA
	Bayes		
	Shibata		
	Hannan-Quinn		
eGARCH (1,1)	Akaike	-5.0584	-5.0587
	Bayes	-5.0485	-5.0469
	Shibata	-5.0584	-5.0587
	Hannan-Quinn	-5.0548	-5.0545
eGARCH (2,1)	Akaike	-5.0853	-5.0859
	Bayes	-5.0716	-5.0702
	Shibata	-5.0853	-5.0859
	Hannan-Quinn	-5.0804	-5.0802
eGARCH (2,2)	Akaike	-5.0196	NA
	Bayes	-5.0039	
	Shibata	-5.0196	
	Hannan-Quinn	-5.0140	
iGARCH (1,1)	Akaike	-5.1474	-5.1498
	Bayes	-5.1415	-5.1420
	Shibata	-5.1474	-5.1498
	Hannan-Quinn	-5.1453	-5.1470
iGARCH (2,1)	Akaike	-5.1527	-5.1526
	Bayes	-5.1449	-5.1428
	Shibata	-5.1527	-5.1527
	Hannan-Quinn	-5.1499	-5.1491
iGARCH (2,2)	Akaike	-5.1496	-5.1547
	Bayes	-5.1397	-5.1429
	Shibata	-5.1496	-5.1547
	Hannan-Quinn	-5.1460	-5.1505
TGARCH(1,1)	Akaike	-5.8914	-5.8920
	Bayes	-5.8815	-5.8803
	Shibata	-5.8914	-5.8921
	Hannan-Quinn	-5.8878	-5.8878
TGARCH(2,1)	Akaike	-5.9253	-5.8819
	Bayes	-5.9115	-5.8662
	Shibata	-5.9253	-5.8819
	Hannan-Quinn	-5.9203	-5.8763

Model	Information criteria	Std t innovation	Skewed stdt innovation
TGARCH(2,2)	Akaike	-5.8908	-5.8752
	Bayes	-5.8751	-5.8575
	Shibata	-5.8908	-5.8752
	Hannan-Quinn	-5.8851	-5.8688
NGARCH(1,1)	Akaike	-15.563	-13.191
	Bayes	-15.553	-13.179
	Shibata	-15.563	-13.191
	Hannan-Quinn	-15.559	-13.187
NGARCH(2,1)	Akaike	-14.470	-16.419
	Bayes	-14.458	-16.405
	Shibata	-14.470	-16.419
	Hannan-Quinn	-14.466	-16.414
NGARCH(2,2)	Akaike	-9.5866	-11.248
	Bayes	-9.5729	-11.232
	Shibata	-9.5866	-11.248
	Hannan-Quinn	-9.5817	-11.242
apARCH(1,1)	Akaike	-7.8258	NA
	Bayes	-7.8140	
	Shibata	-7.8258	
	Hannan-Quinn	-7.8216	
apARCH(2,1)	Akaike	-8.1226	-8.7718
	Bayes	-8.1069	-8.7541
	Shibata	-8.1226	-8.7718
	Hannan-Quinn	-8.1170	-8.7654
apARCH(2,2)	Akaike	-16.904	9.4341
	Bayes	-16.886	9.4538
	Shibata	-16.904	9.4341
	Hannan-Quinn	-16.897	9.4412
NAGARCH(1,1)	Akaike	-5.1428	-5.1402
	Bayes	-5.1330	-5.1285
	Shibata	-5.1428	-5.1403
	Hannan-Quinn	-5.1393	-5.1360
NAGARCH(2,1)	Akaike	-5.1296	-5.1343
	Bayes	-5.1158	-5.1186
	Shibata	-5.1296	-5.1343
	Hannan-Quinn	-5.1246	-5.1286
NAGARCH(2,2)	Akaike	-5.1221	-5.0439
	Bayes	-5.1063	-5.0262
	Shibata	-5.1221	-5.0439
	Hannan-Quinn	-5.1164	-5.0375
AVGARCH(1,1)	Akaike	-5.8467	-5.6004
	Bayes	-5.8349	-5.5866
	Shibata	-5.8467	-5.6004
	Hannan-Quinn	-5.8425	-5.5954
AVGARCH(2,1)	Akaike	-5.6197	-5.9524
	Bayes	-5.6020	-5.9327
	Shibata	-5.6197	-5.9524
	Hannan-Quinn	-5.6134	-5.9453
AVGARCH(2,2)	Akaike	-5.4227	-5.8644
	Bayes	-5.4031	-5.8428
	Shibata	-5.4228	-5.8644
	Hannan-Quinn	-5.4157	-5.8567

