

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE
MONTERREY
CAMPUS ESTADO DE MÉXICO
SCHOOL OF ENGINEERING AND SCIENCES



**TECNOLÓGICO
DE MONTERREY®**

**Framework for Consistent Generation of Linked Data: the Case
of the User's Academic Profile on the Web**

A dissertation presented by

Joanna ALVARADO URIBE

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Computer Science

Advisor: Dr. Miguel GONZÁLEZ MENDOZA
Co-Advisor: Dr. Héctor Gibrán CEBALLOS CANCINO

Thesis Committee:	Dr. Jesús Favela Vara	President
	Dr. María de Lourdes Guadalupe Martínez Villaseñor	Secretary
	Dr. Hugo Estrada Esquivel	Examiner
	Dr. Miguel González Mendoza	Examiner
	Dr. Héctor Gibrán Ceballos Cancino	Examiner

Atizapán de Zaragoza, Estado de México

16th November, 2018

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE
MONTERREY
CAMPUS ESTADO DE MÉXICO
SCHOOL OF ENGINEERING AND SCIENCES

The committee members, hereby, certify that have read the dissertation presented by Joanna ALVARADO URIBE and that it is fully adequate in scope and quality as a partial requirement for the degree of Doctor of Philosophy in Computer Science, with a major in Intelligent Systems.

Dr. Miguel González Mendoza
Tecnológico de Monterrey
Principal Advisor

Dr. Héctor Gibrán Ceballos Cancino
Tecnológico de Monterrey
Co-advisor

Dr. Jesús Favela Vara
Centro de Investigación Científica y
de Educación Superior de Ensenada,
Baja California
Committee Member

Dr. María de Lourdes Guadalupe
Martínez Villaseñor
Universidad Panamericana
Committee Member

Dr. Hugo Estrada Esquivel
Centro de Investigación e Innovación
en Tecnologías de la Información y
Comunicación
Committee Member

Dr. Rubén Morales Menéndez
Dean of Graduate Studies
School of Engineering and Sciences

Atizapán de Zaragoza, Estado de México
16th November, 2018

Declaration of Authorship

I, Joanna ALVARADO URIBE, declare that this thesis titled, 'Framework for Consistent Generation of Linked Data: the Case of the User's Academic Profile on the Web' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- Some parts of this work have been published in different journals [1, 2] and conferences [3, 4].

Joanna ALVARADO URIBE
Atizapán de Zaragoza, Estado de México
16th November, 2018

©2018 by Joanna ALVARADO URIBE
All rights reserved

“One, remember to look up at the stars and not down at your feet. Two, never give up work. Work gives you meaning and purpose and life is empty without it. Three, if you are lucky enough to find love, remember it is there and don’t throw it away.”

Stephen Hawking

To God, my parents, brother, and family

Acknowledgements

Thank you, God, for listening to my prayers and for your blessings. I want to thank my family for all the love, support, and encouragement. Thanks to my parents María Ángela Uribe and Juan Alvarado for everything, for always being the greatest parents ever. To my brother Dilan Alvarado for brightening my life, you are the best! To my grandmother Juana Sánchez for always being aware of me. To my aunt Valentina Uribe, my uncles Alberto Alvarado and Rosalío Hernández, and my cousins Iván Hernández, Gardenia Hernández, Leonardo Alvarado, Omar Uribe, and Reyna Trejo thank you for all the support, understanding, and love. I also want to thank all members of the Alvarado family and the Uribe family for being there.

I want to thank my advisor, Dr. Miguel González, and my co-advisor, Dr. Héctor Ceballos, for all the patience, advice, time, comments, discussions, corrections, guidance, and effort that you invested in my academic training along these years. Dr. Miguel González, thank you for the confidence placed in me to collaborate in a professional way with you. I could not have asked for or had better mentors.

I want to thank Dr. Jesús Favela, Dra. María de Lourdes Guadalupe Martínez, and Dr. Hugo Estrada for all their time, experience, and accurate feedback that contributed to improve and delimit the present work. I also want to thank Dr. Raúl Monroy, Dr. Oscar Herrera, Dr. Noé Castro, and Dra. Fany Eisenberg for all their support, guidance, comments, and assistance. Their experience and motivation clarified and expanded my research work.

I want to thank all the Ph.D. and Master friends and fellows for all the recreational and academic time we spent together, as well as for their words of encouragement during difficult times. Thanks for everything Yair Barrera, Cesar Montiel, Andrei Tapia, Mauricio Martínez, Víctor Ferman, Héctor Sánchez, Marcelo Estrella, Miguel Zabala, Andrés Salas, Benito Camiña, and Nisim Hurst.

I sincerely thank both Tecnológico de Monterrey and Consejo Nacional de Ciencia y Tecnología (CONACYT) for the support provided to accomplish this Degree through the scholarship no. 298597. I also want to thank ©2016 HOP Ubiquitous S.L. (HOPU), Universidad Católica San Antonio de Murcia (UCAM), and the town council and tourism office representatives from Ceutí for the reception, attention, and support given during my research fellowship supported by the SmartSDK project (SmartSDK project is co-funded by the EU's Horizon2020 programme under agreement number 723174 - ©2016 EC and by CONACYT agreement 737373). SmartSDK project thanks for the support.

Framework for Consistent Generation of Linked Data: the Case of the User's Academic Profile on the Web

by

Joanna ALVARADO URIBE

Abstract

Decision management is relevant for high-value decisions that involve multiple types of input data. Since the Web allows users to keep in touch with other users and likewise, share their data (such as features, interests, and preferences) with applications and devices to customize a provided service, the online data related to these users can be collected as input data for a decision-making process. However, these data are usually provided to the application or device used in a given time, causing three major issues: data are isolated when are provided to a specific entity, data are scattered in the network, and data are found in different formats (structured, semi-structured, and unstructured). Therefore, with the aim of supporting decision makers to make better decisions, in a certain scenario, the proposal to automatically unify, align, and integrate the user data concerning this scope into a centralized and standardized structure that allows, at the same time, to model the user's profile on the Web in a consistent and updated manner as well as to generate linked data from the integrated information is addressed. This is where Decision Support Systems, Semantic Web, and context-enriched services become the cornerstones of the computational approach proposed as a solution to these issues. Firstly, given the generality of fields that can constitute a user profile, the definition of a scope that allows validating the proposed approach is emphasized for this research work. Secondly, the proposal, development, and evaluation of the computational solutions that allow dealing with the data modeling, integration, generation, and updating consistently are highlighted in this research. Therefore, a study focused on the academic area is proposed for this work in order to support researchers and data managers at the institutional level in processes and activities concerning this area, specifically at Tecnológico de Monterrey. To achieve this goal, the design of an interdisciplinary, justified, and interoperable meta-schema (called Academic SUP) that allows to model the user's academic profile on the Web, as well as the development of a computational framework (named as AkCeL) that allows to integrate, generate, and update data into such a meta-schema consistently are proposed in this research work. In addition, in order to support researchers in their decision-making processes, the development of a recommendation algorithm (called C-HyRA) that allows providing a research areas list interesting for researchers, as well as the adoption of a visualization platform related to the academic area to present the information generated by AkCeL are put forward in this proposal. As a result, unified, consistent,

reliable, and updated information of the researcher' academic profile is provided on the Web from this approach, in both text and graphics, through the VIVO platform to be consumed primarily by researchers and educational institutions to support their networks and statistics of collaboration/publication and research.

List of Figures

2.1	Graphic: research works addressed as state-of-the-art and related work in this proposal.	34
3.1	Schematic view of AkCeL.	37
3.2	Graphic representation of the schema matching process.	39
3.3	Graphic representation of the data retrieval process.	50
3.4	Graphic representation of the data recommendation process.	61
4.1	Responders' residence profile.	89

List of Tables

2.1	Comparison of the POI Recommendation Approaches	22
2.2	Comparison of the Ontological Approaches	33
3.1	Properties provided by each information source (Part I).	42
3.2	Properties provided by each information source (Part II).	43
3.3	First version of Academic SUP.	44
3.4	Second version of Academic SUP.	44
3.5	Schema matching between raw properties and the ontologies' classes and properties.	46
3.6	Definition of the Academic SUP's subclasses and properties.	48
3.7	Academic SUP's classes and properties	49
3.8	Statistics of the dataset downloaded from dblp (retrieved on 2018-04-12).	51
3.9	Statistics of the dataset downloaded from the service linked to Scopus (retrieved from 2018-03-21 to 2018-03-23).	53
4.1	Number of surveys assigned to each cluster.	80
4.2	Smart POIs dataset.	85
4.3	User preferences dataset: example of structure for each user.	86
4.4	Smart POIs' categories dataset.	86
4.5	User preferences identified by User-CEUTI-1.	91
4.6	Recommendations given by the algorithm to the responder identified by User-CEUTI-1 (I).	91
4.7	Recommendations given by the algorithm to the responder identified by User-CEUTI-1 (II).	92
4.8	Comparison between the recommendations given by the user-based CF against the user-based CF with an average aggregation operator that integrates nine similarity measures.	93
4.9	MSE of each of the nine similarity and distance measures concentrated in the user-based CF algorithm and of user-based CF with an average aggregation operator that integrates the nine measures.	94
4.10	MSE of different combinations of the similarity and distance measures for the user-based CF with an average aggregation operator.	94
4.11	Comparison between the recommendations given by the user-based CF algorithm against the user-based CF with an average aggregation operator that integrates five similarity and distance measures.	95
4.12	MSE of the nine similarity and distance measures concentrated in the user-based CF algorithm and of user-based CF with an average aggregation operator that integrates five similarity and distance measures.	95

4.13	Comparison between the recommendations given by the user-based CF algorithm with an average aggregation operator against the user-based CF with an average aggregation operator and the Smart POIs' categories.	96
4.14	Comparison between the user-based CF with an average aggregation operator + the Smart POIs' categories and HyRA.	97
4.15	Comparison between the user-based CF with an average aggregation operator and HyRA.	97
4.16	Comparison between the UG and HyRA recommendation algorithms.	98
5.1	Statistics of authors for the alignment process.	105
5.2	Comparison of the alignment results between Academic SUP and SIIP.	106
5.3	Alignment results for the researcher identified as Researcher-8.	107
5.4	Alignment results for the researcher identified as Researcher-21.	107
5.5	Comparison of the integration results between Academic SUP and SIIP without data preprocessing and with a modified version of the proposed data alignment.	111
5.6	Concentration of the results of the proposed scenario and the baseline scenario for the data preprocessing and alignment processes described in this work.	113
5.7	The user's explicit preferences dataset: an example of the structure for each user.	116
5.8	The ITESM's Campuses geographical information dataset (Part I)	123
5.9	The ITESM's Campuses geographical information dataset (Part II)	124
5.10	Summary of users with C-HyRA's recommendations through 100 executions.	126
5.11	Results of the comparison between the recommendations given by C-HyRA and the research areas concentrated in the pilot subset.	126
B.1	Subject categorization for the research area "Agricultural and Biological Sciences" proposed by Scopus	138
B.2	Subject categorization for the research area "Arts and Humanities" proposed by Scopus	139
B.3	Subject categorization for the research area "Biochemistry, Genetics and Molecular Biology" proposed by Scopus	139
B.4	Subject categorization for the research area "Business, Management and Accounting" proposed by Scopus	140
B.5	Subject categorization for the research area "Chemical Engineering" proposed by Scopus	140
B.6	Subject categorization for the research area "Chemistry" proposed by Scopus	140
B.7	Subject categorization for the research area "Computer Science" proposed by Scopus	141
B.8	Subject categorization for the research area "Decision Sciences" proposed by Scopus	141
B.9	Subject categorization for the research area "Earth and Planetary Sciences" proposed by Scopus	141
B.10	Subject categorization for the research area "Economics, Econometrics and Finance" proposed by Scopus	141
B.11	Subject categorization for the research area "Energy" proposed by Scopus	142
B.12	Subject categorization for the research area "Engineering" proposed by Scopus	142
B.13	Subject categorization for the research area "Environmental Science" proposed by Scopus	142
B.14	Subject categorization for the research area "Immunology and Microbiology" proposed by Scopus	143

B.15 Subject categorization for the research area “Materials Science” proposed by Scopus	143
B.16 Subject categorization for the research area “Mathematics” proposed by Scopus	143
B.17 Subject categorization for the research area “Medicine” proposed by Scopus	144
B.18 Subject categorization for the research area “Neuroscience” proposed by Scopus	145
B.19 Subject categorization for the research area “Nursing” proposed by Scopus	145
B.20 Subject categorization for the research area “Pharmacology, Toxicology and Pharmaceutics” proposed by Scopus	146
B.21 Subject categorization for the research area “Physics and Astronomy” proposed by Scopus	146
B.22 Subject categorization for the research area “Psychology” proposed by Scopus	146
B.23 Subject categorization for the research area “Social Sciences” proposed by Scopus	147
B.24 Subject categorization for the research area “Veterinary” proposed by Scopus	147
B.25 Subject categorization for the research area “Dentistry” proposed by Scopus	147
B.26 Subject categorization for the research area “Health Professions” proposed by Scopus	148
B.27 Subject categorization for the research area “Arts & Humanities” proposed by Web of Science	149
B.28 Subject categorization for the research area “Life Sciences & Biomedicine” proposed by Web of Science (Part I)	150
B.29 Subject categorization for the research area “Life Sciences & Biomedicine” proposed by Web of Science (Part II)	151
B.30 Subject categorization for the research area “Physical Sciences” proposed by Web of Science	152
B.31 Subject categorization for the research area “Social Sciences” proposed by Web of Science	152
B.32 Subject categorization for the research area “Technology” proposed by Web of Science	153
B.33 Subject categorization for the research area “Agricultural and Biological Sciences” proposed for C-HyRA	154
B.34 Subject categorization for the research area “Arts and Humanities” proposed for C-HyRA	155
B.35 Subject categorization for the research area “Biochemistry, Genetics and Molecular Biology” proposed for C-HyRA	156
B.36 Subject categorization for the research area “Business, Management and Accounting” proposed for C-HyRA	157
B.37 Subject categorization for the research area “Chemical Engineering” proposed for C-HyRA	157
B.38 Subject categorization for the research area “Chemistry” proposed for C-HyRA	158
B.39 Subject categorization for the research area “Computer Science” proposed for C-HyRA	159
B.40 Subject categorization for the research area “Decision Sciences” proposed for C-HyRA	160
B.41 Subject categorization for the research area “Earth and Planetary Sciences” proposed for C-HyRA	160
B.42 Subject categorization for the research area “Economics, Econometrics and Finance” proposed for C-HyRA	161
B.43 Subject categorization for the research area “Energy” proposed for C-HyRA	161
B.44 Subject categorization for the research area “Engineering” proposed for C-HyRA	162

B.45 Subject categorization for the research area “Environmental Science” proposed for C-HyRA	163
B.46 Subject categorization for the research area “Immunology and Microbiology” proposed for C-HyRA	164
B.47 Subject categorization for the research area “Materials Science” proposed for C-HyRA	164
B.48 Subject categorization for the research area “Mathematics” proposed for C-HyRA	165
B.49 Subject categorization for the research area “Medicine” proposed for C-HyRA (Part I)	166
B.50 Subject categorization for the research area “Medicine” proposed for C-HyRA (Part II)	167
B.51 Subject categorization for the research area “Medicine” proposed for C-HyRA (Part III)	168
B.52 Subject categorization for the research area “Neuroscience” proposed for C-HyRA	169
B.53 Subject categorization for the research area “Nursing” proposed for C-HyRA . .	170
B.54 Subject categorization for the research area “Pharmacology, Toxicology and Pharmaceutics” proposed for C-HyRA	171
B.55 Subject categorization for the research area “Physics and Astronomy” proposed for C-HyRA	171
B.56 Subject categorization for the research area “Psychology” proposed for C-HyRA	172
B.57 Subject categorization for the research area “Social Sciences” proposed for C-HyRA	173
B.58 Subject categorization for the research area “Veterinary” proposed for C-HyRA .	174
B.59 Subject categorization for the research area “Dentistry” proposed for C-HyRA .	174
B.60 Subject categorization for the research area “Health Professions” proposed for C-HyRA	175

Contents

Abstract	v
List of Figures	vii
List of Tables	viii
Abbreviations and Acronyms	xv
1 Introduction	1
1.1 Motivation	3
1.2 Justification	4
1.3 Problem Statement and Context	5
1.4 Research Questions	7
1.5 Hypothesis	7
1.6 Contributions	8
1.7 Dissertation Organization	8
2 Related Work	10
2.1 Framework	10
2.1.1 Architectures	11
2.1.2 Systems Addressing Event Processing and Stream Reasoning	13
2.1.2.1 C-SPARQL on S4	14
2.1.2.2 EP-SPARQL	15
2.1.2.3 Answer Set Programming (ASP)	17
2.1.3 Semantic Approaches incorporating Similarity and Relatedness Measures	18
2.2 Traditional and POI Recommendation Algorithms	18
2.2.1 Comparative: POI Recommendation Algorithms	20
2.3 User Profiles and Ontologies	22
2.3.1 User Profile from the Human Resources Approach	23
2.3.2 User Profile from the Computer Science Approach	23
2.3.2.1 Approaches considering Dynamic and Contextual Information	23
2.3.2.2 Approaches considering Academic Resources	29
2.3.3 Comparative: Ontological User Profiles	32
2.4 Summary	34

3	Proposed Framework for Consistent Generation of Linked Data Focused on the User's Academic Profile	35
3.1	A General View of AkCeL	36
3.2	User's Academic Profile Meta-Schema	38
3.2.1	Information Sources Definition and Schema	40
3.2.2	Schema Matching: Ontologies	44
3.2.3	Academic SUP (Meta-Schema User's Profile)	45
3.3	Dynamic Data Enrichment	49
3.3.1	Data Retrieval	49
3.3.1.1	Information Sources Definition: Data	50
3.3.1.1.1	dblp	51
3.3.1.1.2	Scopus	52
3.3.1.2	Data Preprocessing	53
3.3.2	Data Alignment	54
3.4	Data Recommendation	60
3.4.1	User-Based CF: Analysis and Description	62
3.4.2	User-Based CF with the Average Aggregation Operator	64
3.4.3	User-Based CF with the Average Aggregation Operator + Publications' Categories	64
3.4.4	C-HyRA	65
3.5	Data Visualization	70
3.6	Conclusions	71
4	Use Case of HyRA: A Hybrid Recommendation Algorithm Focused on Smart POI. Ceutí as a Study Scenario	73
4.1	The HyRA's Approach	75
4.1.1	User-Based CF: Analysis and Description	75
4.1.2	User-Based CF with the Average Aggregation Operator	77
4.1.3	User-Based CF with the Average Aggregation Operator + Smart POIs' Categories	78
4.1.4	HyRA	78
4.2	Experimental Scenario Based on Surveys	79
4.2.1	Project Background	79
4.2.2	The Surveys and the HyRA Evaluation Scenario	80
4.3	Data Requirements for HyRA	83
4.3.1	Smart POIs Dataset in Ceutí	83
4.3.2	User Preferences Dataset	83
4.3.3	Smart POIs' Categories Dataset	84
4.4	Results and Discussion	87
4.4.1	Surveys: Dissemination and Analysis	87
4.4.2	HyRA: Analysis and Discussion	90
4.4.3	Smart POI Recommendation through User-Based CF with an Average Aggregation Operator	93
4.4.4	Smart POI Recommendation through User-Based CF with an Average Aggregation Operator + Smart POIs' Categories	96

4.4.5	Smart POI Recommendation through Geographical Influence + User-Based CF with an Average Aggregation Operator + the Smart POIs' Categories (HyRA)	96
4.5	Conclusions	98
5	Use Case: Generation of Research Recommendations and Statistics based on a Unified, Updated, and Consistent User's Academic Profile	100
5.1	Scope of the Use Case	101
5.1.1	The Data Alignment Process' Evaluation Scenario	102
5.1.2	The C-HyRA's Evaluation Scenario	103
5.1.3	The Data Visualization's Deployment Scenario	104
5.2	Results and Discussion	104
5.2.1	Data Alignment Process	105
5.2.2	Baseline Scenario: Data Integration without Data Preprocessing and with a modified version of the Proposed Data Alignment	108
5.2.3	Data Requirements for C-HyRA: the User's Explicit Preferences, Research Areas (Categories/Classification), and Geographical Information	115
5.2.3.1	The User's Explicit Preferences about Research Areas	115
5.2.3.2	Research Areas: Categories/Classification	116
5.2.3.2.1	Scopus	117
5.2.3.2.2	Web of Science	118
5.2.3.2.3	The Unified Publications' Areas and Categories Dataset	119
5.2.3.3	Geographical Information	121
5.2.4	Recommendations based on C-HyRA	125
5.2.5	Front-End: Data Visualization	126
6	Conclusions and Further Work	128
6.1	Future Work	133
A	Definitions	134
A.1	Data Retrieval	134
A.2	Data Preprocessing	134
A.3	Data Recommendation	135
A.4	User Profile and Ontologies	136
B	Research Areas: Categories/Classification	138
B.1	Scopus	138
B.2	Web of Science	149
B.3	The Unified Publications' Areas and Categories Dataset	154
	Bibliography	176

Abbreviations and Acronyms

Academic SUP	meta-Schema User's Profile
AkCeL	frAmework for Consistent generation of Linked data
AMIPCI	<i>Asociación Mexicana de Internet</i>
API	Application Programming Interface
ASP	Answer Set Programming
CF	Collaborative Filtering
C-HyRA	researCh Hybrid Recommendation Algorithm
CRF	Conditional Random Fields
CRTCF	Cross-Region Topic-based Collaborative Filtering
C-SPARQL	Continuous SPARQL
CWA	Closed World Assumption
dblp	DataBase systems and Logic Programming
dc	Dublin Core
DoD	Department of Defense
DSS	Decision Support Systems
EP-SPARQL	Event Processing SPARQL
FOAF	Friend of a Friend
Geo-PFM	Geographical Probabilistic Factor Model
GPS	Global Positioning System
GTAG	Geographical-Temporal influences Aware Graph
GT-BNMF	Geographical-Topical Bayesian Non-negative Matrix Factorization
HR	Human Resources
HSWS	Hybrid Semantic Web Search
HTML	HyperText Markup Language
HyRA	Hybrid Recommendation Algorithm

IoT	Internet of Things
IRI	Internationalized Resource Identifier
IT	Information Technology
JDL	Joint Directors of Laboratories
LBSN	Location-Based Social Network
LTA	Long-Term Activation
LTSCR	Location and Time aware Social Collaborative Retrieval model
MIBO	Middle Building Ontology
MSDF	Multisensor Data Fusion
MSE	Mean Squared Error
MTA	Medium-Term Activation
NLP	Natural Language Processing
OA	Open Access
OAI-PMH	Open Archives Initiative - the Protocol for Metadata Harvesting
ONCOR	Ontology- and Evidence-based Context Reasoner
ORCID	Open Researcher and Contributor Identifier
OS	Operating System
OWA	Open World Assumption
OWL	Web Ontology Language
PECO	Personalised Context Ontology
PERSONAF	PERsonalised PERvasive Scrutable ONtologicAl Framework
PIM	Personal Information Management
PIMS	Personal Interaction Management System
POI	Point-Of-Interest
QoL	Quality of Life
RDB	Relational Database
RDF	Resource Description Framework
REDALYC	<i>Red de Revistas Científicas de América Latina y el Caribe, España y Portugal</i>
REST	Representational State Transfer
RUDAR	Roskilde University Digital Archive
SAUPO	Situation-Aware User Profile Ontology
SIIP	<i>Sistema de Información para la Investigación y el Posgrado</i>

SIT	School of Information Technologies
SKOS	Simple Knowledge Organization System or Simple Knowledge Organization for the Web
Smart POI	Smart Point of Interaction
SNS	Social Networking Sites
SNSs	Social Network Sites
SPICE	Service Platform for Innovative Communication Environment
SQL	Structured Query Language
STA	Short-Term Activation
SW	Semantic Web
TIM	Task Information Management
u2m	Ubiquitous User Model
U2MIO	Ubiquitous User Modeling Interoperability Ontology
UG	User preference/Geographical influence based recommendation
UML	Unified Modeling Language
UPOI-Mine	Urban POI-Mine
UPOS	User-Profile Ontology with Situation-Dependent Preferences Support
vCard	Electronic Business Cards
VISIT	Virtual Intelligent System for Informing Tourists
W3C	World Wide Web Consortium
WBPR-FD	Weighted Bayesian Personalized Ranking model with visit Frequency and Distance
XML	Extensible Markup Language

Chapter 1

Introduction

Decisions are determinations of courses of action [5]. A “good decision making” means a person (user) is informed and has relevant and appropriate information on which to base his/her choices among alternatives [6]. Therefore, a decision management is relevant for high-value decisions that involve multiple kinds of input data or where decision logic is frequently modified since this discipline is focused on designing, building, and maintaining systems that make structured decisions. Thereby, decision management can be applied to decision automation and decision support scenarios [5]. Since a decision support scenario is raised for this proposal, Decision Support Systems (DSS) will be studied.

Another discipline that will be addressed in this research work because its conception and vision is Semantic Web (SW). SW has been driven by its applicability to big data analysis [7]. The explosion of Social Networking Sites (SNS) and social tagging is presented as a major advance in the vision of SW, allowing through the semantic tagging the building of a “social graph”. Nevertheless, social graphs are in closed and controlled platforms, representing a challenge for the development of an open and interoperable SW [8, 9].

In the same way, the context-enriched services are relevant in this research given the application of it. These services have been adopted by both independent software vendors and end-user Information Technology (IT) departments, and will be driven by the Internet of Things (IoT) as a source of contextual attributes [9]. Contextual awareness has become a significant competitive differentiator for user engagement since the context-enriched services provide enriched, situation-aware, targeted, personalized, and relevant content, functions, and experiences by combining demographic, psychographic, historical, and environmental data with other information [9].

In addition, notions presented in other areas such as Multisensor Data Fusion (MSDF), Stream Reasoning, Data Mining, and Text Processing are relevant for this proposal due to their key involvement in the development of the proposed framework. These contributions are summarized below:

1. **Big Data.** In order to extract key information from massive volumes of a wide variety of types of data, allowing high-velocity acquire, discovery, and/or analysis such data [10], this term is embraced to this proposal.
2. **Multisensor Data Fusion.** This technology allows to combine information from various sources with the aim of obtaining a unified data [11, 12]. Such combination or fusion is defined through architectures and models designed for this area. Hence, the proposed framework for generation of data, in this work, is based on the centralized data fusion architecture [13].
3. **Stream Reasoning.** This research area unifies stream processing and reasoning over streaming data in order to process data streams (generated by heterogeneous or homogeneous sources), to produce the results requested by users, and at the same time, to provide ways of inferring implicit knowledge [14]. In accordance with the foregoing, this field is essential to determine the system, techniques, or algorithms that will be used to process the information gathered by the proposed framework as well as to carry out the discovery of new data.
4. **Data Mining.** This field is focused on the discovery of interesting, unexpected, or valuable structures in large datasets [15]. Therefore, in this proposal, it takes relevance in the processes incorporated into the proposed framework.
5. **Text Processing.** This discipline is developed to carry out a systematic treatment of all the ways of producing a text based on an existing text, trying to retain the meaning of that text [16]. In this work, it is used to treat the information obtained from the different information sources addressed in this proposal.

1.1 Motivation

The computerized DSS possess capabilities that can facilitate decision support in several ways, such as speedy computations; improved communication, collaboration, and data management; increased productivity of group members; quality, agile, and ubiquitous support; overcoming cognitive limits in processing and storing information; using the Web; among others [17]. Therefore, in accordance with these capabilities and the goals of this research work, the study of the computerized DSS is considered as a basis to support the design and building of the proposed framework.

Similarly, this proposal is based on the definition of SW since it allows adding semantic information to unstructured information to transform it into machine-understandable and linking the unstructured information to formalized concept definitions by using web technologies [9]. In the same way, this research is encouraged under the context of the original vision of SW. That is, the vision of providing more meaning to Web information through the logical connection of terms in order to establish interoperability between systems [18]. Since, on the one hand, SW is part of the Open World Assumption (OWA), also known as the Classical Paradigm [19], stipulating that there may be unspecified information (considered as unknown) that can be inferred [19]. On the other hand, SW has been incorporated into Big Data applications [7], allowing the user to exploit large amounts of information in a given context. Hence, this work starts from these facts with the aim of conceptualizing a possible solution to model data of the users' profile, using information available on the Web as well as in private repositories.

Besides, it is important to highlight that the inference is a centerpiece of SW because therethrough new relationships can be discovered between resources [8, 19, 20]. Therefore, the motivation of using standards and languages engineered for SW, such as Resource Description Framework (RDF) and Web Ontology Language (OWL), increases due to they represent the basis to support various types of inference and to enhance the Web's interoperability [18]. In addition to contributing "to liberate data from silos that are framed by proprietary database schemas", main objective of Linked Data, [21]. Thus, in this work, these principles are considered to model the user's profile and to generate linked data belonging to such users.

Finally, the decision to use as an information source to SNS is encouraged on a brief note published by Springer [22]. This note highlights the importance of being a user of SNSs by mentioning three key benefits that provide the user when he/she is part of them, such as to:

1) connect with affine people and communities, 2) promote work, business, articles, and so on, and 3) support on building of his/her “online image”. Moreover, it also expresses that within web crawlers the sites with higher ranking are SNS. As a result, search engines will show results mainly focused on these sites rather than business or academic sites. In the same way, the *Asociación Mexicana de Internet* (AMIPCI) indicates that 9 out of 10 Internet users access to some social network site in Mexico [23]. These facts also support the election of using SNSs as information sources and of employing web crawlers for collecting data.

1.2 Justification

This research is based on the four-step process that managers usually follow to make decisions: (1) define the problem, *i.e.*, a decision situation that may deal with some difficulty or with an opportunity; (2) construct a model that describes the real-world decision situation; (3) identify possible solutions to the modeled decision situation and evaluate the solutions; and (4) compare, choose, and recommend a potential solution to the decision situation [17]. This in order to conceptualize a framework that allows generating linked data, recommendations, and statistics for researchers affiliated with a certain educational institution as well as providing publication statistics for institutions with the purpose of supporting both researchers and institutions with their decision-making processes related to the academic area.

In addition, another of the goals of this research proposal is the user profile modeling on the Web. On the one hand, the idea of building a user profile model is supported in the envisioning of some researchers, belonging to the user modeling community, that propose to use and to share the user models’ information among applications. By using and sharing this information, the preferences, interests, and characteristics of the user could be integrated into the context of applications in order to enhance the service provided [24]. On the other hand, this goal is based on the conception that “the Web is the enabler of the new era of digitalization”. This idea is supported by the initial sentence of a Gartner analysis [7]. Besides, such fact is justified by two reasons: 1) the Web provides the connective that links up users to contents, applications with another applications, and devices to data; and 2) the Web boosts novel apps, websites, portals, and mobile apps (backbone for digital business). This is possible thanks to different tools, such as browsers, hybrid mobile apps, or RESTful APIs [7].

However, it is necessary to adopt technologies designed for SW in order to allow the knowledge inference (discovery) as stated in OWA. Aspect that, for example, in a Relational Database (RDB) could not be afforded due to RDB works under the Closed World Assumption (CWA) [19]. Also for these reasons, user profile modeling is developed through an ontology since, at the same time, it provides common conceptualizations for data integration [18] and represents one of the two major approaches used to address the lack of interoperability in the user modeling [24].

Moreover, the impact of the proposed framework is justified by noting that it can be adopted easily in other areas and for other schemes when the main objective is to carry out a dependable data integration. This is considered due to the proposed framework is also based on the centralized architecture designed for the Multisensor Data Fusion area [13], a field widely approached by a great variety of scientific, engineering, management, among other subjects [11].

Finally, for the proposed recommendation algorithm approach, it is noteworthy that the traditional recommendation systems have been widely addressed, where the user generally provides ratings to the items, such as books, movies, music, among others [25]. However, the point-of-interest (POI) recommendation systems have just emerged recently [25] as a consequence of the quick development of new location-based technologies. Thus, the study of both types of approaches is considered as an area of opportunity for the scope of this proposal.

1.3 Problem Statement and Context

Decision making is a process of choosing among two or more courses of action for the purpose of attaining one or more goals [17]. Such a process involves several characteristics, of which the following is relevant to this proposal: collecting information and analyzing a problem, which can take time and can be expensive. On the one hand, there may not be sufficient information to make an intelligent decision or on the other hand, too much information may be available (*i.e.*, information overload). Hence, in order to support decision makers for making better decisions, the process and the important issues involved in decision making must be understood. Thereby, appropriate methodologies for assisting decision makers can be proposed since the contributions that information systems can produce are, in the same way, visualized [17]. In accordance with the foregoing, a study related to the academic area is proposed for this work in order to support researchers and data managers at the institutional level in aspects concerning this area.

