An Ontology based Data Model for Iranian Research Information

Marzieh Raoufnezhad
Information and Communication Technology Department
Ferdowsi University of Mashhad, Iran
raoufnezhad@um.ac.ir

Mohsen Kahani
Factuality of Engineering
Ferdowsi University of Mashhad,Iran
kahani@um.ac.ir

Yaghoob Maharati
Faculty of Economics and Business Administration
Ferdowsi University of Mashhad,Iran
maharati@um.ac.ir

Abstract-**As the number of researcher increases, the amount of information related to research activities grows rapidly. As a result, the management of this information for better retrieval and analysis has become an important issue. Many data models abroad and within Iran have been developed to address this issue. In this paper, after comparing some of these models, a new ontology based data model is proposed. The evaluation results show that the proposed method increases the performance and the organization of research information management compared to the existing methods.**

Keywords-Research Information; Ontology; IRInO; CERIF; SEMAT; SAED

I. INTRODUCTION

Importance of utilization effective approach for comprehending and analyzing data as a key strategy of organizations highlighted by large amount of data and their growing trend. These data help organization in collecting ideas and visions, recognizing and employing opportunities and threats to access competitive advantage. Therefore, many of these organizations attempt to keep their value in challenging and competitive world using modern technologies and forming information systems by focusing on semantic processing of data instead data processing[1]. The higher education isn't anticipated in the rule.

Since success in every community depends on effectiveness of higher education system[2] and data and research information can be effective in advancement of knowledge, promotion of scientific methods and macroeconomic policy, so a standard model for managing research data, development research information system and establishing structural knowledge base for more effective decision making is necessary for every society or organization based on comprehensive science and knowledge of day.

European countries, like Slovakia and France have already developed their research information management systems. The supreme council for science, research and technology of Iran [1] has planned the national research information management model[2] (SEMAT), and is further developing it.

As a semantic tool, ontology is defined as structured knowledge in a specific scope, which is developed over the concepts in that area and the relations between them. Ontology is a clear and formal definition for a knowledge base, consisting of concepts (or class), roles (or properties) between instances and restrictions, with a set of elements and individual or instances which define the knowledge base[3]. In other words Ontology states that what elements (class or concept) compose a knowledge collection and what are relations between these classes[4].

Considering the fact, that research is one of the significant parts of ontology of higher education[5], this paper follows two purposes: first one is improving the structure of SEMAT-as Iranian national research information data model- performed based on results of qualitative comparison between descriptive methods including SEMAT and scientific information system of Ferdowsi University [3] (SAED) and Common European Research Information Format (CERIF).

The second purpose is designing an ontology as a standard pattern for describing research data in order to create a structured knowledge base for research area based on the SEMAT data model. This ontology which has not been designed until now, is presented in RDF format and can be a foundation for inferring and extracting the ontology of research areas in research-based organizations, universities in particular.

This paper is organized as follows: Related works in Section 2, the proposed method in Section 3 and evaluation in Section 4. Finally, the conclusion and future works are presented.

[1] http://www.atf.gov.ir

[2] http://semat.ir

[3] http://www.um.ac.ir

978-1-5090-4335-4/16/$31.00 ©2016 IEEE

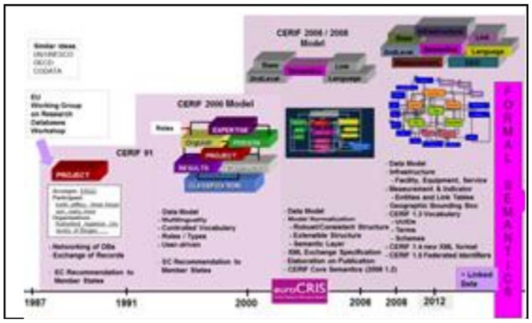


Figure 1. Development process of CERIF from 1987 to 2012[6]

II. Literature Review

Works of the area will be examined by two aspects. Since SEMAT uses CERIF model for descripting research data, first some system based CERIF data model will be examined. Then SAED will be introduced as an internal example of CERIF. SAED data model is not based on CERIF data model, but its study is beneficial because it is implemented and performed in one of credential universities of Iran.

