Dr Winai Nadee

College of Computing
Prince of Songkla University, Phuket Campus
80 Moo 1 Vichitsongkram Road
Kathu, Phuket, Thailand 83120
Tel: +66 76 276732
Mobile: +66 8 8183 7788
Email: winai.na@phuket.psu.ac.th
winain@gmail.com

PROFILE ►

I am a lecturer in Business Informatics teaching in an undergrad E-Business programme. I received my PhD from Henley Business School, University of Reading, UK. My research interest is innovation diffusion and adoption, innovation in organisations, cultural and cognitive research. My PhD research title was *"An Approach for Identifying Conflicts in Technology Adoption at the Informal, Formal and Technical Level"*. My research aimed to investigate innovation adoption patterns and relationship structures concerning individual (cognitive, cultural), organisation and technology perspectives.

I am also interested in implementing data provision and analytics in business environments to help improving business performance and support achieving business goals and strategies by analysing financial activities and optimising business operations. For my 10-year industry experience, I led software development and IT infrastructure implementation projects, which provide technical advice to solves technical problems in IT development and critical IT production environment. Moreover, I was specialised in emerging platform i.e. Google Cloud Platform, Docker, Docker Compose. I also have experience in application development platform (mainframe and open systems) including modern software programming such as Python (Backend), VueJS (Frontend).

EDUCATION ►

Year	Educational Institution	Degree
2017	Business Informatics, Systems & Accounting Henley Business School, University of Reading - UK	PhD in Business Informatics, Systems & Accounting
2008	Mahidol University - TH	MSc in Computer Science
2002	King Mongkut's Institute of Technology North Bangkok - TH	BSc in Applied Statistics

WORK HISTORY ►

Year	Position	Organisation/Company
Jan 2017 - Present	Lecturer in E-Business, E-Business BSc Programme Manager	College of Computing Prince of Songkla University, Phuket Campus
March 2008 – July 2012	IT Specialist/Technical Leader /Solution Architect	IBM Thailand - TH
Dec 2004 – February 2008	Technical Leader/Senior Analyst	IBM Solution Delivery -TH
Apr 2004 – Dec 2004	Java Developer	Advanced Research Group Co., Ltd. - TH
July 2002 – Mar 2004	Cobol Developer	True Corporation Public Co., Ltd. - TH

PUBLICATIONS ►

Prutsachainimmit, K. and **Nadee, W.** (2018) Towards Data Extraction of Dynamic Content from JavaScript Web Applications. Proceedings of the 32nd International Conference on Information Networking (ICOIN2018). Chiang Mai, Thailand. Page: 750-754.

Nadee, W., Gulliver, S., Ali, S. (2017) A Dual Aspect Model: Modeling Systems Alignment. In: Proceedings of the 9th International Conference on Management of Digital EcoSystems (MEDES). ACM, 7 – 9 November 2017, Bangkok, Thailand, pp. 159-163.

Kyritsis, M., Gulliver, S., Feredoes, E. and **Nadee, W.** (2016) Do more pictures mean more effort? Investigating the effects of monocular depth on target detection in a 3D WIMP pictures folder. In: The Ninth International Conference on Advances in Computer-Human Interactions (ACHI 2016), 24 - 28 April 2016, Venice, Italy, pp. 1-5.

Nadee, W., Alhammad, M., Uppal, A. and Gulliver, S. (2015) Dual semiosis analysis model managing customer-focused service innovation. In: 12th International Joint Conference on Computer Science and Software Engineering 2015, 22 - 24 July 2015, Songkla, Thailand, pp. 144-149.

ACADEMIC SERVICES ▶

Subcommittee of SME Financial Analysis – Office of Industry, Phuket City.
Instructor of Curriculum Development of Vocational Education (Scientific Based) – Phang Nga Technical College, Phan Nga Province.
Consultant for IT Development Project – Advance Life Insurance PLC.
Consulting Technologist – Weserve.co.th, Phuket City.

WORKING EXPERIENCES / ACHIEVEMENTS ▶

Prince of Songkla University, Phuket Campus (January 2017 – Present)
January 2017 – Present: Lecturer in E-Business undergraduate programme.

- Teaching Modules: Project management, Strategic Management, Collaborative E-Business, Business Research Method, Enterprise Resource Planning.
- Managing E-Business undergraduate programme/curriculum – curriculum manager (until July 2018).

IBM Thailand Co., Ltd. (March 2008 – July 2012) - Thailand
November 2011 – July 2012: Technical Leader/Solution Architect for Application Management Services (Banking – Electronic Channel Application)

- Application design, development and maintenance (online direct debit Application and internet banking application).
- Support migration for the new core banking system.
- Test environment design and management (CICS) for support application development, system testing, pre-SIT, SIT, UAT and non-functional testing (NFR).

September 2010 – November 2011: Technical Leader/Assistant Service Delivery Manager to operate IT Middleware Operation Service in Kasikorn Bank.

- Mentoring and leading of middleware operation service team (7 staffs)
  - Middleware Implementation - installation/configuration (WAS, WMQ, ITDS).
  - Performance Management: system tuning/capacity forecasting.
  - Production Support: change deployment and incident fix.
  - Defect investigation, preventive actions and system monitoring solution.
  - Security: Health check/Hardening.
  - Assessment of all existing middleware inventory and overall systems performance improvement to support business growth.
  - Apply common/best practices.

April 2010 – August 2010: Technical Leader/Solution Architect for Application Management Services

- Application Deployment and Test Environment Management.
  - Define standard and implementation guideline.
  - Manage test environment to support new core banking implementation.
- Lead application development/maintenance team focusing on Performance Management and Testing plus Security.
  - Analysis and resolving an application performance problem.
  - Support and consult for performance testing.
  - Provide consulting and supporting application development to comply with the web application security standard (OWASP Top 10).

March 2008 – April 2010: Perform as a Technical Leader/Solution Architect (Bank legacy application) of New Branch System Replacement - KTransformation Project for Kasikorn Bank.

- design integration architecture of existing bank's legacy application (CICS) with new branch system.
- Design/Implement CICS Web Service for all legacy banking application (CICS).
- Proof of concept for expose SAFE system (existing core bank system) using CICS web services.
- Test environment design and management (CICS) for support application development, system testing, pre-SIT, SIT, UAT and non-functional testing (NFR).
- Support and consult for security health check policy implementation with WAS and WMQ.

IBM Solution Delivery Co., Ltd. (December 2004 – February 2008) - Thailand

February 2007 – February 2008: Perform as a Technical Specialist focus areas of Application Management Services in KBank IT outsourcing and IBM middleware technology design, implementation and support (non-KBank customer).

- Support IBM middleware product design and implementation e.g. Clustering, High Availability, Service Integration.
- Support and consult best practice application development and performance tuning for many of IBM customers.
- Involve as architect to do assessment and planning of H/W estimation and sizing, operation design for IT infrastructure of KTransformation project in KBank.
- Mentoring staff in project in area of software development to train new staff generation.

April 2006 – January 2007: Perform as a Technical Leader of Internet Banking system refreshment project.

- Lead development team and system admin team in Internet Banking hardware and software refreshment (System P570)
- Conduct proof of concept for research feasibility of system migration.
- Identify and resolve problem of migration task.
- Implement new software version installation on new hardware sets (WebSphere Application Server ND plus HTTP Server, WebSphere MQ, Tivoli Directory Server).
- System configuration including clustering, high availability requirement and resolve remaining problem of old system in the new system.
- Support overall end to end system testing.
- Perform system security hardening and health check.
- Migration system to new environment including due with customer site process.