*Note: NA-Not Available.*

**Table 2.**  
 GARCH models and their performance on the log returns of daily log zenith Bank returns.



the daily stock price of Zenith Bank returns were modeled with nine different GARCH models (sGARCH, gjrGARCH, eGARCH, iGARCH, aPARCH, TGARCH, NGARCH, NAGARCH and AVGARCH) with maximum lag of 2. Most the information criteria for the sGARCH model were not available because the model fails to converge. The lowest information criteria were associated with apARCH (2,2) with Student t-distribution followed by NGARCH(2,1) with skewed student t distribution. The caution here is that GARCH model should not be selected only based on information criteria only but the significance value of the coefficients, goodness-of-fit and backtesting should be considered also [3]. The estimated GARCH models for the zenith bank stock with nine different GARCH models (sGARCH, gjrGARCH, eGARCH, iGARCH, aPARCH, TGARCH, NGARCH, NAGARCH and AVGARCH) shows that most of the coefficients of the fitted GARCH models were

Models	Std		Skewed std	
	Persistence	Half-life volatility	Persistence	Half-life volatility
sGARCH (1,1)	NA	NA	NA	NA
sGARCH (2,1)	NA	NA	NA	NA
sGARCH(2,2)	NA	NA	0.9783281	31.63581
gjrGARCH(1,1)	0.9945289	126.3447	NA	NA
gjrGARCH(2,1)	0.9939226	113.7067	NA	NA
gjrGARCH(2,2)	NA	NA	NA	NA
eGARCH (1,1)	0.9495821	13.39848	0.9501065	13.543
eGARCH (2,1)	0.9799226	34.17597	0.9802555	34.75809
eGARCH (2,2)	0.9775862	30.57714	NA	NA
iGARCH (1,1)	1	infinity	1	infinity
iGARCH (2,1)	1	infinity	1	infinity
iGARCH (2,2)	1	infinity	1	Infinity
TGARCH(1,1)	0.9463794	12.57713	0.9587135	16.43969
TGARCH(2,1)	0.9529079	14.36961	0.9506704	13.70184
TGARCH(2,2)	0.9315479	9.775345	0.9470317	12.73636
NGARCH(1,1)	0.9925847	93.1287	0.9732531	25.56687
NGARCH(2,1)	0.984207	43.54208	0.9888705	61.93282
NGARCH(2,2)	0.9704479	23.10679	0.9971636	244.0279
apARCH(1,1)	0.9759139	28.42987	NA	NA
apARCH(2,1)	0.9829391	40.28021	0.9853317	46.90736
apARCH(2,2)	0.9869038	52.58005	0.9513766	13.90596
NAGARCH(1,1)	0.9933088	103.2444	0.9950269	139.0335
NAGARCH(2,1)	0.9910378	76.99442	0.9942849	120.9365
NAGARCH(2,2)	0.9974602	272.5659	0.9978423	320.8989
AVGARCH(1,1)	0.9579476	16.13387	0.9315018	9.768526
AVGARCH(2,1)	0.9311321	9.714181	0.9513755	13.90564
AVGARCH(2,2)	0.9635552	18.6704	0.9633697	18.57406