A. Systems based on CERIF data model

In the early 80's, Heads of some national research funding organizations, started a project called IDEAS[7] for examining the databases' association with research information. Based on this, from 1987 to 1989 the EXIRPTS[8] project was expanded on G7 countries, and both projects were successfully completed. From this, CERIF emerged, which for the first time was considered as a simple standard and an information exchange format for describing projects; persons, organizations were represented as attributes.

However, since in practice, the CERIF91 was not a suitable standard and was excessively strict, a new group was formed under the coordination of the European commission and from experts who were members of the Europe Union and associate members, and the CERIF2000 which is a set of instructions for dealing with research information systems, was established and is yet being developed. It aims to expand access to ongoing research information, utilize and share them with the research society and also provide opportunities for skills and knowledge exchange and promote both of which together. CERIF uses semantic RDF format for describing data. "Fig. 1", shows the development process of CERIF from 1987 to 2012.

Data description formats for systems based on CERIF data model have been divided into three categories including RDF, XML and based on characteristics that have been investigated as following.

1) Resource Description Framework

Common European Research Information Format – Semantic Web (CERIF-SW) project is the result of development distributed information system from heterogeneous data sources and database harmonization. It makes research information accessible for developing knowledge management by using semantic web solutions. In order to have an information system with advanced capabilities in retrieving information, The CERIF-SW is targeted to develop a tool set for constructing next generations of research information systems and merging them by new technologies.. For this reason, they have considered some goals based on web standards. For example RDF distributed knowledge management system was one of CERIF goals to make research information of European institution accessible[9].

CERIF-SW had some reasons for using semantic web solutions which include ability of merging different resource data such as information banks, digital libraries, repositories etc. and ability of using ontology for retrieving information intelligently[10].

2) XML Format

The Central Registry of Theses and Dissertations (CRTD) of Slovakia[11] has been established with the aim of registering theses, collecting and archiving final exams and doctoral works from colleges and universities, to support science, research and education in Slovak Republic and has been running from 2009. All documents of higher education institutes in Slovakia must be registered in this center considering the legislations, and the validity of the documents must be examined before the defense. These systems operates in two levels: In the first stage the university and college will create their own local repository and in the second stage, meta-data is sent to the global repository. CRTD will match the document with the global repository for Plagiarism detection. Then the output is used as a decision making tool in the examination commission. The format used in CRTD is based on MARC XML and the xsd schema contains a description of several entities used in CERIF[12].

3) Based on characteristics

Islamic Republic of Iran designed and created national information system of SEMAT[13], a cooperative environment for merging metadata in research institutes[12] of country, under the name of Iran 1404 and 20 years vision with the aim of accessibility of output of data for scientific community. Nowadays, higher council of research and technology collect research information in the form of Excel file from universities and higher education institutions[14]. The considered structure for data in Excel file is designed based on approach of characteristics.

Three data description format are compared in table 1. According to table 1, using RDF as a standard for describing data makes unity in restoring information and eases extraction of them and create enrich information base in the area. If information store in information base in a unite manner, they will be manage and query more easily[15, 16].

The aim of this paper is using ontology to extract SEMAT concepts, entities, and relationships and then represent them in RDF. The RDF output of this model can be a standard knowledge model for designing and implementing knowledge base of research information management systems.

B. Scientific information system of Ferdowsi University

In addition to SEMAT, from 2003 Ferdowsi University of

TABLE I. A QUALITATIVE COMPARISON OF CERIF RESEARCH MODELS, SEMAT AND SAED

Dimension		RDF	XML	Excel
Class And Properties	Clear & explicit definition	✓	✗	✗
	Defining relationships between classes and attributes	✓	✗	✗
	Class inheritance	✓	✗	✗
Data Model flexibility		✓	✓	✗
Easy development		✓	✓	✗
URI (*Universal* Resource *Identifiers)*		✓	✗	✗
Infering and Discovering new information		✓	✗	✗
Data Model		Graph	Tree	Flat

Mashhad has designed a system in order to collect and organize scientific university information, titled University Academic Information System (SAED). This system was suggested by the university's vice chancellor for research and technology, with two goals, accelerate the access of up to date information and prepare necessary reports for submission to the ministry of science[1]; and eliminating the paper work flow. This System was designed and implemented as a web based system by Ferdowsi University's Center for Information and Communication[2] (FAVA), and users can access its information through the Internet.