The figures reported by AMIPCI [23] represent a good indicator of inclusion and growth in the Internet usage by Mexican people, which reflects not only the situation in Mexico but also elsewhere in the World. This growth is mainly driven by the fact that the Web allows *users* to keep in touch with other *users* and likewise, share their data (features, interests, and preferences) with applications and devices to customize the service provided. However, these data are usually provided to the application or device used at any given time, causing three major issues [24]:

1. Data are isolated when are provided to a specific entity. That is, data silos are generated because the data are hidden within walled spaces [26].
2. Data are scattered in the network.
3. Data are found in different formats (structured, semi-structured, and unstructured).

As a consequence, there is a marked need to unify, to align, and to integrate these data into a centralized structure that models the *user's* academic profile on the Web in a consistent, up-to-date, and timely manner.

In addition, in the new era of digitalization, it is fundamental to count on browsers, hybrid mobile apps, and RESTful APIs [7] that allow *end-users* to find the searched information in an accurate, quick, and simple way. However, search engines index tens to hundreds of millions of web pages producing a significant amount of different terms [27], entailing a key question: Which of these hundreds of millions of indexed web pages containing accurately what the *user* expects? [28]. The answer to this question depicts an issue for this proposal because to perform the population of the data to the *user's* profile, the data collection coming from both private and public information sources, such as Relational Databases, Web services, applications, and files in a non-proprietary format using open standards from W3C, is proposed.

Thus, according to the above formulation and taking into account that the *end-user* is the main customer of this approach, there are three main issues that are intended to solve in this proposal:

1. **To define the *user's* academic profile schema on the Web.** It is necessary to design a meta-schema that provides a unified and standardized *user's* academic profile from other research schemata and modeling principles that are defined in both the Computer Science field and Human Resources area.

2. **To establish how data are retrieved from the information sources and integrated into the *user*'s academic profile meta-schema dynamically and consistently.** It is necessary to develop a framework that considers different data formats and information sources as well as distinct update rates, and in some cases, redundancy or lack of information, *i.e.*, the same data can be provided in several information sources or cannot be provided by the *user* concerned.
3. **To define the computational solutions for supporting the decision-making processes.** It is necessary to develop a recommendation algorithm that involves both *user* preferences and contextual data about the *user* and his/her publications, as well as adopt a proposed solution focused on the data visualization related to the academic area.

1.4 Research Questions

- I What structure and features must the user's profile meta-schema, that allows modeling representative information of his/her academic field in a unified, standardized, online, and interoperable format, have?
- II How can a framework, that allows integrating, generating, and updating the data of a researcher into a defined schema from information published in different academic information sources consistently way, be built?
- III How can a recommendation algorithm, that allows to embrace the researcher preferences and the contextual information related to the research area, be built?

1.5 Hypothesis

The development of a framework based on DSS and the information integration centralized architecture designed for the MSDF field provides a consistent approach to integrate, generate, and up-to-date data into the user's academic profile meta-schema, which by being modeled through an ontology constituted of features related to user profiling presented in both the Human Resources area and Computer Science contributions, gives an interdisciplinary and justified identification of researchers in addition to a standardized, unified, updated, reliable,

online, and interoperable academic profile that allows researchers and educational institutions to support their decision-making processes related to the academic area.

1.6 Contributions

1. Academic SUP (meta-Schema User's Profile): an interoperable user's profile meta-schema able to model representative information about the academic field of a user on the Web.
2. AkCeL (frAmework for Consistent generation of Linked data): a framework able to generate and integrate data into the user's academic profile meta-schema, as well as able to update this information in a consistent manner.
3. C-HyRA (researCh Hybrid Recommendation Algorithm): a hybrid recommendation algorithm able to suggest a research areas list of interest for a certain researcher.

1.7 Dissertation Organization

This manuscript is organized as follows. In Chapter 2, the basis for this proposal is provided, such rationale is classified according to the three principal contributions from this work: framework, recommendation algorithm, and user profile (ontologies). In Chapter 3, the research and development proposal of the framework to generate data consistently, called AkCeL, is described, such a description is divided into four components: static schema, dynamic data enrichment, data recommendation, and visualization. For the static schema section, the proposal and building of an ontology-based schema focused on the user's academic profile, named as Academic SUP, are presented. Subsequently, the data retrieval, preprocessing, and alignment processes constituting the dynamic data enrichment component of AkCeL are described, from the selection of information sources to the description of aligned information. Later, the proposed data recommendation approach and algorithm, called C-HyRA, are explained, in the same way, the pseudocodes of C-HyRA are provided. Finally, the proposed user interface and visualization for the use case of this research are presented. In Chapter 4, the use case to assess only the proposed recommendation algorithm in the tourism sector, named as HyRA, is described, from the considered approach to the results and discussion of its evaluation. In Chapter 5, the scope, the experimental scenarios concerning the data alignment and recommendation processes, and

the deployment scenario related to visualization of this research are introduced. Then, the results obtained for the proposed use case and discussion about these results are given. Finally, in Chapter 6, the general conclusions of this research proposal, as well as the proposed future work, are presented.

Chapter 2

Related Work

In the literature, two underlying paradigms in SW are mentioned. The first paradigm is known as the Classical Paradigm and the second is the Datalog Paradigm. The Classical Paradigm is based on the open environment that allows the inference between resources and the use of ontologies. While Datalog Paradigm is based on the Closed World Assumption of the Relational Databases with some improvements, including others logical basis. Of these, the paradigm recommended by Patel-Schneider and Horrocks [19] to found SW is the Classical Paradigm, such a paradigm matches the proposed approach in this research. In the same way, the Linked Data area is considered in this proposal because this area presents a set of best practices related to the SW architecture in order to publish, to share, and to interlink structured data on the Web [21].

In this Chapter, a review of the state-of-the-art and related work is provided with the aim of defining, structuring, comparing, and positioning this research proposal. Such a review is presented in accordance with the three major fields of study in this work: computational framework, recommendation algorithm, and user profile modeling. To conclude, a graphic that summarizes all the research addressed by each major field is included.

2.1 Framework

The research works presented in this section constitute the theoretical basis of this proposal focused mainly on architectures and systems related to the Decision Support Systems, Multi-sensor Data Fusion, and Semantic Web areas. Similarly, systems addressing event processing

and stream reasoning, as well as semantic approaches incorporating similarity and relatedness measures are introduced below.

2.1.1 Architectures

Throughout this document, specific concepts related to Decision Support Systems (DSS) and Multisensor Data Fusion (MSDF) are mentioned. However, in this section, a general description of each area is introduced below.

On the one hand, a decision making can be supported on existing, historical data or information collected from different information sources, where such an information can come in the form of facts, numbers, graphics, among others. Therefore, since these data are collected from several sources and possibly with distinct formats, these data must be joined together and organized. The process of organizing and examining the available information is known as the modeling process. The objective of these models is to help decision makers understand the consequences of selecting an option. For example, models can be used to help better understand what customers need from certain processes to improve customer relationship management [6].

In the same way, the adequacy of the available information, the quality of the information, the number of options, and the appropriateness of the modeling effort available at the time of the decision are factors that determine the quality of the decision. Hence, the use of decision support systems (DSS) represents one way to accomplish the goal of bringing together the appropriate information and models for informed decision-making [6]. DSS are interactive computer-based systems that help decision makers use data and models to deal with unstructured and semi-structured problems by allowing them to bring together information from different sources, assist in the organization and analysis of information, and facilitate the evaluation of assumptions underlying the use of specific models [6, 17]. That is, DSS allow decision-makers to [6]:

- Access relevant data across the organization to make choices among alternatives.
- Analyze data generated from other internal information sources.
- Access information external from the organization.
- Analyze the information in a manner that will be helpful to that particular decision and will provide that support interactively.

Summarizing, DSS provide decision-makers the ability to explore business intelligence in an effective and timely manner [6].

In addition, three main characteristics of a DSS are listed below, which are maintained in the proposed framework:

1. A DSS must access data from different sources.
2. A DSS facilitates the development and evaluation of a model of the choice process. In other words, the DSS must allow users to transform data into information, which helps them make a good decision.
3. A DSS must provide a good user interface through which users can navigate and interact.

On the other hand, Multisensor Data Fusion (MSDF) is a technology that allows combining information from various sources with the objective of obtaining a unified data [11, 12]. Such a combination or fusion is defined through architectures and models designed for this area [13]. For this proposal, specifically, the architectures were studied. Therefore, according to the above, the following categorization of the architectures was considered: centralized fusion, autonomous fusion, and hybrid fusion [13]. Where, for this work, the use of the centralized data fusion architecture is highlighted because it represents the most accurate approach to fuse data theoretically [13]. Furthermore, it is the most appropriate architecture for the proposed framework.

A brief overview of the MSDF area is presented by Hall and Llinas [13]. In this work, the authors provide a guide on data fusion and an assessment of the state-of-the-art and state-of-practice in this field. On the one hand, the JDL (Joint Directors of Laboratories) Data Fusion Process model, as well as architectures proposed and used in this area, are explained. On the other hand, an introduction to techniques applied to data fusion with a discussion of some fundamental issues is given. In addition, they describe and summarize its applications both for the Department of Defense (DoD) area and non-military applications (non-DoD).

Another relevant work for this proposal, mentioned by Hall and Llinas [13], is the Data Fusion Lexicon built by the Joint Directors of Laboratories (JDL) Data Fusion Working Group, established in 1986 [29]. This document aims to provide a common terminology for theoreticians, developers, and users involved in the data fusion area, in order to improve communications

within the data fusion community and thereby, facilitate the exchange of information and cooperation [13, 29]. Such a lexicon is intended to be used to define one of the modules of the proposed framework: Data Alignment.

Finally, in recent years, an updated review of the state-of-the-art in the MSDF area is proposed by Khaleghi *et al.* [11]. They aim to provide a generic and comprehensive view of existing low-level data fusion methodologies, as well as the most recent developments and emerging trends in the area. Likewise, popular definitions, conceptualizations, purposes, potential advantages, challenging aspects of the data fusion, and recent advances are discussed. In addition, they also present less-studied issues in this area and provide new avenues of research, aside from four ongoing research trends in the data fusion community, such as automated fusion, belief reliability, secure fusion, and fusion evaluation.

Once DSS and MSDF were introduced, some systems focused on the event processing and stream reasoning are presented in the following section in order to review the proposed approaches for these issues, mainly considering systems addressing semantic technology, temporal operators, and historical data.

2.1.2 Systems Addressing Event Processing and Stream Reasoning

With the aim of identifying the systems proposed in the literature to deal with event processing and stream reasoning, the research work presented by Margara *et al.* [14] was examined.

Margara *et al.* [14] presented a review about Stream Reasoning, an area emerged in the last few years in order “to bridge the gap between reasoning and stream processing”. They begin describing the changing nature of the Web, presenting some estimated figures of the information provided by applications such as Facebook, YouTube, and Twitter per minute. Likewise, they highlight that these data streams, coming from distinct sources, present a heterogeneous structural and semantic level. Therefore, they consider necessary to count on approaches responsible for operating on-the-fly on several of these streams simultaneously, in order to allow the implementation of real-time services for end-users. With the aim of demonstrating its relevance, they selected and described some scenarios that can benefit from stream reasoning through the solvency of the requirements identified and analyzed in each of these. On the other hand, they survey existing approaches in the area, classifying them into three groups (Semantic Stream Processing, Semantic Event Processing, and Time-Aware Reasoning), and underlining both

their advantages and disadvantages via defined and aligned features with the requirements of the scenarios. It is noteworthy that this alignment was carried out to identify whether there is currently at least one system that can fulfill all the requirements established in the application areas, checking that such approaches are not yet complete. Thereby, considering these two analyzes, they proposed a research agenda with a set of possible theoretical and implementation solutions to fill these gaps and thus, to further progress in this field.

For this section, only the most relevant systems addressing semantic technology, temporal operators, and historical data are described below, presenting its main characteristics as well as its possible advantages and disadvantages.

2.1.2.1 C-SPARQL on S4

Although the solution is an implementation of the RDFS reasoning and the C-SPARQL query language on the S4 streaming platform, this analysis is focused only on the query language because it is considered a new language for supporting continuous queries over streams of RDF data [14, 30]. It is important to note that this system covers the parallel/distributed processing requirement [14].

Continuous SPARQL (also called C-SPARQL) is an extended and registered language of SPARQL to support continuous queries. It is continuously executed over RDF data streams, taking into account windows of such streams [30]. The characteristics, advantages, and disadvantages listed below are based on the work of Barbieri *et al.* [30].

- Characteristics
 1. Extended version of SPARQL.
 2. It supports continuous queries.
 3. It supports data aggregation.
 4. It considers windows of data streams.
 5. It allows streams reasoning.
 6. RDF streams are identified by an IRI (Internationalized Resource Identifier), which indicates the location of the actual streaming data source.
 7. RDF streams are continuously produced and annotated with a timestamp.

8. RDF streams are selected through windows. A window extracts the last data stream elements.
 9. Physical extraction. It is defined via a specified number of triplets.
 10. Logical extraction. It is defined through a given time interval (the number of triplets is variable).
 11. The types of outputs of C-SPARQL are the same as in SPARQL: Boolean answers, variable bindings, new RDF triples, or RDF descriptions of resources.
- Advantages
 1. It guarantees interoperability by supporting streams in RDF format.
 2. It allows dealing with updated knowledge.
 3. It follows the same syntax of SPARQL, *i.e.*, a regular SPARQL query is also a C-SPARQL query.
 4. It allows multiple independent aggregations within the same query. It is noteworthy that this capability is not provided in SQL (Structured Query Language).
 5. It allows performing both continuous and statics queries.
 6. It allows the construction of new RDF data streams.
 - Disadvantages
 1. It works only with RDF streams.

2.1.2.2 EP-SPARQL

Event Processing SPARQL (EP-SPARQL) is a unified high-level language for event processing and stream reasoning [31]. It is important to note that this system covers the temporal operators and historical data requirements [14]. The characteristics, advantages, and disadvantages listed below are based on the work of Anicic *et al.* [31].

- Characteristics
 1. New high-level language for Event Processing and Stream Reasoning:
 - (a) Event processing. It detects compound events within streams of simple events opportunely.

- i. An event can be a tweet, a geospatial data sent by a GPS-enabled device, a newly updated status of a user from a Web application or the status of an object in that application.
- ii. A complex event is the action of combining simpler events into compound ones through different event operators and temporal relationships. For its description is necessary both semantics and temporal relationships.
- iii. It provides on-the-fly analysis of event streams.
- iv. It cannot combine streams with background knowledge, which describes the context or domain in which streaming data are interpreted, and cannot perform reasoning tasks.

(b) Stream reasoning.

2. This language extends the SPARQL language with its event processing and stream reasoning capabilities.
3. The execution model is grounded on logic programming. An approach based on event-driven backward chaining rules that perform an effective event-driven inference.
4. It counts on event processing and inference capabilities over temporal and static knowledge.
5. It uses event patterns to detect complex events.

- Advantages

1. It counts on event processing and inference capabilities over temporal and static knowledge.
2. There is an open-source prototype implemented in Prolog.
3. Both event processing and inference capabilities are applied to temporal and static knowledge.
4. It exists a Prolog's library to transform an RDFS ontology into Prolog rules and facts.

- Disadvantages

1. It is implemented in Prolog language (technical aspect and coding capability).

2.1.2.3 Answer Set Programming (ASP)

ASP-based approach for stream reasoning founded on the sliding window model. This approach reads an “offline” encoding just once and keeps only the n last entries of an “online” data stream. By integrating ASP into this approach is dealt with time-decaying program parts. Thereby, they proposed a novel language that allows specifying and reasoning about time-decaying logic programs in an effective way [32, 33]. It is important to note that this system covers the temporal operators requirement [14]. The characteristics, advantages, and disadvantages listed below are based on the works of Gebser *et al.* [32, 33].

- Characteristics

1. It provides new techniques to deal with emerging and expiring data in a seamless way.
2. A time-decaying logic program is a logic program with a lifespan.
3. This approach can be implemented in agent technology, belief revision and update, cognitive robots, among others.
4. It was implemented within the reactive ASP solver oclingo.

- Advantages

1. It handles time-decaying data and programs within the reasoning methodology of ASP.
2. It integrates modeling techniques to handle changing data without continuous reprocessing or increasing memory demands.
3. This approach is general purpose.

- Disadvantages

1. For new reasoning scenarios, a re-engineering of the underlying ASP systems is necessary.
2. It lacks a reactive optimization feature.

2.1.3 Semantic Approaches incorporating Similarity and Relatedness Measures

Similarity measures represent a key concept for the proposed framework since they are integrated into the recommendation algorithm described in this proposal. Therefore, the introduction of some approaches addressing both semantic technologies and similarity measures is presented in this last section in order to present different, but related, application and revision approaches to the research work in the next section.

Firstly, a basis for reviewing the similarity and relatedness measures used in the Computational Linguistics area, such as the *tf-idf* measure, is the Sidorov's book [34]. Subsequently, a project whose approach integrated both semantic search techniques and Natural Language Processing (NLP) concepts is AquaLog [35]. AquaLog was presented as a portable solution since it can be adapted to any given ontology, using NLP technologies to map a natural language query to semantic markup [35]. Later, another work that addressed the relatedness measure is introduced by Youssif *et al.* [36]. They proposed a technique, called Hybrid Semantic Web Search (HSWS), to build a semantic graph. This technique allows indicating the relatedness between the concepts extracted from articles representing a certain topic. In addition, HSWS could be a relevant technique since it is an unsupervised technique and, therefore, does not require any type of training. Finally, a recent approach, called intelliSearch, is presented by Mehta *et al.* [37]. They introduced an implementation of a semantic web search engine connected to Wordnet and based on the semantic relatedness, applying the *tf-idf* measure to calculate the term frequency of each web page.

2.2 Traditional and POI Recommendation Algorithms

Recommendation systems are based on personalization systems. Amoretti *et al.* [38] defined a personalization system as a computer-based application that learns the behavior of a person to generate and manage his/her profile. Specifically, when the personalization system can provide suggestions to a user according to his/her profile, then these systems are called as recommendation systems. Such recommendations can be of any type of product or interest, such as places, technology, entertainment, food, and so on. For this reason, the recommendation systems can support other applications and services in adapting to the specific preferences of each user. Netflix, YouTube, and Spotify are a few examples of applications and services that

make use of recommendation systems. Consequently, an active and challenging research area is the development of algorithms capable of giving accurate recommendations to users based on their individual preferences.

As previously mentioned, recommendation systems can be used in several contexts. Because one of the main goals of this research work is to propose and develop a recommendation algorithm that suggests a research areas list according to the contextual information concentrated in the user's academic profile, this proposal is focused on two types of systems: the traditional recommendation systems and the POI recommendation systems. On the one hand, the traditional recommendation approaches commonly obtain user preferences through ratings that he/she provides to certain items in an application or service, such as books, movies, or music [25, 39]. On the other hand, the POI recommendation systems model the users' visiting preferences in order to recommend POIs that the user never visited before but could be interested in [40–42]. Therefore, according to these definitions and the goal of this algorithm, this research is mainly focused on the related work to the POI recommendation algorithms.

Some approaches about POI recommendation algorithms are briefly described below. Firstly, Kang *et al.* [43] proposed a Personalized POI Recommendation Method for the tourist POI recommendation as well as the POI and user data that can be exploited for this task. Specifically, they used the user's explicit preferences and POI categories to carry out the tourist recommendations. Around five years later, Ye *et al.* [44] proposed a unified POI recommendation framework to provide a POI recommendation service for location-based social networks (LBSNs), exploring user preference, social influence, and geographical influence. Later, Ying *et al.* [45] proposed an Urban POI-Mine (UPOI-Mine) approach to suggest urban POIs based on the users' check-ins, POI categories and popularity, along with social influence. Subsequently, Zheng *et al.* [46] proposed the cross-region topic-based collaborative filtering (CRTCF) method based on hidden topics mined from user check-in records with the aim of recommending new POIs to a user in regions where he/she has rarely been before. In the same year, Liu *et al.* [42] proposed a Geographical-Topical Bayesian Non-negative Matrix Factorization (GT-BNMF) model that allows capturing the geographical influences on user's check-in behaviors, as well as integrating the POIs' regional popularity. Similarly, Liu *et al.* [47] proposed a two-stage category-aware POI recommendation model to suggest a personalized POI based on user's check-ins, geographical influences, POI categories, and temporal information. Then as well, Meehan *et al.* [48] proposed a work in progress to deal with problems of inappropriate suggestions arisen information overload and inadequate content filtering by means algorithms implemented in their application

in development called as VISIT (Virtual Intelligent System for Informing Tourists), a context-aware tourist app. Later, Yuan *et al.* [41] proposed Geographical-Temporal influences Aware Graph (GTAG) to deal with the problem of the time-aware POI recommendation; with GTAG, they intended to model check-in records as well as to exploit both geographical and temporal influences of these records for the time-aware POI recommendation. Afterwards, Liu *et al.* [49] proposed a general geographical probabilistic factor model (Geo-PFM) framework which can capture the geographical influence on a user's check-in behavior. In the same year, Zhang and Wang [50] proposed a location and time aware social collaborative retrieval model (LTSCR) for the successive POI recommendation task considering the user's location, time, and social information simultaneously. Subsequently, Yu *et al.* [51] proposed a recommender of personalized travel packages with multiple POIs based on crowd-sourced user footprints to help users find interesting locations as well as to generate travel packages consisting of different types of locations and visiting sequences. To carry out the recommendations, crowd-sourced check-in records, ratings, POI categories, geographical influence, and temporal information are considered. Finally, Guo *et al.* [40] proposed a weighted Bayesian personalized ranking model with visit frequency and distance (WBPR-FD) to give POI recommendations using user's check-ins and geographical distance.

2.2.1 Comparative: POI Recommendation Algorithms

A comparative table summarizing the previously mentioned works is presented to highlight the contributions of this research. The aspects considered are described below:

- Year (Y). It refers to the year of publication of the approach.
- Rating (R). Data that considers the recommendation algorithm to address the user's explicit preference on POIs.
- Check-in (CI). Data that considers the recommendation algorithm to address the user's implicit preference on POIs.
- Geographical Influence (GI). Factor that is examined in the POI recommendation approach.
- Social Influence (SI). Factor that is explored in the POI recommendation approach.

- Category (C). Data that considers the recommendation algorithm to address the POI tags, categories, or topics.
- Another Context Data (ACD). Some other data that the POI recommendation algorithm considers different from the data and factors mentioned in this comparison.
- Information Source (IS). Source that is employed in the collection of the data used to evaluate the POI recommender.
- Similarity and Distance Measures (SDM). Measure that is applied in the POI recommendation algorithm.
- User-based CF with Aggregation Operator (UCF+), where \mathcal{X}^* indicates that the approach works with the user-based CF algorithm without an aggregation operator. Algorithm that is implemented as a POI recommender using an aggregation operator as a similarity measure.
- Scope (S). Field of application of the approach.

In summary, according to Table 2.1, it is concluded that only one approach [43] addresses ratings, nine of them address check-ins [40–42, 44–47, 49, 50], and one addresses both ratings and check-ins [51]. In the same way, geographical influence is the most used factor in the literature explored with eight works [40–42, 44, 47, 49–51] while social influence is examined only in four works [44–46, 50]. Furthermore, the POI categories are also more explored than social influence with six approaches [42, 43, 45–47, 51]. Other characteristics that have been analyzed are temporal influence by four approaches [41, 47, 50, 51] and POI popularity by one approach [45]. From the information sources used to collect data for the recommendation algorithm, most approaches used LBSN [40–42, 44–47, 49–51] and only two approaches [43, 48] used other information sources. Regarding the similarity and distance measures applied in the recommendation algorithms, cosine similarity is the most common measure in the recommendation approaches with six works [43–46, 50, 51] and Euclidean distance is also used in [45]. Lastly, the user-based CF algorithm is employed in three research works [43, 44, 51] using the Cosine similarity as a similarity measure. Therefore, according to this review, a recommendation algorithm that allows dealing with the user preferences as well as contextual information, considering both geographical influence and categories (by being the most discussed aspects in the literature), and incorporating different similarity and distance measures into the recommendation process is considered as a relevant approach to explore.

TABLE 2.1: Comparison of the POI Recommendation Approaches

Author	Y	R	CI	GI	SI	C	ACD	IS	SDM	UCF+	S
Kang <i>et al.</i> [43]	2006	✓	✗	✗	✗	✓	✗	Jeju-do Tourist Association from Republic of Korea and surveys from the Internet	✓	✗*	Tourism
Ye <i>et al.</i> [44]	2011	✗	✓	✓	✓	✗	✗	LBSN	✓	✗*	LBSN
Ying <i>et al.</i> [45]	2012	✗	✓	✗	✓	✓	✓	LBSN	✓	✗	Urban areas
Zheng <i>et al.</i> [46]	2013	✗	✓	✗	✓	✓	✗	LBSN	✓	✗	LBSN
Liu <i>et al.</i> [42]	2013	✗	✓	✓	✗	✓	✗	LBSN	✗	✗	LBSN
Liu <i>et al.</i> [47]	2013	✗	✓	✓	✗	✓	✓	LBSN	✗	✗	LBSN
Meehan <i>et al.</i> [48]	2013	✗	✗	✗	✗	✗	✓	WorldWeatherOnline API, Twitter, and users themselves	✗	✗	Tourism
Yuan <i>et al.</i> [41]	2014	✗	✓	✓	✗	✗	✓	LBSN	✗	✗	LBSN
Liu <i>et al.</i> [49]	2015	✗	✓	✓	✗	✗	✗	LBSN	✗	✗	LBSN
Zhang and Wang [50]	2015	✗	✓	✓	✓	✗	✓	LBSN	✓	✗	LBSN
Yu <i>et al.</i> [51]	2016	✓	✓	✓	✗	✓	✓	LBSN	✓	✗*	Tourism
Guo <i>et al.</i> [40]	2017	✗	✓	✓	✗	✗	✗	LBSN	✗	✗	LBSN

2.3 User Profiles and Ontologies

Firstly, the user profile definition from the perspective of the Human Resources area is presented. Secondly, the user profile modeling through an ontological structure from the computational point of view is described. For then concluding with a comparison of the user profile approaches in the Computer Science area.

2.3.1 User Profile from the Human Resources Approach

From the Human Resources area, Alles [52] described a technical and practical methodology for selecting people based on competencies in order to incorporate new staff within an organization. She clearly defined the activities and characteristics that a human resources specialist or any stakeholder responsible for choosing these employees may consider taking into account during this process at each stage planned. For this proposal, Chapter four called Profile Definition stands out since the basis for discerning between a profile and an anti-profile, a concept underlying the definition of a profile, is presented. Also, a guidance constituted by seven steps for building a profile, describing best practices as well as examples based on her experience, is provided. Additionally, some application forms that help to relieve a profile are given. Although the application context is not the same, this reference could be considered as a support to design a profile in the Computer Science area.

2.3.2 User Profile from the Computer Science Approach

Henceforth, the works carried out under the computational approach are described. A key work for this area was written by Golemati *et al.* [53], building a user profile ontology, which incorporates concepts and properties that have been used in existing literature, applications, and ontologies related to this domain: user context and profiling. Similarly, they are based on design criteria proposed for the process of building an ontology, such as clarity, coherence, extensibility, and so on. This with the aim of obtaining a general, customizable, and extensible user ontology. Such an ontology was applied to two areas: adaptive and personalized visualization, and Personal Information Management (PIM). It is worth mentioning that this ontology was modeled only with user static and semi-permanent features, considering the dynamic features as further work. In the same way, they proposed as future work to use questionnaires for obtaining the user profiles' properties, that is, as a means to populate the proposed user profile ontology.

2.3.2.1 Approaches considering Dynamic and Contextual Information

Some approaches addressing the user's dynamic and context information, in a field of application different from the academic area, are described below. Sutterer *et al.* [54] presented a user profile ontology dedicated to describe situation-dependent user sub-profiles, called User Profile Ontology with Situation-Dependent Preferences Support (UPOS). UPOS can be used by

context-aware adaptive service platforms for mobile communication and information services to automatically trigger a situation-dependent personalization of services, based on a single condition or a conjunction of conditions belonging to a certain situation of the user. On the one hand, the UPOS' intuitive design approach and simple understanding are accomplished by considering design issues described in the literature as well as recommendations from the Human Factors research area. On the other hand, the authors proposed a modeling methodology to enable multiple sub-profiles, provide a means to attach meaningful conditions to sub-profiles, and include top-level user context as well as user attribute vocabulary concepts (extensibility). As an example of its extensibility, some FOAF and vCard (Electronic Business Cards) classes are proposed as concrete user attribute vocabularies, and as a demonstration of its applicability, a slightly adapted version of UPOS is used in the Service Platform for Innovative Communication Environment (SPICE) project. Although the authors provide a UPOS download site, it is not available now. However, a version of UPOS can be carried out since the descriptions about its classes, object properties, and datatype properties, and a visualization of its relationships are given in this work. In the same year, Katifori *et al.* [55] proposed an ontology-based user profiler in the context of a Personal Interaction Management System (PIMS). The profiler allowed users to create their personal ontology firstly starting from one of the available template ontologies. Subsequently, the profiler allowed populating and customizing these personal ontologies through forms in a web interface. This work presented a proposal to develop a non-expert user-oriented ontology presentation method that can be integrated into a PIMS prototype and into any application that intends to model the user's personal information using an ontology. Hence, they introduced an ontology, called Personal Ontology, to store the user's knowledge base in the proposed system. The Personal Ontology model encompassed a basic core of general concepts that may be extended and enriched to accommodate several user stereotypes or individual profiles. To create the set of upper-level concepts for Personal Ontology, profile information models managed by several applications and proposed by researchers, as well as general ontologies were considered. Thereby, Personal Ontology comprised a wide range of user characteristics, including personal information as well as relationships with other people (both personal and employment), preferences, and interests. In addition, since the ontology schema and the ontology instances, as well as their relevant slot values, were operating together, the profiler presented a dynamic nature by immediately becoming aware of any changes made to the ontology schema, *e.g.*, when adding new classes or types of relationships. Consequently, the profiler forms were updated by ontology designers each time the ontology schema was modified. Finally, although the authors provide a Personal Ontology download site, it is not available now.