July 2005 – March 2006: Perform as a Technical Leader of eBooth project which support Foreign Exchange and Bill Payment business via booth channel of KBank.

- Analysis of the business requirements provided by the customer.
- Design web application base on object-oriented concept and best practice framework.
- Implementation using Struts Framework, EJB, Web Services, Hibernate (ORM).
- Server configuration: Clustering, Security hardening.
- Database design, implementation and tuning.
- Testing - unit testing & integration testing.
- Perform system security hardening and health check.

December 2004 – June 2005: Perform as a Technical Leader of Internet Banking project which supports bank business service via internet channel of KBank.

- Analyse current development environment problem.
- Establish practical development environment e.g. launch use of Eclipse IDE, re-arrange unmanaged source code.
- Develop enhancement feature of Internet Banking application e.g. Automated Customer Registration via KBank branch.

Advanced Research Group Co., Ltd. (April 2004 – December 2004) - Thailand

JAVA developer in Credit Bureau System for Hong Kong CCRA project.

- Analyse, design and develop Data Load System, Batch Request System, Web-based Report System.
- JBoss Application Server Cluster: configuration (EJB, Messaging) and performance tuning.
- Application deployment (package installation kit and deployment plan).

True Corporation Public Co., Ltd. (July 2002 – March 2004) - Thailand

COBOL developer to support production operation and feature enhancement of the telephone customer service system.

- Develop application using COBOL on IBM (OS/390) --> CA-IDMS and other tools such as ADS, TSO
- Develop JAVA application using JDBC, Servlets, JSP to integrate IBM mainframe with other system such as Oracle, Informix.

CERTIFICATIONS / AWARD RECOGNITIONS ▶

Received Date	Certification/Recognition	Type	Awarded By
2010	Service Excellent (Team)	Award	IBM Global Business Service
2008	Bravo Award	Award	IBM Global Business Service
2007	SYS Star Award (Team)	Award	Kasikornbank PLC
2007	Service Excellent (Team)	Award	IBM Global Service
2006	Service Excellent	Award	IBM Solution Delivery

TRAININGS ▶

Date	Course	Skills/Type	Venue
15-Nov-2005	Administration of WebSphere Application Server V6	IT	IBM TH
13-Jul-2006	Building High Performance Teams	Soft skill	IBM TH
22-Aug-2006	IBM SOA Technical Briefing and Bootcamp	IT	IBM TH
01-Sep-2006	J2EE Architecting and Documenting Program	IT	Software Park - TH
18-Apr-2007	Developing Enterprise IT Architecture	IT	Software Park - TH
31-Jul-2007	WebSphere Enterprise Service bus – Implement ESB	IT	IBM TH
16-Oct-2007	IT Architectural Thinking	IT	IBM TH
19-Aug-2008	WebSphere Technical Conference	IT	IBM SG
6-April-2011	IBM Smart Biz Cloud - Enterprise: Deep Dive Enablement	IT	IBM SG

SKILL ▶

z/OS: MVS, TSO, ISPF, SDSF and OMVS
CICS: Enterprise COBOL, DB2 and CICS Web Services
Middleware: IBM WebSphere – Application Server and MQ
Rational: RAD for System z, Rational Software Architect for Java, Rational AppScan
Programming Languages:
- **Java:** ***Expertise and Experience in Java EE platform*** – Web Application Development Servlet/JSP, Struts Framework, EJB, Messaging, Web Services, Spring Framework, ORM tools (Hibernate).
- **Python:** Django Framework
- **Javascript:** VueJS Framework

Database: DB2, Oracle, MySQL/MariaDB, PostgreSQL, MongoDB, NoSQL
UNIX: AIX (System Admin, System Trace, Monitoring), Linux, Shell Script i.e. bash
Markup Language: XML, XSL, WSDL, CSS, JSON
Security Consult: OWASP Top 10 Web Application Security Risks
DevOps: Docker, Docker Compose, Google Cloud Platform, Nginx
Data Science: Jupyter Lab/Hub, Numpy, Pandas, Xarray

PERSONALITY ▶

I am a male, Thai nationality. I was born at November 6th, 1980. My personal interests are sports - running, reading, and baking.

Towards Data Extraction of Dynamic Content from JavaScript Web Applications

Korawit Prutsachainimmit
College of Computing
Prince of Songkla University
Phuket, Thailand
korawit.p@phuket.psu.ac.th

Winai Nadee
College of Computing
Prince of Songkla University
Phuket, Thailand
winai.na@phuket.psu.ac.th

Abstract— **An enormous data in World Wide Web and social media has open opportunities for business and organization to get the significant value that leads to efficient operations. As a result, Web Data Extraction has become an important tool for gathering and translating semi-structured documents into valuable information. However, one of the major challenges is dealing with changes from Web documents, especially emerging of JavaScript Web development technology that has significantly affected the way to embed and rendering data of Web pages. In this paper, we propose a design and implementation of a new Web Data Extraction system that aims for extracting data from JavaScript Web applications. The proposed system enables users to select valuable data from online Web documents by defining data extraction rules and data transformation patterns. The extraction engine automatically scrapes and transforms semi-structure data into relational data. The preliminary evaluation results showed that our proposed system has successfully extract data from modern JavaScript Web applications.**

Keywords— Information retrieval; Web Data Extraction; JavaScript Web application; JSON;

I. INTRODUCTION

With the explosive growth of the World Wide Web and social media, a tremendous of data has become available online in the form of text, photos, videos, etc. This situation opens the opportunity for users to benefit from the available information in many interesting ways, especially data analytics. With analytics of big data from World Wide Web and social media, business and organization can offer significant value that leads to efficient operations, higher profits and better customer's experience. However, the information on the World Wide Web and social media is mainly designed for human browsing, not in a structured form that can be used by other applications. Though many data sources are publicly available as Web services, many others data sources are still not accessible through a programming interface. As a result, the technologies for extracting data from Web documents, so-called Web data extraction, have become an important tool for gathering data [1].

Web Data Extraction systems are software applications aiming at extracting data from Web documents and translate them into structured data. The common process involves fetching and extracting. Fetching is a process of downloading a Web page, which is similar to what browser does when a user browses a Web page. Extracting is done later when the target documents are downloaded. The downloaded documents may be searched, parsed, formatted and then transformed into structured data. The process of fetching and extracting is usually done automatically and repeatedly in order to deliver extracted data into a spreadsheet, database or some other application [2].

One of the major challenges in Web Data Extraction is dealing with changes of Web document over time. Evolving of Web development technologies is a key factor that affects the structure of Web documents. For instance, emerging of JavaScript frameworks in Web development has significantly changed the way of embedding data and the rendering process of Web pages.

In cases of extracting data from a traditional Web application, the extraction process begins with sending a request to the target server. In every request, the server renders data, embed and formatted in an HTML document, and responses it back to the client. With this method, a Web Data Extraction tool can normally process and extract data from the downloaded documents. In contrast to the traditional Web applications, modern JavaScript Web applications fetch data from Web servers through asynchronous JavaScript calls. The server returns only data (no HTML markup) to the client in an asynchronous manner. JavaScript code in client browsers uses the received data to construct the page dynamically. As a result, the client browser or Web Data Extraction process cannot receive the HTML documents and accesses DOM of the target Web pages with the ordinary method. Thus, extracting data from modern JavaScript Web applications is a new challenge for designing and developing a Web Data Extraction tool.