Although the basic intellectual infrastructure in the systems design, was collecting and making the university's intellectual capital available, but from 2005, it was also considered as a platform for processing research services.

In addition, table 2 is the qualitative comparison between important classes of CERIF, SEMAT and SAED. Generally all three data models formed from three entries or classes such as 1) main entries 2) infrastructure entries 3) additional entries.

These models are the same in some aspects such as main entries, research result and relationships between entries. However, they differ in abstract level. CERIF and SEMAT are conceptual but SAED use physical database. Format of data description in CERIF, is based on RDF but in SEMAT is based on Data Description Format.

As shown in table 2, SEMAT does not cover some defined entries of CERIF and SAED. Hence for having more effective national research information model it is required that SEMAT be optimized. Thus the classes and concepts in CERIF and SAED that does not exist in SEMAT should be added to SEMAT.

III. THE PROPOSED MODEL

This paper propose a model through optimization SEMAT data model and description in RDF using ontology.

TABLE II. QUALITATIVE COMPARISON BETWEEN RESEARCH DATA OF CERIF, SEMAT AND SAED

Class		*CERIF*	*SEMAT*	*SAED*
Basic Entetics	Organization	✓	✓	✓
	Person	✓	✓	✓
	Project	✓	✓	✓
Infrastructure Entities	Equipment	✓	✓	✓
	Facility	✓	✓	✓
	Service	✓	✗	✓
Additional Entities	Event	✓	✓	✓
	Funding	✓	✓	✓
	Prize	✓	✓	✓
	Address	✓	✓	✓
	Curriculum Vitae	✓	✓	✓
	Research Location	✓	✗	✓
	Expertise and Skill	✓	✓	✓
	Measurement and Indicator	✓	✗	✗
	Currency	✓	✓	✓
	MoU	✗	✓	✓
	Citation	✓	✗	✓
	Qualification	✓	✗	✗
Result Entities	Product	✓	✓	✓
	Publication	✓	✓	✓
	Patent	✓	✓	✓
Relationship		✓	✓	✓
Classification		✓	✓	✗

These optimization will be explained as follow:

A. *Optimization of National Research Information, SEMAT Data Model*

The following new classes has been proposed for improving the model:

- **Monitoring and evaluating.** This class is significant from the point that plagiarism, repetitive projects and non-valuating researches can be prevented by supervising and recording research projects of country and their result in a unite manner.
- **Multimedia.** SEMAT doesn’t considered voice record, films, and digital contents of the lessons, pamphlets, PPT files and visual conferences in its worksheet. Indeed these

[1] http://www.msrt.ir

[2] http://its.um.ac.ir

cases are kind of research data and need to be added to SEMAT under the name of multimedia.

- **Research locations.** Nature of research locations can be different from organization. In an organization like universities, many research locations such as labs, research institutes, R&D centers, and centers of excellence exist that each of them has its professional area. For example, knowledge and communication of researchers with places. Being aware of the research Location, allows to be created a network of experts and specialists in different fields of science-research in order to generate knowledge in all over the country.

- **Granting credit of research.** According to significant distance between Iran and industry and developed countries in terms of science and technology, creating motivation should be the main goal of responsive executors. Therefore granting can create motivation, accelerate official affairs, and utilize the optimized ability of researchers.

- **Research activities.** Like speeches or theorizing seats which include scientific criticizes, innovations and theorizing, research activity has been considered as one of scientific-technologic activities, according to promotion bill of research and technology of November 24, 2010

B. Describing SEMAT Data Model in the RDF Format Based on Ontology

Designing method of Iranian Research Information Ontology (IRInO) is explained in this part, which includes methodology of designing ontology and its implementation. Ontology developers believe METHONTOLOGY[17] methodology is suitable for different scopes and use it. Based on METHONTOLOGY methodology, designing IRInO includes 5 following phases:

1) Specification

The aim of designing IRInO is construction of reference ontology, representation of a knowledge model in research area especially in research area of universities. Therefore, ontology must be designed to cover different aspects of research area. In addition, scope of the ontology includes Iranian national research information and research area of Ferdowsi university of Mashhad as a research centered organization. 13 competency questions are presented which include 3 aspects: (a) which persons and organizations do research? (b) What are the research projects and products and what aspect do they have? (c) What are the required resources? These questions are based on opinion of experts in research area of Ferdowsi university of Mashhad, examination of SEMAT instruction completion[14] and Technical document of national research information model[18]. IRInO should give response to these qualification questions. .