However, an overview of the Personal Ontology's classes can be reviewed in this work. Another approach that describes an ontology-based user profile model is the proposed work by Stan *et al.* [56]. This model allowed users to generate a situation-aware social network in order to control how specific categories of people could reach them when they were in a given situation, such as at work or at home. That is, this model aimed to provide full control to a certain person for managing both the groups of people who can contact him/her and the means of communication that these groups can use in determined social situations, and to reduce the time consuming and human interaction that the user can spend on this task of management. Hence, a user profile was proposed to characterize the current situation of the user and to provide the reachability preferences established for his/her social network according to situational sub-profiles. Here, the authors defined as a social network the list of people who have a relationship with a specific user and as a sub-profile a subset of the profile where a set of communication preferences, related to a certain person or group of people in a given situation, are contained. The proposed user profile model, called Situation-Aware User Profile Ontology (SAUPO), is based on UPOS (an ontology mentioned earlier in this section) to represent the current situation of the user, extending such an approach with the conjunction of context dimensions to achieve a better identification of the user's situation in real-time. In the same way, changes related to the situation concept were incorporated into the original UPOS model, such as the replacement of the concept "Condition" with "Situation" and the definition of reachability preferences in a situation-aware sub-profile. A description of the application of SAUPO to fill the instances of the user profile with information learned and extracted from user-habits is introduced. For this case, log files containing context data of users, such as location, environment, time, and accessible devices, were used. Finally, the authors did not provide a SAUPO download site, however, an overview of SAUPO can be reviewed in this work. Subsequently, Niu and Kay [57] dealt with the issue of building an ontological framework that can be integrated into applications which aim to provide personalized location information, about the people inside a building, in a pervasive computing environment. Therefore, they developed a framework, called PERSONAF (Personalised Pervasive Scrutable Ontological Framework), to support pervasive ontological reasoning and to aid users in understanding and controlling pervasive applications. In other words, the objectives of PERSONAF are to support the personalization in a pervasive computing environment, to provide a scrutable representation and reasoning, and to be an ontological approach to achieve these goals. The main elements of PERSONAF are the ontology representation, called PECO (Personalised Context Ontology), that can adapt location information to different contexts and users in pervasive computing, and a reasoning engine, named ONCOR (Ontology-

and Evidence-based Context Reasoner), that can interpret PECO and provide the programming interface used by applications. On the one hand, PECO has three layers: a handcrafted ontology, called Middle Ontology, that describes the key concepts about buildings and relationships between them (stable, foundation layer of PERSONAF); an Application Ontology that defines a particular building, which must be created afresh for each new building and can be static if there are no changes to the building; and an Accretion Ontology that represents the fast-changing aspects as well as the personal ontologies. On the other hand, ONCOR interprets the PECO ontology through appropriate mechanisms for the representation used in each layer, that is, standard ontology representations and reasoning are used to interpret the middle and application ontology, and a reasoning mechanism, which uses a runtime selection of the resolver tool, is used to deal with the accretion ontology. As a case study, an implementation of PERSONAF focused on the building housing the School of Information Technologies (SIT) was carried out. The implementation of the three layers of PECO was MIBO (Middle Building Ontology) for Middle Ontology, the SIT building application ontology for Application Ontology, and the dynamic accretion ontology for Accretion Ontology. To assess the implementation of PERSONAF, a version of PERSONAF was implemented in a demonstration application called Adaptive Locator, which delivered personalized information about the people's location in the SIT building and allowed the user to examine the personalization to know the evidence and processes underlying the system reasoning through an interface. This approach used available resources to populate the ontology, presenting two important features: the usage of multiple resources and the maintenance of the details of its propositional beliefs, the set of evidence supporting each one, and the details of the source of each piece of evidence. Finally, the authors did not provide a PECO download site, however, an overview of PECO can be reviewed in this work. In the same year, Dix *et al.* [58] proposed methods to use spreading activation on web-scale information resources with the aim of allowing web-scale reasoning, based on dynamically identifying a relatively small, but appropriate, "working set" of entities and relations, to support users in having the right information available at the right time. They extended existing works on spreading activation by developing methods that linked a user's personal ontology to large external repositories, such as corporate information and the Web, in order to model context in personal ontologies and allow automatic enrichment through the entire Web of data. That is, the population of the personal ontology with cached data from external repositories, where the choice of what data to fetch or discard is related to the level of activation of the entities already in the personal ontology or the cached data. They proposed using ontological type tags to automatically fill in the fields in an unconstrained personal ontology when certain user data

have already been entered, trying to deal with the situation of not having previous information or several alternatives through spreading activation, as a means to predict the context of the user's actions (tasks) and provide the appropriate data as well as the possible actions on this data. For this end, they proposed an ontology called Personal Ontology, which comprised a wide range of user characteristics, including personal information as well as relationships with other people (both personal and employment), preferences, and interests, based on standards, such as FOAF and vCard, as well as proprietary profiles, such as Facebook. This ontology supported the spreading activation process, to identify the entities of interest to the user in a specific context and the relationships among these entities, as follows: firstly, when a certain activity is performed by a user, some entities in Personal Ontology could be related to the context of this activity. Therefore, these entities receive an "immediate activation", which can be spread to other connected entities within Personal Ontology. Finally, entities that exceed a certain threshold of activation can be considered as better candidates for the user to perform a subsequent activity. In this work, the spreading activation process was used in Task Information Management (TIM) to provide context inference to tools that support TIM, applying this process on personal ontologies. Hence, in order to consider the aspects related to the spreading activation process within the ontology components, the Personal Ontology model was extended to include the following properties: Short-Term Activation (STA), Medium-Term Activation (MTA), Long-Term Activation (LTA), Immediate Activation (IA), incoming activation (IN), maximum LTA (MAXLTA), and Long-Term Weight (LTW). Similarly, rules to update the activation levels of the entities within Personal Ontology were incorporated. Finally, although the authors provided an extended Personal Ontology download site, it is not available now. However, an overview of the extended Personal Ontology's classes and properties can be reviewed in this work. Later, an approach that considers addressing the syntactic and semantic heterogeneity of user models was presented by Martinez-Villaseñor *et al.* [24]. They proposed to deal with user modeling interoperability using a two-tier matching strategy for concept schemata alignment and to provide a knowledge representation for a ubiquitous user model, called u2m, through a dynamic user profile structure based on Simple Knowledge Organization for the Web / Simple Knowledge Organization System (SKOS). The concept alignment process, on the one hand, established the mappings from the individual concepts of the sources with the corresponding u2m concepts (organized in an SKOS concept scheme), and on the other hand, this process determined the evolution of u2m over time, defining when a new concept should be added. The concept alignment process is carried out in two phases, an element level matching and a structure level matching, using Dice coefficient, Longest common substring, and the

semantic similarity based on WordNet as matching techniques as well as sets of neighbors of the target concepts in the hierarchy (which define the context of each concept in the source) and if-then rules. U2m supported the personalization of web services by sharing and reusing profile information with semantic web technologies, being (u2m) the mediation between profile suppliers and consumers (sources). To define the user profile structure, some international initiatives towards the standardization of user profile structure, the nature and relevance of the information of user profiles, and some applications and devices were considered. Therefore, the user profile structure integrated static, semi-static, and dynamic concepts, such as demographic and general user data, device and service profiles, and user preferences and interests, into the start-up configuration of u2m. Furthermore, to provide the semantic representation of u2m and the profile instances, an ontology was proposed, called ubiquitous user modeling interoperability ontology (U2MIO). This ontology had the user model concept scheme, the concept schemata for each source with the most recent instance, and semantic mappings. That is, U2MIO stored only one instance, the most recent information extracted, from each source and u2m had only one concept of each similar attribute, allowing only the addition of new concepts, sub-collections, and collections from the concept alignment process. Their application scenario is focused on sharing and reusing data from different user profiles, to improve u2m and determine parameters of web services, in order to deal with overweight and obesity. Specifically, U2MIO was set up with data from one profile of a specialized web application to monitor a person's diet and physical activity (polarpersonaltrainer.com website), collected from Polar devices (supplier), and a personalization was carried out by reusing the supplier's information for the LogWeight Web Service from TrainingPeaks (consumer). Finally, the authors did not provide a U2MIO download site, however, an overview of U2MIO can be reviewed in this work. Subsequently, a user profile ontology for customizing context-aware applications within mobile environments was proposed by Skillen *et al.* [59]. They were mainly focused on analyzing user behavior and characterizing his/her contextual needs for context-aware applications, as well as the ontological modeling of dynamic components used within these applications. For this purpose, an ontology, called User Profile Ontology, was developed. User Profile Ontology addressed the changing behavior of the user, combining static and dynamic user's concepts, providing an adaptable model across several application environments. For example, the personalization of a user's smartphone settings during a meeting or a trip. In this ontology, a top-down design approach was used, where the top-level embraced general user's concepts, such as Activity and Interest, and the lower-levels encompassed more specialized concepts, such as Activity_Type and Interest_Level. Therefore, the ontology development process consisted of firstly identifying key terms to describe a user,

and then modeling them as ontological classes. For instance, the `User_Profile`, `Preference`, `Capability`, and `Interest` classes were introduced to support the customization of the service according to the user's information. After all the classes were defined, object and data properties were determined for each class in order to relate one class to another through specific objects or relationships (data-types). These properties have an important role in relating key concepts and inferring new information. As a case study, the ontology model was adopted within the MobileSage project in order to provide context-aware personalized services for elderly people, allowing to improve their Quality of Life (QoL), through mobile-based technologies. Finally, the authors did not provide a User Profile Ontology download site, however, an overview of the User Profile Ontology's classes and properties can be reviewed in this work.

2.3.2.2 Approaches considering Academic Resources

Up to this point, research works proposing an ontology to model the user's profile in application domains related to PIM, TIM, among others have been described. These ontologies are relevant for the definition and modeling of the ontology proposed in this work; however, research related to the modeling of users in the academic area, such as researcher profiles, is also necessary to address since the scope of this proposal is focused on modeling the user's academic profile. Therefore, approaches dealing with academic environments are introduced below.

A first approach was proposed by Yao *et al.* [60]. They proposed a unified tagging approach to develop a semantic profile of an academic researcher based on information available on the Web, mainly on personal homepages. Such an approach supported the assignment of tags to the text obtained from the homepages, where one tag corresponded to one property of the user's profile, using Conditional Random Fields (CRF) as the tagging model. To achieve this goal, a researcher profile schema was defined and a process consisting of the following three steps was carried out: find, preprocess, and tag the researcher's homepage. On the one hand, the researcher profile schema, called Researcher Profile Ontology, turned out of extending the FOAF ontology, containing the following aspects: basic information (*e.g.* affiliation), contact information (*e.g.* address), educational history (*e.g.* graduated university), and publications. On the other hand, the process began with identifying the homepage using a web search engine and a classifier; secondly, the text gathered from the selected homepage was segmented into tokens and for each token, a possible tag was assigned; finally, for a sequence of tokens, a sequence of

tags was determined through CRF. For their experimentation, they used the ArnetMiner application to obtain the information from researchers, as well as the dblp (DataBase systems and Logic Programming) bibliography to extract the publication data. From ArnetMiner, 448,291 records were collected, of which 1,000 records of researchers were randomly chosen for their application scenarios. To integrate the researcher information and publication data into the same profile, they adapted a name-reconciliation algorithm since different researchers shared the same name. This algorithm used a semantic similarity based on the researcher/author's context to find coincidences in the researcher/author's possible names between both information sources. From this work, another outstanding aspect mentioned is the observation that the researcher's research interests can be inferred from the researcher's publications, an observation that was integrated into the proposed recommendation algorithm approach in this research (C-HyRA). In addition to sharing a similar aspect by integrating publication data from dblp. However, they did not include an annotation of research interest, a property considered within the ontology proposed in this research. Finally, the authors did not provide a Researcher Profile Ontology download site, however, an overview of the Researcher Profile Ontology's class and properties can be reviewed in this work. Another approach was proposed by Katifori *et al.* [61]. They proposed the use of ontologies as a long-term knowledge repository for PIM-related information as well as the use of spreading activation on ontologies to provide context inference to tools that support TIM. Therefore, they explored the application of the spreading activation theory of the human memory on ontologies with the aim of developing a context inference model to be used by any ontology-based PIM/TIM prototype system. To this end, a Personal Ontology was proposed for the domain of the user's personal collection, as a basis of an intelligent mechanism to support these systems. This ontology was an extension of one already published since it was enriched with more user-related classes for the stereotype of a researcher, which was used for the evaluation of the spreading activation algorithm. The Personal Ontology model incorporated a basic core of general concepts, such as personal data, relationships (both personal and employment), preferences, and interests, which could be enriched to accommodate various user stereotypes or individual profiles. To generate this ontology, a top-down approach as well as the Gruber's design criteria, such as clarity, coherence, extensibility, among others, were adopted. In the same way, to establish a simple and comprehensive set of upper-level concepts, profile information models existing in applications and related work to the profiling research area as well as ontologies (for example, vCard) were considered. As an application scenario, they proposed Personal Ontology as a basis for the spreading activation framework focused on supporting the

PIM and TIM systems. Specifically, a preliminary evaluation of the spreading activation algorithm was carried out using a test platform developed by them as a plug-in, called ActiveOnto. This plug-in aimed at supporting the setting of the spreading activation algorithm parameters as well as partially testing the algorithm's functionality within a PIMS. Two relevant aspects of their work related to the field of application of this proposal are the researcher profile modeling and the inclusion of the spreading activation algorithm in the academic approach. Finally, although the authors provided an extended Personal Ontology download site, it is not available now. However, an overview of the extended Personal Ontology's classes and properties can be reviewed in this work. Lastly, in recent years, Becerril *et al.* [62] presented a semantic approach to discover knowledge into structured information with the OAI-PMH (Open Archives Initiative - the Protocol for Metadata Harvesting) protocol considering ontological representations and the user's context. They proposed a methodology to collect, transform, merge (enrich), and store the information initially structured under the OAI-PMH protocol, and at the same time, to retrieve and provide the information requested by a user through a query, considering the contextual information of the same user. Therefore, firstly a batch processing was carried out to collect the information from the metadata repositories based on OAI-PMH, where XML files are obtained containing data described under the simple set of Dublin Core (dc) metadata. Subsequently, the metadata collected from the repositories based on OAI-PMH were transformed to the RDF/XML format in order to allow the enrichment of these records. For instance, to enrich the author's data, such as establishing co-authorship relationships, using both the FOAF ontology and an extended version of the dc ontology. For this, a semi-automatic merge process was performed to integrate the schemata between these two ontologies (FOAF and dc) into a single ontology. Later, once the merged ontology was obtained and the collected metadata were stored in RDF (knowledge base), in order to provide the information requested by the user, an information retrieval engine as well as a reasoner based on rules were developed to process such a query on the triplets previously transformed. Specifically, the proposed inference process considers only exact matches between the values of the triplets and the parameters of the query (same property) as well as with the user's contextual information to establish an association and deliver a result. In the same way, the user's contextual information was modeled and entered through an ontological profile, called *OntoOAIEstudiante*, based on a student model proposed in the literature since the academic-scientific nature of the information addressed in this work. This student model consists of four classes that cover information related to his/her educational process, academic activities, and personal information. To validate this approach, two repositories incorporating OAI-PMH were used: REDALYC (Red de Revistas Científicas de América

Latina y el Caribe, España y Portugal) and RUDAR (Roskilde University Digital Archive). The experimental scenario consisted in discovering academic resources from the Sociology area relevant to the student according to his/her query and profile. Finally, the authors did not provide an OntoOAIEstudiante download site, however, an overview of the OntoOAIEstudiante's classes and properties can be reviewed in this work.

2.3.3 Comparative: Ontological User Profiles

To highlight the contributions provided by the proposed ontology (Academic SUP) as well as by each related work, considering only the user profile approaches described in the computational area, Table 2.2 is designed. This comparative table is composed of the following aspects:

- Year (Y). It refers to the year of publication of the approach.
- Ontology (O). Data that indicates the name given to the ontology used in the approach.
- Static Information (SI). Approach that addresses the user's static information.
- Dynamic Information (DI). Approach that deals with the user's dynamic information.
- Contextual Information (CI). Approach that explores the user's contextual information.
- Ontology available on the Web (OAW). Field that indicates the availability of ontology on the Web.
- Human Resources (HR), where \mathcal{X}^* indicates that the approach addresses aspects related to this area. Approach that considers the user's modeling principles and definitions from the Human Resources area.
- Scope (S). Field of application of the approach.

TABLE 2.2: Comparison of the Ontological Approaches

Author	Y	O	SI	DI	CI	OAW	HR	S
Golemati et al. [53]	2007	User Profile Ontology version 1	✓	✗	✓	✓	✗	Adaptive and personalized visualization, and PIM
Sutterer et al. [54]	2008	UPOS	✓	✓	✓	✗	✗*	Context-aware adaptive service platforms for mobile communication and information services (SPICE project)
Katifori et al. [55]	2008	Personal Ontology	✓	✓	✓	✗	✗	PIMS
Stan et al. [56]	2008	SAUPO	✓	✓	✓	✗	✗	Functional settings of mobile devices according to the user situation
Niu and Kay [57]	2010	PECO	✓	✓	✓	✗	✗	Personalization in a pervasive computing environment (Adaptive Locator application)
Dix et al. [58]	2010	Personal Ontology (extended)	✓	✓	✓	✗	✗	TIM
Martinez-Villaseñor et al. [24]	2012	U2MIO	✓	✓	✓	✗	✗*	User-adaptive systems (overweight and obesity)
Skillen et al. [59]	2012	User Profile Ontology	✓	✓	✓	✗	✗	Personalization of context-aware applications within mobile environments (MobileSage project)
Yao et al. [60]	2007	Researcher Profile Ontology	✓	✗	✓	✗	✗	Researcher's profiling (expert finding)
Katifori et al. [61]	2010	Personal Ontology (extended)	✓	✓	✓	✗	✗	Ontology-based PIM and TIM systems (PIMS)
Becerril et al. [62]	2016	OntoOAI-Estudiante	✓	✓	✓	✗	✗	Discovery of academic-scientific knowledge from structured repositories under the OAI-PMH protocol (student's profiling)

According to the information presented in Table 2.2, all of the approaches [24, 53–62] dealt with both static and contextual information related to the user while nine [24, 54–59, 61, 62] of the 11 research works explored the user's dynamic information. Regarding the online availability of the ontology, only one [53] of the reviewed approaches presented this characteristic. Furthermore, two works [24, 54] incorporated Human Factors into their approaches. Finally, from the scope

is concluded that the user’s researcher profile has been scarcely studied by the user’s modeling community since only two research works [60, 61] presented the researcher’s profile modeling and one [62] presented the student’s profile modeling as the application domain. In the same way, only two [61, 62] of these three approaches addressed the user’s dynamic information and none of them considered the user’s modeling principles and definitions from HR. Therefore, the need to propose a user’s academic profile, incorporating the user’s modeling principles and definitions from both the HR area and the Computer Science area, that allows consistent integration of static and dynamic user information as well as contextual information both the user and his/her publications is highlighted as an area of opportunity.

2.4 Summary

This Chapter concludes with the visualization presented in Figure 2.1, which concentrates all the research works addressed as state-of-the-art and related work in this proposal.

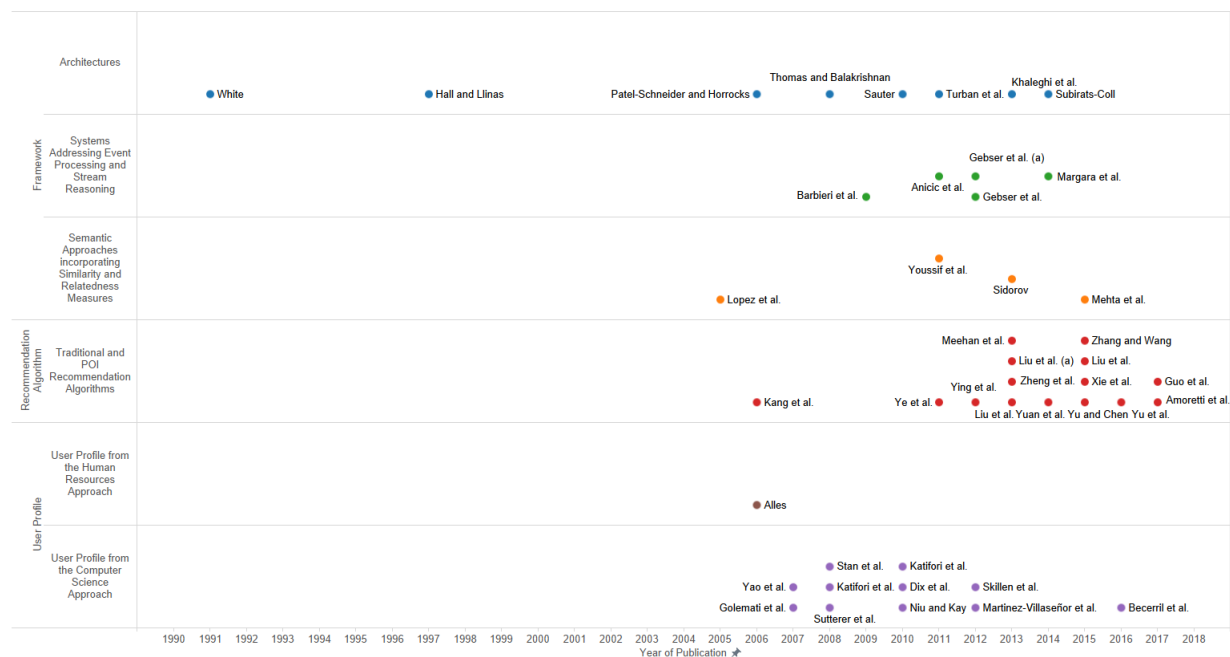


FIGURE 2.1: All research works addressed as state-of-the-art and related work in this proposal are presented in accordance with the classification given in this Chapter. Where, the x-axis represents the year of publication of the contribution and the y-axis the areas wherein the state-of-the-art and related work are divided. The reference to the author is included as a label.

Chapter 3

Proposed Framework for Consistent Generation of Linked Data Focused on the User's Academic Profile

This research work is based mainly on the principles of inference and consistency with the aim of proposing a framework focused on the consistent generation of data for a certain user's academic profile schema on the Web. Since, on the one hand, inference is a centerpiece in the development of the Semantic Web because therethrough new relationships between resources can be discovered [8, 19, 20], a notion that stands out to a greater extent when the Classical Paradigm is approached [19], as in this proposal. On the other hand, consistency is a central criterion for the data quality in a database because it ensures that the integrated data are clean of inconsistencies, errors, and conflicts [63]. Thus, the issues that arise when the data are integrated into a single repository from heterogeneous information sources, with redundant or lacking information, are overcome in a consistent way.

In accordance with the foregoing, the design and building of the proposed framework, called AkCeL (frAMework for Consistent generation of Linked data), is founded on the architecture of DSS since its methodology recognizes the need for data to solve problems [17], as mentioned in this proposal. In addition, DSS presents characteristics and capabilities that AkCeL intends to provide. For instance, semi-structured or unstructured problems; support managers (users) at all levels; support intelligence design, choice, and implementation; modeling and analysis; data access; stand-alone, integration, and Web-based; among others [17]. Consequently, AkCeL will

be constituted by similar components to those that integrate a DSS, *i.e.*, a data management subsystem, a model management subsystem, a user interface subsystem, and a knowledge-based management subsystem [17].

In addition, AkCeL aims to populate and to update a users' academic profile schema modeled and developed through an ontology for this research. Hence, in order to achieve the stated objectives, research and academic information sources available on the Web are used. These information sources can be Open Access (OA) or private. For the OA information sources, the following aspects are considered: widely used by researchers and focused on both academic and research information. For the private information sources, the same considerations for OA are followed, in addition to having the permits for their exploitation since, for this research work, an information source is private if authentication is required to access the data. As a result, according to the above statements, the selected OA information sources are dblp (DataBase systems and Logic Programming) and SIIP (Sistema de Información para la Investigación y el Posgrado), and the allowed private information source is Scopus.

Similarly, because a consistent generation of data is proposed, processes to carry out the data retrieval, preprocessing, alignment, recommendation, and visualization are incorporated into AkCeL. Therefore, in order to explain and to structure the entire proposed framework (AkCeL), this Chapter is organized as follows. In Section 3.1, a general view of AkCeL is introduced; in Section 3.2, the static schema related to the scope is described; in Section 3.3, the data retrieval, preprocessing, and alignment processes are explained; in Section 3.4, the data recommendation process is presented; and in Section 3.5, the user interface and statistics proposed for this use case are given.

3.1 A General View of AkCeL

The architecture of DSS can consider five components: data, models, knowledge (or intelligence), user interface, and users. Data are the first component of the DSS architecture, where data related to a specific situation are manipulated by using models. These models are the second component of the DSS architecture, which can be standard or customized. A knowledge or intelligence component is the third component of the DSS architecture, which can support any of the other components or act as an independent component. A user interface is the fourth

component of the architecture, which allows users to interact with DSS. Finally, users are the fifth component of the architecture [17].

AkCeL is based on these five components to define its components as follows: static user's academic profile meta-schema (data), Section 3.2; dynamic data enrichment (knowledge), Section 3.3; data recommendation (model), Section 3.4; data visualization (user interface), Section 3.5; and researchers as well as data managers at the institutional level (users). The schematic view of AkCeL is presented in Figure 3.1.

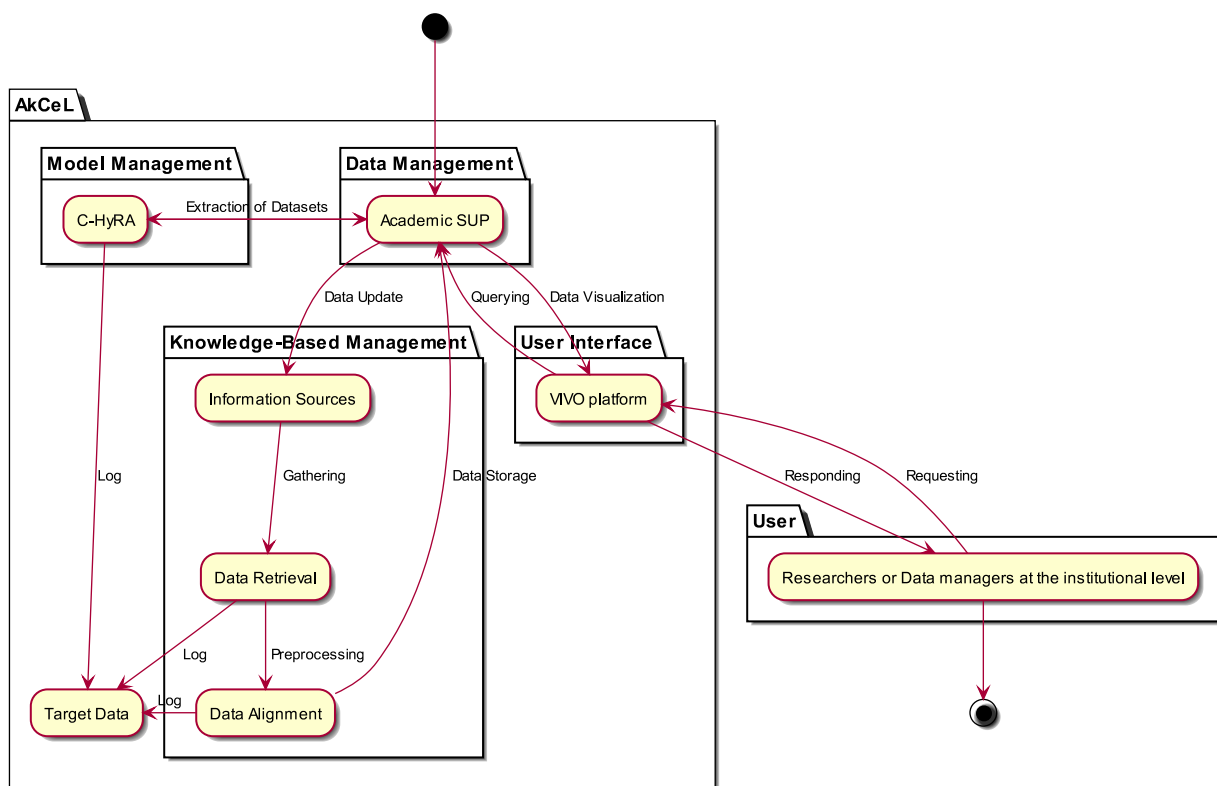


FIGURE 3.1: Schematic view of AkCeL.

Furthermore, two implicit processes are incorporated into AkCeL: “Target Data” and “Data Update”. “Target Data” is responsible for providing a log of the workflow carried out on the knowledge and model components of AkCeL while “Data Update” is responsible for populating and updating information in the user’s academic profile meta-schema, currently, this process is executed manually. Therefore, the log file will be constituted of all the data used in each of the processes of AkCeL. As a result, this file can support other users, such as information managers, to perform future queries of this information as well as to improve the processes of AkCeL. For example, this log can support the data update process to adjust the update periods of each feature of the user’s profile or to know the origin of the data.

Such a log represents one of the outputs of this proposal. Hence, for each process (retrieval, preprocessing, alignment, or recommendation), the data that will be registered are listed below.

1. Update date. Data obtained from the OS when new information is integrated into the user's academic profile meta-schema.
2. Process. Process' name used to carry out the generation or integration of a new data. For instance, retrieval, preprocessing, alignment, or recommendation.
3. Property updated. Property's name that is updated.
4. Last update. Date gathered from this log related to the last update of the property value.
5. Last value. Data retrieved from this log related to the updated previous value.
6. Last information source. Information source' name where the previous value was collected.
7. Current value. Data gathered, preprocessed, aligned, or recommended.
8. Current information source. Information source' name where the current value is collected.

Finally, a set of assumptions are established according to the processes defined and shown in Figure 3.1.

1. AkCeL is compound of a static schema, *i.e.*, a schema that is invariable in time.
2. Because the publications' research areas and categories are required for the recommendation process, the information sources that do not use a subject categorization scheme will not be considered for this process.
3. After the integration of a new data into the proposed ontology, the temporal data of each process will be stored in a log represented by the process "Target Data".

3.2 User's Academic Profile Meta-Schema

Many authors [24, 53–62] have addressed user profile modeling using an ontology. These works have been focused on developing a general user ontology with applications related to Personal Information Management (PIM), context-aware adaptive service platforms for mobile communication and information services, functional settings of mobile devices according to the user

situation, among others. However, ontologies focused on addressing academic issues, such as researcher profiles, have been scarcely tackled [60–62]. Therefore, the proposal and development of an ontology consisting of unified, consistent, and updated information about the user’s academic profile on the Web is proposed as a field of study for this research work.

Thereby, for the first component of AkCeL, the design and development of a user’s academic profile meta-schema are contemplated, which will be established as a static schema once its constitution is defined. Firstly, this definition is based on the individual schemata of the information sources selected in this research work as well as on the characteristics related to the user profiling proposed in both the Human Resources area and the Computer Science contributions. Secondly, in order to allow the integration of this meta-schema into other semantic platform/s/systems, a schema matching is carried out manually between the raw properties constituting this meta-schema and the ontologies’ classes and properties chosen for this process. To finally generate the user’s academic profile meta-schema named as Academic SUP (meta-schema user’s profile). The phases comprising this process are shown in Figure 3.2 and are described below.

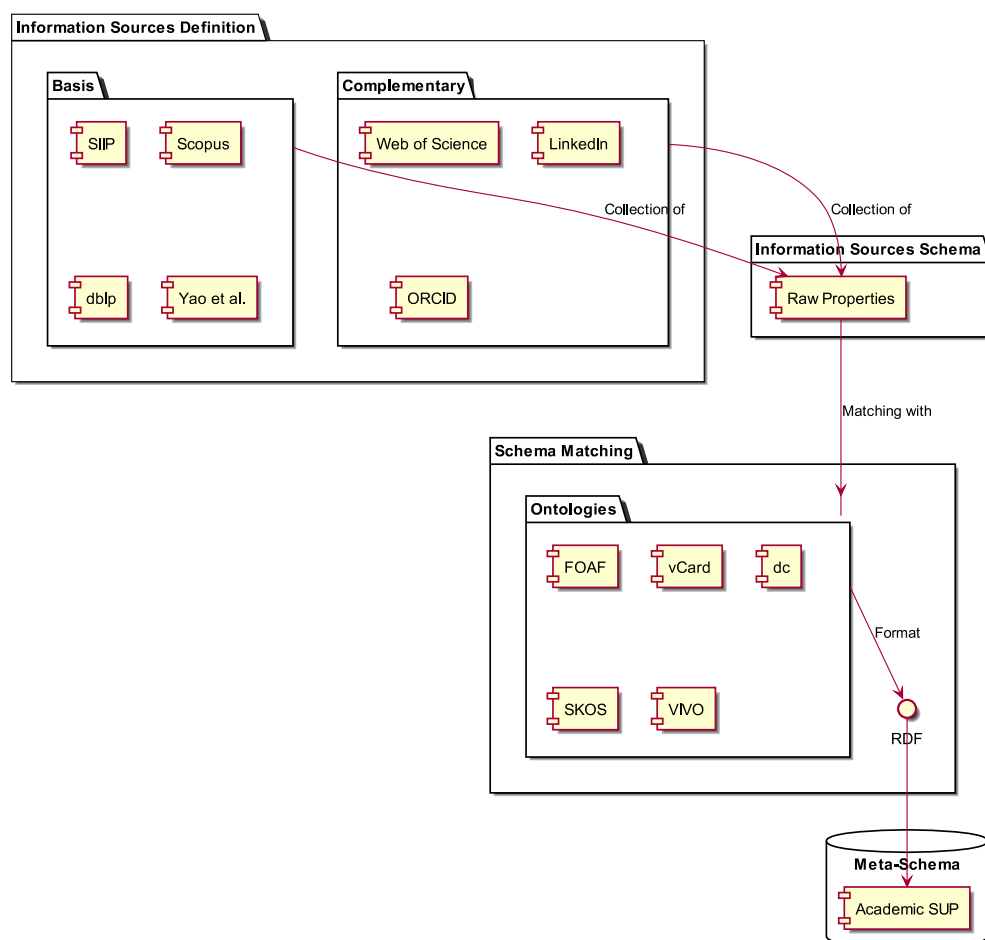


FIGURE 3.2: Graphic representation of the schema matching process.

3.2.1 Information Sources Definition and Schema

This first phase aims to establish the information sources that will be examined to define the user's academic profile meta-schema. These sources are classified according to the main information sources addressed in this research work (basis) to retrieve and validate the information used in this proposal as well as the complementary information sources supplementing this research.

Therefore, the information sources that are defined as the base sources of this phase are SIIP, Scopus, dblp, and Yao *et al.* [60]. As previously mentioned, such a definition is based on the sources used to retrieve and validate the information handled in the processes of this proposal in order to concentrate the available properties/fields in each information source. The SIIP repository is considered as the main basis for this meta-schema since this system consolidates different databases with information belonging to the Tecnológico de Monterrey related to academic issues [64]. In the same way, Scopus and dblp are established as a basis for this meta-schema because the datasets used to carry out the experiments proposed in this research are collected from these sources. Similarly, Yao *et al.* [60]'s work is comprehended in these main sources as these researchers also focused on the definition of a researcher profile schema, extending the FOAF ontology, one of the ontologies considered in this proposal for the schema matching process.