In this paper, we propose a semi-automatic Web Data Extraction system that aims for extracting data from modern JavaScript Web applications. The proposed system is designed to enable end-users to select data from existing Web documents by defining data extraction rules and data transformation patterns. The extraction process is invoked automatically following the user-defined schedule. We have tested our proposed system by extracting product information from an online shopping Website.

II. BACKGROUND AND RELATED WORK

A. Web Data Extraction

Web Data Extraction system has been used in a variety of applications including document analysis, business intelligence, social media, analytics, etc. Systems and tools are developed for extracting data from unstructured documents (emails, business forms or technical papers) and semi-structured documents i.e. Web documents [3]. In this research, we focus on extracting data from Web documents which are massive of semi-structured information presented in HTML formats. To efficiently pull out data from HTML documents, the existing system adopts the broad class of techniques that is text processing, DOM parsing, natural language processing and machine learning [4, 5]. The high-efficiency algorithms and state of the art approaches are implemented as open-source libraries and commercial software.

The designing of Web Data Extraction consists of two parts that apply two different techniques. The first part is algorithms for extracting data from HTML documents. Since a Web document is a hierarchy of HTML elements that are usually represented as DOM (Document Object Model), most of data extraction systems rely on tree-based techniques i.e. addressing, matching and weighing of tree nodes [6]. The second part is called Wrapper, which is a procedure that implements one or multiple data extraction algorithms [7]. To carry on with data extraction process, wrapper continuously runs the data extraction algorithms, transforms and merges them into a structured format. Obviously, exploiting DOM of Web documents is the key mechanism in existing Web Data Extraction approaches.

B. JavaScript Web Application

The growing popularity of JavaScript has changed the way of developing Web applications. JavaScript Web applications evolve the process of rendering Web page and DOM manipulation to offer better Web browsing experience. One of the widely adopted architecture for JavaScript Web application is Single Page Application (SPA). SPA is web application that fully loads all resources in the initial request. The individual components, including DOM, can be replaced or updated independently depending on user's interaction [8]. The other technique that adds the interactive capability to SPA is Asynchronous JavaScript and XML (AJAX). Instead of making a new request and update the whole page with new data every time. SPA acquires only data, usually in JSON format, from the server by creating a background process for sending asynchronous requests to the server. When the requested data has arrived, SPA injects the data into HTML elements by dynamically manipulating DOM.

Since asynchronous data transfer and dynamic DOM manipulation is the key features of SPA and modern JavaScript Web applications, extracting data from this kind of Web applications has become a new limitation of existing Web Data Extraction system, which mainly relies on DOM processing. Therefore, this research aims to find an optimal solution which enables data extraction from modern JavaScript Web applications.

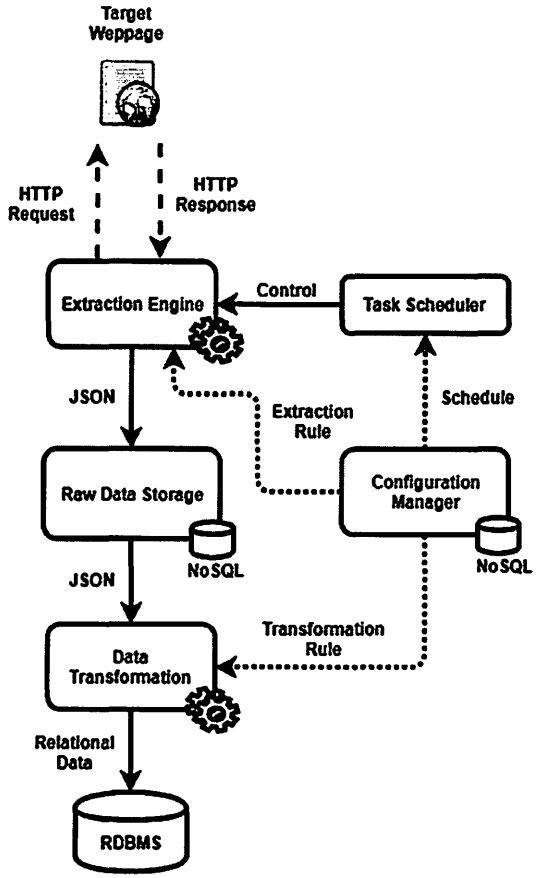

Fig. 1. An overview of the proposed architecture

III. PROPOSED APPROACH

The key idea of our approach is the new design and implementation of data extraction engine. We propose a new Web Data Extraction engine that utilizes the headless browser for fetching Web documents and dealing with dynamically generated DOM.

Rather than performing tree-based techniques on DOM, our approach focuses on extracting JSON data that is asynchronously transferred and cached inside the target Web documents. To accommodate end-users in the data extraction tasks, our approach implements a task scheduler for semi-automatic repeating the data extraction process. A configuration system is used to allow users to define extraction rules and other configurations. Data transformation is designed to translate semi-structured data into structured data that can be an input of data analytics process. In the following subsections, we describe the overview architecture of our approach. We also highlight the key ideas and techniques that we have applied to deal with challenges in enabling data extraction of JavaScript Web application.

A. Overview of Architecture

An overview of the proposed architecture shows in Fig 1. The proposed system consists of 5 major modules as the following:

Extraction Engine. Once the extraction process starts, Extraction Engine is responsible for fetching and extracting data corresponding to the extraction rules provided by Configuration Manager.

Task Scheduler. To control schedule and frequency of running Extraction Engine, Task Scheduler reads configuration from Configuration Manager and invokes Extraction Engine following the user-defined schedules.

Raw Data Storage. This module is a NoSQL data storage designed to keep the original JSON data which is the output of Extraction Engine. Once JSON data is extracted from a target Web document, Extraction Engine stores original version of JSON data in this storage.

Data Transformation. With pre-defined rules from Configuration Manager, JSON data from Raw Data Storage is transformed into relational data and stored in a relational database management system by this module. Since the transformation rules can be added or changed later, keeping raw data in a separated storage gives flexibility and benefits to data transformation process and future data analytics.

Configuration Manager. To enable semi-automatic data extraction and transformation, Configuration Manager provides user-defined configurations which include data extraction rules, invocation schedule and data transformation rules for supporting execution of other modules.

B. Extraction Engine

To deal with dynamically generated DOM used in modern JavaScript Websites, the extraction engine employs headless browser technique for fetching the target Web documents. In this way, other processes of Extraction Engine can extract data from DOM of the fetched documents as ordinary HTML documents that contain a static DOM. The detail process of Extraction Engine is illustrated in Fig 2.

During our experiment, by analyzing DOM of tested Websites, we found that modern JavaScript Websites acquire data from back-end Web services using JSON format and store the received JSON data in some part of its DOM as cached data. As a result, our approach leverages the embed JSON data for the data extraction process by using the DOM parser to extract the embed JSON data instead of getting text value inside HTML tags as implemented in the existing approaches. By extracting cached JSON data from the target Web documents, our proposed method can overcome the common problem of Web extraction engine, i.e. dealing with changing of DOM elements, because JSON data are rarely changed. Thus, the major task of Extraction Engine is finding and extracting JSON chunk of data that represents required information from the target Web pages.

The major challenge of Extraction Engine is finding cached JSON data in DOM of the documents. Our solution is allowing users to specify target data in the form of data extraction rules. An example of data extraction process is shown in Fig. 3. The processing flow of the example can be described as following:

Step 1. Read the DOM extraction rules from Configuration Manager.