2) Conceptualization

In this class, the concepts extracted and classified from SEMAT instruction completion and technical document of national research information model. After initial design of IRInO, it was adopted with SAED by non-cooperative observing method to complete IRInO. The concepts and classes extracted from SAED which did not exist in SEMAT, added to desired ontology.

3) Formalization

The phase is conceptual explanation of ontology in phase 2 into a formal model. Relationships between classes, infrastructure classes, and concepts specified. These relationships determined based on literature, opinion of experts in research area and designing ontology field through interview and examination of nearest existing ontologies of the area. For example, relationships such as hasAuthor, hasPublisher,… defined for book or paper.

4) Implementation

In the phase, formalization model of phase 3 changes into knowledge model, using Protégé software. Designing method of IRInO is based on using existing classifications in desired scope, because there is no ontology for SEMAT now. Combined method is used for determining hierarchy of ontology classes[19]. "Fig 2" is a view of overall IRInO structure in Protégé software.

5) Maintenance

In the phase, defects of ontology will be fixed and implemented ontology will be updated. The phase not only performs in final step of designing and implementing IRInO, but defects and problems of IRInO were detected and fixed using Protégé reasoner tools such as Hermit 1.3.8, opinion of experts (include research experts and designing ontology-

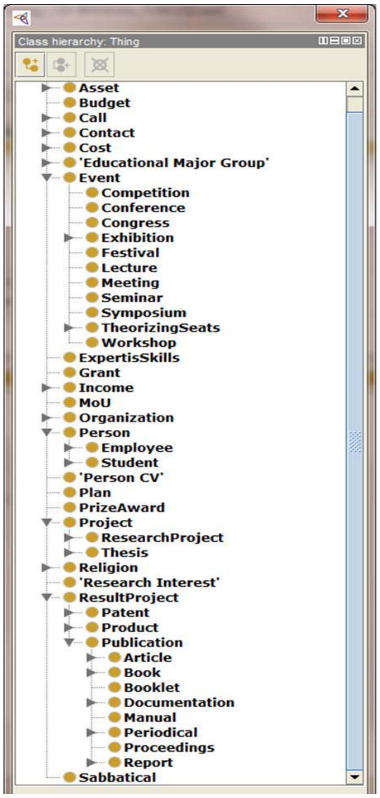


Figure 2. Overall Structure in IRInO

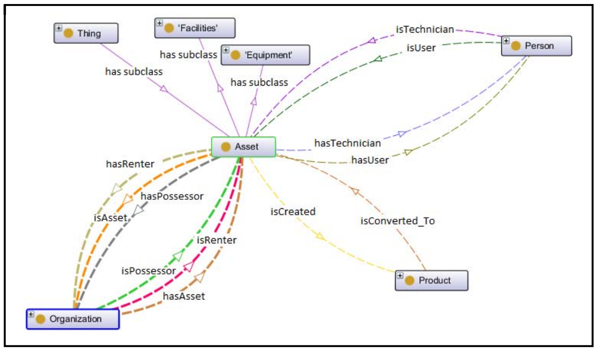


Figure 3. Asset class, subclasses, and their relationship with other classes in IRInO

experts of Web Technology Lab of Ferdowsi university[1]), in 4 previous phases cyclic. After it ontology was updated.

A small sample of IRInO classes, subclasses and their relationships is depicted in “Fig 3”. The small sample is chosen to show a better comprehension of concepts, classes, and relationships. Asset class corresponds with infrastructure class in CERIF equipment and facility classes are its subclasses.

IV. Evaluation

IRInO Ontology was evaluated with the three conceptual, structural and usability aspects. The evaluation results will be explained in follow.

A. *Conceptual Evaluation: Assessing IRInO based on opinion of experts.*

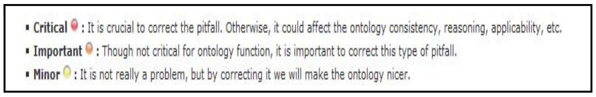
Opinion of Experts used through phases of designing ontology up to designing final version of IRInO ontology as noted in phase 5.