**Table 4.**  
*Persistence and half-life volatility of the GARCH models of daily log zenith Bank stock returns.*

Model	Distributions	Alpha	Expected Exceed	Actual VaR Exceed	Unconditional Coverage (Kupiec) H <sub>0</sub> : Correct Exceedances	Conditional Coverage (Christoffersen) H <sub>0</sub> : Correct Exceedances and Independence of Failure	
eGARCH (1,1)	Student t	1%	10.7	10	LR.uc Statistic: 0.047 LR.uc Critical: 6.635 LR.uc p-value: 0.828 Reject Null: NO	LR.cc Statistic: 0.236 LR.cc Critical: 9.21 LR.cc p-value: 0.889 Reject Null: NO	
		5%	53.5	67	LR.uc Statistic: 3.332 LR.uc Critical: 3.841 LR.uc p-value: 0.068 Reject Null: NO	LR.cc Statistic: 3.497 LR.cc Critical: 5.991 LR.cc p-value: 0.174 Reject Null: NO	
	Skewed student t	1%	10.7	10	LR.uc Statistic: 0.047 LR.uc Critical: 6.635 LR.uc p-value: 0.828 Reject Null: NO	LR.cc Statistic: 0.236 LR.cc Critical: 9.21 LR.cc p-value: 0.889 Reject Null: NO	
		5%	53.5	74	LR.uc Statistic: 7.425 LR.uc Critical: 3.841 LR.uc p-value: 0.006 Reject Null: YES	LR.cc Statistic: 7.428 LR.cc Critical: 5.991 LR.cc p-value: 0.024 Reject Null: YES	
	NGARCH (1,1)	Student t	1%	10.7	76	LR.uc Statistic: 171.505 LR.uc Critical: 6.635 LR.uc p-value: 0 Reject Null: YES	LR.cc Statistic: 175.258 LR.cc Critical: 9.21 LR.cc p-value: 0 Reject Null: YES
			5%	53.5	135	LR.uc Statistic: 93.627 LR.uc Critical: 3.841 LR.uc p-value: 0 Reject Null: YES	LR.cc Statistic: 101.753 LR.cc Critical: 5.991 LR.cc p-value: 0 Reject Null: YES
NGARCH (2,1)	Skewed student t	1%	10.7	74	LR.uc Statistic: 163.466 LR.uc Critical: 6.635 LR.uc p-value: 0 Reject Null: YES	LR.cc Statistic: 171.614 LR.cc Critical: 9.21 LR.cc p-value: 0 Reject Null: YES	
		5%	53.5	141	LR.uc Statistic: 106.038 LR.uc Critical: 3.841 LR.uc p-value: 0 Reject Null: YES	LR.cc Statistic: 111.739 LR.cc Critical: 5.991 LR.cc p-value: 0 Reject Null: YES	
apARCH (2,2)	Student t	1%	NA	NA	NA	NA	
		5%	NA	NA	NA	NA	
TGARCH (2,1)	Student t	1%	10.7	31	LR.uc Statistic: 25.744 LR.uc Critical: 6.635 LR.uc p-value: 0 Reject Null: YES	LR.cc Statistic: 25.755 LR.cc Critical: 9.21 LR.cc p-value: 0 Reject Null: YES	
		5%	53.5	92	LR.uc Statistic: 24.225 LR.uc Critical: 3.841 LR.uc p-value: 0 Reject Null: YES	LR.cc Statistic: 24.823 LR.cc Critical: 5.991 LR.cc p-value: 0 Reject Null: YES	

*Note:* uc.LRstat: the unconditional coverage test likelihood-ratio statistic; uc.critical: the unconditional coverage test critical value; uc.LRp: the unconditional coverage test p-value; cc.LRstat: the conditional coverage test likelihood-ratio statistic; cc.critical: the conditional coverage test critical value; cc.LRp: the conditional coverage test p-value; NA: not available.

**Table 5.** Backtesting of the GARCH models: GARCH roll forecast (backtest length: 1070) for the log daily zenith Bank stock returns.

not significant at 5% level except for eGARCH (1,1) model that provided significant coefficients in most cases. In the overall, most of the estimated GARCH models revealed absence of serial correlation in the error terms and absence of ARCH effects in the residuals. Because of limited space, we presented only the result of eGARCH (1,1) model in **Table 3** above.

Persistence of GARCH model measure whether the estimated GARCH model is stable or not as shown in **Table 4** above. In financial time series literature it should be less than 1 [3, 36]. Most of the models are stable except for iGARCH model. The half-life measure how long it will take for mean-reversion of the stock returns. The result revealed an average of 10 days for mean-reversion to take place.

The **Table 5** above presented the backtesting test of some selected GARCH model. The backtesting result of the apARCH (2,2) was not available while eGARCH(1,1) with Skewed student t-distribution, NGARCH(1,1), NGARCH(2,1), and TGARCH (2,1) failed the backtesting but eGARCH (1,1) with student t-distribution passed the backtesting approach which is supported by the results in **Table 5** above. Therefore with the backtesting approach, eGARCH(1,1) with student t-distribution emerged the superior model for modeling Zenith Bank stock returns in Nigeria [30, 31]. This chapter recommended the backtesting approach to selecting reliable GARCH model for estimating stock returns in Nigeria.

## 6. Conclusions

This book chapter investigated the place of backtesting approach in financial time series analysis in choosing a reliable GARCH Model for analyzing stock returns. To achieve this, The chapter used a secondary data that was collected from [www.cashcraft.com](http://www.cashcraft.com) under stock trend and analysis. Daily stock price was collected on Zenith bank stock price from October 21st 2004 to May 8th 2017. The chapter used nine different GARCH models (sGARCH, gjrGARCH, eGARCH, iGARCH, aPARCH, TGARCH, NGARCH, NAGARCH and AVGARCH) with maximum lag of 2. Most the information criteria for the sGARCH model were not available because the model could not converge. The lowest information criteria were associated with apARCH (2,2) with Student t distribution followed by NGARCH(2,1) with skewed student t distribution. The caution here is that GARCH model should not be selected only based on information criteria only but the significance value of the coefficients, goodness-of-fit and backtesting should be considered also [3].