Regarding the complementary information sources, Web of Science, LinkedIn, and ORCID (Open Researcher and Contributor Identifier) are chosen to supplement the meta-schema proposed in this work since initially, Web of Science and LinkedIn had been considered as information sources for the data retrieval process. However, for Web of Science, an institutional permission to access the data is required [65] and for LinkedIn, a user's explicit permission to access his/her profile is required [66], to mention a few barriers described in Section 3.3.1. Therefore, the study of the schemata of these sources is contemplated in the design of the proposed meta-schema, considering the inclusion of the data provided by these information sources as future work. In the case of ORCID, this source was explored because this identifier is included in both dblp [67] and Web of Science [65], also considering its possible addition to the schema in a future version.

Hence, the properties included in the SIIP repository, retrieved from a VIVO's deployment [68], are presented in Table 3.1 while in Table 3.2, other properties obtained from the information available in the selected information sources are introduced in order to consider its incorporation

into the proposed schema. Thereby, the fields obtained from the dataset belonging to Scopus, the dataset extracted through a service linked to an institutional RDB from ITESM [69]; the properties identified from the whole dblp dataset downloaded on 2018-04-12 [67] and from the publication authored by Ley [70]; as well as the properties provide by Yao *et al.* [60] in their researcher profile schema proposed were manifested in both Table 3.1 and Table 3.2. Similarly, the properties that can be gathered from Web of Science through its web service named as Web Services Expanded [65], the properties shown in LinkedIn by accessing from a registered account [71], and the properties contained in the ORCID public data file downloaded on 2018-10-11 [72] were incorporated into these Tables.

For both Tables (3.1 and 3.2), the occurrences of each property according to the specified information source as well as its total frequency of occurrence are provided to support the selection of the properties that will be included in the proposed meta-schema. This decision is determined according to the total frequency of occurrence of each property and is delimited by the information that can be retrieved or inferred from the information sources that are used for this purpose. That is, firstly, the definition of the schema for this research work is based on the total frequency of a field or property. Thus, considering that seven information sources were selected to build this meta-schema, the properties having a total frequency of occurrence equal to or greater than four are integrated into Academic SUP. As a result, according to the previous statement and Tables 3.1 and 3.2, the properties taken into account to constitute the proposed meta-schema are shown in Table 3.3.

Secondly, by following additional considerations to avoid the redundancy of data and allow both data inference and data recommendation, a second version of this meta-schema is provided in Table 3.4. That is, on the one hand, the “AuthorFamilyName” and “AuthorGivenName” properties were not incorporated into the second version of Academic SUP since these properties cannot be retrieved from the base information sources for this work and the “AuthorFullName” property contains the same data. In the same way, the “doi” property was not maintained in this version because three of four base sources did not include it; however, this property can be contemplated for the third version of Academic SUP as further work. On the other hand, the “Languages” property was added to the second version of Academic SUP because this property could be inferred from the publications’ data related to each researcher, although it only has three occurrences. Similarly, “AffiliationAddress” and “AffiliationCity” were integrated into this second version as the values of these properties support the data recommendation process.

TABLE 3.1: Properties provided by each information source (Part I).

Field	SIIP [68]	Scopus [69]	dblp [67, 70]	Yao <i>et al.</i> [60]	Web Of Science [65]	LinkedIn [71]	ORCID [72]	Frequency
Position	✓			✓		✓	✓	4
Email	✓			✓	✓	✓		4
AuthorFullName	✓	✓	✓	✓	✓		✓	6
AuthorFamilyName	✓				✓	✓	✓	4
AuthorGivenName	✓				✓	✓	✓	4
AffiliationName	✓	✓		✓	✓	✓	✓	6
ResearchInterests	✓	✓		✓	✓		✓	5
PublicationTitle	✓	✓	✓		✓	✓	✓	6
SourceTitle	✓	✓	✓		✓	✓	✓	6
ArticlePublicationYear	✓	✓	✓		✓	✓	✓	6
BsDate	✓			✓		✓	✓	4
BsUniversity	✓			✓		✓	✓	4
BsMajor	✓			✓		✓	✓	4
MsMajor	✓			✓		✓	✓	4
MsUniversity	✓			✓		✓	✓	4
MsDate	✓			✓		✓	✓	4
PhdMajor	✓			✓		✓	✓	4
PhdUniversity	✓			✓		✓	✓	4
PhdDate	✓			✓		✓	✓	4
NumberOfPublications	✓							1
AuthorPhotography	✓			✓		✓		3

TABLE 3.2: Properties provided by each information source (Part II).

Field	SIIP [68]	Scopus [69]	dblp [67, 70]	Yao <i>et al.</i> [60]	Web Of Science [65]	LinkedIn [71]	ORCID [72]	Frequency
AffiliationAddress				✓	✓	✓		3
AffiliationCountry		✓			✓	✓	✓	4
AffiliationPhone				✓		✓		2
Homepage				✓		✓		2
AuthorDescription						✓		1
Role					✓		✓	2
BirthDate						✓		1
AffiliationFax				✓				1
Citations		✓						1
AuthorOtherIds		✓			✓		✓	4
Employment						✓	✓	2
Languages					✓	✓	✓	3
AffiliationZipCode					✓	✓		2
AffiliationCity		✓			✓		✓	3
PublicationKeywords					✓			1
doi					✓	✓	✓	4
Funding			✓		✓			1

TABLE 3.3: First version of Academic SUP.

Field	Frequency	Field	Frequency
AuthorFullName	6	AffiliationName	6
PublicationTitle	6	SourceTitle	6
ArticlePublicationYear	6	ResearchInterests	5
Position	4	Email	4
AuthorFamilyName	4	AuthorGivenName	4
BsDate	4	BsUniversity	4
BsMajor	4	MsMajor	4
MsUniversity	4	MsDate	4
PhdMajor	4	PhdUniversity	4
PhdDate	4	AffiliationCountry	4
AuthorOtherIds	4	doi	4

TABLE 3.4: Second version of Academic SUP.

Field	Frequency	Field	Frequency
AuthorFullName	6	AffiliationName	6
PublicationTitle	6	SourceTitle	6
ArticlePublicationYear	6	ResearchInterests	5
Position	4	Email	4
BsDate	4	BsUniversity	4
BsMajor	4	MsMajor	4
MsUniversity	4	MsDate	4
PhdMajor	4	PhdUniversity	4
PhdDate	4	AffiliationCountry	4
AuthorOtherIds	4	AffiliationAddress	3
Languages	3	AffiliationCity	3

Finally, in order to enrich the properties defined for the second version of Academic SUP and at the same time, to support the integration of this meta-schema into other semantic developments, a schema matching process is carried out in this component as described in the next Section.

3.2.2 Schema Matching: Ontologies

A critical issue in many applications aimed at data integration is schematic heterogeneity since it leads to the lack of interoperability. Therefore, the main objective of a schema matching process is to find the mapping and matching (relationship) between the elements of two schemata or ontologies, one of them as the source schema and the other as the target schema [73].

Due to the naming conflicts, levels of abstraction, among other aspects presented in schemata, a schema matching process cannot be completely done automatically. For this reason, manual

approaches to a specific domain as well as more general semi-automatic models, methods, and prototypes have been proposed in the literature [73]. For this work, a manual schema matching process was carried out since a set of well-defined properties were previously identified in the academic domain and the schemata of the source ontologies for VIVO [74, 75] were considered as a guide to establish the relationships between the properties.

Then, to carry out the schema matching process, firstly, ontologies widely used and associated with academic resources, such as scientific publications, as well as with general user profiles and researcher profiles were reviewed. The objective of this review is to select a set of ontologies (target schema) that allow to standardize and enrich the raw properties defined for Academic SUP (source schema). Hence, from this step, the resulting ontologies are FOAF, vCard, dc, SKOS, and VIVO, which are briefly defined in Appendix A.4. In the same way, to validate the selection of these ontologies, the source ontologies for VIVO [74, 75] were consulted, verifying the use of these ontologies in this platform.

Once the target schema was established, the raw properties defined in Table 3.4 were divided, extended, and organized with other related properties and used in the processes defined in Ak-CeL to establish the final source schema for this process, such properties are shown in Table 3.5. For example, the “ResearchInterests” property was identified by “Areas_interest” and supplemented by “Areas_disinterest”. Subsequently, each raw property was manually mapped with the respective class or property belonging to the ontologies selected for this schema matching process, starting with the schemata of the source ontologies for VIVO [74, 75]. Finally, the matches provided in Table 3.5 were determined.

Lastly, in order to describe the latest version of Academic SUP in this work, the definition of the Academic SUP’s classes and properties, as well as the hierarchy between these classes and properties are provided in the next Section.

3.2.3 Academic SUP (Meta-Schema User’s Profile)

Academic SUP (meta-Schema User’s Profile) is designed and developed as an ontology in order to be an interoperable structure of the user’s academic profile on the Web and, at the same time, support the storage of the user’s known (initial) information, as well as new data in accordance with this knowledge representation. For these reasons, the definition of Academic SUP was based on both the individual schemata of the information sources selected in this research work

TABLE 3.5: Schema matching between raw properties and the ontologies' classes and properties.

Raw Property	Ontologies' Classes and Properties
Identification	foaf:Person
Name_author	vcard:Name
Email_affiliation	vcard:Email
Id_SNS	
Scopus	vivo:scopusId
Language_researcher	vcard:Language
hasEducation	vivo:EducationalProcess
hasProfession	vivo:Relationship
hasOrganization	foaf:Organization
hasInterest	skos:Collection
Education	vivo:EducationalProcess
Degree	vivo:AwardedDegree
Name_degree	vivo:AcademicDegree
Bs	
Ms	
Phd	
Date_degree	vivo:DateTimeInterval
hasOrganization	foaf:Organization
Interest	skos:Collection
Type_interest	skos:OrderedCollection
Research	skos:memberList
hasPublication	dcterms:BibliographicResource
hasPreference	skos:Collection
Publication	dcterms:BibliographicResource
Title_publication	dcterms:title
Name_source	dcterms:source
Date_publication	dcterms:date
Preference	skos:Collection
Type_preference	skos:OrderedCollection
Areas_interest	skos:memberList
Areas_disinterest	skos:memberList
Organization	foaf:Organization
Name_organization	vcard:organization-name
Address_organization	vcard:Address
City_affiliation	vcard:locality
Country_affiliation	vcard:country
Date_affiliation	vivo:DateTimeInterval
Profession	vivo:Relationship
Name_position	vivo:Position

and the user profiling standards and contributions stated in Social Science as well as Computer Science.

According to Alles [52], a profile or an anti-profile can be described in the profile definition

process. For the Human Resources area, the anti-profile defines a profile focused on the description of a person while the profile addresses the information related to a work position as well as the model of competencies defined by the organization [52]. However, for the purposes of this research, it is proposed that the profile is considered as the description of a researcher (the anti-profile definition), including his/her preferences, while the anti-profile is focused on the researcher's disinterests.

As a result, the first consolidated representation of Academic SUP is proposed in accordance with the works provided by the Human Resources area [52] as well as by the Computer Science area [24, 53–62], the information sources consulted for its definition [65, 67–72], and the ontologies enriching this meta-schema [74–82]. The description of Academic SUP is provided below. Firstly, the definition of the classes constituting Academic SUP is presented. Secondly, the properties composing these classes are defined in Table 3.6. Finally, the hierarchy's representation of all the Academic SUP's classes and properties is shown in Table 3.7.

Description of the Academic SUP's classes

- SuperClass: Researcher.
 1. Class: Identification. Data related to the identification and description of the researcher. For instance, the Scopus identifier assigned to the researcher in this information source.
 2. Class: Education. Issues concerning the educational history of the researcher. For example, the degree's name obtained by the researcher.
 3. Class: Interest. Data related to the types of interest of the researcher, such as a hobby or work-related interests. Specifically, research as a type of interest for the researcher.
 4. Class: Publication. Entity concerning with the publications' data made by the researcher. For example, the publication's title.
 5. Class: Preference. Data related to the types of preference of the researcher, such as the areas of interest.
 6. Class: Organization. Entity concerning with the organization's data where the researcher is affiliated, such as the organization's name.
 7. Class: Profession. Data related to the researcher's position within an organization, such as the position's name.

TABLE 3.6: Definition of the Academic SUP's subclasses and properties.

Property	Domain	Range
Name_author	SubClass of Identification	String
Email_affiliation	SubClass of Identification	String
Id_SNS	SubClass of Identification	Enumerate [Scopus]
Scopus	Property of Id_SNS	Number
Language_researcher	SubClass of Identification	String
hasEducation	SubClass of Identification	Instance [Education]
hasProfession	SubClass of Identification	Instance [Profession]
hasOrganization	SubClass of Identification	Instance [Organization]
hasInterest	SubClass of Identification	Instance [Interest]
Degree	SubClass of Education	Boolean
Name_degree	Property of Degree	Enumerate [Bs, Ms, Phd]
Bs	SubProperty of Name_degree	String
Ms	SubProperty of Name_degree	String
Phd	SubProperty of Name_degree	String
Date_degree	Property of Degree	DateTime
hasOrganization	Property of Degree	Instance [Organization]
Type_interest	SubClass of Interest	Enumerate [Research]
Research	Property of Type_interest	Instance [Publication], Instance [Preference]
hasPublication	SubProperty of Research	Instance [Publication]
hasPreference	SubProperty of Research	Instance [Preference]
Title_publication	SubClass of Publication	String
Name_source	SubClass of Publication	String
Date_publication	SubClass of Publication	DateTime
Type_preference	SubClass of Preference	Enumerate [Areas_interest, Areas_disinterest]
Areas_interest	Property of Type_preference	String
Areas_disinterest	Property of Type_preference	String
Name_organization	SubClass of Organization	String
Address_organization	SubClass of Organization	String
City_affiliation	SubClass of Organization	String
Country_affiliation	SubClass of Organization	String
Date_affiliation	SubClass of Organization	DateTime
Name_position	SubClass of Profession	String

These areas, information sources, and ontologies were considered aim to achieve the goal of developing a unified, interoperable, and standardized ontology, incorporating representative features that allow users to be described in an academic environment and that support educational institutions to carry out their research statistics.

TABLE 3.7: Academic SUP's classes and properties

Super-Class	Class	SubClass	Property	SubProperty	
Researcher	Identification	Name_author			
		Email_affiliation			
		Id_SNS	Scopus		
		Language_researcher			
		hasEducation			
		hasProfession			
		hasOrganization			
	Education	Degree	Name_degree		Bs
					Ms
					Phd
			Date_degree		
		hasOrganization			
	Interest	Type_interest	Research	hasPublication	hasPreference
	Publication	Title_publication			
		Name_source			
		Date_publication			
	Preference	Type_preference	Areas_interest		
			Areas_disinterest		
	Organization	Name_organization			
		Address_organization			
		City_affiliation			
		Country_affiliation			
		Date_affiliation			
	Profession	Name_position			

3.3 Dynamic Data Enrichment

The knowledge component of AkCel is focused on the consistent collection and integration of data into the user's academic profile schema. Therefore, a workflow consisting of the data retrieval, preprocessing, and alignment processes has been proposed in order to carry out this objective. All of the processes that make up this workflow are described below.

3.3.1 Data Retrieval

The data retrieval process aims to collect the data related to the classes and properties defined in Academic SUP from the information sources selected for this research work. Therefore, on the one hand, a selection of information sources as well as computational solutions for extracting

data must be performed. On the other hand, since the information retrieved can come in different encoding formats or standards, a data preprocessing process must be considered. Such activities are shown in Figure 3.3.

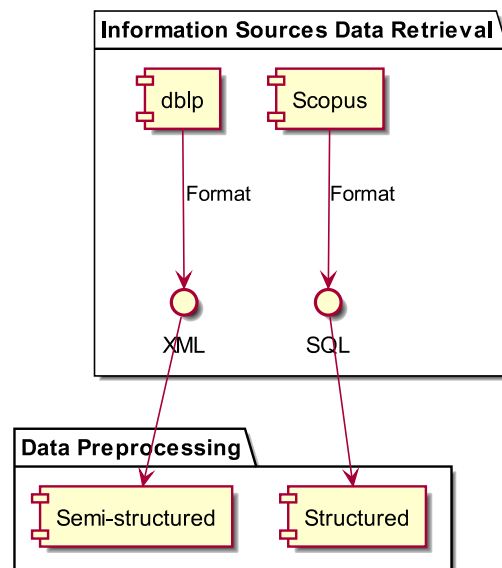


FIGURE 3.3: Graphic representation of the data retrieval process.

Firstly, the description of the selected information sources and the statistics of the information collected are provided. Subsequently, the explanation of the data preprocessing carried out on the retrieved information is presented.

3.3.1.1 Information Sources Definition: Data

To select the information sources used in this process, the following specifications must be taken into account: the information sources must be OA (preferably), in case of being private, the permits of accessing the data must be provided; the information sources must be focused on academic or research information; the information sources must be reliable; and the information sources must be widely used by the academic area on the Web. At first, LinkedIn [22], dblp, Web of Science, and Scopus were choice according to the aspects mentioned above. However, in the case of LinkedIn, since the number of calls to extract information is limited and a user's explicit permission to access his/her profile is required [66], the information collected from this information source could be incomplete. Similarly, in the case of Web of Science, an institutional permission to access the data is required [65], which is beyond the scope of the author of this proposal at this time. Consequently, LinkedIn and Web of Science are not included in this

process, which leads to only two information sources are selected for this process: dblp and Scopus.

3.3.1.1.1 dblp

This information source is a joint data service of the University of Trier and Schloss Dagstuhl focused on providing online open bibliographic information on major computer science journals and proceedings. That is, the scope of dblp are publications from computer science only, addressing mainly international publications, *i.e.*, publications in English language (although there are exceptions to this rule). Some of the topics covered by dblp are Artificial Intelligence, Computational Biology, Data Mining, Decision Support, Machine Learning, Semantic Web, among others [67].

In addition, since the dblp's data are released under the Open Data Commons ODC-BY 1.0 license, a free access to copy, distribute, use, modify, transform, build upon, and produce derived works from them is provided [67]. Therefore, the data retrieval process from dblp is can carried out in the manner specified in the license. On the one hand, the whole dblp dataset can be downloaded as a XML (Extensible Markup Language) file [67], a brief definition of XML is provided in Appendix A.1. On the other hand, certain information from dblp can be extracted through a primitive dblp query API based on requests [67, 83]. In the same way, dblp allows that its website can be crawled until a certain number of requests and following their recommendations [67].

Therefore, firstly, to obtain whole dblp dataset, the XML file (dblp.xml) containing all bibliographic records is downloaded manually. Such a file will be used to carry out the experiments proposed in this proposal and is constituted of the statistics shown in Table 3.8 when downloading it (retrieved on 2018-04-12) [67]. Afterwards, a crawler developed to collect the information of only a specified user (researcher) is provided to support the data update process. Thereby, only the information required is retrieved.

TABLE 3.8: Statistics of the dataset downloaded from dblp (retrieved on 2018-04-12).

Statistics	Number
Publications	4,121,310
Authors	2,069,135
Conferences	5,386
Journals	1,571

3.3.1.1.2 Scopus

This information source is an abstract and citation database of peer-reviewed literature, such as scientific journals, books, and conference proceedings, in the fields of science, technology, medicine, social sciences, and arts and humanities. Scopus covers publications from all geographical regions, including non-English titles as long as English abstracts can be provided with the articles. In addition, Scopus provides smart tools to track, analyze, and visualize research [84].

Scopus APIs expose curated abstracts and citation data from all scholarly journals indexed by the Elsevier's abstract and citation database: Scopus. The use of these APIs is tied to specific use cases, such as Academic Research, allowing to have a direct access to real-time Scopus data, to perform an easy integration with client applications and/or directly with client websites, among others benefits [85]. However, the user of these APIs has a limited access to a basic metadata for most citation records, as well as to basic search functionality; and for academic research use case, no mining of the entire Scopus dataset is permitted [85], which can be an issue for some fields of application. Although, a full Scopus APIs access is granted to customers with Scopus subscription [85].

To use Scopus APIs is required to request an API Key that can be obtained from Elsevier Developer Portal as well as to respect the policies for using APIs. Some examples of these APIs are Abstract Retrieval API, Affiliation Search API, and Author Search API [85]. For this research work, the dataset belonging to Scopus is extracted through a service linked to an institutional RDB from ITESM [69]. Hence, SQL is the language used, a brief definition of SQL is provided in Appendix A.1. Because confidentiality and privacy agreements some statistics and information related to researchers belonging to this information source are not explicitly provided. Therefore, the information presented in this proposal has been reviewed and allowed for publication. Consequently, the statistics shown in Table 3.9 (retrieved from 2018-03-21 to 2018-03-23) represent only the information downloaded to carry out the experiments described in this proposal since a larger number of records has been stored in the institutional RDB queried.

TABLE 3.9: Statistics of the dataset downloaded from the service linked to Scopus (retrieved from 2018-03-21 to 2018-03-23).

Statistics	Number
Publications	10,074
Authors	39,162
Journals, books, and conference proceedings	1,258
Retrieved records	1,915,934

3.3.1.2 Data Preprocessing

The data preprocessing process aims to transform the data collected into the format required by the data alignment process as well as the data recommendation process in order to support the execution of the proposed steps for these processes. In addition to transforming the gathered data, other types of preprocessing are also considered at this phase, such as data cleaning, integration, and reduction [86]. A brief definition of the types of preprocessing is provided in Appendix A.2. Hence, the next data preprocessing is carried out semi-automatically on the downloaded datasets, when the files are handled as plain text files.

dblp

1. UTF-8 format. The file resulting is saved again in a new XML file with UTF-8 format because special characters were presented in the retrieved information.
2. Special characters. To transform special characters found in some records, since this file contains characters in HTML (HyperText Markup Language), an HTML code dictionary, available in [87], is integrated into this process. For example, the words “Schönhage” and “über” are transform into “Schonhage” and “uber”, respectively.
3. Record identified by “Computer Science Curricula 2013”. This record is made up of all the authors’ names who have a publication indexed in dblp. Therefore, the elimination of this record provides benefits for the subsequent processes of AkCeL since, for example, it reduces the time of the data alignment process.

Scopus

1. UTF-8 format. Although the UTF-8 format was specified when downloading the dataset, the file resulting is saved again in a new csv file with UTF-8 format because special characters were presented in the retrieved information.

2. Duplicated records. To remove the duplicated records a condition composed by user id, publication id related to user id, research area id related to publication, and category id related to publication is integrated into this process. Therefore, a searching for coincidences is performed through all the dataset in order to keep only one record by each duplicated record. As a result, a dataset constituted by 39,162 records is obtained (from 1,915,934 records retrieved initially).
3. Special characters. To transform special characters found in some records, the Python's `ftfy` library [88, 89] is added to this process. For example, the words "Verduzco-MartÃ­nez" and "GonzÃ¡lez-SÃ¡nchez" are transformed into "Verduzco-Martínez" and "González-Sánchez", respectively.
4. Orthographic accent. To fix orthographic accents found in some words, the Python's `ftfy` library [88, 89] is used again in this process. For example, the words "Verduzco-Martínez" and "González-Sánchez" are fixed into "Verduzco-Martinez" and "Gonzalez-Sanchez", respectively.
5. Down exclamation mark. To remove down exclamation marks found in some words, a replacement process is incorporated to replace ¡ with an empty space. For example, the words "Gonzalez-Sanchez" are transformed into "Gonzalez-Sanchez", respectively.
6. File delimiter. To keep consistency the datasets used in the recommendation process, the comma delimiter is replaced with the semicolon delimiter.

3.3.2 Data Alignment

The data alignment process is responsible for achieving the consistency in the preprocessed data of the information sources defined for this proposal when they are integrated into Academic SUP. For this purpose, some data alignment techniques are applied to resolve the redundancy or ambiguity in the data; for example, from setting the publication date to a common time format to determine whether the data gathered belongs to both the query context and the specified user [13, 29]. For this proposal, in order to achieve consistency in the integration of the collected and preprocessed data from `dbpl` and `Scopus`, a set of production rules of the type "if-then" is proposed. Thereby, the aligned information can be stored in Academic SUP, according to the structure mentioned in Section 3.2.3, and then, it can be visualized through a visualization platform adopted for this research.

For this alignment process, the information to be aligned is concerning publications. That is, equal publications are integrated as one instance (redundancy of information), while different publications are directly integrated as a new instance into Academic SUP (lack of information). For this purpose, the author's name, the publication's title, and the publication's source are key properties to carry out this alignment. Since the author's name could vary in each publication made by the same researcher, these variants have been related to the corresponding researcher. Therefore, although the publications had a different author's name for the same researcher, these can be aligned with the specific researcher according to this relationship.

The description and pseudocodes of the data alignment process proposed for this research are provided below.

1. Extraction of researchers affiliated with a certain institution. For this research, the researchers' names affiliated with Tecnológico de Monterrey, who are in Scopus dataset, are retrieved. Scopus is the information source considered as a basis for this alignment process because it contains the largest amount of information related to the researchers. Therefore, the researchers' names as well as their identifiers are extracted. These names are gathered with the original format presented in the Scopus dataset, *i.e.*, without preprocessing.
2. Filtering of the researchers by a specified research area. For this use case, the researchers who have published in the "Computer Science" research area are selected from the Scopus dataset. The filtering of the records belonging to these researchers is based on the research areas related to their publications. This step is omitted when the alignment process is carried out for all researchers in all research areas, that is, without a constraint of research area.
3. Extraction of the different researcher's names as author for the first information source. In this step, the researchers' names, selected in Step 1, are congregated. This activity allows to add the preprocessed researcher's name with his/her respective names without preprocessing, collected in Step 1, through his/her Scopus identifier. This is with the objective of concentrating all the possible names with which one same researcher has published and in this way, contributing to the alignment of his/her publications indexed in Scopus with his/her publications indexed in dblp.
4. Generation of a file containing the Academic SUP's properties and the values for each property according to the first information source selected. Firstly, the Scopus dataset is

formatted according to the properties defined in Academic SUP with the aim of generating a final file of this information source that supports the final file of this alignment process. Secondly, the instances of these properties are retrieved in accordance with the following activities:

- (a) Gathering of the researcher's names with which he/she has published. This activity is based on the names' list resulting from Step 3 in order to obtain the names' list with which a researcher has published. Since the format of the researcher' name is as surname and name, to standardize the names with the format presented in the dblp dataset, the names are reordered as name and surname.
- (b) Collection of interesting research areas for the researcher. For this activity, the research areas of interest for the researchers are gathered from the results obtained in the data recommendation process using the proposed algorithm: C-HyRA.
- (c) Collection of research areas not interesting for the researcher. For this activity, the file containing all the research areas present in the Scopus dataset, Appendix B.1, and the research areas specified as preferred for the researcher are used. Thereby, the research areas different to the areas of interest of the researcher are retrieved as the research areas in which the researcher has no preference.
- (d) Retrieving of information related to the affiliation's organization of the researcher. In this activity, the Scopus dataset and the dataset built for C-HyRA concerning geographical influence are used in order to filter the researchers affiliated with Tecnológico de Monterrey, have the relationship between the researcher' name and the organization's name (to which the researcher is affiliated), and finally, concentrate the information of this organization, such as address, city, and country.
- (e) Assembly of the instances for the properties that can be obtained from the first information source. The Scopus identifier, publication's title, and publication's source (in which the article has been published), as well as the aforementioned data for this step are collected from Scopus.
- (f) Formatting and validation of information retrieved. Finally, once all the data are obtained, these data are arranged according to the properties defined in Academic SUP. To do this, when carrying out the assembly of all the instances, it is validated that the researcher's name and the publication's title do not coincide with another existing record to avoid the redundancy of data. In case a record matches one already stored, the second instance is omitted.

5. Generation of a file containing the Academic SUP's properties and the values for each property according to the second information source chosen. For this activity, the file resulting from the data preprocessing process for dblp is used. The data collected are the date of publication, researcher's name, publication's title, and publication's source. If the publication has more than one author, the authors' names are kept in the same record. In addition, it is important to mention that the publication's title in dblp contains a full stop by default, hence, such full stop was removed to standardize the publication's title with the publication's title in Scopus.
6. Filtering of the authors for the second information source. Since only the records of the researchers affiliated with Tecnológico de Monterrey are required for this research, a search for these researchers is performed in dblp. However, dblp does not have a property related to the affiliation of the author. Therefore, the researchers' names collected from Scopus, in Step 3, are used to filter the records of these researchers in dblp. In this way, only the information of the researchers contained in the names' list obtained from Scopus, who are affiliated with Tecnológico de Monterrey, are retained.
7. Generation of the final file constituted of aligned information for researchers from the defined information sources. For this work, the data to align come from Scopus and dblp, as mentioned throughout the process. Firstly, the researcher's name and the publication's title contained in the dblp dataset are compared with the Scopus dataset. Then, two cases are presented: (1) the researcher's name and the publication's title contained in the dblp dataset do not match one instance of the Scopus dataset, thus, this instance is added to the final file resulting from the data alignment process; and (2) the researcher's name and the publication's title contained in the dblp dataset match one instance of the Scopus dataset, therefore, the record of Scopus is updated with the date of publication and publication's source of the corresponding publication, in their respective fields of the final file. This update does not remove the information in the Scopus dataset but it adds the new information in the corresponding properties, in case the existing information is different from this one. Finally, the update and alignment changes resulting from this process are stored in the final file. In the same way, the original information of both Scopus and dblp are stored as separate files (one file for each information source) in order to support possible future activities (log).

The general procedure corresponding to the preprocessing and filtering steps of the information contained in the Scopus and dblp datasets described in the data alignment process is provided in Algorithm 1.

Algorithm 1 Preprocessing and filtering during alignment.

function *PreprocessDataForAlignment*

$S \leftarrow \text{ScopusDataSet}$

$D \leftarrow \text{DblpDataSet}$

$\text{FilteredScopusDataset} \leftarrow \{\}$

$\text{FormattedScopusDataset} \leftarrow \{\}$

$\text{FormattedDblpDataset} \leftarrow \{\}$

$\text{allResearchersNames} \leftarrow \{\}$

for researcher in S **do**

$\text{researcherNames} \leftarrow \text{NamesOfResearcher}[\text{researcher}]$

$\text{allResearchersNames} \leftarrow \text{allResearchersNames} \cup \{(\text{researcherId}, \text{researcherNames})\}$

if researcher[publicationArea] == “ComputerScience” **then**

$\text{researcherData} \leftarrow S[\text{researcher}]$

$\text{researcherData} \leftarrow \text{researcherData} \cup \{\text{allResearchersNames}[\text{researcher}]\}$

$\text{FilteredScopusDataset} \leftarrow \text{FilteredScopusDataset} \cup \{\text{researcherData}\}$

end if

end for

for data in $\text{FilteredScopusDataset}$ **do**

$\text{formattedData} \leftarrow \text{FormatData}\{\text{data}\}$

$\text{FormattedScopusDataset} \leftarrow \text{FormattedScopusDataset} \cup \{\text{formattedData}\}$

end for

for data in D **do**

if data[user] in $\text{FormattedScopusDataset}$ **then**

$\text{formattedData} \leftarrow \text{FormatData}\{\text{data}\}$

$\text{FormattedDblpDataset} \leftarrow \text{FormattedDblpDataset} \cup \{\text{formattedData}\}$

end if

end for

return $\text{FormattedScopusDataset}, \text{FormattedDblpDataset}$

end function

The general procedure corresponding to the alignment step of the information contained in the Scopus and dblp datasets described in the data alignment process is provided in Algorithm 2.

Algorithm 2 Alignment process.

```

function DataAlignment(FormattedScopusDataset, FormattedDblpDataset)
  ScopusUpdatedDataset  $\leftarrow$  {}
  ScopusDblpDataset  $\leftarrow$  {}
  for dblpData in FormattedDblpDataset do
    for scopusData in FormattedScopusDataset do
      if dblpData[researcherName] == scopusData[researcherName] then
        if dblpData[publicationTitle] == scopusData[publicationTitle] then
          updatedScopusData  $\leftarrow$  scopusData  $\cup$  {dblpData}
          ScopusUpdatedDataset  $\leftarrow$  ScopusUpdatedDataset  $\cup$  {updatedScopusData}
        end if
      end if
    end for
  end for
  for data in FormattedScopusDataset do
    if data not in ScopusUpdatedDataset then
      ScopusDblpDataset  $\leftarrow$  ScopusDblpDataset  $\cup$  {data}
    else
      ScopusDblpDataset  $\leftarrow$  ScopusDblpDataset  $\cup$  {ScopusUpdatedDataset[data]}
    end if
  end for
  for data in FormattedDblpDataset do
    if data not in ScopusDblpDataset then
      ScopusDblpDataset  $\leftarrow$  ScopusDblpDataset  $\cup$  {data}
    end if
  end for
end function

```

3.4 Data Recommendation

The recommendation process intends to suggest interesting data but unknown for a researcher from his/her academic profile and through the information available in the previously defined information sources. Therefore, one of the main goals of this research work is to propose and to develop an algorithm that suggests a research areas list of relevance for a researcher based on both the information concentrated in Academic SUP and the features considered in the traditional and POI recommendation algorithms. Hence, the proposed recommendation algorithm, called researCh Hybrid Recommendation Algorithm (C-HyRA), considers the user's explicit preferences, the publications' research areas and subject categories, and the affiliation's geographical information to carried out a new suggestion.