B. *Structural Evaluation: Assessing IRInO based on criteria of assessing ontology*

In the method, criterias such as consistency, completeness, conciseness and clarity[20] and dimensions such as Structural Dimension ،Functional Dimension and Usability-Profiling Dimension assessed using Oops! tool[21] which is An On-line Tool for Ontology Evaluation.

The tool shows defects and errors in three levels including critical, important and monitor, after entering an OWL file. The guide for significance of error is shown “Fig 4”. The method has been repeated to the time that Oops! Showed no error in evaluation of all mentioned dimensions. Oops! Showed "Congratulations!" as successful final result of evaluation after fixing all errors. Details of evaluation result of IRInO using Oops! Tool is shown in table 3 and table 4.

[1] http://wtlab.um.ac.ir



- **Critical** ● **:** It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.
- **Important** ● **:** Though not critical for ontology function, it is important to correct this type of pitfall.
- **Minor** ● **:** It is not really a problem, but by correcting it we will make the ontology nicer.

Figure 4. Guide for significance of error in Oops!

TABLE III. Evaluation Criteria of IRInO using Oops!

Classification by Evaluation Criteria	
Consistency	✓
Completeness	✓
Conciseness	✓

TABLE IV. Evaluation Dimension of IRInO using Oops!

Classification by Dimension		
Structural Dimension	Modelling Decisions	✓
	Wrong Inference	✓
	No Inference	✓
	Real World Modelling or Common Sense	✓
	Ontology language	✓
Functional Dimension	Requirements Completeness	✓
	Application context	✓
Usability-Profiling Dimension	Ontology Clarity	✓
	Ontology Understanding	✓

Since there hasn't been any ontology for the SEMAT data model until now, evaluation of the IRInO based on coverage criteria [20], has been done by considering the SEMAT data model. The evaluation results are given in table 5.

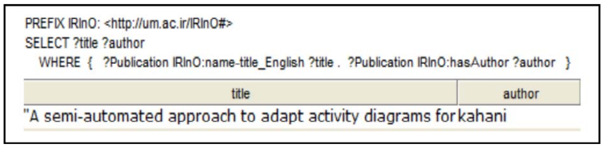
TABLE V. Evaluation of IRInO for Coverage Criteria

Dimension	***SEMAT***	***IRInO***
Number of Classes	63	267
Number of Properties	522	358

* In table 5 The number of classes in the SEMAT column points to the number of tables in SEMAT, and the number of classes in the IRInO column points to the number of Ontology entities. On the other hand, the number of properties for SEMAT and IRInO point to the number of all table fields and the total number of object and data properties, accordingly.

As shown in table 5, number of properties for IRInO is lower than SEMAT, because in protégé properties be calculated logically. Also, the axiom count for IRInO is 3317 and the logical axiom count is 1202.

The result of the evaluation shows that not only IRInO covers National research information model, but also more dimensions of research area is considered in it.



```
PREFIX IRInO: <http://um.ac.ir/IRInO#>
SELECT ?title ?author
  WHERE {  ?Publication IRInO:name-title_English ?title .  ?Publication IRInO:hasAuthor ?author  }
```

title	author
"A semi-automated approach to adapt activity diagrams for	kahani

Figure 5. A simple SPARQL query

C. Usability Evaluation: Assessing IRInO using DL Query and SPARQL

The aim of this evaluation is examination accuracy and effectiveness of IRInO. For this, first data of SAED system mapped to RDF using IRInO, then some SPARQL query and DL query were designed according to competency questions and at the end they were implemented in protégé software. Designed ontology could answer to competency questions. The result of queries show that mapping data of SAED system are extractable and usable correctly. A simple SPARQL query is shown "Fig 5".

V. Conclusion and future works

Different models of describing research data is examined in the paper. Existing challenges of National research information model (SEMAT) were extracted by comparing and evaluating these models. These challenges include ineffectiveness of described model based on characteristic rather than descriptive models, which are based on XML and RDF, not considering classes, supervising concepts, evaluation, multimedia, research locations, research activities etc.