Step 2. Parse the fetched documents to create DOM tree

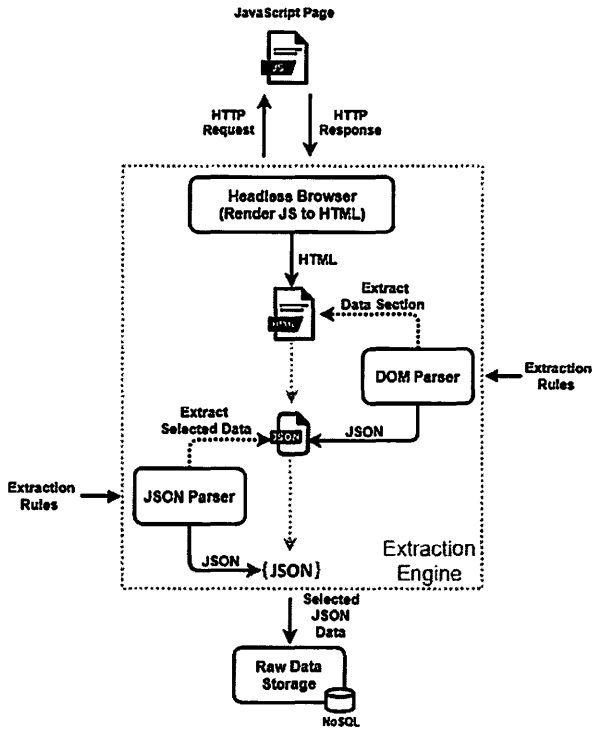

Fig. 2. The detail process of Extraction Engine

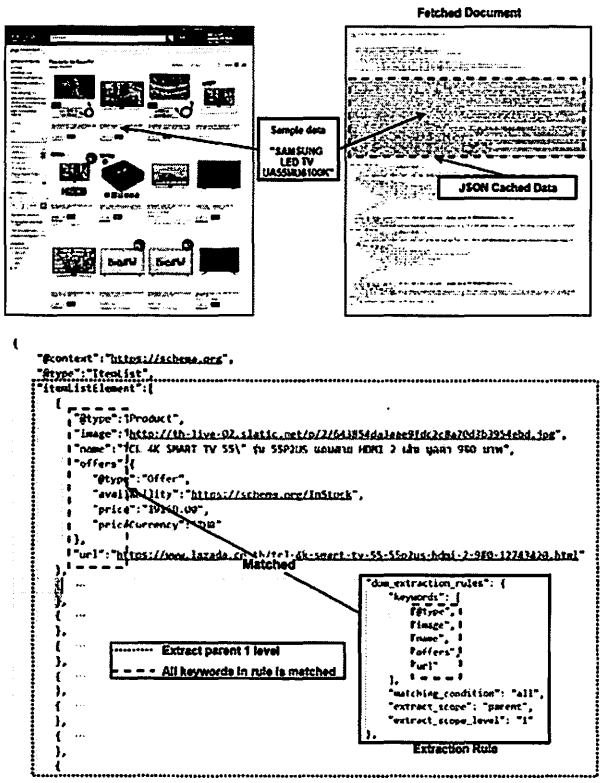

Fig. 3. An example of data extraction process

Step 3. Traverse each element in DOM tree.

Step 4. Compare the current DOM element with the value specified in the extraction rules.

According to the "matching_condition: all", DOM Parser traverses each element of the fetched documents and finds the element that contains all the values specified in the extraction rules i.e. "@type", "image", "name", "offers", and "url". If all the specific elements are found in a specific DOM element, DOM Parser extracts one parent level of JSON data from the target document, i.e. "itemListElements", which is the JSON collection that contains data of all products of the target document. Otherwise, move to the next DOM elements (Step 3).

Since the output JSON chuck may contain unrelated or uninterested data, the JSON parser is responsible for selecting data according to the criteria specified in the extraction rules. Finally, the selected JSON data is stored in the Raw Data Storage for the data transformation process.

C. Configuration File

The configuration file is defined as a well-formed JSON document that contains setting parameters for each module in a separated configuration section. The template of the configuration file and short explanation of configuration values in each section is illustrated in TABLE. I.

TABLE. I. TEMPLATE OF THE CONFIGURATION FILE

Configuration Sections

Target Web Document

```
"target": [
    {
        "base_url": "[URL of target Web document]",
        "parameters": [
            "[Parameter for completing URL of target Web document]"
        ],
        "params_matching": "[Parameter matching pattern (one-many / one-one)]"
    }
],
```

Task Scheduler

```
"schedule": {
    "interval": "[Frequency of extraction]",
    "unit": "[Unit of frequency]",
    "start_time": "[Time to start the extraction process]",
    "end_time": "[Time to stop the extraction process]",
    "no_limit": "[Stop the extraction process when this value is reached]"
},
```

DOM Parser

```
"dom_extraction_rules": {
    "keywords": [
        "[Sample data for searching JSON chunk]"
    ],
    "matching_condition": "[Matching condition for keywords (all / one)]",
    "extract_scope": "[Scope of JSON chunk extraction (parent / top)]",
    "extract_scope_level": "[No of extraction level]"
},
```

Data Transformation

```
"transfomation_rules": {
    "db_connection_name": "[RDBMS connection name]",
    "table_name": "[RDBMS table name]",
    "source": [
        "[JSON data fields]"
    ],
    "destination": [
        "[RDBMS columns]"
    ]
}
```

IV. IMPLEMENTATION

A. System Development

In order to enable users to easily extract data from Web documents and to experiment our solutions with the real-world JavaScript Web applications, we have developed our proposed system as a command-line application using Node.js. We have selected PhantomJS [9] for implementing the headless browser in Extraction Engine. MongoDB and MySQL are applied for NoSQL data storage and RDBMS respectively. A user can start the extraction process by calling the system via a command-line tool with a valid configuration file as a parameter. The output of the system is the relational data that is stored in a relational database corresponding to the configuration.

B. Experimental

We have conducted a preliminary evaluation of our system using a data extraction scenarios.

Scenario. *Extract product information from a shopping Websites:* This scenario simulates a data gathering task that aims to collect product information, including name, description, and prices, from a shopping Website. We selected Lazada [10], the top shopping online Website of Thailand, as the target Website for this scenario. The extraction rules and configurations were set to retrieve name, description and price of selected product categories. The expected extraction result is a time series of data that can express pricing trend and support price prediction in the future. The frequency of extraction was set to one time per day due to the discount campaign of Lazada which not commonly change within a day. An example of target Web document of this scenario shows in Fig. 4.