The recommendation is performed when a user's new publication is indexed in the information sources. Thus, the proposal of the C-HyRA approach is described as follows: the adaptation of a user-based Collaborative Filtering (CF) algorithm integrating an average aggregation operator constituted of different similarity and distance measures allows dealing with the user preferences; the incorporation of the geographical influence factor into the modified user-based CF algorithm allows addressing the affiliation's geographical information; and the merging of a content-based approach allows handling the publications' subject categories to carried out the recommendation process.

In accordance with the foregoing, on the one hand, this proposal could be classified into the user-based CF systems belonging to the memory-based category from the CF systems [41, 44], as well as into a content-based system [44]. On the other hand, it could be considered as a POI recommendation system [41, 42, 44]. It means, firstly, the proposed approach addresses the user's explicit preferences (ratings) through a user-based CF algorithm that embraces an average aggregation operator integrated by five similarity and distance measures. Secondly, the proposal becomes a content-based system by incorporating the publications' subject categories into the modified user-based CF algorithm. Thirdly, the proposed approach is considered a POI recommendation system by integrating the geographical influence factor into the modified user-based CF algorithm. As a consequence, C-HyRA is referred to as a hybrid recommendation approach by embracing the traditional recommendation approach as well as the POI recommendation approach in its suggestion process, such as presented in Figure 3.4.

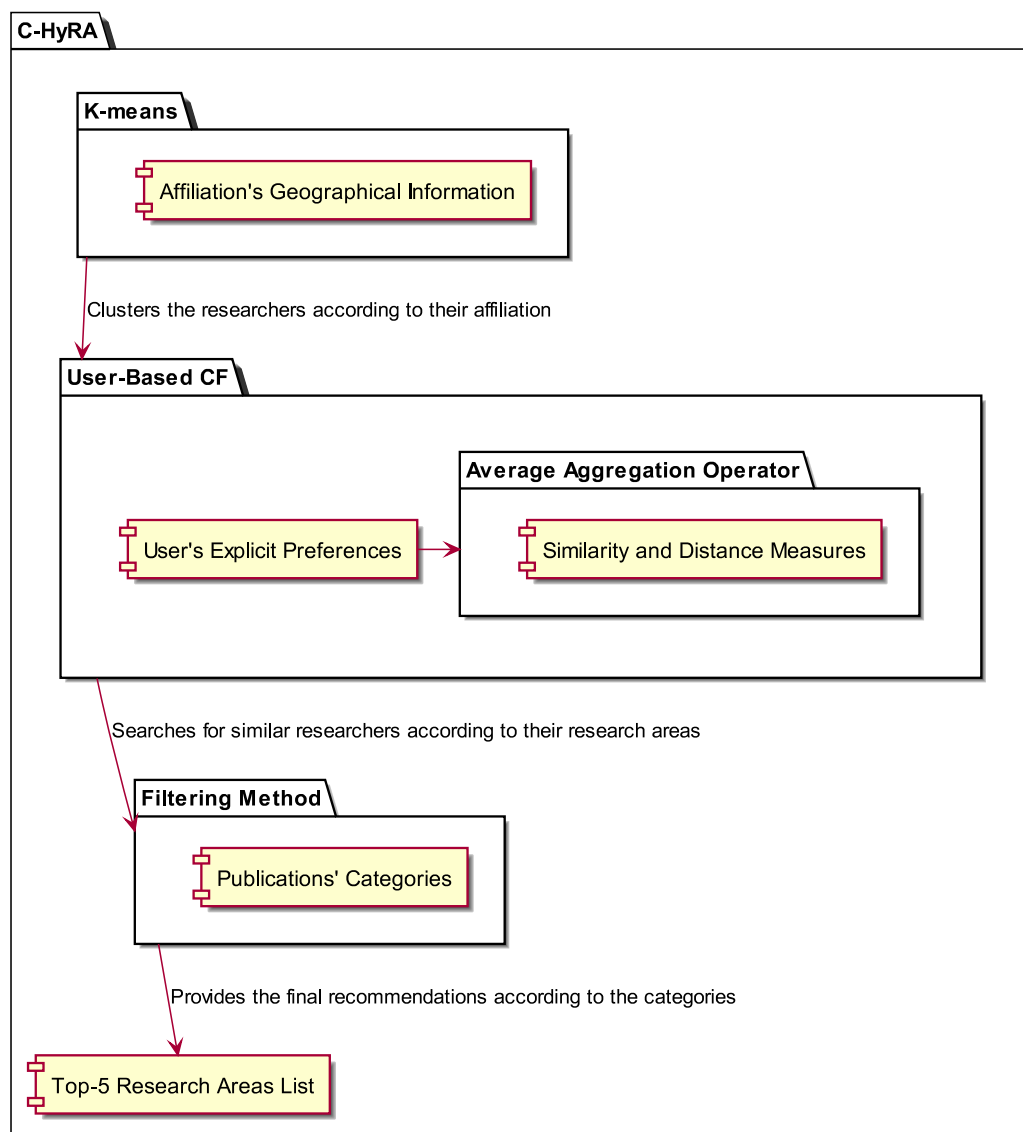


FIGURE 3.4: Graphic representation of the data recommendation process.

The description and pseudocode of C-HyRA are provided as follows. Firstly, the user-based CF algorithm is introduced; secondly, the user-based CF algorithm with the average aggregation operator is described. Then, the user-based CF algorithm with the average aggregation operator complemented with the publications' subject categories is presented; lastly, the C-HyRA's approach, the user-based CF algorithm with the average aggregation operator complemented with the publications' subject categories as well as with geographical influence, is explained.

3.4.1 User-Based CF: Analysis and Description

A traditional recommendation approach is defined for dealing with the user's explicit preferences since in the Academic SUP proposal, the interest and disinterest research areas for a researcher, according to his/her publications, are considered. Therefore, the researcher's preference towards a certain research area is represented by three for "little interesting", four for "interesting", and five for "very interesting", while the "disinterest" is identified by the value one. As a result, an approach based on the user-based CF algorithm is chosen and adapted because it is the one most used by researchers to address the recommendation based on ratings [25]. For this purpose, on the one hand, a function that allows to infer the values of the preferences to the research areas is incorporated. On the other hand, the development reported by Caraciolo [90] is used as a basis for the implementation of the user-based CF algorithm. Additionally, in order to develop C-HyRA, the NumPy package [91], the SciPy library [92], and the Scikit-learn library [93] are used. The description and pseudocode of the user-based CF algorithm are given below.

1. Rating deduction process. Since users (researchers) related to the scope of this research work do not explicitly provide ratings to the research areas explored by them, a function that calculates such ratings is incorporated into the proposed recommendation algorithm. This function is mainly focused on deducting the user's explicit preferences, *i.e.*, the ratings whose values are higher than two. Therefore, the calculated ratings have a range from three to five, being three "little interesting", four "interesting", and five "very interesting". In order to allocate these values, for each user, firstly, the research areas and their categories, where the researcher has published, are extracted. Subsequently, the frequency of occurrence of each research area related to the user's publications is calculated. Then, the ratings are distributed for each area according to its frequency of occurrence. That is, the value five is given to the area whose frequency of occurrence is the highest, the value three is provided to the area whose frequency of occurrence is the lowest, and the value four is assigned to the area whose frequency of occurrence is between the highest frequency and the lowest frequency. In case the frequency of occurrence of several areas is the same, the corresponding value is assigned to each one of them. Finally, to denote the research areas not explored (or "not interesting") for a certain researcher, this function also allocates the value one to these areas. Algorithm 3 provides the pseudocode of this phase.

2. Data gathering process. The dataset that serve as input for C-HyRA is loaded: the ratings calculated for each research area according to the specified user (dataset described in Section 5.2.3.1). Subsequently, a ground-truth subset is built from the dataset that concentrates all ratings calculated for each research area. For this purpose, from one to seven rated research areas from each user are randomly extracted to compose the ground-truth subset, where only research areas whose rating values oscillate between three to five are conserved. This with the aim of leaving behind research areas that are not interesting for a user and that are represented by a rating of one. By setting to seven the maximum number of research areas that can be extracted, a total maximum of 58.33% of the rated research areas of each user is retained to represent their preferences. This ground-truth subset is taken as if the user had only rated this number of research areas. The remaining research areas of each user are used as not rated research areas that can be recommended by C-HyRA. These remaining research areas and their rated values are kept in a separate pilot subset to compare such research areas against the research areas recommended by C-HyRA. Algorithm 4 provides the pseudocode of this phase.
3. Compute similarities between users. A comparison between a user with the rest of them is performed to obtain the N users who have most similar preferences with him/her. The rationale behind this is that users who have similar values to a certain user share similar preferences [25]. Thus, it is more likely that the research areas recommended by these similar users matches the preferences of the specified user. To find those users that share analogous preferences with a specific user, a paired comparison of their ratings of research areas is carried out. This comparison iterates through each available user in the dataset to retain all research areas that are presented in the preferences of both users. Then, the ratings of the two users' research areas are compared by using one distance or similarity measure. Independent experiments are carried out using the following measures: Euclidean distance, Pearson correlation, Cosine similarity, Manhattan distance, Chebyshev distance, Spearman correlation, Bray–Curtis distance, Canberra metric, and Squared Euclidean distance. After the paired comparison, a descending list of similarity values among users is obtained by each distance or similarity measure. The similarity values closer to 1 indicate that both users share more preferences in common, while similarity values closer to 0 express the opposite. Algorithm 5 provides the pseudocode of this process.
4. Recommend research areas. For each available user in the dataset, excluding the user that is selected for giving recommendations, the research areas that the selected user has

not explored are extracted. Then, each research area not explored is ranked through a weighted mean. The weighted mean contemplates the rating of research area and the similarity value of the user that has been compared with the selected user. Consequently, a descending list of N ranked research areas is obtained. From this list, the Top-5 research areas are recommended to the specific user. As a result, only the research areas that could be interesting for the specific user are recommended.

3.4.2 User-Based CF with the Average Aggregation Operator

1. Rating deduction process. Same process as described in Section 3.4.1 and presented in Algorithm 3.
2. Data gathering process. Same process as described in Section 3.4.1 and presented in Algorithm 4.
3. Compute similarities between users. Same process as described in Section 3.4.1 and presented in Algorithm 5.
4. Recommend research areas. For each available user in the dataset, excluding the user that is selected for giving recommendations, the research areas that the selected user has not explored are extracted. Then, each research area not explored is ranked through a weighted mean. The weighted mean contemplates the rating of research area and the similarity value of the user that has been compared with the selected user. Consequently, a descending list of N ranked research areas is obtained. This process is carried out for all similarity and distance measures described in Appendix A.3. Thus, five descending lists of N ranked research areas are computed. Afterwards, the frequency of appearance of all research areas embraced in these descending lists is calculated with the objective that all frequencies of the research areas are averaged by the total number of measures used. Lastly, the Top-5 research areas from the final descending list are recommended to the specific user.

3.4.3 User-Based CF with the Average Aggregation Operator + Publications' Categories

1. Rating deduction process. Same process as described in Section 3.4.1 and presented in Algorithm 3.

2. Data gathering process. Same process as described in Section 3.4.1 and presented in Algorithm 4. Furthermore, the publications' categories dataset is loaded. Such a dataset is described in Section 5.2.3.2.3.
3. Compute similarities between users. Same process as described in Section 3.4.1 and presented in Algorithm 5.
4. Recommend research areas. Firstly, all categories of the research areas explored by the user selected to give recommendations are extracted. Then, the research areas' categories are ranked according to their frequency of appearance. Thus, a descending frequency list of the research areas' categories is obtained. Thereupon, for each similarity or distance measure, a list of similar users is obtained, who best resemble the specified user according to the procedure described in the previous step. Afterwards, the categories of each research area present in the preferences of each similar user are ranked according to the descending frequency list of the research areas' categories of the specified user. Finally, all research areas of each similar user are sorted to obtain those that better resemble the specified user preferences. That is, research areas whose categories are closer to the rated research areas' categories of the specified user are more likely to be recommended. Consequently, a descending list of N research areas ranked by their categories is obtained. From this list, the Top-5 research areas are recommended to the specific user.

3.4.4 C-HyRA

1. Rating deduction process. Same process as described in Section 3.4.1 and presented in Algorithm 3.
2. Data gathering process. Same process as described in Section 3.4.3 and presented in Algorithm 4. Furthermore, the ITESM's Campuses geographical information dataset is loaded. Such a dataset is described in Section 5.2.3.3.
3. Address geographical influence. All Campuses loaded in the data gathering process are clustered using K-means with the Euclidean distance. Because the geographical distribution of Campuses, only three clusters are enough to embrace them all. The calculation of the optimal number of clusters is beyond the scope of this paper. As a result, a list containing the cluster's number to which each Campus belongs is obtained. Then, the

cluster' number of each Campus associated with the specific user is extracted. Subsequently, the clusters' numbers associated with the chosen user are compared against the clusters' numbers of the Campuses of the rest of users. As a result, the users who share one cluster in common with the specific user are retained. This metric has two purposes: first, to ensure that the Campuses geographically closer to user's location are retained for supporting the recommendation process; and second, to decrease the computational calculations that the recommendation algorithm has to perform. Consequently, a list of users that are affiliated with Campuses geographically closer to the Campus of a given user is obtained.

4. Recommend research areas. Finally, the procedure described in Section 3.4.3 is performed to obtain the Top-5 research areas that are going to be recommended, except step 1. As a result, a research areas list that could be interesting for the specific user is recommended. Algorithm 6 provides the pseudocode of this process.

Algorithm 3 C-HyRA—Rating deduction process.

function *RateResearchersPreferences*
 $R \leftarrow \text{ResearchersInformationDataset}$
 $\text{ratedPublishedAreas} \leftarrow \{\}$
for user in R **do**
 $\text{researchersPublishedAreas} \leftarrow \text{ExtractResearcherAreas}(\text{user})$
 $\text{researchersNotPublishedAreas} \leftarrow \text{ExtractResearcherNoAreas}(\text{user})$
 $\text{frequencyOfEachResearcherPublishedArea} \leftarrow \text{FrequencyresearchersPublishedAreas}$
 $\text{ratedAreasPerUser} \leftarrow \{\}$
for area in $\text{frequencyOfEachResearcherPublishedArea}$ **do**
if $\text{areaFrequency} == \text{MaxValue}(\text{frequencyOfEachResearcherPublishedArea})$
then
 $\text{areaRatedValue} = 5$
 $\text{ratedAreasPerUser} \leftarrow \text{ratedAreasPerUser} \cup \{(\text{area}, \text{areaRatedValue})\}$
else if $\text{areaFrequency} == \text{MinValue}(\text{frequencyOfEachResearcherPublishedArea})$
then
 $\text{areaRatedValue} = 3$
 $\text{ratedAreasPerUser} \leftarrow \text{ratedAreasPerUser} \cup \{(\text{area}, \text{areaRatedValue})\}$
else
 $\text{areaRatedValue} = 4$
 $\text{ratedAreasPerUser} \leftarrow \text{ratedAreasPerUser} \cup \{(\text{area}, \text{areaRatedValue})\}$
end if
 $\text{ratedPublishedAreas} \leftarrow \text{ratedPublishedAreas} \cup \{(\text{user}, \text{ratedAreasPerUser})\}$
end for
for area in $\text{researchersNotPublishedAreas}$ **do**
 $\text{areaRatedValue} = 1$
 $\text{ratedAreasPerUser} \leftarrow \text{ratedAreasPerUser} \cup \{(\text{area}, \text{areaRatedValue})\}$
end for
end for
return $\text{ratedPublishedAreas}$
end function

Algorithm 4 C-HyRA—Data gathering process.

```

groundTruthSubset ← {}
pilotSubSet ← {}
function LoadDatasets(ratedPublishedAreas)
  areas ← AreasOfEachResearchAreaDataset
  categories ← CategoriesOfEachResearchAreaDataset
  locations ← LocationsOfEachTecCampusDataset
  areasCategories ← areas ∪ categories
  for user in ratedPublishedAreas do
    researchersSelectedAreas ← RandomAreasAboveRate3(ratedPublishedAreas(user))
    groundTruthSubset ← groundTruthSubset ∪ {(user, researchersSelectedAreas)}
    pilotSubSet ← pilotSubSet ∪ {user, (ratedPublishedAreas \ researchersSelectedAreas)}
  end for
  return groundTruthSubset, pilotSubSet, areasCategories, locations
end function

```

Algorithm 5 C-HyRA—Compute similarities between users.

```

descListOfSimilarResearchers ← {}
similarityDistances ← {Pearson, Euclidean, Cosine, Manhattan, Chebyshev}
function GetSimilarUsers(specificResearcher, otherResearchers, groundTruthSubset,
  similarityDistances)
  for distance in similarityDistances do
    listOfSimilarResearchers ← {}
    for user in otherResearchers do
      sharedAreas ← GetCommonAreas(specificResearcher, user, groundTruthSubset)
      userSimilarityValue ← CalculateSimilarityMetric(specificResearcher, user,
        sharedAreas, distance)
      listOfSimilarResearchers ← listOfSimilarResearchers ∪ {(user,
        userSimilarityValue)}
    end for
    descendingList ← DescendingSort(listOfSimilarResearchers)
    descListOfSimilarResearchers ← descListOfSimilarResearchers
      ∪ {descendingList}
  end for
  return descListOfSimilarResearchers
end function

```

Algorithm 6 C-HyRA—Recommend research areas.

 $clusterDistances \leftarrow \{Euclidean\}$
function $Recommend(specificResearcher, otherResearchers, groundTruthSubset,$
 $similarityDistances, areasCategories, locations, descListOfSimilarResearchers)$
 $clusters \leftarrow GetClusters(locations)$
 $clusterAffiliationsBySpecificResearcher \leftarrow GetAffiliatedClusters$
 $(groundTruthSubset[specificResearcher], clusters)$
 $mustSharedClusters \leftarrow 1$
 $possibleSimilarResearchers \leftarrow \{\}$
for user in $otherResearchers$ **do**
 $clustersOfAffiliationsOfOtherUsers \leftarrow GetAffiliatedClusters$
 $(groundTruthSubset[user], clusters)$
 $sharedClusters \leftarrow CompareClusters(clusterAffiliationsBySpecificResearcher,$
 $clustersOfAffiliationsOfOtherUsers)$
if $sharedClusters \geq mustSharedClusters$ **then**
 $possibleSimilarResearchers \leftarrow possibleSimilarResearchers \cup \{user\}$
end if
end for
 $ratedCategoriesOfSpecificResearcher \leftarrow$
 $FrequencyRatedCategoriesOfSpecificResearcher(areasCategories[specificResearcher])$
 $descListOfSimilarResearchers \leftarrow GetSimilarUsers\{specificResearcher,$
 $otherResearchers, groundTruthSubset, similarityDistances\}$
 $listOfAreas \leftarrow \{\}$
for user in $descListOfSimilarResearchers$ **do**
for area in user **do**
 $categoriesOfAreas \leftarrow GetCategoryOfArea(area[user])$
 $ratedCategories \leftarrow RateCategories(ratedCategoriesOfSpecificResearcher,$
 $categoriesOfAreas)$
 $listOfAreas \leftarrow listOfAreas \cup \{ratedCategories\}$
end for
end for
 $descendingListOfAreas \leftarrow SortDescending(listOfAreas)$
 $recommendationResults \leftarrow Top5(descendingListOfAreas)$
return $recommendationResults$
end function

3.5 Data Visualization

The goal of this component is allow the visualization and interaction with the information generated by AkCeL. This visualization consists of providing both the user's individual statistics and the research statistics of a specific institution. For this purpose, a platform that matches the aforementioned purposes is embraced. Thereby, the VIVO platform is adopted because the characteristics and capabilities that this platform provides, some of them are mentioned below.

VIVO is a member-supported, open source software and an ontology that supports the recording, editing, searching, browsing, and visualizing of research data [94]. This platform is widely used by institutions around the world, developed to support domains of scholarly activity, and integrated by a collection of ontologies [74, 94]. For instance, VIVO is currently deployed in Tecnológico de Monterrey to facilitate the identification of experts affiliated with the Tecnológico de Monterrey [95], scope of this scenario.

For this use case, the proposed statistics are listed below.

- The researcher's individual profile.
- The researchers' profile classified by the city of affiliation.
- The researchers' profile classified by the research areas as well as by the subject categories.
- Publication statistics per researcher and subject category.
- Publication statistics by institution and subject category as well as by institution and research area.
- Relationships between researchers (author - coauthor).
- Rankings among researchers according to the research area and subject category.
- The most representative research areas for each institution (Campus).
- The least representative research areas for each institution (Campus).

3.6 Conclusions

In this Chapter, the framework (AkCeL) and its components proposed in this research work were described. Firstly, a brief introduction to AkCeL, the UML activity diagram of the AkCeL's workflow, and the assumptions considered to achieve an appropriate functioning were provided. Subsequently, two implicit processes, called "Target Data" and "Data Update", were introduced since these processes support the appropriate functioning of AkCeL by recording specific data of the processes and by allowing the population and updating of the Academic SUP's properties, respectively. Later, each component constituting AkCeL was explained in different Sections as follows.

As the first component of AkCeL, the development of a user's academic profile meta-schema, named as Academic SUP, was described. In this Section, the selection of the information sources and properties used to design Academic SUP, the schema matching process carried out between the raw properties corresponding to the proposed meta-schema and the ontologies' classes and properties defined for this process, and the consolidated version of Academic SUP were presented.

In the knowledge component of AkCeL, called Dynamic Data Enrichment, the data retrieval, preprocessing, and alignment processes were explained. For the data retrieval process, the selection of the information sources used to retrieve the data related to the processes proposed in this research work was described. In the same way, a description of the selected information sources, the computational extraction solutions used, and the number of data collected were provided. Subsequently, an explanation of the data preprocessing performed on the collected data was presented. Lastly, for the data alignment process, the description and pseudocodes of such a process were provided.

In the model management component of AkCeL, the data recommendation process, as well as the research Hybrid Recommendation Algorithm (C-HyRA) proposed to carry out such a process, were described. Then, firstly, the scope and approach of C-HyRA were presented. Subsequently, a graphic representation of the phases constituting C-HyRA was introduced. Finally, an explanation of these phases along with their pseudocodes was provided.

Lastly, the data visualization through the VIVO platform was introduced. In this Section, the VIVO platform, as well as the statistics proposed for the use case addressing the generation of research recommendations and statistics, were described.

In the next Chapter, a use case addressing a proposed recommendation algorithm for the tourism sector is described. This use case is introduced with the objective of validating the approach addressed to C-HyRA from the scope examined for this proposal: POI recommendation systems. Thus, an introduction to this use case and the recommendation algorithm's approach proposed for this sector are presented as well as the experimental scenario, the datasets used, and the results of this scenario.

Chapter 4

Use Case of HyRA: A Hybrid Recommendation Algorithm Focused on Smart POI. Ceutí as a Study Scenario

Nowadays, Physical Web [96] together with the increase in the use of mobile devices, Global Positioning System (GPS), and SNS have caused users to share enriched information on the Web such as their tourist experiences [25]. Nevertheless, generally, tourist guide applications are based on information heavily related to the location, disregarding other types of context information, which can be provided by the user. Consequently, all information is available to all users in touch, leading to the issue known as “information overload” [48] as well as problems of inappropriate suggestions [97]. Such facts entail the need to enhance the user’s individual tourist experience according to his/her preferences and context information.

Hence, the main issue to be addressed is recommending a new point-of-interest (POI) where users might be interested based on their personal preferences and contextual information. On the one hand, such a statement matches with the problem of POI recommendation, *i.e.*, the difficulty of suggesting personalized recommendations of places of interest, such as restaurants and movie theaters, for users [25, 42, 49]. On the other hand, this approach also coincides with one of the benefits of POI recommendation: to help both residents and visitors to explore new and interesting places in a certain area [41].

The POI recommendation systems have just emerged recently [25] as a consequence of the quick development of new location-based technologies. Specifically, this approach will deal with Google's Physical Web technology [96] integrated into a device called Smart Spot, and the concept of Smart Point of Interaction (Smart POI) defined as a smart point of interaction between users (citizens and visitors) and a Smart Spot [98, 99], a technology and a device that will be used for the first time in a research study related to the formalization of a recommendation algorithm in the tourism sector.

Therefore, the objective of the proposed algorithm, called Hybrid Recommendation Algorithm (HyRA), is to recommend a Smart POI list for a user according to the user preferences, the Smart POIs' contextual information such as categories and geographical information, and the characteristics of Smart Spot in conjunction with the definition of Smart POI. The definition of this use case arises in cooperation with ©2016 HOP Ubiquitous S.L. (HOPU) and Universidad Católica San Antonio de Murcia (UCAM) as well as the town council and tourism office representatives from Ceutí within a research fellowship supported by the SmartSDK project. The purpose of this use case is to validate the proposed recommendation algorithm in the tourism sector, since HyRA encodes similar aspects to C-HyRA but with different application area, tourism and research sectors, respectively, such as the modified user-based CF algorithm along with the Smart POIs' categories and the geographical influence factor.

To test HyRA, data belonging to both users and Smart POIs are required. Thus, to collect information related to the POIs from Ceutí, a town belonging to the Región de Murcia in Spain, a research study [4] has been considered. In addition, with the aim of generating a dataset with the users' explicit preferences, two surveys designed in a previous study [4], one in Spanish and the other in English, have been redesigned and disseminated. Lastly, to define the POIs' categories, the All Categories section from the Yahoo! Answers website [100] as well as the description of the POIs from Ceutí have been reviewed. As a result, three datasets have been produced: one dataset composed of 16 Smart POIs, another constituted by the 16 preferences of 200 respondents, and another consisting of 13 categories.

Firstly, in Section 4.1, the description of HyRA is provide. Subsequently, in Section 4.2, the experimental scenarios for this use case are described. Then, in Section 4.3, the datasets used and generated in this approach are addressed. Finally, in Sections 4.4 and 4.5, the results and discussion of the experiments, and conclusions, respectively, are presented.

4.1 The HyRA's Approach

The main objective of HyRA is to recommend a Smart POI list for a user according to the user preferences, the Smart POIs' contextual information such as categories and geographical information, and the characteristics of Smart Spot in conjunction with the definition of Smart POI. Therefore, on the one hand, the proposal of the HyRA's approach is based on both the concept of Smart POI and the Smart Spot device. On the other hand, it is established that user preferences are obtained through ratings given by a user to Smart POIs in Ceutí due to the traditional recommendation algorithm chosen to address them. In addition, the similarity and distance measures that will be used in the average aggregation operator have also been defined. These measures are Euclidean distance, Cosine similarity, Spearman correlation, Pearson correlation, Manhattan distance, Bray–Curtis distance, Canberra metric, Chebyshev distance, and Squared Euclidean distance.

4.1.1 User-Based CF: Analysis and Description

Initially, the HyRA's approach was oriented entirely to the POI recommender systems since the Smart POIs defined for this research are POIs of the heritage from Ceutí. However, considering the characteristics of the technology implemented into a Smart Spot, one main assumption was established.

- In the POI recommendation systems, user preferences are reflected and inferred by the frequency of check-in at locations [25, 49]. For this scenario, such preferences can be obtained from Smart Spot through interaction between it and the user's smartphone. Nevertheless, since Smart Spot constantly emits signals to the mobile devices of the users [99], the user's smartphone can receive all the signals that any Smart Spot emits. Therefore, the interaction between a user and a Smart POI (check-in) does not necessarily indicate interest on that Smart POI, but only that the user is close to it. Consequently, to get the user preferences using Smart Spot, a solution based on the traditional recommendation systems approach was proposed. This is, to have explicitly the ratings for the items [25] considering to Smart POIs as items.

Accordingly, a traditional recommendation approach was defined for dealing with the user's explicit preferences: the user-based CF algorithm. For this end, the five-star rating system

was incorporated into the surveys since it is the online explicit feedback mechanism that allows collecting more feedback from users [101].

Firstly, the user-based CF algorithm is described. Secondly, the user-based CF algorithm with the average aggregation operator is described. Then, the user-based CF algorithm with the average aggregation operator complemented with Smart POIs' categories is introduced. Lastly, the HyRA's approach, the user-based CF algorithm with the average aggregation operator complemented with Smart POIs' categories and with geographical influence, is presented.

1. Data gathering process. The datasets that serve as input for HyRA are loaded: the Smart POIs located in Ceutí (4.3.1) and the users' ratings for each Smart POI (4.3.2). Subsequently, a ground-truth subset was built from the dataset that concentrates all users' ratings for each Smart POI. There are randomly extracted from one to 11 rated Smart POIs from each user to compose the ground-truth subset, where only Smart POIs whose rating values oscillate between three to five are conserved. This with the aim of leaving behind Smart POIs that are not interesting for a user and that are represented with a rating below three. By setting to 11 the maximum number of Smart POIs that can be extracted, a total maximum of 70% of the rated Smart POIs of each user is retained to represent their preferences. This ground-truth subset is taken as if the user had only rated this number of Smart POIs. The remaining Smart POIs of each user are used as not visited (not rated) Smart POIs that can be recommended by HyRA. In addition, these remaining Smart POIs and their rated values were preserved in a separate subset to compare the true rated Smart POIs against the Smart POIs recommended by HyRA.
2. Compute similarities between users. A comparison between a user with the rest of them is performed to obtain the N users who have most similar preferences with him/her. The rationale behind this is that users who have similar values to a certain user share similar preferences [25]. Thus, it is more likely that the Smart POIs recommended by these similar users matches the preferences of the specified user. To find those users that share analogous preferences with a specific user, a paired comparison of their ratings of Smart POIs is carried out. This comparison iterates through each available user in the dataset to retain all Smart POIs that are presented in the preferences of both users. Then, the ratings of the two users' Smart POIs are compared by using one distance or similarity measure. Independent experiments are carried out using the following measures: Euclidean distance,

Pearson correlation, Cosine similarity, Manhattan distance, and Chebyshev distance. After the paired comparison, a descending list of similarity values among users is obtained per each distance or similarity measure. The similarity values closer to 1 indicate that both users share more preferences in common, while similarity values closer to 0 express the opposite.

3. Recommend Smart POIs. For each available user in the dataset, excluding the user that is selected for giving recommendations, are extracted the Smart POIs that the selected user has not visited. Then, each Smart POI not visited is ranked through a weighted mean. The weighted mean contemplates the rating of Smart POI and the similarity value of the user that has been compared to the selected user. Consequently, a descending list of N ranked Smart POIs is obtained. From this list, the Top-5 Smart POIs are recommended to the specific user. As a result, only the Smart POIs that could be interesting for the specific user are recommended.

4.1.2 User-Based CF with the Average Aggregation Operator

1. Data gathering process. Same process as described in Section 4.1.1.
2. Compute similarities between users. Same process as described in Section 4.1.1.
3. Recommend Smart POIs. For each available user in the dataset excluding the user that is selected for giving recommendations, the Smart POIs that the selected user has not visited are extracted. Then, each Smart POI not visited is ranked through a weighted mean. The weighted mean contemplates the rating of Smart POI and the similarity value of the user that has been compared to the selected user. Consequently, a descending list of N ranked Smart POIs is obtained. This process is carried out for all similarity and distance measures described in Appendix A.3 as well as for the Spearman correlation, Bray–Curtis distance, Canberra metric, and Squared Euclidean distance. Thus, nine descending lists of N ranked Smart POIs are computed. Afterwards, the frequency of appearance of all Smart POIs embraced in these descending lists is calculated with the objective that all frequencies of the Smart POIs are averaged by the total number of measures used. Lastly, the Top-5 Smart POIs from the final descending list are recommended to the specific user.

4.1.3 User-Based CF with the Average Aggregation Operator + Smart POIs' Categories

1. Data gathering process. Same process as described in Section 4.1.1. Furthermore, the Smart POIs' categories dataset is loaded. Such a dataset is described in Section 4.3.3.
2. Compute similarities between users. Same process as described in Section 4.1.1.
3. Recommend Smart POIs. Firstly, all categories of the Smart POIs visited by the user selected to give recommendations are extracted. Then, the Smart POIs' categories are ranked according to their frequency of appearance. Thus, a descending frequency list of the Smart POIs' categories is obtained. Thereupon, for each similarity or distance measure, a list of similar users is obtained, who best resemble the specified user according to the procedure described in the previous step. Afterwards, the categories of each Smart POI present in the preferences of each similar user are ranked according to the descending frequency list of the Smart POIs' categories of the specified user. Finally, all Smart POIs of each similar user are sorted to obtain those that better resemble the specified user preferences. That is, Smart POIs whose categories are closer to the rated Smart POIs' categories of the specified user are more likely to be recommended. Consequently, a descending list of N Smart POIs ranked by their categories is obtained. From this list, the Top-5 Smart POIs are recommended to the specific user.