Suggestions of the paper for promotion of national SEMAT include fixing defects, adding mentioned cases in research area, designing promoted SEMAT ontology and distribution of described model in the RDF format. Evaluation results show that the ontology can be considered as an overall schema of research information of country and can be a base for deduction and extraction of ontology in research area in a way that be suitable for research centered organizations such as universities.

Designing and implementing data warehouse on IRInO ontology is one of future works of the research.

References

[1] J. Brank, M. Grobelnic, and D. Mladenic, "A survey of Ontology evaluation techniques," in In Proceedings of the Conference on Data Mining and Data Warehouses(SiKDD 2005), Ljubljana, Slovenia, 2005.

[2] T. Hasan, A. Ramaprasad, and C. Singai, "Rethinking higher education research: Ontologymapping of higher education systems," in 37th HERDSA Annual International Conference, Hong Kong, 2014.

[3] Q. Lu, "OntoKBEval: a support tool for OWL Ontology evaluation," Unpublished master's thesis, Concordia University, Montreal, Quebec, Canada, 2006.

[4] TJ. Bright, EY. Furuya, GJ. Kuperman, JJ. Cimino, and S. Bakken, "Development and evaluation of an ontology for guiding appropriate antibiotic prescribing," *Journal of Biomedical Informatics,* vol. 45, no. 1, 2012.

[5] A. Ramaprasad, "Envisioning a world-class university system for India," *International Journal of TechnologyManagement and Sustainable Development,* vol. 10, no. 1, pp. 45-54, 2011.

[6] J. Dvořák, *CERIF 1.5 Tutorial*, Research & Development & Innovation Information System (the national CRIS for CZ), 2013.

[7] K. G. Jeffery, J. O. Lay, J. Miquel, S. Zardan, F. Naldi, and I. V. Parenti, "IDEAS: a system for international data exchange and access for science," *Science Information Processing and Management,* vol. 25, no. 6, pp. 703-711, 1989.

[8] F . Naldi, K. Jeffery, G. Bordogna, and J. Lay, "A Distributed Architecture to Provide Uniform Access to Pre-Existing Independent," *Heterogeneous Information Systems RAL Report 92-003*, 1992.

[9] S. Buswell, D. Brickley, and B. Matthews, *SWAD-Europe Deliverable 5.1: Schema Technology Survey*, 2004.

[10] A. Lopatenko, A. Asserson, UiB, and K. G. Jeffery, *CERIF - Information Retrieval of Research Information in a Distributed Heterogeneous Environment*.

[11] J. Turňa, J. Noge, and D. Zendulková, "The system SK CRIS, scientific publications and theses – mirror of Slovak science," in CRIS2012: 11th International Conference on Current Research Information Systems, 2012.

[12] J. Schöpfel, D. Zendulkova, and O. Fatemi, "Electronic theses and dissertations in CRIS," in 12th International Conference on Current Research Information Systems, Rome, Italy, 2014.

[13] M. J. Khoshroo, and O. Fatemi, "SEMAT, National Current Research Information System for IRAN," in 10th International Conference on Current Research Information Systems(CRIS), Aalborg, Denmark, 2010.

[14] National Samat Steering Committee, "information complement Style of paragraph "D" of the budget plans," Winter, 2011.

[15] M. Schuhmacher, and S. P. Ponzetto, "Knowledge-based graph document modeling," in 7th ACM international conference on Web search and data mining, 2014, pp. 543-552.

[16] C. H. Chang, M. Kayed, R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 18, no. 10, 2006.

[17] M. Fernández, A. Gómez-Pérez, and N. Juristo, "METHONTOLOGY: from Ontological Art towards Ontological Engineering," *In Proceedings of the AAAI97 Spring Symposium*, pp. 30-40, 1997.

[18] Iranian Research Institute for Information Science and Technology, "Technical document of National research information model ", Autumn, 2012.

[19] F. N. Noy, and L. D. McGuinness, *Ontology Development 101: A Guide to Creating Your First Ontology*, 2001.

[20] D. Vrandecic, "Ontology Evaluation " PhD thesis, Karlsruhe Institute of Technology, Karlsruhe, Germany, 2010.

[21] M. Poveda-Villalón, A. Gómez-Pérez, and M. Suárez-Figueroa, "OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation," *International Journal on Semantic Web and Information Systems (IJSWIS),* vol. 10, no. 2, pp. 7-34, 2014.