The backtesting result of the apARCH (2,2) was not available while eGARCH (1,1) with Skewed student t distribution, NGARCH(1,1), NGARCH(2,1), and TGARCH (2,1) failed the backtesting but eGARCH (1,1) with student t distribution passed the backtesting approach. Therefore with the backtesting approach, eGARCH(1,1) with student distribution emerged the superior model for modeling Zenith Bank stock returns in Nigeria [30, 31]. This chapter recommended the backtesting approach to selecting reliable GARCH model.

## Acknowledgements

I wish to acknowledge my PhD student and M.Sc. students that have worked under my supervision in the area of financial time series analysis.

## Conflict of interest

The Author declares no conflict of interest.



## **Author details**

Monday Osagie Adenomon  
Department of Statistics and NSUK-LISA Stat Lab, Nasarawa State University,  
Keffi, Nigeria

\*Address all correspondence to: adenomonmo@nsuk.edu.ng

## **IntechOpen**

---

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Adenomon, M. O. (2017): Introduction to Univariate and Multivariate Time Series Analysis with Examples in R. Nigeria: University Press Plc.
- [2] Adenomon, M. O. & Emenogu, N. E. (2020): Double-Edged Sword of Global Financial Crisis and COVID-19 Pandemic on Crude Oil Futures Returns. doi:10.20944/preprints202005.0501.v1
- [3] Emenogu, N. G.; Adenomon, M. O. and Nweze, N. O. (2019): Modeling and forecasting Daily stock Returns of Guaranty Trust Bank Nigeria Plc Using ARMA-GARCH Models, Persistence, Half-life Volatility and Backtesting. Science World Journal, 14(3):1–22.
- [4] Ruppert, D. (2011): Statistics and Data Analysis for Financial Engineering. New York: Springer Science + Business Media
- [5] Lawrance, A. J. (2013): Exploration Graphics for Financial Time Series Volatility. Journal of the Royal Statistical Society Series C (Applied Statistics), 62 (5): 669–686.
- [6] Asemota, O. J. and Ekejiuba, U. C. (2017): An Application of Asymmetric GARCH Models on Volatility of Banks Equity in Nigeria's Stock Market. CBN Journal of Applied Statistics 8(1): 73–99.
- [7] Adigwe P.K; Nwanna I.O and Amala A. (2015): Stock Market Development And Economic Growth In Nigeria: An Empirical Examination (1985–2014). Journal of Policy and Development Studies 9(5): 134–154.
- [8] Yaya, O. S.; Akinlana, D. M and Shittu, O. I. (2016): Modelling Nigerian Banks' Share Prices Using Smooth Transition GARCH Models. CBN Journal of Applied Statistics 7(2):137–158.
- [9] Emenike, K.O. and Aleke, S.F (2012). Modelling Asymmetric Volatility in the Nigerian Stock Exchange. European Journal of Business and Management, 4: 52–62.
- [10] Arowolo, W.B. (2013). Predicting Stock Prices Returns Using GARCH Model. The International Journal of Engineering and Science, 2(5): 32–37.
- [11] Emenike, K. O. and Ani, W. U. (2014): Volatility of the Banking Sector Returns in Nigeria. Ruhuna Journal of Management and Finance, 1(1):73–82.
- [12] Abubakar, A. and Gani, I. M. (2013): Impact of Banking Sector Development on Economic Growth: Another Look at the Evidence from Nigeria. Journal of Business Management & Social Sciences Research (JBM&SSR), 2(4): 47–57.
- [13] Atoi, N. V. (2014) “Testing Volatility in Nigeria Stock market using GARCH Models”. *CBN journal of applied statistics*, 5: 65–93.
- [14] Tsay RS (2005) *Analysis of Financial Time Series*, 2nd Edition. New Jersey: John Wiley & Sons.
- [15] Rossi, E. (2004) “Lecture notes on GARCH models”. University of Pavia, March.
- [16] Ding, Z, Granger, C. W. J. & Engle, R. F. (1993) “A long memory property of stock market returns and a new model”, *Journal of Empirical Finance*, 1, 83–106.
- [17] Jiang, W. (2012) “Using the GARCH model to analyse and predict the different stock markets” *Master Thesis in Statistics*, Department of Statistics, Uppsala University Sweden.
- [18] Grek A (2014) Forecasting accuracy for ARCH models and GARCH(1,1) family which model does best capture the volatility of the Swedish stock market? *Statistics Advance Level Theses 15hp*; Örebro University.