4.1.4 HyRA

1. Data gathering process. Same process as described in Section 4.1.3. Furthermore, the Smart POIs' geographical location dataset is loaded. Such a dataset is described in Section 4.3.1.
2. Address geographical influence. All Smart POIs loaded in the data gathering process are clustered using K-means with the Euclidean distance. Due to the geographical distribution of Smart POIs, only three clusters are enough to embrace them all. The calculation of the optimal number of clusters is beyond the scope of this paper. As a result, a list containing the cluster' number to which each Smart POI belongs is obtained. Then, the cluster' number of each Smart POI visited by the specific user is extracted. Subsequently, the clusters' numbers visited by the chosen user are compared against the clusters' numbers

of the Smart POIs of the rest of users. As a result, the users that share at least N Smart POIs visited in common with the specific user are retained. Here, it is important to mention that the value of N is calculated as follows: one plus the result of the number of clusters visited by the specific user divided by two. This metric has two purposes: first, to ensure that the Smart POIs geographically closer to the users' location preferences are retained for a possible recommendation; and, second, to decrease the computational calculations that the recommendation algorithm has to perform. Consequently, a list of users that have visited Smart POIs geographically closer to the Smart POIs of a given user is obtained.

3. Recommend Smart POIs. Finally, the procedure described in Section 4.1.3 is performed to obtain the Top-5 Smart POIs that are going to be recommended, except step 1. As a result, a Smart POI list that could be interesting for the specific user is recommended.

4.2 Experimental Scenario Based on Surveys

A background about figures and data defined for conducting the test phase of HyRA is introduced. Later, the experiments defined for assessing this approach are described.

4.2.1 Project Background

This section aims to present the project's main background since the methodology proposed and used to study the application scenario in Ceutí was addressed and discussed in [4]. Such work describes the selection of POIs in Ceutí, the definition of the target audience as well as the sampling methods for this scenario, and the design of the survey. Therefore, only brief statements and key figures to introduce the experimental scenario are provided below.

- Selection of POIs. 16 POIs in Ceutí were defined as Smart POIs. Information about these Smart POIs is presented in Section 4.3.1.
- Definition of the target audience. Two types of tourists were included in the total target audience: Residents in Spain (86.4%) and Non-residents in Spain (13.6%). On the one hand, the resident target audience was the population of the Región de Murcia ≥ 18 years old. On the other hand, the non-resident target audience was defined as non-resident travelers in Spain.

- Definition of representative sampling (surveys). The conditions to ensure the building of a database representative of the target audience were defined as follows:
 - The non-probabilistic and cluster-based sampling methods were selected to conduct the surveys. This decision was based on the target audience is hard to identify and the sample is a pilot study [102].
 - The 6.75% margin of error was defined to ensure a representative sample of the target audience. Therefore, the number of surveys to be collected was estimated at 200, of which 173 people must be resident travelers in Spain (86.4%) and 27 people must be non-resident travelers in Spain (13.6%).
 - The 27 surveys for non-resident travelers in Spain were collected globally while the 173 surveys for resident travelers in Spain were divided into clusters. That is, three clusters were considered for this scenario, i.e. 18–30, 31–50, and >50, which also were divided into women and men. Hence, the number of surveys per cluster is shown in Table 4.1.
 - The survey was designed and managed online via Google Forms (<https://www.google.com/intl/en/forms/about/>), and structured in both Spanish (<https://goo.gl/VrC0ve>) and English (<https://lnkd.in/dzqVyJD>) language to facilitate its dissemination.

TABLE 4.1: Number of surveys assigned to each cluster.

Cluster	Age Range	Men	% Men	Women	% Women
1	18–30	17	9.8266	16	9.2486
2	31–50	37	21.3873	35	20.2312
3	>50	32	18.4971	36	20.8092
Total		86	49.7110	87	50.2890

4.2.2 The Surveys and the HyRA Evaluation Scenario

The experimental scenario will be divided into two phases: survey evaluation and HyRA test. On the one hand, to know the effectiveness of the surveys, a pilot dissemination phase is considered. In this phase, the respondents will be encouraged to provide an explicit feedback about their appreciation regarding the surveys' design and subject-matter, since the implicit feedback will be given by their answers. Subsequently, a period of up-to-date of both surveys is proposed for finally disseminating them to the target audience. On the other hand, with the aim of

evaluating the Smart POI recommendations given by HyRA, a scenario constituted of different tests is designed. These tests include the use of diverse distance and similarity measures, the Smart POIs' categories, and the geographical influence factor. For this purpose, the following steps are proposed.

1. Extraction of a ground-truth subset of ratings on Smart POIs of each user. With the aim of counting on a ground-truth to assess the recommendation algorithm, the Smart POIs dataset is divided into two. The ground-truth subset is obtained by randomly select up to 11 Smart POIs from each user whose rates vary from three to five stars. By doing this, approximately 70% from the 16 Smart POIs ratings given by the users can be captured. The aim of this subset is to serve as a ground-truth dataset that allows the recommendation algorithm to have a representation of the preferences of each user. The remaining Smart POIs of each user are used as not visited (not rated) Smart POIs that can be recommended by the recommendation algorithm. The original ratings that each user gives to each Smart POI, which belong to this last subset, are preserved to later compare the recommendations provided by the recommendation algorithm.
2. Selection and implementation of a set of similarity and distance measures to provide the Smart POI recommendation. The objective of this activity is to calculate the first recommendations for this scenario. Experiments are carried out by using each similarity and distance measure described in Appendix A.3. Furthermore, the following measures were also tested: Spearman correlation, Bray–Curtis distance, Canberra metric, and Squared Euclidean distance. First, the ground-truth subset is obtained as described above. Then, for each user, his/her recommendations are calculated with each similarity and distance measure. The procedure and description of the algorithm is found in Section 4.1.1.
3. Incorporation of the validated similarity and distance measures into the average aggregation operator. The aim of this activity is to increase the proposed recommendation algorithm precision. For this experiment, all similarity and distance measures described in Appendix A.3 are concentrated into an average aggregation operator as described in Section 4.1.2. In this experimental phase, one hundred executions are performed in order to compare the user-based CF algorithm with the average aggregation operator against its counterpart with one similarity or distance measure at a time. Each execution is independent of the others, that is, each execution calculated its own random ground-truth subset that is used at that time in both versions of the proposed algorithm.

4. Definition and integration of the Smart POIs' categories to the proposed recommendation algorithm. The aim of this activity is to increase the proposed recommendation algorithm precision. In this test scenario, the Smart POIs' categories are taken into account and added to the recommendation algorithm supplemented with the average aggregation operator as described in Section 4.1.3. In addition, one hundred executions are performed in order to compare the proposed recommendation algorithm supplemented with the average aggregation operator against its counterpart that adds Smart POIs' categories. Each execution is independent of the others, that is, each execution calculated its own random ground-truth subset that is used at that time in both versions of the proposed algorithm.
5. Implementation of the geographical influence factor in the proposed recommendation algorithm. The aim of this activity is to increase the proposed recommendation algorithm precision. In this phase, the Smart POIs' locations are integrated into the proposed recommendation algorithm that considers the Smart POIs' categories. The algorithm description can be reviewed in Section 4.1.4. Consistently, one hundred executions are performed to compare the proposed recommendation algorithm supplemented with both the average aggregation operator and the Smart POIs' categories against the recommendation algorithm that adds the geographical influence factor (HyRA). In addition to carrying out the same executions, the results of the recommendation algorithm supplemented with the average aggregation operator against the results of HyRA are compared. Each execution is independent of the others, that is, each execution calculated its own random ground-truth subset that is used at that time in the three versions of the proposed algorithm.
6. Compare the different approaches of the recommendation algorithm. To provide the version of the recommendation algorithm that delivers better recommendations to all users, the results of all implementations previously described are compared. The first step is to sort in descending order the Smart POIs preferences of each user contained in the not visited (not rated) dataset, this with the purpose of obtaining the preferences of each user from the highest to the lowest. Subsequently, the original rating that users granted to each Smart POI recommended by each algorithm per each user is extracted. Consequently, a list that concentrates the Smart POI recommendations with the original ratings for each version of the recommendation algorithm is obtained. Thus, the algorithm whose lists of recommendations deliver the Smart POIs with higher ratings for all users stands as the best approach for this study.

4.3 Data Requirements for HyRA

Firstly, information related to the Smart POIs selected for the experimental phase is shown. Subsequently, the dataset composed of the user's explicit preferences is introduced. Finally, the dataset constituted of the Smart POIs' categories is presented. The datasets generated in this research are freely available for download through a GitHub© repository called HyRA datasets (<https://github.com/JoAlvaradoU/HyRA-datasets.git>).

4.3.1 Smart POIs Dataset in Ceutí

The proposed recommendation algorithm requires knowing information about the Smart POIs that will be considered to carry out the recommendations. Therefore, a dataset composed of 16 Smart POIs previously defined for this work has been generated. The structure and records of this dataset are presented in Table 4.2. A brief description of the information contained in the fields of this dataset is provided below.

- **Smart POI Identifier.** The field that identifies the Smart POI and allows establishing a relationship with the user dataset to extract the ratings assigned by each user to these Smart POIs as well as with the Smart POI's categories dataset to obtain the tags that describe them.
- **Name.** The Smart POI's title in both English and Spanish language.
- **Location.** The column that indicates the Smart POI's coordinate in decimal degrees, whose format is [latitude, longitude].

4.3.2 User Preferences Dataset

Regarding the user explicit feedback, a subset of the information collected through the surveys has been extracted to build the user preferences dataset, which is used by the proposed recommendation algorithm in its traditional part. This dataset is made up of the 16 ratings of 200 people. Because the dataset has 3200 records, only the structure and data of one of the respondents are shown in Table 4.3. A brief description of its fields is provided below.

- **User Identifier.** The field that identifies the respondent, solely for purposes of the algorithm because no personal information was collected.
- **Smart POI Identifier.** The key to extracting the information from the Smart POIs dataset.
- **Rating.** The given numerical value by the respondent to the Smart POI according to his/her preferences. This value is within the range from 1 to 5, being 1 not interesting and 5 very interesting.

4.3.3 Smart POIs' Categories Dataset

Concerning the definition of the Smart POIs' categories, the All Categories section from the Yahoo! Answers website [100] as well as the description of these places were used to build the categories' dataset for this scenario. As a result, a dataset composed of 13 different categories was structured, where each Smart POI has three or four of the 13 categories already defined, as shown in Table 4.4. Such categories are Sculpture, Outdoors, Human, Mural, Museum, Church, Noria, Building, Architecture, Nature, Art, Square, and Park. A brief description of the dataset fields is presented below.

- **Smart POI Identifier.** The key to extracting the information from the Smart POIs dataset.
- **Category-X.** The fields that indicate the category name.

TABLE 4.2: Smart POIs dataset.

Smart POI Identifier	Name	Location
Heritage-ES-Ceuti-1	Stepping Strong Original: Pisando fuerte	38.078472,-1.270139
Heritage-ES-Ceuti-2	Allegory of Life Original: Alegoría de la Vida	38.078889,-1.271444
Heritage-ES-Ceuti-3	“7 Chimneys” Museum Original: Museo “7 Chimeneas”	38.079417,-1.272889
Heritage-ES-Ceuti-4	“La Conservera” Contemporary Art Museum Original: Museo de Arte Contemporáneo “La Conservera”	38.079194,-1.269000
Heritage-ES-Ceuti-5	“Santa Maria Magdalena” Church Original: Iglesia “Santa María Magdalena”	38.079056,-1.269528
Heritage-ES-Ceuti-6	Arabic Ruins of Ceuti Original: Ruinas Árabes de Ceutí	38.078417,-1.27016
Heritage-ES-Ceuti-7	Hermitage of San Roque Original: Ermita de San Roque	38.082111,-1.28466
Heritage-ES-Ceuti-8	My Metaphysical Garden Original: Mi Jardín Metafísico	38.080722,-1.276806
Heritage-ES-Ceuti-9	Apothecary’s Noria Original: Noria del Boticario	38.100167,-1.287722
Heritage-ES-Ceuti-10	Children bathing in La Acequia of Ceuti Original: Niños Bañándose en La Acequia de Ceutí	38.079389,-1.270056
Heritage-ES-Ceuti-11	The Mural of San Roque Original: El Mural de San Roque	38.079833,-1.273306
Heritage-ES-Ceuti-12	Queen Mariana Original: Reina Mariana	38.077806,-1.274861
Heritage-ES-Ceuti-13	The Canning Woman Original: La Mujer Conservera	38.077778,-1.274194
Heritage-ES-Ceuti-14	“Miguel de Cervantes” Sculpture Original: Escultura “Miguel de Cervantes”	38.077306,-1.271722
Heritage-ES-Ceuti-15	Tribute to the Emigrant Original: Homenaje al Emigrante	38.079472,-1.271917
Heritage-ES-Ceuti-16	Torso Original: Torso	38.081444,-1.276944

TABLE 4.3: User preferences dataset: example of structure for each user.

User Identifier	Smart POI Identifier	Rating
User-CEUTI-1	Heritage-ES-Ceuti-1	4
User-CEUTI-1	Heritage-ES-Ceuti-2	2
User-CEUTI-1	Heritage-ES-Ceuti-3	4
User-CEUTI-1	Heritage-ES-Ceuti-4	4
User-CEUTI-1	Heritage-ES-Ceuti-5	4
User-CEUTI-1	Heritage-ES-Ceuti-6	3
User-CEUTI-1	Heritage-ES-Ceuti-7	5
User-CEUTI-1	Heritage-ES-Ceuti-8	4
User-CEUTI-1	Heritage-ES-Ceuti-9	3
User-CEUTI-1	Heritage-ES-Ceuti-10	4
User-CEUTI-1	Heritage-ES-Ceuti-11	3
User-CEUTI-1	Heritage-ES-Ceuti-12	4
User-CEUTI-1	Heritage-ES-Ceuti-13	4
User-CEUTI-1	Heritage-ES-Ceuti-14	3
User-CEUTI-1	Heritage-ES-Ceuti-15	3
User-CEUTI-1	Heritage-ES-Ceuti-16	5

TABLE 4.4: Smart POIs' categories dataset.

Smart POI Identifier	Category-1	Category-2	Category-3	Category-4
Heritage-ES-Ceuti-1	Sculpture	Outdoors	Human	
Heritage-ES-Ceuti-2	Mural	Outdoors	Human	Art
Heritage-ES-Ceuti-3	Museum	Building	Architecture	Art
Heritage-ES-Ceuti-4	Museum	Building	Architecture	Art
Heritage-ES-Ceuti-5	Church	Building	Architecture	Art
Heritage-ES-Ceuti-6	Museum	Building	Architecture	Outdoors
Heritage-ES-Ceuti-7	Church	Building	Architecture	Outdoors
Heritage-ES-Ceuti-8	Mural	Outdoors	Nature	Art
Heritage-ES-Ceuti-9	Noria	Outdoors	Architecture	Nature
Heritage-ES-Ceuti-10	Mural	Outdoors	Human	Art
Heritage-ES-Ceuti-11	Mural	Outdoors	Human	Art
Heritage-ES-Ceuti-12	Sculpture	Outdoors	Human	
Heritage-ES-Ceuti-13	Sculpture	Outdoors	Human	Square
Heritage-ES-Ceuti-14	Sculpture	Outdoors	Human	
Heritage-ES-Ceuti-15	Sculpture	Outdoors	Human	Square
Heritage-ES-Ceuti-16	Sculpture	Outdoors	Human	Park

4.4 Results and Discussion

The results obtained from the proposed experimental phase as well as an analysis of the same from both the point of view of the user experience and the recommendation algorithm approach are described.

4.4.1 Surveys: Dissemination and Analysis

A pilot dissemination of the survey in Spanish was carried out with 10 residents in Spain and two foreign people to gather feedback about the design and subject-matter, mainly. Once the survey in Spanish was improved, the survey in English was carried out from the final survey in Spanish. Likewise, three foreign people performed an analysis of the subject-matter to ensure the clarity of the questions in this language. During these reviews, several changes were suggested.

- Spanish version
 - “Age” question. In the first surveys, the birthdate was asked to the respondents. However, this field was changed to the four age ranges established (<18, 18–30, 31–50, and >50) to directly do the clustering of each participant.
 - “Residence” question. The type of format to introduce this answer was specified since sometimes, only the city, country, or locality was typed by the respondent, entailing possible issues to determine the residence of the participant.
 - Sort the questions. The questions related to the tourism in the Región de Murcia and Ceutí “Do you usually tour the Region of Murcia (Spain)?”, “Have you ever visited Ceutí?”, and “if you visited Ceutí, what was the reason for the visit?” were realigned. Firstly, these questions were located between the personal information questions and the SNS questions; thus, some respondents asked if the questions related to the tourism in the Región de Murcia as well as Ceutí and the questions about the usage of SNS were associated, due to their answers could change according to this condition. Hence, to clarify that questions corresponding to the usage of SNS were formulated to know the user preferences in general, these three questions were located after the SNS questions.
 - New options for the answers. Two situations were presented: people from Ceutí and people who had never visited Ceutí answered the survey. Therefore, respondents

- suggested incorporating “I am from Ceutí” for the “have you ever visited Ceutí?” and “what was the reason for the visit?” questions, as well as “I have not visited Ceutí” for the last question. In addition, in the “what social networks do you use to publish your location during your travels or visits?” question was proposed to add the Twitter option. Such suggestions were integrated into the survey.
- Information about Ceutí. A brief introduction about Ceutí was described in the have you ever visited Ceutí? question to contextualize foreign respondents.
 - English version
 - “Residence” question. The type of format was modified to indicate to the user only writing his/her country.
 - Re-formulated question. “Do you usually tour the Region of Murcia (Spain)?” was rephrased to have you ever visited the Region of Murcia?”
 - Points of tourist attraction. The names of these points were translated for their identification, although the original name was also maintained.

After incorporating the changes suggested by the pilot target audience, both surveys were disseminated. These surveys were delivered from 6 April 2017 to 21 April 2017 through the following media:

- SNS: HOP Ubiquitous, town council of Ceutí, and Tecnológico de Monterrey.
- Instant messaging (WhatsApp): people involved in the project (HOP Ubiquitous and Tecnológico de Monterrey).
- E-mail: people involved in the project (HOP Ubiquitous, Tecnológico de Monterrey, and town council of Ceutí).

Considering that the target audience should be composed of residents from the Región de Murcia as well as foreign people, some groups were selected to distribute the surveys.

- People related to the Spanish members of the project located in different geographical locations from the Región de Murcia.
- People related to the Mexican members of the project located in Mexico.

- People identified by the town council of Ceutí.
 - Members of transnational meetings of the town council of Ceutí’s European projects.
 - Members of transnational meetings of the European projects in which the Ceutí’s IES is involved.
 - Members of the relations between families with the St Berthevin City in France.

The total amount of established surveys (200) was surpassed and the respondents’ locations confirm that the aim of surveying people belonging to the Región de Murcia was, mostly, achieved. Consequently, the study provides a global vision about the target audience preferences, where such respondents can be potential visitors to the town of Ceutí. However, although more than 200 surveys were collected, when building the clusters defined for each age range, the >50 age range clusters could not be completed successfully with only residents from the Región de Murcia. Hence, taking into account that people resident from Spain (not belonging to the Región de Murcia) also participated in the study, the missing user profiles were obtained of this group of respondents. Therefore, three Spanish profiles non-resident in the Región de Murcia were introduced to these clusters. This fact can be appreciated in Figure 4.1.

To conclude, these surveys, in addition to supporting the building of the dataset related to the target audience preferences, also contributing to one of the capacities that the Internet of Things (IoT) presents to improve any sector: the data collection about the user [103]. As a result, two datasets are provided from this approach: one dataset consisting of the preferences of 200 people and one dataset composed of information corresponding to 16 Smart POIs.

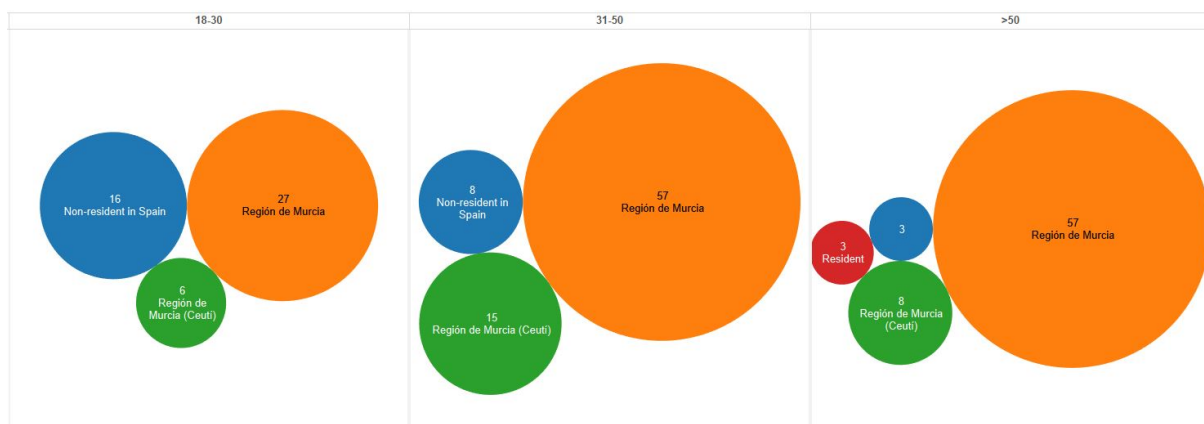


FIGURE 4.1: Responders’ residence profile.

4.4.2 HyRA: Analysis and Discussion

To have a varied set of similarity and distance measures with which to search for better recommendations, the following measures are implemented: Pearson correlation, Euclidean distance, Cosine similarity, Spearman correlation, Manhattan distance, Bray–Curtis distance, Canberra metric, Chebyshev distance, and Squared Euclidean distance.

Afterwards, a function to divide the set of ratings given by a user has been introduced. That is, two subsets of the global set (16 ratings) are generated, one with 70% of the ratings and another with 30%, 11 ratings and 5 ratings, respectively. The largest set is assigned to the similarity calculation among users while the smallest set will be maintained to make the comparison between the recommendations provided by the algorithm and this set. However, these parameters were modified because the number of Smart POIs recommended was always the same: five. Then, the only observed change was the similarity value among the same Smart POIs since others Smart POIs could not be included as only five Smart POIs were available to recommend. Hence, a variation in the process of producing these subsets was introduced: the number of ratings to form the subset assigned to the similarity calculation would be random from 1 to 11. In this way, new Smart POI recommendations were ensured.

An example of the results obtained from this analysis is provided below. In Table 4.5 are presented the ratings given by the responder identified by User-CEUTI-1 for each Smart POI and in Table 4.6 as well as Table 4.7 are shown the recommendations suggested for this user. Such recommendations are labeled from 1 if it is the least recommended to 5 if it is the most recommended. Where:

1. Smart POIs used for the similarity calculation = {Stepping Strong, Allegory of Life, Arabic Ruins of Ceuti, Hermitage of San Roque, My Metaphysical Garden, Queen Mariana, The Canning Woman, Torso}—eight Smart POIs
2. Smart POIs available for the recommendations = {Apothecary’s Noria, “7 Chimneys” Museum, Tribute to the Emigrant, The Mural of San Roque, “Santa Maria Magdalena” Church, Children bathing in La Acequia of Ceuti, “La Conservera” Contemporary Art Museum, “Miguel de Cervantes” Sculpture}—eight Smart POIs

TABLE 4.5: User preferences identified by User-CEUTI-1.

Smart POI Name	Rating
Stepping Strong	4
Allegory of Life	2
“7 Chimneys” Museum	4
“La Conservera” Contemporary Art Museum	4
“Santa Maria Magdalena” Church	4
Arabic Ruins of Ceuti	3
Hermitage of San Roque	5
My Metaphysical Garden	4
Apothecary’s Noria	3
Children bathing in La Acequia of Ceuti	4
The Mural of San Roque	3
Queen Mariana	4
The Canning Woman	4
“Miguel de Cervantes” Sculpture	3
Tribute to the Emigrant	3
Torso	5

TABLE 4.6: Recommendations given by the algorithm to the responder identified by User-CEUTI-1 (I).

Smart POI	Pearson	Euclidean	Cosine	Spearman	Manhattan
Stepping Strong					
Allegory of Life					
“7 Chimneys” Museum		4	3		3
“La Conservera” Contemporary Art Museum	1	5	5	1	5
“Santa Maria Magdalena” Church	4	2	4	2	4
Arabic Ruins of Ceuti					
Hermitage of San Roque					
My Metaphysical Garden					
Apothecary’s Noria	3	3	2		2
Children bathing in La Acequia of Ceuti	5	1	1	3	1
The Mural of San Roque	2			4	
Queen Mariana					
The Canning Woman					
“Miguel de Cervantes” Sculpture				5	
Tribute to the Emigrant					
Torso					

TABLE 4.7: Recommendations given by the algorithm to the responder identified by User-CEUTI-1 (II).

Smart POI	Bray–Curtis	Canberra	Chebyshev	Squared Euclidean
Stepping Strong				
Allegory of Life				
“7 Chimneys” Museum	3	3	3	3
“La Conservera” Contemporary Art Museum	5	5	5	5
“Santa Maria Magdalena” Church	4	4	4	4
Arabic Ruins of Ceuti				
Hermitage of San Roque				
My Metaphysical Garden				
Apothecary’s Noria	2	2	2	2
Children bathing in La Acequia of Ceuti	1	1	1	1
The Mural of San Roque				
Queen Mariana				
The Canning Woman				
“Miguel de Cervantes” Sculpture				
Tribute to the Emigrant				
Torso				

Accordingly, for this example, the Cosine, Manhattan, Bray–Curtis, Canberra, Chebyshev, and Squared Euclidean measures provided the same recommendations.

1. 5—“La Conservera” Contemporary Art Museum
2. 4—“Santa Maria Magdalena” Church
3. 3—“7 Chimneys” Museum
4. 2— Apothecary’s Noria
5. 1—Children bathing in La Acequia of Ceuti

Subsequently, the results obtained by the user-based CF algorithm against the results obtained by the user-based CF with an average aggregation operator are compared to detect the algorithm that provides best Smart POI recommendations. To perform the comparison, both algorithms are executed one hundred times with their independent ground-truth subset as described in Section 4.2.2.

4.4.3 Smart POI Recommendation through User-Based CF with an Average Aggregation Operator

After performing the experiments described in Section 4.2.2, the results of all executions of user-based CF and the user-based CF with an average aggregation operator (CF + AO) are compared.

Firstly, experiments with nine similarity and distance measures are performed. Table 4.8 shows the total counts in which each algorithm wins over the other, the total counts in which the recommendation results end up in a tie, and the counts in which each individual distance wins across all users. Additionally, the count of no comparisons across all users is presented. If it is not possible to calculate any similar user for a given distance, then it is not possible to recommend any Smart POI. Thus, in the absence of recommendations, the comparison of results is not performed.

TABLE 4.8: Comparison between the recommendations given by the user-based CF against the user-based CF with an average aggregation operator that integrates nine similarity measures.

	Counts of Winning Comparisons	% Winning Comparisons
Wins of CF across all executions	72,968	44.00
Winning distance of CF across all executions	Euclidean distance with 68/100 executions	NA
Wins of CF + AO across all executions	55,445	33.43
Draws across all executions	35,848	21.61
No comparisons across all executions	1599	0.96

In Table 4.8, it can be noticed that the recommendations of Smart POIs made through the user-based CF algorithm best resemble, in general, the preferences of all available users. It is worth mentioning that the Euclidean distance brings results that better resemble the users' preferences for more than a half of executions of the experiments. Furthermore, and strictly speaking, if it is not possible to perform a comparison due to the lack of Smart POI recommendations through the user-based CF, then the user-based CF + AO algorithm stands as the recommendation algorithm that must be used due to its faculty of always deliver a recommendation to the user. Afterwards, the Mean Squared Error (MSE) for each recommendation given by each similarity and distance measure as well as by the user-based CF + AO are computed. Results are concentrated in Table 4.9.

TABLE 4.9: MSE of each of the nine similarity and distance measures concentrated in the user-based CF algorithm and of user-based CF with an average aggregation operator that integrates the nine measures.

Similarity and Distance Measures, or Algorithm	MSE
Euclidean distance	0.85
Cosine similarity	1.07
Chebyshev distance	1.37
Pearson correlation	1.54
Manhattan distance	1.62
Bray–Curtis distance	1.63
Canberra metric	1.80
Squared Euclidean distance	1.82
Spearman correlation	14.68
User-based CF + AO	2.16

Results in Table 4.9 show that the user-based CF with Euclidean distance has the lowest MSE of all distances, and is even lower than the user-based CF + AO algorithm. In contrast, the Spearman correlation has the highest MSE of all measures. Furthermore, it can be noticed that the inclusion of the Spearman correlation into the user-based CF + AO increases the value of its MSE. Thus, additional experiments are performed to obtain a better combination of the similarity and distance measures for the user-based CF + AO algorithm. The experiments are performed by taking out the similarity metric that has the highest MSE value each one hundred executions of the algorithm. Table 4.10 concentrates the MSE values of these experiments.

TABLE 4.10: MSE of different combinations of the similarity and distance measures for the user-based CF with an average aggregation operator.

Number of Measures	MSE
Nine	2.16
Eight	1.69
Seven	1.67
Six	1.60
Five	1.57
Four	1.60
Three	1.22
Two	1.13

In Table 4.10, it can be noticed that removing one measure ensures the reduction of the MSE. However, the purpose of the aggregation operator is to provide diversity in the recommendations made by the algorithm. Thus, MSE values from each previous version of the user-based CF + AO algorithm are compared to the MSE values of each similarity or distance measure in the respective experiment. Due to lack of space, the results of the best version of the user-based

CF + AO algorithm is presented. The best version of the user-based CF + AO algorithm is obtained by using the following five similarity and distance measures: Euclidean distance, Cosine similarity, Chebyshev distance, Pearson correlation, and Manhattan distance. Table 4.11 shows the results of one hundred executions of the best combination obtained from such experiments, while Table 4.12 presents the MSE computed for the same experiments.

TABLE 4.11: Comparison between the recommendations given by the user-based CF algorithm against the user-based CF with an average aggregation operator that integrates five similarity and distance measures.

	Counts of Winning Comparisons	% Winning Comparisons
Wins of CF across all executions	65,636	42.53
Winning distance of CF across all executions	Euclidean distance with 60/100 executions	NA
Wins of CF + AO across all executions	55,159	35.74
Draws across all executions	31,977	20.72
No comparisons across all executions	1562	1.01

TABLE 4.12: MSE of the nine similarity and distance measures concentrated in the user-based CF algorithm and of user-based CF with an average aggregation operator that integrates five similarity and distance measures.

Similarity and Distance Measures, or Algorithm	MSE
Euclidean distance	0.84
Cosine similarity	1.08
Chebyshev distance	1.38
Pearson correlation	1.55
Manhattan distance	1.62
Bray–Curtis distance	1.64
Canberra metric	1.80
Squared Euclidean distance	1.81
Spearman correlation	14.70
CF + AO	1.57

Table 4.11 shows that the winning counts difference between the user-based CF algorithm and the user-based CF + AO algorithm is reduced. Furthermore, Table 4.12 indicates that the MSE of the user-based CF + AO algorithm is significantly decreased, positioning it before the Manhattan distance. Thus, the decision of keeping those five similarity and distance measures is due to: (1) decrement of the MSE value obtained using the nine similarity and distance measures; (2) retaining more than the half of the available similarity and distance measures; and (3) always delivering a recommendation. For these reasons, the user-based CF + AO algorithm with five similarity and distance measures is selected as the basis of the proposed recommendation algorithm.

4.4.4 Smart POI Recommendation through User-Based CF with an Average Aggregation Operator + Smart POIs' Categories

The inclusion of categories, tags, or topics is another frequent approach used in the literature in order to improve the recommendations performed by various algorithms. Consequently, the Smart POIs' categories are included in the algorithm that obtained the best recommendation results from the previous experiment. Then, the experimentation phases described in Section 4.2.2 are performed. Table 4.13 summarizes the results obtained in the comparison of the user-based CF + AO algorithm against the user-based CF with an average aggregation operator and the Smart POIs' categories (CF + AO + C). Moreover, a comparison among the user-based CF + AO + C algorithm with the five selected similarity and distance measures and with only the Euclidean distance is included.

TABLE 4.13: Comparison between the recommendations given by the user-based CF algorithm with an average aggregation operator against the user-based CF with an average aggregation operator and the Smart POIs' categories.

	Counts of Winning Comparisons	% Winning Comparisons
Wins of CF + AO across all executions	6164	22.06
Wins of CF + AO + C across all executions	9170	32.82
Wins of CF + AO + C across all executions (Euclidean)	8849	31.67
Draws across all executions	3760	13.45

Results show that the addition of categories into the recommendation algorithm improves the general resemble of the users' preferences. It is noteworthy that these two versions of the recommendation algorithm do not present the lack of results; therefore, it is possible to carry out a comparison. Additionally, another finding is presented by using the five similarity and distance measures to generate new recommendations since the proposed algorithm provides better recommendations than only using the distance with the lowest MSE.