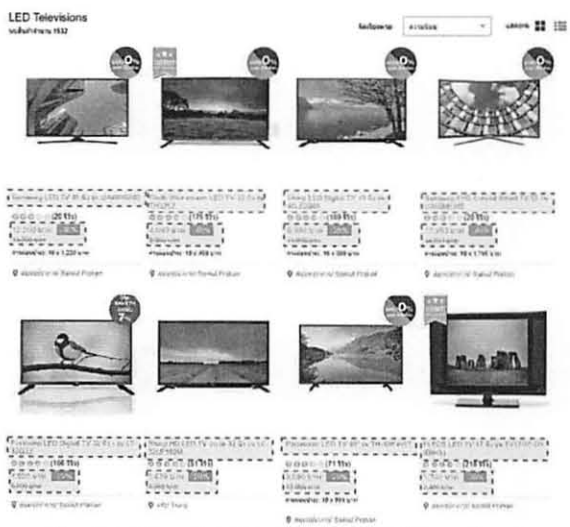

Fig. 4. An Example of target document of the experiment

TABLE. II. EVALUATION RESULT

Evaluation Result of Scenario I				
URL Parameter	DOM extraction keywords	Target	Result	Invalid
http://www.lazada.co.th/shop-led-tv/	name	30	74	44
http://www.lazada.co.th/shop-led-tv/	name,image	30	58	28
http://www.lazada.co.th/shop-led-tv/	image,name,offers	30	30	0
http://www.lazada.co.th/shop-led-tv/	@type	30	30	0
http://www.lazada.co.th/shop-led-tv/	@type,image,name,offers	30	30	0
http://www.lazada.co.th/shop-monitors/	name	30	65	35
http://www.lazada.co.th/shop-monitors/	name,image	30	52	22
http://www.lazada.co.th/shop-monitors/	image,name,offers	30	30	0
http://www.lazada.co.th/shop-monitors/	@type	30	30	0
http://www.lazada.co.th/shop-monitors/	@type,image,name,offers	30	30	0
http://www.lazada.co.th/shop-kitchen-and-dining/	name	30	69	39
http://www.lazada.co.th/shop-kitchen-and-dining/	name,image	30	55	25
http://www.lazada.co.th/shop-kitchen-and-dining/	image,name,offers	30	30	0
http://www.lazada.co.th/shop-kitchen-and-dining/	@type	30	30	0
http://www.lazada.co.th/shop-kitchen-and-dining/	@type,image,name,offers	30	30	0

C. Result and Discussion

The experiment was designed to let our system extracts 30 product information from 3 different groups of the product represented by 3 URL Parameters. Each URL Parameter was tested with 5 different DOM extraction configurations. The result column represents the number of the record that is extracted and stored in the relational database. The extracted items were compared with original information to determine accuracy and invalidity, as displayed in the invalid column. The evaluation result is illustrated in TABLE. II.

The result shows that our system has successfully extracted the interested data from the target Web documents. However, invalid items were found when the DOM extraction configurations are not correctly set. We found that JSON data of the target Website is distributed in many locations of DOM. Consequently, the first and second try of our experiment has failed because the number of matching keywords in the extraction rule is not enough to distinguish the cached JSON data. Thus, selecting efficient keywords for DOM extraction rule is an important factor for our approach. Moreover, there are some considerations and limitations that should be discussed.

1) JavaScript page navigation: In some scenario, modern Javascript Web applications implement its navigation system using lazy loading technique. The first request will receive only first set of data. The remaining data will be fetched by clicking a button, which is sending another request to request next set of data from the back-end system. As a result, we have to add more URL parameter to the configuration to simulate a request that fetches least recent data. In order to automatically navigate through all data, this feature should be added to Extraction Engine.

2) Duplicate data: Since we cannot perfectly set the frequency of extraction to match the changes of data in the target Web sites, repeatedly extracting the same documents has caused duplicate records in both scenarios. Thus, we have added a configuration setting, i.e. "allow_dupplicate", to enable the user to specify wheater duplication is allow or not.

3) Limitation: Our proposed system is designed to extract the JSON data from the target Web documents. The target Web sites are required to expose JSON data in their DOM. However, to be seen and ranked by the search engines, some modern JavaScript applications render their pages from server-side as ordinary client-server Web applications. Thus, we believe that extracting data from DOM is still an important technique for the Web data extraction. As for the future work, we have planned to integrate state of the art DOM extraction techniques with our proposed method to allow universal Web data extraction.

V. CONCLUSION

This paper has presented a semi-automatic Web data extraction system that aims for extracting data from modern JavaScript Web applications. The proposed system enables users to select data from existing Web documents by defining extraction rules, schedules, storage and data transformation logic. The design of Web data extraction engine that utilizes the headless browser for fetching dynamic Web documents and the application of DOM parser techniques to extract the embed JSON data have presented and implemented as a command-line tool. The preliminary evaluation results have shown that our proposed system has successfully extract data from modern JavaScript Web applications i.e. online shopping. The limitations and future work have also been discussed.

REFERENCES

[1] C. Chang, M. Kayed, M. R. Girgis, K. F. Shaalan, C. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A Survey of Web Information Extraction Systems," IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1411–1428, 2006.

[2] E. Ferrara, G. Fiumara, and R. Baumgartner, "Web Data Extraction , Applications and Techniques : A Survey," ACM Trans. Comput. Log., vol. V, no. June, pp. 1–20, 2010.

[3] N. Negm, P. Elkafrawy, and A. B. Salem, "A Survey of Web Information Extraction Tools," Int. J. Comput. Appl., vol. 43, no. 7, pp. 975–8887, 2012.

[4] T. Gogar, O. Hubacek, and J. Sedivy, "Deep neural networks for web page information extraction," in IFIP Advances in Information and Communication Technology, 2016, vol. 475, pp. 154–163.

[5] D. Laishram and M. Sebastian, "Extraction of Web News from Web Pages Using a Ternary Tree Approach," in Proceedings - 2015 2nd IEEE International Conference on Advances in Computing and Communication Engineering, ICACCE 2015, 2015, pp. 628–633.

[6] S. López, J. Silva, and D. Insa, "Using the DOM Tree for Content Extraction," Electron. Proc. Theor. Comput. Sci., vol. 98, pp. 46–59, 2012.

[7] S. Flesca, G. Manco, E. Masciari, E. Rende, and A. Tagarelli, "Web wrapper induction: a brief survey," AI Commun., vol. 17, no. 2, pp. 57–61, 2004.

[8] M. A. Jadhav, B. R. Sawant, and A. Deshmukh, "Single Page Application using AngularJS," in International Journal of Computer Science and Information Technologies, 2015, vol. 6, no. 3, pp. 2876–2879.

[9] PhantomJS, "Full web stack No browser required." [Online]. Available: http://phantomjs.org/.

[10] Lazada, "Lazada." [Online]. Available at: http://www.lazada.co.th/.

A Dual Aspect Model: Modeling Systems Alignment

Winai Nadee
College of Computing
Prince of Songkla University
Kathu, Phuket, Thailand
+66 (0) 76 276732
winai.na@phuket.psu.ac.th

Stephen R. Gulliver
Henley Business School
University of Reading
Reading, UK
+44 (0) 118 3784422
s.r.gulliver@henley.ac.uk

Sarah Ali
Henley Business School
University of Reading
Reading, UK
+44 (0) 118 378 7768
sarah.ali@pgr.reading.ac.uk

ABSTRACT
In literature, individuals, organizations, and technology can all been modelled as systems. This paper justifies the need for, and describes the development of a dual aspect model, which can be used to model and manage the interaction, alignment, and conflict between systems. Our initial dual aspect model consisted of overlapping semiotic onions [14], however model results did not match empirical data. When onion layer orders were adapted to reflect Hall's Major Triad definitions [6], system layer interaction was found to fit well with empirical data collected within modern organizations. Decomposition of the model allows academics and / or practitioners to consider systems alignment stages. Moreover, correction of layer definitions facilitates consideration of norms at the core belief / conceptual level, which supports analysis of human intention. In summary, this paper proposes a newly validated four layered onion structure and the dual alignment pathways, which can be used together to consistently represent and manage alignment between systems (i.e. individuals, technologies and/or organizational systems). The overlapping of these onions, can also be used to study the technical, formal, and informal interplay between individual, organizational and technology aspects.