- [19] Hsieh, K. C. & Ritchken, P. (2005) "An Empirical Comparison of GARCH Option Pricing Models". *Review of Derivatives Research*, 8 (3): 129–150.
- [20] Lanne, M. & Saikkonen, P. (2005) "Nonlinear GARCH Models for Highly Persistent Volatility". *Econometrics Journal*, 8 (2): 251–276.
- [21] Malecka, M. (2014) "GARCH Class Models Performance in Context of High Market Volatility". *ACTA Universitatis Lodzianensis Folia Oeconomica*, 3: 253–266
- [22] Nelson D (1991) Conditional heteroskedasticity in asset pricing: A new approach. *Econometrica* 59, 347–370.
- [23] Enocksson D, Skoog J (2012) *Evaluating VaR (Value-at-Risk): with the ARCH/GARCH class models via, European Union. Lambert Academic Publishing*
- [24] Dhamija and Bhalla (2010) "Financial time series forecasting: comparison of neural networks and ARCH models", *International Research Journal of Finance and Management*, 49 (1), 159–172
- [25] Ali, G. (2013) "EGARCH, GJR-GARCH, TGARCH, AVGARCH, NGARCH, IGARCH, and APARCH Models for Pathogens at Marine Recreational Sites". *Journal of Statistical and Econometric Methods*, 2 (3): 57–73.
- [26] Banerjee, A., & Sarkar, S. (2006). Modeling daily volatility of the Indian stock market using intraday data. Working Paper No. 588, IIM, Calcutta. Retrieved March 1, 2017, from <http://www.iimcal.ac.in/res/upd%5CWPS%20588.pdf>
- [27] Ahmed, R. R.; Vveinhardt, J.; Streimikiene, D. and Channar, Z. A. (2018): Mean Reversion in International Markets: Evidence from GARCH and Half-Life Volatility Models. *Economic Research*, 31(1):1198–1217.
- [28] Kuhe, D. A. (2018): Modeling Volatility Persistence and Asymmetry with Exogenous Breaks in the Nigerian Stock Returns. *CBN Journal of Applied Statistics*, 9(1):167–196.
- [29] Christoffersen P, Pelletier D (2004) Backtesting value-at-risk: A duration-based approach, *Journal of Financial Econometrics*, 2(1), 84–108.
- [30] Nieppola O (2009) Backtesting Value-at-Risk Models. M.Sc. Thesis, Helsinki School of Economics, Finland.
- [31] Christoffersen P (1998) Evaluating Interval Forecasts, *International Economic Review*, 39, 841–862.
- [32] Christoffersen P, Hahn J, Inoue A (2001) Testing and Comparing Value-at-Risk Measures, *Journal of Empirical Finance*, 8, 325–342.
- [33] Wilhelmsson, A. (2006): GARCH Forecasting Performance under Different Distribution Assumptions. *Journal of Forecasting*, 25:561–578.
- [34] Chen, C. Y-H. (2014): Does Fear Spill over? *Asia-Pacific Journal of Financial Studies*, 43:465–491
- [35] Abdulkareem A, Abdulkareem KA (2016) Analyzing Oil Price-Macroeconomic Volatility in Nigeria. *CBN Journal of Applied Statistics*, 7(1a): 1–22.
- [36] Emenogu, G. N.; Adenomon, M. O and Nweze, N. O.(2020); On the Volatility of Daily Stock Returns of Total Petroleum Company of Nigeria: Evidence from GARCH Models, Value-at-Risk and Backtesting, *Financial Innovation*, Springer Publisher.





*Edited by Kingsley Okoye*

This book intends to provide the reader with a comprehensive knowledge of the latest developments within the Linked Open Data (LOD) framework and the benefits of supported systems. The book covers the entire spectrum of “*Linked Open Data - Applications, Trends and Future Developments*” with six chapters. Each of the chapters provides an all-inclusive conceptualization of the LOD concepts, methodological approaches, case studies, and the main applications both in theory and practice. This book is a reference and educational book targeted to data scientists, software developers, semantic web engineers, information system designers, process managers, teachers, and researchers, and general consumers in application of LOD methods within various contexts.

Published in London, UK

© 2020 IntechOpen  
© brenkee / pixabay

**IntechOpen**

ISBN 978-1-83962-673-9



9 781839 626739