4.4.5 Smart POI Recommendation through Geographical Influence + User-Based CF with an Average Aggregation Operator + the Smart POIs' Categories (HyRA)

In addition to the Smart POI's categories, the use of geographical influence is also one approach handled in the literature for improving recommendations. Therefore, the geographical influence

factor is added to the user-based CF + AO + C recommendation algorithm as described in Section 4.1.4. Table 4.14 shows the results of the comparisons between the user-based CF + AO + C and the GI + user-based CF + AO + C (HyRA). Furthermore, Table 4.15 shows the results of the comparisons between the user-based CF + AO and HyRA.

TABLE 4.14: Comparison between the user-based CF with an average aggregation operator + the Smart POIs' categories and HyRA.

	Counts of Winning Comparisons	% Winning Comparisons
Wins of CF + AO + C across all executions	27	0.14
Wins of HyRA across all executions	36	0.18
Draws across all executions	19,630	99.68

TABLE 4.15: Comparison between the user-based CF with an average aggregation operator and HyRA.

	Counts of Winning Comparisons	% Winning Comparisons
Wins of CF + AO across all executions	6129	31.12
Wins of HyRA across all executions	9653	49.02
Draws across all executions	3911	19.86

Results in Table 4.14 show that for most of the cases to use the recommendation algorithm with or without the geographical influence is indifferent. However, it can notice that integrating the geographical influence factor resembles slightly better the general users' preferences than the algorithm that does not include it. Additionally, the results of Table 4.15 corroborate that the use of the geographical influence factor and the Smart POIs' categories are favorable for the recommendation results. Even though the Smart POIs encompassed in the dataset are located geographically close one to another, the inclusion of geographical influence can provide Smart POI recommendations that suit better the users' preferences.

Finally, HyRA is compared with another POI recommendation algorithm in the literature that embraces both user-based CF and geographical influence. In Ye et al. [44], unified collaborative recommendation algorithm (USG) and the user preference/geographical influence based recommendation (UG) algorithm are the two algorithms with the best performances. However, the USG algorithm is not chosen to be compared with HyRA because USG comprises a Friend-based Collaborative Filtering, an approach that is not addressed in this research. Therefore, UG is implemented and compared with HyRA by presenting an approach closer to the HyRA approach. Table 4.16 shows the comparative results between the UG and HyRA algorithms.

TABLE 4.16: Comparison between the UG and HyRA recommendation algorithms.

	Counts of Winning Comparisons	% Winning Comparisons
Wins of UG [44] across all executions	5707	28.99
Wins of HyRA across all executions	11,099	56.38
Draws across all executions	2879	14.63

According to the results obtained in Table 4.16, HyRA resembles better the users' preferences in the dataset. In addition, it is noteworthy that the inclusion of the Smart POIs' categories and the integration of an average aggregation operator into a Smart POI recommendation algorithm allow providing better recommendations than approaches that only consider the user-based CF algorithm and the geographical influence factor. Thus, the geographical influence + user-based CF with an average aggregation operator + the Smart POIs' categories (HyRA) stands so far as the best recommendation algorithm for this research approach by surpassing all the approaches included in these experiments in at least 0.04% and 27.39% of the user-based CF with an average aggregation operator + the Smart POIs categories and UG [44] algorithms, respectively.

4.5 Conclusions

In this Chapter, the first use case to validate the proposed recommendation algorithm is presented. This approach uses a novel device and technology called Smart Spot and Physical Web, respectively, for the tourism sector.

HyRA is based on both traditional and POI recommendation approaches as it incorporates the user's explicit preferences (ratings), the Smart POIs' categories, the geographical influence factor, as well as the characteristics of Smart Spot and Smart POI when suggesting to the user a new Smart POI list to visit. Specifically, with the aim of dealing with the user's explicit preferences along with the characteristics of Smart Spot and Smart POI, a modified user-based CF algorithm, which consists of merging an average aggregation operator integrated by five similarity and distance measures as a single measure into the user-based CF algorithm, is proposed and validated. In the same way, to encode the Smart POIs categories, a filtering method is incorporated, and to unify the geographical influence factor, the K-means algorithm using an Euclidean distance are assembled.

Additionally, with the aim of carrying out the evaluation of HyRA, one survey in Spanish and another in English were disseminated to collect information related to the general user preferences and profiles as well as the user's specific preferences on the defined Smart POIs in this scenario. As a result, two datasets have been structured and generated according to the real-world scenario in Ceutí: one dataset constituted of the 16 Smart POIs, and the other composed of the ratings provided at 16 Smart POIs per 200 people. In addition, an experimental dataset consisting of 13 categories was built.

These three datasets are used in the experimental cases defined in this approach and are published for reuse. The experimental results indicate that HyRA recommends a Smart POI list closer to the user preferences than the approaches included in this evaluation. This research and results were published in [2, 4].

In the next Chapter, a use case addressing the evaluation of the processes proposed in AkCeL for the academic sector is described. This use case is focused mainly on assessing the data pre-processing, alignment, and recommendation processes, and presenting the data visualization's deployment scenario. Hence, the introduction and scope of this use case are presented as well as the evaluation scenarios, the datasets used, and the results of these scenarios.

Chapter 5

Use Case: Generation of Research Recommendations and Statistics based on a Unified, Updated, and Consistent User's Academic Profile

This proposal aims to generate unified, consistent, and updated linked data about the user's academic profile based on Academic SUP (meta-Schema User's Profile). For achieving this goal, AkCeL (frAmework for Consistent generation of Linked data) is proposed. Likewise, AkCeL intends to generate research recommendations and statistics that support the decision-making processes of researchers and data managers at the institutional level.

To carry out the unified, consistent, and updated generation of linked data, a data alignment process is proposed and assessed. In addition, to generate the recommendations, the proposed algorithm in this work (C-HyRA) is used and validated. Finally, to visualize both the resulting information and the research statistics produced with the information generated by AkCeL, the VIVO platform is adopted.

This Chapter is organized as follows. Firstly, in Section 5.1, the scope of this research work, as well as the experimental and deployment scenarios concerning to the data alignment, recommendation, and visualization are defined. Subsequently, in Section 5.2, the results of these scenarios and their respective discussions are provided.

5.1 Scope of the Use Case

For the proposed use case, information about the user's academic profiles and contextual information related to his/her research are considered in order to integrate and generate new information with a better accuracy and in a consistent manner. That is, to integrate new data into Academic SUP according to the defined update and alignment processes, to generate a research areas list of interest for a researcher using C-HyRA, and to visualize research statistics for a researcher or an institution through the VIVO platform.

Therefore, information related to the user's identification, interests, preferences, and publications, belonging to research and academic environments, is collected. Thereby, this proposal is focused on supporting Tecnológico de Monterrey in its reporting activities by allowing it to search for information in Academic SUP and to visualize its research statistics, as well as the researchers affiliated with Tecnológico de Monterrey by allowing them to automatically integrate their academic information and to identify new research and collaboration areas of interest.

With the aim of highlighting the scope of the use case defined for this research work, the context of the scenario is listed below.

1. dblp and Scopus are the information sources selected to collect data for this study.
2. The data retrieval, preprocessing, alignment, recommendation, and visualization are the main processes of this work. The aggregation or inference of new relations in Academic SUP is beyond the scope of this research.
3. All features of the user's profile defined in this proposal are concentrated in Academic SUP.
4. Information strictly related to the academic profile of a person (user) is collected.
5. The development of the Academic SUP ontology is only performed in the English language.
6. The user's profiles specified (written) in the Spanish or English language are selected for this study.
7. The Tecnológico de Monterrey's researchers community is defined as the users for this research work.
8. Computer Science is the research area chosen for this use case.

5.1.1 The Data Alignment Process' Evaluation Scenario

To validate the information aligned and provided by AkCeL, the SIIP (Sistema de Información para la Investigación y el Posgrado) repository is consulted. This source is chosen since is the system that consolidates different databases with information belonging to the Tecnológico de Monterrey, which can be interesting for the researchers, such as theses, publications, patents, among others [64]. Therefore, the following activities are proposed.

1. Selection of a set of researchers in both the SIIP repository and Academic SUP. The objective of this activity is to obtain a dataset consisting of at least 20 researchers affiliated with the Tecnológico de Monterrey whose research area is Computer Science. In addition, another requirement to choose these researchers is that the information about them must be integrated into both sources since this dataset is defined to carry out comparisons between the information provided by AkCeL and the information provided by SIIP.
2. Searching for matches between the information stored in SIIP and the information integrated into Academic SUP. In order to calculate the number of matches between both information sources, the search for matches between the information stored in SIIP (retrieved from a VIVO's deployment [68]) and the aligned information stored in Academic SUP is carried out. Firstly, information with exact matches is provided for the statistics of the comparative results. Subsequently, cases exemplifying the mismatch between SIIP and Academic SUP are presented to analyze the differences between these two sources.
3. Comparison of the results obtained from the proposed scenario and a baseline scenario. The goal of this activity is to describe a baseline scenario of data integration between Scopus and dblp into Academic SUP without incorporating the proposed data preprocessing process and with a modified version of the proposed data alignment process to compare the results of both scenarios. That is, to carry out comparisons between the scenario following the data preprocessing and alignment processes proposed in this work, named as the proposed scenario, and the scenario without data preprocessing and with a modified version of the proposed data alignment called the baseline scenario. For this, firstly, a description of the baseline scenario is presented to introduce the steps followed by it. Subsequently, the results obtained from the baseline scenario are given individually as well as in conjunction with the results obtained in Step 2 of this evaluation scenario (proposed scenario) to compare and analyze the figures calculated from the proposed and baseline

scenarios. Finally, cases exemplifying the mismatch between the proposed scenario and the baseline scenario are presented to examine the differences between these two scenarios.

5.1.2 The C-HyRA's Evaluation Scenario

With the aim of assessing the research areas recommendations given by C-HyRA, a scenario consisting of different steps is designed. These steps include the use of a ground-truth subset constituted by research areas explored by each user. For this purpose, the following steps are proposed.

1. Collection and building of a dataset constituted by ratings deduced on the research areas of each user. With the aim of counting on a dataset to assess the recommendation algorithm, a dataset constituted by a researcher identifier, the research area's name, and a calculated rating for this area is built. The dataset is obtained by randomly collecting information about the researchers, their publications, and the publications' research areas and categories from Scopus®. Since ratings are not explicitly provided, a function to deduce the levels of interest of each researcher on each research area is proposed.
2. Extraction of a ground-truth subset and a pilot subset of the research areas explored of each user. This activity aims to provide a ground-truth subset and a pilot subset to evaluate the recommendations given by the proposed algorithm (C-HyRA); thus, the dataset built in step 1 is divided into two. The ground-truth subset is obtained by randomly selecting up to seven research areas of each user whose ratings vary from three to five stars. By doing this, approximately 58.33% of the 12 research areas covered by the user's publications can be captured, being 12 the highest number of research areas where researchers have published, according to the dataset built in step 1. Thereby, this subset serves as a ground-truth dataset that allows the recommendation algorithm to have a representation of the preferences of each user. The remaining research areas of each user are used as not rated research areas that can be recommended by C-HyRA. These remaining research areas constitute the pilot subset.
3. Comparison of the recommendations provided by C-HyRA against the research areas contained in the pilot subset. In order to validate the recommendations given by C-HyRA, firstly, a list containing the research areas recommended by C-HyRA for each user is extracted. Subsequently, a list concentrating the research areas present in the

pilot subset for each user is obtained. Then, these lists (research areas recommended by C-HyRA and research areas present in the pilot subset) are compared to determine the number of research areas in common for both lists, taking into account that the maximum number of recommendations provided by C-HyRA are five. Afterwards, the percentage of coincidence of the research areas in common is calculated. Finally, the number of research areas recommended by C-HyRA that are present in the pilot subset against the total number of research areas that are present in the pilot subset is provided.

5.1.3 The Data Visualization's Deployment Scenario

To visualize and interact with the data generated by AkCeL, the VIVO platform is adopted. Therefore, with the aim of achieving these purposes, the following activities are proposed.

1. Selection, installation, and configuration of a VIVO's version. The objective of this activity is to choose the VIVO's version best documented for its deployment on Windows Operating System (OS) in order to install and configure this version properly.
2. Incorporation of data and ontology. In order to customize the deployment of VIVO, the structure of the information related to the use case defined for this proposal, *i.e.*, Academic SUP (ontology), as well as the data themselves, are loaded.
3. Visualization of statistics. This activity aims to set up VIVO for visualizing the statistics defined in this use case.

5.2 Results and Discussion

The results of each scenario proposed for this research are presented and discussed in the following sequence: data alignment process in Section 5.2.1, baseline scenario of data integration without data preprocessing and with a modified version of the proposed data alignment in Section 5.2.2, data requirements for C-HyRA in Section 5.2.3, recommendations based on C-HyRA in Section 5.2.4, and data visualization in Section 5.2.5.

5.2.1 Data Alignment Process

For this use case, firstly, a filtering of the preprocessed data for each information source with the parameters established for this use case is carried out. That is, data belonging to researchers affiliated with ITESM and whose research area is Computer Science. Table 5.1 provides the statistics of the authors resulting from the filtering. For the dblp dataset, the filtering by affiliation was not carried out since this dataset contains mostly information related to publications, such as title, pages, year of publication, volume, journal, number, among others [70]; while the data concerning the authors is only the author's name [70] and currently, their ORCID [67]. Therefore, the same researchers filtered from Scopus are selected in dblp to perform the experiments of this process.

TABLE 5.1: Statistics of authors for the alignment process.

Source	Total Authors	Authors for Alignment	% Authors for Alignment
Scopus	39,162	286	0.73
dblp	2,069,135	286	0.01

Thereby, the data alignment process is applied to the information related to the publications of these 286 filtered researchers. Finding, at the time the data alignment process was carried out, an issue for dblp: the absence of the author's name. That is, there are some publications that do not have a related author's name. Therefore, when a publication does not have a related author's name, this record is omitted. Once this rule was integrated into this process, the alignment results were provided.

For evaluation purposes, only the researchers present in both the SIIP repository and Academic SUP are selected. As a result, 21 researchers, introduced in Table 5.2, are chosen to perform comparisons between both information sources: SIIP against Academic SUP (Scopus-dblp). The results of the data alignment process related to the number of publications constituting the final dataset (Scopus-dblp) for each researcher, the count of publications stored in SIIP (retrieved from a VIVO's deployment [68]) for each researcher, and the number of publications that SIIP and Academic SUP (Scopus-dblp dataset) share for each researcher are also provided in Table 5.2.

From the results reported in Table 5.2, it can be observed that the data alignment process is promising. This is based on the fact that AkCeL integrates both publications stored in SIIP

TABLE 5.2: Comparison of the alignment results between Academic SUP and SIIP.

Researcher Identifier	Number of Publications in Academic SUP	Number of Publications in SIIP	Number of Publications in Common
Researcher-1	48	31	5
Researcher-2	78	137	42
Researcher-3	53	81	23
Researcher-4	2	11	0
Researcher-5	1	29	1
Researcher-6	1	4	0
Researcher-7	11	3	0
Researcher-8	3	43	2
Researcher-9	46	3	0
Researcher-10	60	185	38
Researcher-11	2	5	0
Researcher-12	8	10	1
Researcher-13	9	11	1
Researcher-14	46	202	30
Researcher-15	36	2	0
Researcher-16	5	98	1
Researcher-17	2	1	0
Researcher-18	2	30	2
Researcher-19	13	34	7
Researcher-20	5	15	2
Researcher-21	1	2	0

(a means of validation for this alignment) and publications that can be incorporated into the researchers' specific profile, for most of the researchers. These results are achieved despite the fact that, in certain cases, the aligned information provided by AkCeL and stored in Academic SUP (proposed Scopus-dblp dataset) presents fewer publications than SIIP. Considering that the Scopus-dblp dataset is integrated by all the publications belonging to the 286 researchers selected for this alignment from dblp and by an extract of the publications belonging to the same 286 researchers from Scopus, in accordance with what is mentioned in Section 3.3.1. Furthermore, the publications that can be added to each researcher's specific profile may have the same publication date, an earlier publication date, or a publication date after the last update date recorded by the research works covered in the VIVO's deployment [68]. Such as the cases related to the researchers identified as Researcher-8 and Researcher-21.

Firstly, the case of the researcher identified as Researcher-8 is presented in Table 5.3. From the information aligned and provided by AkCeL (Scopus-dblp dataset), three publications are reported while 43 publications are stored in SIIP. Nevertheless, only two publications are shared

between them, obtaining as one new instance for SIIP the publication entitled “Refining semantically annotated business process diagrams”. When looking for this publication in the researcher’s profile in the VIVO platform [68] to verify that it is not integrated into SIIP, two facts were observed: (1) the publication is not integrated into SIIP and (2) the last year of update obtained from the publications registered in SIIP is the year 2013. Consequently, the mentioned publication provided by AkCeL can be added to SIIP, highlighting that its year of publication is 2016, which establishes that AkCeL allows to keep the researchers’ profile up-to-date.

TABLE 5.3: Alignment results for the researcher identified as Researcher-8.

Publications in Academic SUP and in the VIVO’s Deployment [68]	Publications in Academic SUP but not in the VIVO’s Deployment [68]
Integrating semantic annotations in bayesian causal models	Refining semantically annotated business process diagrams
Mapping relational databases through ontology matching: a case study on information migration	

Secondly, the case of the researcher identified as Researcher-21 is presented in Table 5.4. From the information aligned and provided by AkCeL (Scopus-dblp dataset), one publication is reported while two publications are retrieved from the VIVO’s deployment [68]. However, no publication is shared between them, gaining as a new instance for SIIP the publication entitled “Morphological segmentation and digital image processing to retrieve geometric characteristics of fabric filaments”. When looking for this publication in the researcher’s profile in the VIVO platform [68] to check that it is not integrated into SIIP, two facts were observed: (1) the publication is not integrated into SIIP and (2) the last year of update obtained from the publications registered in SIIP is the year 2006. Consequently, the aforementioned publication provided by AkCeL can be added to SIIP, highlighting that its year of publication is 2005, thus completing the researcher’s publications that were also registered in the same year in SIIP.

TABLE 5.4: Alignment results for the researcher identified as Researcher-21.

Publications in Academic SUP and in the VIVO’s Deployment [68]	Publications in Academic SUP but not in the VIVO’s Deployment [68]
	Morphological segmentation and digital image processing to retrieve geometric characteristics of fabric filaments

Therefore, according to the results and the cases described, the proposed data alignment process in this research work allows to keep updated the publications of a certain researcher in different periods of his/her academic activity in a consistent manner. In this way, AkCeL can provide aligned information for each researcher automatically, which is stored in Academic SUP and visualized in the proposed VIVO's deployment.

5.2.2 Baseline Scenario: Data Integration without Data Preprocessing and with a modified version of the Proposed Data Alignment

In order to highlight the importance of preprocessing and aligning the information collected from Scopus and dblp to carry out a consistent data integration in Academic SUP, a baseline scenario, where data preprocessing is not performed and in addition, the data alignment proposed in this work is modified, is described in this Section.

The steps of the baseline scenario carried out with the same information collected from Scopus and dblp, omitting the data preprocessing introduced in Section 3.3.1.2 and modifying the data alignment proposed in Section 3.3.2, are the following:

1. Conversion of the Scopus and dblp datasets to the UTF-8 format. The files generated by the data retrieval process are saved with the UTF-8 format for both Scopus and dblp in new CSV and XML files, respectively. This aims to standardize as much as possible the data used in this scenario.
2. Extraction of researchers affiliated with a certain institution. For this research, the researchers' names affiliated with Tecnológico de Monterrey, who are in Scopus dataset, are retrieved. Scopus is the information source considered as a basis for the proposed alignment process because it contains the largest amount of information related to the researchers. Therefore, the researchers' names as well as their identifiers are extracted. These names are gathered with the original format presented in the Scopus dataset, *i.e.*, without preprocessing.
3. Filtering of the researchers by a specified research area. For this use case, the researchers who have published in the "Computer Science" research area are selected from the Scopus dataset. The filtering of the records belonging to these researchers is based on the research areas related to their publications.

4. Extraction of the different researcher's names as author for the first information source. In this step, the researchers' names, selected in Step 2, are congregated. This activity allows concentrating all the researcher's names without preprocessing, collected in Step 2, through his/her Scopus identifier. This is with the objective of concentrating all the possible names with which one same researcher has published and in this way, contributing to the alignment of his/her publications indexed in Scopus with his/her publications indexed in dblp.
5. Generation of a file containing the Academic SUP's properties and the values for each property according to the first information source selected. Firstly, the Scopus dataset is formatted according to the properties defined in Academic SUP with the aim of generating a final file of this information source that supports the final file of this alignment process. Secondly, the instances of these properties are retrieved in accordance with the following activities:
 - (a) Gathering of the researcher's names with which he/she has published. This activity is based on the names' list resulting from Step 4 in order to obtain the names' list with which a researcher has published. Since the format of the researcher' name is as surname and name, to standardize the names with the format presented in the dblp dataset, the names are reordered as name and surname.
 - (b) Collection of interesting research areas for the researcher. For this activity, the research areas of interest for the researchers are gathered from the results obtained in the data recommendation process using the proposed algorithm: C-HyRA.
 - (c) Collection of research areas not interesting for the researcher. For this activity, the file containing all the research areas present in the Scopus dataset, Appendix B.1, and the research areas specified as preferred for the researcher are used. Thereby, the research areas different to the areas of interest of the researcher are retrieved as the research areas in which the researcher has no preference.
 - (d) Retrieving of information related to the affiliation's organization of the researcher. In this activity, the Scopus dataset and the dataset built for C-HyRA concerning geographical influence are used in order to filter the researchers affiliated with Tecnológico de Monterrey, have the relationship between the researcher' name and the organization's name (to which the researcher is affiliated), and finally, concentrate the information of this organization, such as address, city, and country.

- (e) Assembly of the instances for the properties that can be obtained from the first information source. The Scopus identifier, publication's title, and publication's source (in which the article has been published), as well as the aforementioned data for this step are collected from Scopus.
 - (f) Formatting and validation of information retrieved. Finally, once all the data are obtained, these data are arranged according to the properties defined in Academic SUP. In this step, the assembly of all the instances is carried out allowing the redundant records.
6. Generation of a file containing the Academic SUP's properties and the values for each property according to the second information source chosen. For this activity, the data collected from dblp are the date of publication, researcher's name, publication's title, and publication's source. If the publication has more than one author, the authors' names are kept in the same record. In addition, it is important to mention that the publication's title in dblp contains a full stop by default, which is maintained given the purpose of this scenario.
7. Filtering of the authors for the second information source. Since only the records of the researchers affiliated with Tecnológico de Monterrey are required for this research, a search for these researchers is performed in dblp. However, dblp does not have a property related to the affiliation of the author. Therefore, the researchers' names collected from Scopus, in Step 4, are used to filter the records of these researchers in dblp. In this way, only the information of the researchers contained in the names' list obtained from Scopus, who are affiliated with Tecnológico de Monterrey, are retained.
8. Generation of the final file constituted of aligned information for researchers from the defined information sources. For this work, the data to align come from Scopus and dblp, as mentioned throughout the process. Firstly, the researcher's name and the publication's title contained in the dblp dataset are compared with the Scopus dataset. Then, two cases are presented: (1) the researcher's name and the publication's title contained in the dblp dataset do not match one instance of the Scopus dataset, thus, this instance is added to the final file resulting from the data alignment process; and (2) the researcher's name and the publication's title contained in the dblp dataset match one instance of the Scopus dataset, therefore, the record of Scopus is updated with the date of publication and publication's source of the corresponding publication, in their respective fields of the

final file. This update does not remove the information in the Scopus dataset but it adds the new information in the corresponding properties, in case the existing information is different from this one. Finally, the update and alignment changes resulting from this process are stored in the final file. In the same way, the original information of both Scopus and dblp are stored as separate files (one file for each information source) in order to support possible future activities (log).

As a result, the same 21 researchers presented in Table 5.2 are chosen to perform comparisons with this baseline scenario between both information sources: SIIP against Academic SUP (Scopus-dblp). The results of the baseline scenario related to the number of publications constituting the final dataset (Scopus-dblp) for each researcher, the count of publications stored in SIIP (retrieved from a VIVO's deployment [68]) for each researcher, and the number of publications that SIIP and Academic SUP (Scopus-dblp dataset) share for each researcher are provided in Table 5.5.

TABLE 5.5: Comparison of the integration results between Academic SUP and SIIP without data preprocessing and with a modified version of the proposed data alignment.

Researcher Identifier	Number of Publications in Academic SUP	Number of Publications in SIIP	Number of Publications in Common
Researcher-1	48	31	0
Researcher-2	34	137	32
Researcher-3	315	81	0
Researcher-4	126	11	0
Researcher-5	85	29	85
Researcher-6	0	4	0
Researcher-7	14	3	0
Researcher-8	12	43	8
Researcher-9	0	3	0
Researcher-10	2096	185	1080
Researcher-11	36	5	0
Researcher-12	75	10	0
Researcher-13	0	11	0
Researcher-14	8602	202	5610
Researcher-15	24	2	0
Researcher-16	720	98	170
Researcher-17	0	1	0
Researcher-18	306	30	306
Researcher-19	663	34	357
Researcher-20	272	15	136
Researcher-21	18	2	0

From the results reported in Table 5.5, it can be observed that the data integration process without prior data preprocessing leads to a negative impact on this process. This fact is based on the comparison of the figures obtained from the scenario (proposed scenario) including both data preprocessing and data alignment proposed in this work (Section 5.2.1) with the baseline scenario excluding data preprocessing and modifying the proposed data alignment process, as previously mentioned. Therefore, in order to more easily compare and analyze these figures, Table 5.6 is built with the results of Table 5.2 and Table 5.5. The columns of Table 5.6 are structured as follows:

- Researcher Identifier. It indicates the researcher identifier proposed in this work in order to maintain the data privacy.
- Publications in Academic SUP (#ASUP). Data that specifies the number of publications concentrated in Academic SUP following the data preprocessing and alignment processes proposed in this work (proposed scenario).
- Baseline Publications in Academic SUP (#BL-ASUP). Data that indicates the number of publications concentrated in Academic SUP following the baseline scenario.
- Publications in SIIP (#SIIP). Data that specifies the number of publications retrieved from SIIP following the data preprocessing and alignment processes proposed in this work (proposed scenario).
- Baseline Publications in SIIP (#BL-SIIP). Data that indicates the number of publications retrieved from SIIP following the baseline scenario.
- Publications in Common (#C). Data that specifies the number of publications in common between Academic SUP and SIIP following the data preprocessing and alignment processes proposed in this work (proposed scenario).
- Baseline Publications in Common (#BL-C). Data that indicates the number of publications in common between Academic SUP and SIIP following the baseline scenario.

Then, regarding Table 5.6, several specific conclusions can be given according to the following three comparisons:

1. #ASUP and #BL-ASUP. This examination presents changes in the number of publications reported in Academic SUP in the two scenarios. Firstly, the number of publications

TABLE 5.6: Concentration of the results of the proposed scenario and the baseline scenario for the data preprocessing and alignment processes described in this work.

Researcher Identifier	#ASUP	#BL-ASUP	#SIIP	#BL-SIIP	#C	#BL-C
Researcher-1	48	48	31	31	5	0
Researcher-2	78	34	137	137	42	32
Researcher-3	53	315	81	81	23	0
Researcher-4	2	126	11	11	0	0
Researcher-5	1	85	29	29	1	85
Researcher-6	1	0	4	4	0	0
Researcher-7	11	14	3	3	0	0
Researcher-8	3	12	43	43	2	8
Researcher-9	46	0	3	3	0	0
Researcher-10	60	2096	185	185	38	1080
Researcher-11	2	36	5	5	0	0
Researcher-12	8	75	10	10	1	0
Researcher-13	9	0	11	11	1	0
Researcher-14	46	8602	202	202	30	5610
Researcher-15	36	24	2	2	0	0
Researcher-16	5	720	98	98	1	170
Researcher-17	2	0	1	1	0	0
Researcher-18	2	306	30	30	2	306
Researcher-19	13	663	34	34	7	357
Researcher-20	5	272	15	15	2	136
Researcher-21	1	18	2	2	0	0

concentrated in Academic SUP for both scenarios is the same. For example, the researcher identified as Researcher-1 has 48 publications in both the proposed scenario and the baseline scenario. Secondly, the number of publications in the baseline scenario is less than the number of publications in the proposed scenario since the missing publications were not associated with the researcher in this baseline scenario. For instance, the researcher identified as Researcher-2 has 78 publications in the proposed scenario while he/she has 34 publications in the baseline scenario. Lastly, the number of publications in the baseline scenario is greater than the number of publications in the proposed scenario since the publications associated with the researcher were not validated to avoid repeated records of the same publication in this baseline scenario. For example, the researcher identified as Researcher-10 has 60 publications in the proposed scenario while he/she has 2096 publications in the baseline scenario.

2. #SIIP and #BL-SIIP. This comparison does not change in the two scenarios since the number of publications retrieved from SIIP for both scenarios is the same for all researchers.

Such a result is obtained because these publications come from the same information source (SIIP) and different processes were not carried out to treat this dataset in the baseline scenario.

3. #C and #BL-C. As the comparison of Academic SUP, this examination presents changes in the number of publications reported in common between Academic SUP and SIIP in the two scenarios. Firstly, the records of researchers who do not have coincidences with the publications concentrated in Academic SUP and SIIP increased in the baseline scenario, although the results provided by the proposed scenario show that there are shared publications. For example, the researcher identified as Researcher-1 has five publications in common in the proposed scenario while he/she has zero publications in the baseline scenario. Secondly, the number of publications in common in the baseline scenario is less than the number of publications in common in the proposed scenario since the missing publications were not associated with the researcher in Academic SUP (in the baseline scenario) or the publication's title is different in Academic SUP for both scenarios even if it is the same publication. For instance, the researcher identified as Researcher-2 has 42 publications in common in the proposed scenario while he/she has 32 publications in the baseline scenario, which are repeated records of the same publication. Lastly, the number of publications in common in the baseline scenario is greater than the number of publications in common in the proposed scenario since the publications associated with the researcher in Academic SUP were not validated to avoid repeated records of the same publication in this baseline scenario. For example, the researcher identified as Researcher-5 has one publication in common in the proposed scenario while he/she has 85 publications in the baseline scenario, which are repeated records of the same publication.

As a general conclusion, these variations are caused by special characters, orthographic accents, down exclamation marks, and other aspects not transformed into the baseline scenario by not including the data preprocessing proposed in this work (Section 3.3.1.2). In the same way, redundancy and lack of information are also presented in this baseline scenario by modifying certain rules in the proposed data alignment process, as mentioned at the beginning of this Section, such as not validating the repeated records. Therefore, this baseline scenario demonstrates that the data preprocessing and alignment processes proposed in this research are key parts to achieve the main objective of this research: a consistent data integration and generation.

5.2.3 Data Requirements for C-HyRA: the User's Explicit Preferences, Research Areas (Categories/Classification), and Geographical Information

As mentioned, C-HyRA aims to suggest a research areas list of interest for a researcher based on the information concentrated in Academic SUP as well as the features considered in the traditional and POI recommendation algorithms, such as categories (also named as tags or topics) and geographical influence. For the proposed use case, the user's explicit preferences (ratings), the publication's categories, and the ITESM' Campuses geographical influence are addressed.

Firstly, the dataset composed of the ratings for the publications' research areas is shown. Subsequently, the dataset constituted by research areas and subject categories for the publications (documents) is described. Finally, the dataset integrated by the ITESM' Campuses geographical information is explained.

5.2.3.1 The User's Explicit Preferences about Research Areas

Regarding the user's explicit feedback, a subset of the information collected through the data retrieval process is extracted to build the dataset. Likewise, the ratings calculated by the function described in Section 3.4.1 are integrated into this dataset, according to the user and the research area, in order to provide the user's preferences explicitly, which are used by the proposed recommendation algorithm in its traditional part. This dataset is made up of the 27 ratings of 852 people. Because the dataset has 23,004 records, only the structure and data of one of the users are shown in Table 5.7. In addition, a brief description of its columns is provided below.

- **User Identifier.** The data that identifies the researcher, solely for purposes of this work since personal information is not collected.
- **Research Area Name.** The name that distinguishes each research area.
- **Rating.** The numerical value calculated for each research area according to the specified user's preferences. This value is within the range from three to five, being three "little interesting", four "interesting", and five "very interesting". The value one is assigned to the research areas not explored (or "not interesting") by the user.

TABLE 5.7: The user's explicit preferences dataset: an example of the structure for each user.