CCS Concepts
Knowledge Representation Formalisms and Methods → Representations (procedural and rule-based) → *Modeling Systems Alignment.*

General Terms
Management, Measurement, Documentation, Design, Human Factors, Standardization, Languages, Verification.

Keywords
Dual Aspect; Individual; Organization; Technology; Systems Alignment[1]

1. 'CRUCIAL TRIO' AND SYSTEM ASPECTS
All systems have an input phase, a process capability, an output phase, feedback, and a boundary [11]; where the definition of the boundary defines the scope of the focal system. Accordingly, business success depends upon the alignment of multiple systems/aspects, i.e. technology, individuals (i.e. business stakeholders), and/or technology. Literature implies, however, that individuals, organizations, and technologies can all been modelled as systems [10]. Hall [6], whist studying culture, and was the first person to propose the "crucial trio", which divides a system into three different layers, i.e. formal, informal and technical. Hall defined formal norms as being at the core layer, which relates to the concepts/reasons behind the existence of the system; i.e. the system's purpose. Informal adaptation, the middle layer, defines the ways in which the formal concept will be practically expressed in society. The technical layer, the outer concentric circle, relates to the physical enactment of the expression of the concept [6], via use of tools, rules, processes, etc. Hall's concentric 'crucial trio' layers, allowed Hall to consider the internal alignment of human activity (e.g. play, learning, interaction, etc.) in order to assess the conflict between individuals with different learnt backgrounds and/or cultures. To consider the interaction of organizational systems, Stamper [13] changed the order and definition of Hall's "crucial trio", and developed the 'organizational onion', which has been widely adopted within the semiotic community (figure 1). Wiafe et al. [16], for example, adapted the semiotic onion to consider the factors influencing the selection of persuasive technologies. Jacobs and Nakata [7], applied the organizational onion to analyze social media usage within an organization. Chai-Arayalert and Nakata [3] adopted the organizational onion in the knowledge management domain, where they used the three layers to classify the context of knowledge transfer between two organizations, the source and the recipient. Li et al. [9] used the organizational onion's three layers, to develop integrated health clinical pathways, and proposed a systems architecture by classifying individual practical treatments.

Interestingly when the definitions used by Hall's and Stamper's were critically assessed, despite use of the same words, there were key differences in the definitions and order/layout of layers (see table 1).

The Technical dimension in Hall's crucial trio, unlike Stamper's definition, included consideration of written information (i.e. rules, processes); and did not just consider the technology developed to automate the system. Formal written information was dealt separately within Stamper's 'Formal layer' definition,

MEDES '17, November 7–10, 2017, Bangkok, Thailand
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4895-9/17/11...$15.00
https://doi.org/10.1145/3167020.3167044

yet this resulted in a change in order of layers – see table 2. Informal was found to be consistent between the Hall and Stamper. The Formal layer, as defined in Hall's model, which relates to the individual's beliefs, does not have an equivalent layer in Stamper's semiotic onion. Table 2 shows alignment of the Stamper's onion and Hall's crucial trio layers.

Table 1: Comparison of Crucial Trio and Organizational Onion

	Hall's crucial trio	Stamper's onion
Layer 1 (core)	formal - beliefs	Technical – technology, software systems
Layer 2	Informal – behavior, action	Formal – written rules, processes
Layer 3 (outer)	Technical – logics, tools, rules, processes	Informal – meanings, intentions, beliefs

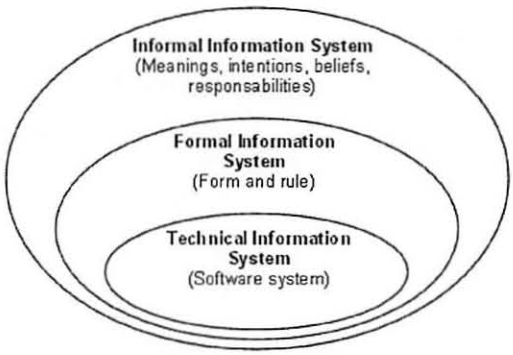

Figure 1: Organizational Onion adapted from Stamper [14]

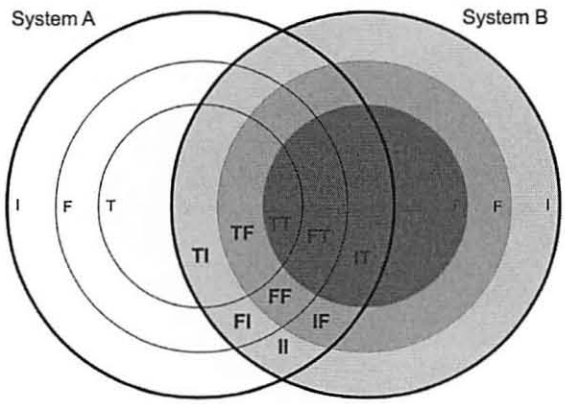

Figure 2: Stamper's Interaction Model - modeling the Interaction of Aspects

Table 2: Mapping of Crucial Trio and Organizational Onion

Hall's crucial trio	Stamper's onion
Formal (F)	
Informal (I)	Informal (I)
Technical (T)	Formal – written rules, processes (F)
Technical (T)	Technical – technology (T)

2. CONSIDERING SYSTEMS INTERACTION

When two systems interact, alignment of key norms is essential if the systems are going to work together. Failure to achieve alignment is likely to result in systems either conflicting, or being rejected [1]. If, for example, a business was seeking the introduction of an ERP system, the business would seek to minimize informal, formal, and technical conflict between the organization and the new software [4, 12]. Alignment between the organization and the new system limits failure risk and/or unnecessary cost faced by the organization in technology customization or reengineering of existing business processes [15]. Alignment between norms is therefore critical to systems alignment, yet analyzing conflicting norms is difficult. We need to compare system layers, yet since the definition and flow between layers is fundamentally different in the work of Hall and Stamper (table 2), which model should be used in context of system conflict analysis?

To investigate the interaction of systems we investigated the interplay of two overlapping semiotic onions (see figure 2), which we termed 'Stampers Interaction Model'. To allow experimentation, this model directly adopted Stampers definitions of terms and/or the flow of dependencies. 'I', the outer concentric circle, represents informal activity. 'F', the middle concentric circle, represents formal rule-based activity. 'T', the core concentric circle, represents technical tools. When two systems fully interact, nine possible interaction points are identified between the two systems; highlighted by an alphabetical code (i.e. TI TF, TT, FT, IT, FI, FF, IF, II). The left alphabetical letter represents the norm layer of System A, and the right alphabetical letter represents the norm layer of System B. TF, for example, represents technology in System A that is not currently supported by and/or used in System B.

3. QUESTIONNAIRE DESIGN

To collect empirical data, we designed a questionnaire that consisted of two parts. **Part one** asked participants in industry about the technologies currently being adopted within their business. Respondents answered part two for each technology considered in part one. **Part two** has three sub sections: i) technology perception; ii) cognitive dissonance; iii) dual aspect. **Section 1** uses the Kano et al. [8] model to compare a positive question and a negative question, on a 5-point Likert scale, in order to identify how respondents, perceive the particular technology. The Kano et al. model, depending on responses, defines technology as being either: must-be, one-dimensional, attractive, indifferent, and reversal [16]. Must-be (M) means that the technology fulfils basic needs. Removing the technology will cause user dissatisfaction, as the technology is critical to basic needs. One-dimensional (O) means that an increased use of the

technology results in increased satisfaction. Although not essential to basic functional needs, there is a lack of fulfilment when it is not there. Attractive (A) means that use of the technology satisfies the individuals, though individuals do not actually need it to meet any functional needs. Removal of an attractive technology will not cause dissatisfaction, as use of the technology was desirable, but not expected. Indifferent (I) means that individuals do not respond to, or care about, the presence of the technology. Reversal (R) means that technology fulfils the individual needs, yet causes dissatisfaction when used.

System A			System B	
TI = CNT+CCB	TF = CNT+CCP	TT = CNT+CCT	FT = CNP+CCT	IT = CNB+CCT
	FI = CNP+CCB	FF = CNP+CCP	IF = CNB+CCP	
		II = CNB+CCB		

Figure 3: Mapping of the questionnaire items with Dual Alignment Model

Section 2 used 3D-RAB instrument questions [15] to measure the respondent's cognitive dissonance state concerning use of the technology. Wiafe et al. [16] designed a model explaining the relationship between attitude and behavior, which includes eleven operationalized items to measure: current behavior (CB, 1 item); attitude towards target behavior (ATTB, 3 items); attitude towards changing behavior (ATCB, 3 items) and the attitude towards maintaining behavior (ATMB, 4 items). In **section 3**, eight statements were used to evaluate informal, formal, and technical change within the dual aspect model.

4. DATA COLLECTION AND ANALYSIS

Data was collected from 217 respondents who worked in technology companies (in Thailand). Valid data contained information about 251 different technologies. Sample numbers were sufficient to confirm the reliability and validity of the study [2]. To support validation, respondent feedback was assigned into two groups: inexperienced user group (CB-) and experienced user group (CB+). The Cronbach's alpha and exploratory factor analysis showed strong reliability, i.e. with Cronbach's alphas 0.899, 0.788 for respectively ATTB and ATCMB of the inexperienced group, and 0.917, 0.687 for the experienced group. Exploratory factor analysis (EFA) showed that factor loading for ATMB03 was below the recommended value of 0.45 (as recommended by Hair et al. [5]), however the CFA fit index decreased when the variable was removed (CFI reduced from .990 to .985, and GFI from .968 to .967). Accordingly, although low, ATMB03 was left in the model for further analysis due to evidence meanings to the theoretical construct [5].

4.1 Dual Aspect Factors

Part 2, section 3, of the questionnaire consisted of eight statements. Statement 2 (CNT: The new technology is required to be customized as it doesn't fit well at the first place.), statement 4 (CNP: The new process needs to change to fit with the current business system.) and statement 7 (CNB: Interaction with the adopting technology is required to be customized to minimize impacts to people's behavior.) were used to collect information about attitude toward changing the new system. Statement 3 (CCT: The existing technology is required to be customized to be compatible with the adopting technology.), statement 5 (CCP: The existing process is required to change to support the new adopting process.) and statement 6 (CCB: People will need to change their way they work once the technology is adopted in place.) were used to represent attitude towards changing the current system). Statements were mapped to nine 'dummy' dual-aspect interaction points (TI, TF, TT, FT, IT, FF, FI, II, and IF) – see figure 3.

Confirmatory factor analysis (CFA) was used to test the loading between dual-aspect interaction points and three latent variables, i.e. Technology Conflict (Technology) = TT+TI+TF, Process Conflict (Process) = FT+FF+FI, and People's behavior Conflict (People Behavior) = IT+IF+II. SEM analysis for dual aspect factors (see figure 4) supports use of the SEM latent variables, yet when tested the dependence of model layers did not match empirical data.

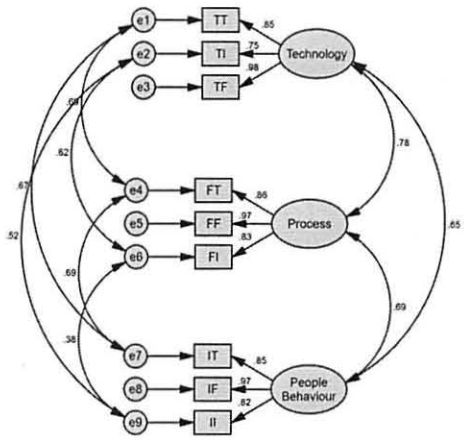

Figure 4: SEM Measurement Model (Confirmatory Factor Analysis) for Dual Aspects – including latent variables.

4.2 Reconsidering Layer Order

In table 2, we showed that the Formal layer in Hall's model, which relates to the individual's background beliefs, does not have an equivalent layer in Stamper's model. This study argues that internal core dimension, i.e. that relate to internal individual beliefs/concepts, should not be seen as informal, yet should instead be allocated a separate layer in the model. In our study, and for the sake of analysis, cognitive dissonance was used to express the central (belief) cognitive dimensions of an individual, which we

will have termed "Concept (C)" in line with Hall's definition (see table 3).

Table 3: Mapping of Crucial Trio and Organizational Onion

Hall's crucial trio	Stamper's onion
Formal (F)	**Concept (C)**
Informal (I)	Informal (I)
Technical (T)	Formal – written rules, processes (F)
Technical (T)	Technical – technology (T)

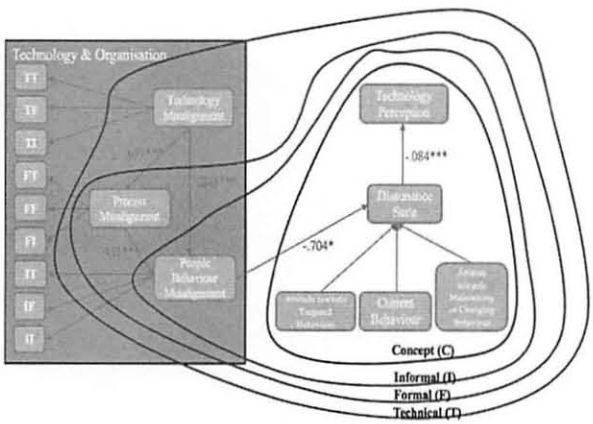

Figure 5: SEM Model

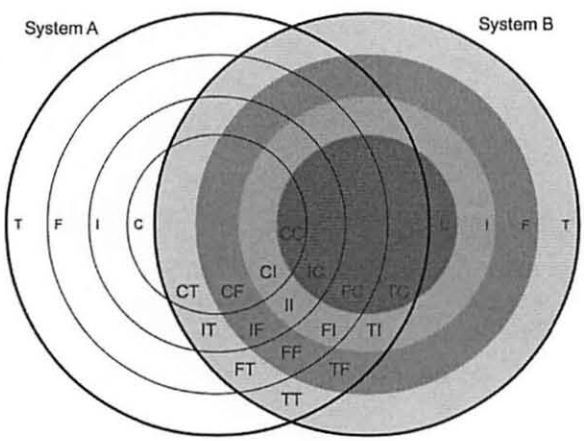

Figure 6: Dual Aspect Model

A SEM structural model was developed and tested to investigate the flow and relationships between informal, formal, informal, conceptual layers (see figure 5). Results show that 'Technology misalignment' affects both 'Process misalignment' (0.245***) and 'People behavior misalignment' (0.697***). 'Process misalignment', influences 'People behavior alignment' (0.441***), which in turn has a relationship with individual 'Dissonance state' (-0.704*). Lastly 'Cognitive dissonance' affects 'Technology perception' (-.084***).

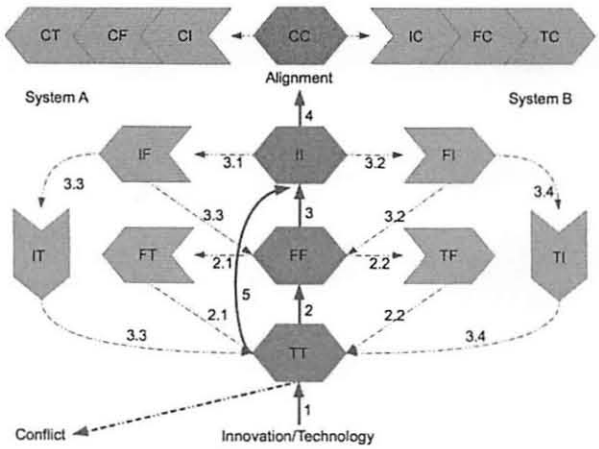

Figure 7: Dual Alignment Pathways

Although Stamper's semiotic onion implies that technology is dependent on informal rules, which is dependent on formal structures, empirical results actually suggest that technical conflict can influence process conflict, which in turn can result in behavioral conflict. A direct link was also identified between technology conflict and informal behavior, implying that technology use directly influences informal behavior. Data implies that, in context of system adoption, that conflict in behavior is driven by technical conflict, which is at odds with Stamper's Semiotic Onion. Results support the use of Stamper's layer definitions [13]; assuming inclusion of an additional concept layer, and the reversal of layer orders (see figure 6).

The empirically validated Dual Aspect Model (figure 6), allows us to consider the interaction of norms in two systems (either/or technical, human, and/or organizational), by identifying whether conflict exist within 16 interaction points (see figure 6 and 7). To achieve informal alignment, empirical results imply that technical and formal alignment must first be achieved. Alignment can be achieved by changing either system A to align to system B – left hand paths in figure 7 (2.1, 3.1, 3.3) – or by changing system B to align to system A– right hand paths in figure 7 (2.2, 3.2, 3.4). If neither system can, or will not, change to facilitate alignment then the conflict will not be resolved unless a workaround can be formally defined.

The authors believe that the proposed Dual Aspect Model, can be practically adopted to assess conflict between any systems (i.e. people, organizations, and/or technology systems). If, for example, a company were to implement an ERP, then the company could use the Dual Aspect Model, supported by the Dual Alignment pathway, to systematically assess potential norm conflicts, between the company and new technology, and/or iteratively manage the pathway of customization and/or Business Process Reengineering in advance of system roll-out. The proposed model, by considering the human dimension, also facilitates consideration of potential conflict with human stakeholders; something that has been shown to be critical to IS implementation success [13]. Although the authors believe that conceptual alignment in business may not be required to support business use (see figure 7), misalignment with people can result in damaging cognitive dissonance, resulting in systems misuse.

5. CONCLUSION AND DISCUSSION

Literature talks about individuals, organizations, and technology as being systems, but top date there has been no effective way to model the conflict and/or guide practitioners toward systems alignment. This paper introduced the work of both Stamper [14] and Hall [6] and proposed a need to model and manage the interaction, alignment, and conflict between systems. Empirical results supported the use of Stamper's layers definitions (i.e. T, I, F), yet suggested the additional inclusion of a concept layer, which was present in Hall's original definition; and facilitates consideration of the core beliefs / concepts. Moreover, findings imply was required in the order of layer dependencies, which implies it is possible gain technology alignment between systems without having to ensure formal process and/or informal alignment first, which was implied by Stamper.

In summary, this paper proposes a validated four-layered onion structure, and the resultant Dual Alignment Pathways, which can be used to consistently model and management system conflict. The overlapping of these onions, can also be used to study the technical, formal, and informal interplay between systems, and the dual alignment pathways, can be used to practically manage system change in order to support practitioners in management of conflict.

6. REFERENCES

[1] Avgerou, C. (2008) 'Information Systems in Developing Countries: A Critical Research Review', Journal of Information Technology, vol. 23, no. 3, pp. 133–146.

[2] Bagozzi, R. P. and Yi, Y. (2012) 'Specification, Evaluation, and Interpretation of Structural Equation Models', Academy of Marketing Science Journal, vol. 40, no. 1, pp. 8–34.

[3] Chai-Arayalert, S. and Nakata, K. (2013) 'Semiotic Approach to A Practice-Oriented Knowledge Transfer', In Liu, K., Li, W., and Gulliver, S. R. (eds.), 14th International Conference on Informatics and Semiotics in Organization (ICISO 2013), Sweden.

[4] Fichman, R. G. and Melville, N. P. (2014) 'How Posture-Profile Misalignment in IT Innovation Diminishes Returns: Conceptual Development and Empirical Demonstration', Journal of Management Information Systems, vol. 31, no. 1, pp. 203–240.

[5] Hair, J. F., Black, W. C., Babin, B. J. and Anderson, R. E. (2009) Multivariate Data Analysis, 7th ed. Upper Saddle River, New Jersey, Pearson Prentice Hall.

[6] Hall, E. T. (1959) The Silent Language, New York, Doubleday.

[7] Jacobs, A. and Nakata, K. (2012) 'Organizational Semiotics Methods to Assess Organizational Readiness for Internal Use of Social Media', In AMCIS 2012 Proceedings, Seattle, Washington.

[8] Kano, N., Seraku, N., Takahashi, F. and Tsuji, S. (1984) 'Attractive Quality and Must-Be Quality', Journal of the Japanese Society for Quality Control, vol. 14, no. 2, pp. 147–156.

[9] Li, W., Liu, K., Yang, H. and Yu, C. (2014) 'Integrated Clinical Pathway Management for Medical Quality Improvement-Based on a Semiotically Inspired Systems Architecture', European Journal of Information Systems, vol. 23, no. 4, pp. 400–417.

[10] Liu, K. (2000). Semiotics in information systems engineering. Cambridge University Press.

[11] Laudon, K. C., & Laudon, J. P. (2010). Management Information Systems: Managing the Digital Firm. 2011.

[12] Onita, C. and Dhaliwal, J. (2011) 'Alignment within the Corporate IT Unit: An Analysis of Software Testing and Development', European Journal of Information Systems, vol. 20, no. 1, pp. 48–68.

[13] Pankratz, O., and Basten, D. (2013). Eliminating Failure by Learning from It–Systematic Review of IS Project Failure. Thirty Fourth International Conference on Information Systems, 1–20.

[14] Stamper, R. K. (1993) 'A Semiotic Theory of Information and Information Systems', In Invited papers for the ICL/University of Newcastle Seminar on 'Information'.

[15] Wastell, D. G., McMaster, T. and Kawalek, P. (2007) 'The Rise of the Phoenix: Methodological Innovation as a Discourse of Renewal', Journal of Information Technology, vol. 22, no. 1, 59–68.

[16] Wiafe, I., Nakata, K., Moran, S. and Gulliver, S. (2011) 'Considering User Attitude and Behavior in Persuasive Systems Design: The 3D-RAB Model', In ECIS, Helsinki, Finland [Online]. Available at: http://aisel.aisnet.org/ecis2011/186.

[17] Xu, Q., Jiao, R. J., Yang, X., Helander, M., Khalid, H. M. and Opperud, A. (2009) 'An Analytical Kano Model for Customer Need Analysis', Design Studies, vol. 30, no. 1, pp. 87–110.