User Identifier	Research Area Name	Rating
User1	Multidisciplinary	1
	Agricultural and Biological Sciences	1
	Arts and Humanities	1
	Biochemistry, Genetics and Molecular Biology	1
	Business, Management and Accounting	1
	Chemical Engineering	1
	Chemistry	1
	Computer Science	5
	Decision Sciences	1
	Earth and Planetary Sciences	1
	Economics, Econometrics and Finance	1
	Energy	1
	Engineering	5
	Environmental Science	1
	Immunology and Microbiology	1
	Materials Science	1
	Mathematics	5
	Medicine	1
	Neuroscience	1
	Nursing	1
	Pharmacology, Toxicology and Pharmaceutics	1
	Physics and Astronomy	1
	Psychology	1
	Social Sciences	1
Veterinary	1	
Dentistry	1	
Health Professions	1	

5.2.3.2 Research Areas: Categories/Classification

Concerning the publications' categories dataset, all of the information sources used to retrieve data and to build the Academic SUP schema were consulted in order to select the information sources that contain a subject categorization scheme. As a result, Scopus® and Web of Science™ are chosen.

However, because two schemata are collected and these schemata contain both similar and different research areas and subject categories, then it is necessary to carry out a unification process. Therefore, firstly, the two schemata will be presented to subsequently introduce the unified publications' areas and categories dataset as well as the description of the unification process carried out.

5.2.3.2.1 Scopus

Scopus offers the broadest coverage of peer-reviewed literature and quality web sources in different research fields [84]. Such a content is classified under four broad subject clusters: life sciences, physical sciences, health sciences, and social sciences and humanities. These clusters in turn are divided into 27 research areas and more than 300 categories (minor subject areas) [84], areas that will be used in this proposal. In addition, titles on Scopus can belong to more than one subject area.

The Scopus' major subject areas are listed below and their minor subject areas are introduced in Appendix B.1. For this use case, these areas are retrieved from records stored in the ITESM's institutional database come from Scopus [69]. Therefore, this dataset is constituted by 27 (major) research areas and 310 categories.

1. "Multidisciplinary" has one category "Multidisciplinary"; therefore, a table is not necessary.
2. "Agricultural and Biological Sciences", Table B.1.
3. "Arts and Humanities", Table B.2.
4. "Biochemistry, Genetics and Molecular Biology", Table B.3.
5. "Business, Management and Accounting", Table B.4.
6. "Chemical Engineering", Table B.5.
7. "Chemistry", Table B.6.
8. "Computer Science", Table B.7.
9. "Decision Sciences", Table B.8.
10. "Earth and Planetary Sciences", Table B.9.
11. "Economics, Econometrics and Finance", Table B.10.
12. "Energy", Table B.11.
13. "Engineering", Table B.12.
14. "Environmental Science", Table B.13.

15. “Immunology and Microbiology”, Table [B.14](#).
16. “Materials Science”, Table [B.15](#).
17. “Mathematics”, Table [B.16](#).
18. “Medicine”, Table [B.17](#).
19. “Neuroscience”, Table [B.18](#).
20. “Nursing”, Table [B.19](#).
21. “Pharmacology, Toxicology and Pharmaceutics”, Table [B.20](#).
22. “Physics and Astronomy”, Table [B.21](#).
23. “Psychology”, Table [B.22](#).
24. “Social Sciences”, Table [B.23](#).
25. “Veterinary”, Table [B.24](#).
26. “Dentistry”, Table [B.25](#).
27. “Health Professions”, Table [B.26](#).

5.2.3.2.2 Web of Science

Web of Science proposes a subject categorization scheme that is used by all Web of Science product databases. This scheme is constituted by five research areas and 151 subject categories. Thereby, journals and books covered by Web of Science Core Collection are assigned to at least one Web of Science’s subject category, which, in turn is mapped to one research area [\[104\]](#).

The Web of Science’s research areas are listed below and their categories are introduced in [Appendix B.2](#).

1. Arts & Humanities, Table [B.27](#).
2. Life Sciences & Biomedicine, Tables [B.28](#) and [B.29](#).
3. Physical Sciences, Table [B.30](#).
4. Social Sciences, Table [B.31](#).
5. Technology, Table [B.32](#).

5.2.3.2.3 The Unified Publications' Areas and Categories Dataset

The unification process is carried out semi-automatically from the Web of Science's categories to Scopus's areas and categories, according to the activities described below.

1. Punctuation marks. The Web of Science's areas and categories include the ampersand sign to join two names while the Scopus's areas and categories use the word "and". Thus, the ampersand sign is replaced by the word "and".
2. The same area/category's name. Some of the Web of Science's areas and categories have names equal to the Scopus's areas and categories, respectively. Hence, when an area or category's name is repeated, one of them is removed. For example, the category "Communication" belonging to the area "Social Sciences" has the same name and area in both Web of Science and Scopus.
3. The category's name contained in a research area. Some of the Web of Science's categories have names that are included in the Scopus's areas. Therefore, when a category's name is contained in a research area, both names are maintained. For example, the category "Veterinary Sciences" in Web of Science is aligned with the area "Veterinary" in Scopus.
4. The category's name contained in another category. Some of the Web of Science's categories have names that are included in the Scopus's categories. Hence, when a category's name is contained in another category, both names are maintained. For example, the category "International Relations" in Web of Science is aligned with the category "Political Science and International Relations" in Scopus.
5. The same category's name and the different area's name. Some of the Web of Science's categories have names equal to the Scopus's categories but with different research areas. When a category's name is the same for both information sources but the area's name is different, one of the category's names is removed and the areas' names are maintained. For example, the category "Biophysics" has the same name for both information sources; however, it belongs to the area "Life Sciences and Biomedicine" in Web of Science and the area "Biochemistry, Genetics and Molecular Biology" in Scopus.
6. The different area/category's name. Some of the Web of Science's areas and categories do not match the Scopus's areas and categories by name. Therefore, when an area and category are not found in the Scopus's areas and categories, this pair of names are added

to the Scopus's areas and categories in similar or related areas and categories, such as branches. For this purpose, a superficial research about the Web of Science's categories on the Web is carried out. For example, the category "Entomology" and area "Life Sciences and Biomedicine" in Web of Science is aligned with the category "Insect Science" and area "Agricultural and Biological Sciences" in Scopus.

As a result, a dataset consisting of 27 research areas (the same as in Scopus) and 670 subject categories, considering duplicate categories, is generated for C-HyRA. The research areas are listed below and their categories are introduced in Appendix B.3.

1. "Multidisciplinary" has one category "Multidisciplinary"; therefore, a table is not necessary.
2. "Agricultural and Biological Sciences", Table B.33.
3. "Arts and Humanities", Table B.34.
4. "Biochemistry, Genetics and Molecular Biology", Table B.35.
5. "Business, Management and Accounting", Table B.36.
6. "Chemical Engineering", Table B.37.
7. "Chemistry", Table B.38.
8. "Computer Science", Table B.39.
9. "Decision Sciences", Table B.40.
10. "Earth and Planetary Sciences", Table B.41.
11. "Economics, Econometrics and Finance", Table B.42.
12. "Energy", Table B.43.
13. "Engineering", Table B.44.
14. "Environmental Science", Table B.45.
15. "Immunology and Microbiology", Table B.46.
16. "Materials Science", Table B.47.

17. “Mathematics”, Table B.48.
18. “Medicine”, Tables B.49, B.50, and B.51.
19. “Neuroscience”, Table B.52.
20. “Nursing”, Table B.53.
21. “Pharmacology, Toxicology and Pharmaceutics”, Table B.54.
22. “Physics and Astronomy”, Table B.55.
23. “Psychology”, Table B.56.
24. “Social Sciences”, Table B.57.
25. “Veterinary”, Table B.58.
26. “Dentistry”, Table B.59.
27. “Health Professions”, Table B.60.

5.2.3.3 Geographical Information

In accordance with the aim of incorporating the geographical influence factor into C-HyRA, a dataset constituted of geographical information related to the 31 ITESM’s Campuses was built. To this end, firstly, information of each Campus was manually collected from the official website of each one of them [105]. Secondly, addresses and geographical coordinates were complemented and obtained, respectively, through Google Maps. Finally, the consolidated information was manually preprocessed to standardize the format of presentation. This preprocessing mainly consisted in adding “ITESM” before the Campus name (*e.g.*, ITESM Estado de Mexico), homogenizing the abbreviations for the street and settlement type (*e.g.*, Carretera Lago de Guadalupe Km. 3.5 and Col. Margarita Maza de Juarez, respectively), adding “No.” before the street number (*e.g.*, Eugenio Garza Sada No. 2501), deleting “C.P.” in the postal code field, replacing the abbreviations given for the Federal Entity name by the Federal Entity full name (*e.g.*, Nuevo Leon instead of N.L.), and removing accents since the information is in the Spanish language (*e.g.*, Nuevo Leon instead of Nuevo León).

As a result, the structure and records of this dataset are presented in Table 5.8 and Table 5.9. In addition, a brief description of the information contained in each field is provided below.

- Id. Field that comprises a Campus's identifier added for purposes of this research work.
- CampusName. Column that comprehends the Campus's name.
- StreetType. Field that stores the street type, name, and number.
- SettlementType. Column that concentrates the settlement type and name.
- Code. Column that comprises the postal code.
- FederalEntity. Field that encompasses the Municipality and Federal Entity name.
- Country. Column that includes the Country name.
- Location. Field that indicates the Campus' approximate coordinate in decimal degrees, whose format is [latitude, longitude].

TABLE 5.8: The ITESM's Campuses geographical information dataset (Part I)

Id	CampusName	StreetType	SettlementType	Code	FederalEntity	Country	Location
01	ITESM Aguascalientes	Av. Eugenio Garza Sada No. 1500	Los Pocitos	20328	Aguascalientes, Aguascalientes	Mexico	21.9336263,- 102.3399448
02	ITESM Central de Veracruz	Av. Eugenio Garza Sada No. 1	Las Quintas	94500	Cordoba, Veracruz	Mexico	18.8919169,- 96.9792207
03	ITESM Chiapas	Carretera a Tapanatepec Km. 149 + 746	Col. Juan Crispin	29020	Tuxtla Gutierrez, Chiapas	Mexico	16.7652652,- 93.2009195
04	ITESM Chihuahua	Av. Heroico Colegio Militar No. 4700	Col. Nombre de Dios	31300	Chihuahua, Chihuahua	Mexico	28.6738206,- 106.0778187
05	ITESM Ciudad de Mexico	Calle del Puente No. 222	Col. Ejidos de Huipulco	14380	Tlalpan, Ciudad de Mexico	Mexico	19.2835373,- 99.1352117
06	ITESM Ciudad Juarez	Blvd. Tomas Fernandez Campos No. 8945	Parque Industrial Antonio J. Bermudez	32470	Ciudad Juarez, Chihuahua	Mexico	31.718026,- 106.3937024
07	ITESM Ciudad Obregon	California No. 2100	Col. Obregon Norte	85010	Ciudad Obregon, Sonora	Mexico	27.5316293,- 109.9450577
08	ITESM Cuernavaca	Autopista del Sol Km. 104	Col. Real del Puente	62790	Xochitepec, Morelos	Mexico	18.805788,- 99.2217638
09	ITESM Cumbres (Preparatoria)	Linces No. 1000	Col. Cumbres Elite	64349	Monterrey, Nuevo Leon	Mexico	25.7340182,- 100.4162218
10	ITESM Estado de Mexico	Carretera al Lago de Guadalupe Km. 3.5	Col. Margarita Maza de Juarez	52926	Atizapan de Zaragoza, Estado de Mexico	Mexico	19.5932342,- 99.2292182
11	ITESM Eugenio Garza Laguera (Preparatoria)	Topolobampo No. 4603	Valle de las Brisas	64790	Monterrey, Nuevo Leon	Mexico	25.6175375,- 100.2756982
12	ITESM Eugenio Garza Sada (Preparatoria)	Dinamarca No. 451	Del Carmen	64710	Monterrey, Nuevo Leon	Mexico	25.6698533,- 100.35707
13	ITESM Guadalajara	Av. General Ramon Corona No. 2514	Nuevo Mexico	45138	Zapopan, Jalisco	Mexico	20.7350183,- 103.4550059
14	ITESM Hidalgo	Blvd. Felipe Angeles No. 2003	Col. Venta Prieta	42083	Pachuca de Soto, Hidalgo	Mexico	20.0961826,- 98.7674321
15	ITESM Irapuato	Paseo Mirador del Valle No. 445	Villas de Irapuato	36670	Irapuato, Guanajuato	Mexico	20.6867749,- 101.3948987

TABLE 5.9: The ITESM's Campuses geographical information dataset (Part II)

Id	CampusName	StreetType	SettlementType	Code	FederalEntity	Country	Location
16	ITESM Laguna	Paseo del Tecnológico No. 751	Col. Ampliación La Rosita	27250	Torreón, Coahuila	Mexico	25.5173638,- 103.3983336
17	ITESM León	Av. Eugenio Garza Sada S/N	Col. Cerro Gordo	37190	León, Guanajuato	Mexico	21.1672807,- 101.7144521
18	ITESM Monterrey	Av. Eugenio Garza Sada No. 2501 Sur	Col. Tecnológico	64849	Monterrey, Nuevo León	Mexico	25.6507023,- 100.2898074
19	ITESM Morelia	Av. Montaa Monarca No. 1340	Ejido Jesus del Monte	58350	Morelia, Michoacan	Mexico	19.6563803,- 101.1641237
20	ITESM Puebla	Via Atlixcayotl No. 2301	Reserva Territorial Atlixcayotl	72453	Puebla, Puebla	Mexico	19.0182913,- 98.2413654
21	ITESM Querétaro	Epigmenio Gonzalez No. 500	Fracc. San Pablo	76130	Santiago de Querétaro, Querétaro	Mexico	20.6130674,- 100.4085385
22	ITESM Saltillo	Prolongacion Juan de La Barrera No. 1241	Las Cumbres	25270	Saltillo, Coahuila	Mexico	25.4485119,- 100.9748456
23	ITESM San Luis Potosí	Av. Eugenio Garza Sada No. 300	Lomas del Tecnológico	78211	San Luis Potosí, San Luis Potosí	Mexico	22.1273754,- 101.0386315
24	ITESM Santa Catarina (Preparatoria)	Av. Dr. Ignacio Morones Prieto No. 290	Sin Nombre de Col. No. 11	66180	Santa Catarina, Nuevo León	Mexico	25.6620159,- 100.4305262
25	ITESM Santa Fe	Av. Carlos Lazo No. 100	Col. Santa Fe	01389	Delegacion Alvaro Obregon, Ciudad de Mexico	Mexico	19.3597741,- 99.2581498
26	ITESM Sinaloa	Bldv. Pedro Infante No. 3773 Pte.	Recursos Hidraulicos	80100	Culiacan, Sinaloa	Mexico	24.8011206,- 107.4214074
27	ITESM Sonora Norte	Bldv. Enrique Mazon Lopez No. 965		83000	Hermosillo, Sonora	Mexico	29.1696785,- 110.9113319
28	ITESM Tampico	Bldv. Petrocel Km. 1.3	Puerto Industrial	89600	Altamira, Tamaulipas	Mexico	22.380393,- 97.9013296
29	ITESM Toluca	Av. Eduardo Monroy Cardenas No. 2000	San Antonio Buenavista	50110	Toluca de Lerdo, Estado de Mexico	Mexico	19.2682249,- 99.7056808
30	ITESM Valle Alto (Preparatoria)	Carretera Nacional Km. 267.7	Col. La Estanzuela	64986	Monterrey, Nuevo León	Mexico	25.5712835,- 100.2495774
31	ITESM Zacatecas	Av. Pedro Coronel No. 16	Col. Dependencias Federales	98000	Guadalupe, Zacatecas	Mexico	22.7475198,- 102.521187

5.2.4 Recommendations based on C-HyRA

Once the datasets related to the user's explicit preferences (ratings), the publication's categories, and the ITESM' Campuses geographical influence are built, the extraction of the ground-truth subset and the pilot subset consisting of the explored research areas of each user is carried out.

However, another step is performed before starting with the extraction, the definition of the number of clusters for the K-means algorithm used in the geographical influence part. Although this definition is beyond the scope of this proposal, a test on the number of clusters is introduced. That is, initially, the number of clusters was defined in three, but, in order to obtain new clusters, this number was increased. This new number of clusters was established in four. Nevertheless, when the new clusters were visualized (four), the institutions composing the first three clusters were maintained, except for the Campus "ITESM Chihuahua", which constituted the fourth cluster. Therefore, since the fourth cluster is composed of only one institution, the recommendations will be limited to only this Campus. Consequently, the number of clusters is again adjusted to three.

Afterwards, the ground-truth subset is extracted from the dataset corresponding to the user's explicit preferences in order to provide C-HyRA with such preferences, as well as the pilot subset with the aim of comparing the recommendations given by C-HyRA against the research areas contained in this pilot subset to validate such recommendations. For this purpose, 100 executions of C-HyRA are carried out, obtaining the results reported in Tables 5.10 and 5.11.

Table 5.10 concentrates the number of users to whom C-HyRA gives a recommendation, as well as the number of users, of the total number of users, who are concentrated in the pilot subset and the number of users, of the total number of users, who do not belong to the pilot subset for 100 executions of the experiments. That is, in 100 runs of the experiments, the number of users to whom C-HyRA gives recommendations, the number of users in the pilot subset, and the number of users not present in the pilot subset is always the same. Although C-HyRA gives recommendations to users who are not present in the pilot subset, these users were not taken into account for the results of the comparison of the recommendations due to this fact. The results of the comparisons were calculated for each experiment carried out, resulting in a percentage of coincidence of 100% in each one and with at least half of the research areas of interest recommended by C-HyRA by each user. Because the repetitiveness of the results, in Table 5.11 is shown a result of the comparison of 100 results of the experiments.

TABLE 5.10: Summary of users with C-HyRA's recommendations through 100 executions.

Number of users with recommendations	Number of users present in the pilot subset	Number of users not present in the pilot subset
23	12	11

TABLE 5.11: Results of the comparison between the recommendations given by C-HyRA and the research areas concentrated in the pilot subset.

User Identifier	% Coincidence	Recommended Research Areas / Research Areas in the Pilot Subset
User1	100	4/5
User2	100	3/5
User3	100	5/5
User4	100	3/5
User5	100	4/5
User6	100	4/5
User7	100	4/5
User8	100	4/5
User9	100	4/5
User10	100	4/5
User11	100	3/5

From Table 5.11, it can be noticed that all the recommendations provided by C-HyRA match the user's preference areas. On the one hand, the second column concentrates the percentage of coincidence of the research areas recommended to the user that are present in the pilot subset of the same user. For these experiments, the percentage obtained by the recommendation algorithm in providing the user's preferences is 100% in all cases. This percentage must be interpreted as that C-HyRA is able to recommend the research areas of interest for a specific user, and not that C-HyRA achieves 100% accuracy. This, since the ratings for the research areas and categories were deduced by C-HyRA for both the pilot subset and the ground-truth subset. On the other hand, in the third column, it can be noticed that C-HyRA can recommend at least half of the research areas of interest that are contained in the users' pilot subset. As a consequence, these results are promising in showing that C-HyRA can be adopted as a hybrid recommendation algorithm in the field of research.

5.2.5 Front-End: Data Visualization

A VIVO's version has been installed on the Windows OS to visualize the results of this research work. The complete installation and configuration process is carried out as specified

in the VIVO's official documentation page [106]. However, complementary information that in this VIVO deployment is consulted and performed, such as information about the software prerequisites and configuration to install VIVO on Windows OS, is provided below.

On the one hand, the VIVO platform requires that the following software is pre-installed in order to successfully run on any OS.

- Java (SE). Version 1.7.x or higher is required, available from its official website [107]. A special installation or configuration process is not required.
- Apache Tomcat. Version 7.x or higher is required, available from its official website [108]. The configuration process followed to work properly is described in [109].
- Apache Ant. Version 1.8 or higher is required, available from its official website [110]. The configuration process followed to work properly is described in [111].
- MySQL. Version 5.1 or higher is required, available from its official website [112]. The configuration process followed to work properly is described in [113].

On the other hand, in this research work, the VIVO's version 1.8.1 is chosen for its deployment because the documentation for this version is more detailed when this deployment is carried out. The VIVO's version 1.8.1 is downloaded from the official GitHub website [114].

Chapter 6

Conclusions and Further Work

To highlight the contributions and findings of this research work, firstly, the research questions are introduced and discussed. Subsequently, some issues and findings, as well as the advantages and disadvantages of this proposal, are described. Finally, future work is presented.

The research questions proposed in this research were achieved as follows:

- I *What structure and features must the user's profile meta-schema, that allows modeling representative information of his/her academic field in a unified, standardized, online, and interoperable format, have?*

This research work provides a user's academic profile meta-schema, called Academic SUP, based on schemata related to the Human Resources and Computer Science areas as well as schemata related to academic information sources. Consequently, Academic SUP allows researchers to have an interdisciplinary and justified representation of information representative of his/her academic field. In addition, since Academic SUP was built as an ontology, this meta-schema allows modeling such information in a structured and interoperable format. Thereby, Academic SUP is a meta-schema that follows the principles and conception of SW, entailing the benefits that it gives. However, the characteristics of unification and standardization were not fully achieved since the consulted schemata can be modified over time to be adjusted to new requirements or fields of application. Academic SUP will be available on the Web in order to be reused by other semantic approaches and support other standardization efforts related to the researcher's profile.

II *How can a framework, that allows integrating, generating, and updating the data of a researcher into a defined schema from information published in different academic information sources consistently way, be built?*

The proposed framework for consistent generation of linked data, called AkCeL, is considered as a contribution for the Computer Science discipline since it allows to carry out a consistent generation of linked data coming from different types of information sources, and therefore, with distinct formats, *i.e.*, semi-structured (dblp) and structured data (Scopus). This goal was achieved by the principles of design and built taken into account for its structure from DSS and the information integration centralized architecture from MSDF field. In the same way, the processes integrated into AkCeL allow to contribute to the data integration and generation, consistently. On the one hand, AkCeL performs data retrieval, the preprocessing of these data, and then, its alignment. On the other hand, AkCeL makes the recommendation of a research areas list of interest for researchers as well as the visualization of the data generated by this framework and certain research statistics produced with the same information. From the evaluation scenario presented in the use case related to the academic sector, the results are promising since AkCeL allows accomplishing the purpose of supporting both researchers and data managers at the institutional level in their decision-making processes concerning the Computer Science field. Although these results need to be validated in medium and large scale in different research areas.

III *How can a recommendation algorithm, that allows to embrace the researcher preferences and the contextual information related to the research area, be built?*

The proposed recommendation algorithm's approach allows achieving the objective of this question. This approach incorporates the user's explicit preferences (ratings) and contextual information related to categories and geographical information. Moreover, the proposed recommendation algorithm integrates an average aggregation operator consisting of five similarity and distance measures as a single measure into the user-based CF algorithm. This conclusion is based on the results presented in this proposal since the proposed recommendation algorithm approach for both the use case related to the research area and the use case related to the tourism sector, C-HyRA and HyRA, respectively, has shown promising recommendations to user profiles of each scenario. Therefore, C-HyRA and HyRA are also

considered as contributions for these fields given the results shown in this proposal. Similarly, the evaluations of these algorithms have provided datasets that can be used to assess other recommendation approaches in the same areas. In addition to proposing functions or processes to build such datasets. For example, to integrate the publications' areas and categories into C-HyRA, a proposal to unify the research areas and categories of Web of Science and Scopus was described.

Furthermore, regarding the use case related to the research area, the Tecnológico de Monterrey's researchers community is introduced as the first users of AkCeL. To carry out the validation of the initial generation of their academic profiles concerning their publications, the information provided by the SIIP repository about them, through a VIVO's implementation, is used. Such validation gives encouraging results.

Further considerations, as well as some issues and findings concerning this proposal and its use case, are described below.

- Although the data collection was carried out for all the research areas mentioned in this proposal, only the information related to Computer Science was used by the data alignment process in order to have a controlled evaluation scenario covered by the dblp approach.
- Since the conception of Academic SUP, the aggregation or inference of new relationships in Academic SUP has not been studied because it is not part of the research objectives. Therefore, Academic SUP is mentioned as a static schema, that is, the Academic SUP's structure will remain unchanged once it is defined.
- Regarding the data retrieval process, the number of information sources that AkCeL addressed was limited to two, dblp and Scopus. An issue that affected the incorporation of another private information source is due to the necessary permits for exploitation.
- Considering that the language of the information sources chosen for this proposal is mainly English and that the proposed ontology is also specified in English, the information collected, if it is in a different language, is transformed into a standardized format with respect to the English language without carrying out a translation of it. For instance, for the Spanish language, the orthographic accents are removed. Therefore, to transform the data into a standard format, a data preprocessing process was integrated into AkCeL.

- Due to the diverse nature of the information sources, different presentations of data from the same information source are provided, *i.e.*, the “X” information source can be exploited through its unstructured or semi-structured format. For example, the “dblp” information can be retrieved through its website or using its XML file. For this proposal, the first, second, and third extraction option is established as the structured, semi-structured, and unstructured format, respectively.
- The information available in the information sources was incomplete for the recommendation process. However, computational solutions were proposed to deal with these information gaps.
- Considering the continuous improvement of AkCeL as well as another support tool for the decision-making processes of data managers at the institutional level, a log file was proposed to record the workflow of each process. Thereby, the tracking of the collected data can be carried out. As a result, updates to these processes can be based on this log file.
- The VIVO platform is adopted since the purpose of this proposal is to provide statistics and to allow the interaction with the information generated by AkCeL rather than the development of a visualization platform.
- As a recommendation, an information source pertaining to the researcher’s affiliation institution that provides information related to that affiliation and his/her position is needed to supplement his/her information. In this way, the processes of AkCeL could be better supported and validated.

Finally, this research work brings more advantages than disadvantages according to the benefits and issues that have been identified, which are listed below.

Pros

1. The integrated information is more representative and complete by coming from different information sources.
2. Reliable information is visualized by coming from reliable information sources.
3. Updated information could be provided by integrating new data almost simultaneously as they are indexed.

4. A new researcher could more quickly identify researchers interested in a particular research area of their educational institution.
5. AkCeL could allow researchers to identify prominent people as well as new researchers in a specific research area of their educational institution.
6. Researchers could be in contact with other researchers more effectively by having updated information, *e.g.*, via affiliation's email.
7. The visualization of the researcher's public profile, according to the geographical area, could allow establishing new connections with other research groups that are in the same city for possible collaborations.
8. The researcher could more easily query and visualize his/her publication's statistics by having his/her data unified consistently from different information sources.
9. The researcher's information will not be visible to all people interacting with the VIVO platform because only the authenticated researcher will be able to query all his/her information.
10. Educational institutions could generate their research statistics more quickly, consistently, and representatively.
11. Educational institutions and researchers could support their decision-making processes in the research statistics and information generated by AkCeL.

Cons

1. A researcher will not have access to all of his/her information generated by AkCeL if he/she is not a registered user. Only a partial view of his/her information will be displayed.
2. It is necessary for researchers or data managers to make an initial registration in the VIVO platform in order that AkCeL has more accurate data of the initial user profiles.
3. It is necessary for the researcher to carry out a validation of the information generated by AkCeL, preferably whenever updates are shown, with the aim of supporting the data consistency.

6.1 Future Work

The FIWARE platform [115] can be a solution to develop a productive application related to AkCeL using its components (Generic/Specific Enablers). For example, to carry out the authentication of a user, the Identity Manager component of the FIWARE platform [116] is considered since can allow to quickly incorporate this activity into the application and therefore, it will reduce the development time of this authentication component. For this authentication process, two options are proposed: authentication through one of the information sources considered in this proposal or registration via a questionnaire linked to his/her institutional email account.

In the same way for AkCeL, in order to develop a customized update, a schema of update periods can be incorporated for all the properties defined into Academic SUP. As a consequence, all the processes established in this framework will be executed only when an update notification is released. From the evaluation side of AkCeL, a validation of the results obtained from AkCeL can be carried out by expert users (researchers). For example, to collect feedback from the researchers reported in these experiments about the usefulness and impact of the information generated by AkCeL.

For the use case of HyRA, the analysis of all the results obtained from the surveys is considered to design an application for the tourism sector in Ceutí oriented to the target audience's preferences. In addition, the standardization of both the hierarchy of the categories and the categories for Smart POIs is proposed. Furthermore, the incorporation of more contextual information, such as the time factor, can be integrated into HyRA and C-HyRA. For example, to consider the museums' opening hours before recommending them.

Finally, to expand the input data concerning the publications' areas and categories for C-HyRA, the allocation of research areas and categories for the publications concentrated in dblp can be performed.

Appendix A

Definitions

A.1 Data Retrieval

- Extensible Markup Language (XML) [117]. A W3C open standard focused on describing data using embedded tags. That is, XML defines what the elements contain on a page.
- Structured Query Language (SQL) [118, 119]. A relational data language composed of a data definition language (DDL), a data manipulation language (DML), and a data control language (DCL).

A.2 Data Preprocessing

- Data transformation. It occurs when data is not found in the desired format. Namely, all data from a defined field are transformed into a common format via powerful tools, such as normalization, data discretization, and concept hierarchy generation [86].
- Data cleaning. It is applied to detect and/or to eliminate noise and outliers, to fill missing values, and to resolve inconsistencies [86].
- Data integration. It is used to merge data coming from multiple sources, such as databases, into a coherent data store. For example, a data warehouse [86].
- Data reduction. Just as its name implies, it is employed to reduce the data or dataset size through other processes, such as removing redundant features or irrelevant attributes, data aggregation, and clustering [86].

A.3 Data Recommendation

Similarity and Distance Measures:

- **Euclidean.** Euclid stated that a line is the shortest distance between two points. Euclidean distance is represented in Equation (A.1) [120].

$$Euclidean = \sqrt{\sum_{i=1}^n |P_i - Q_i|^2} \quad (A.1)$$

where P_i and Q_i are components of an Euclidean vector indexed with i ; and n is the sample size.

- **Pearson.** It is a measure of the strength of a linear association between two variables. In a broad sense, the Pearson correlation coefficient returns the distance of all data points that best fit through data. Its representation is given by Equation (A.2) [121].

$$Pearson = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \quad (A.2)$$

where x_i and y_i are single samples indexed with i ; n is the sample size; \bar{x} and \bar{y} are the sample mean of x and y , respectively; and S_x and S_y are the sample standard deviation of x and y , respectively.

- **Cosine.** It is also called the angular metric. It measures the angle between two vectors, *i.e.*, it is the normalized inner product. The cosine similarity metric is represented in Equation (A.3) [120].

$$Cosine = \frac{\sum_{i=1}^n P_i Q_i}{\sqrt{\sum_{i=1}^n P_i^2} \sqrt{\sum_{i=1}^n Q_i^2}} \quad (A.3)$$

where P_i and Q_i are components of a vector indexed with i ; and n is the sample size.

- **Manhattan.** It is also known as rectilinear distance and taxicab norm. It calculates several projections in the mathematical space, where the size of blocks does not affect the

distances. The Manhattan distance is represented in Equation (A.4) [120, 122].

$$Manhattan = \sum_{i=1}^n |P_i - Q_i| \quad (A.4)$$

where P_i and Q_i are components of a vector indexed with i ; and n is the sample size.

- **Chebyshev.** It is also called the chessboard distance in 2-D or minimax approximation. It was derived by Pafnuty Lvovich Chebyshev. This distance is used when the value of P tends to infinity. Its representation is given by Equation (A.5) [120].

$$Chebyshev = \max_i |P_i - Q_i| \quad (A.5)$$

where P_i and Q_i are components of a vector indexed with i .

A.4 User Profile and Ontologies

For the Human Resources area [52]:

- Anti-profile defines a profile focused on the description of a person.
- Profile addresses the information related to a work position as well as the model of competencies defined by the organization.

For the purposes of this Thesis:

- Profile defines the description of a user in the research area.
- Anti-profile addresses the user's disinterests in the research area.

Ontologies:

1. **FOAF (Friend of a Friend).** It is used to model users and their personal relationships [76].
2. **vCard (Electronic Business Cards).** It is used to model people and their business relationships. This also includes information about their organizations [77, 78].

3. **dc (Dublin Core)**. It is used to describe users as authors of websites, articles, books, publications, and so on [79, 80].
4. **SKOS (Simple Knowledge Organization System)**. It is used to model information on the Web. This ontology is included in most ontologies [81].
5. **VIVO**. It is used to represent the expertise of people involved in the generation, transmission, and preservation of knowledge and creative works [82].

Appendix B

Research Areas:

Categories/Classification

B.1 Scopus

The Scopus' minor subject areas are presented according to the major subject areas below.

TABLE B.1: Subject categorization for the research area "Agricultural and Biological Sciences" proposed by Scopus

Research Area	Category
Agricultural and Biological Sciences	Agricultural and Biological Sciences (miscellaneous)
	Agronomy and Crop Science
	Animal Science and Zoology
	Aquatic Science
	Ecology, Evolution, Behavior and Systematics
	Food Science
	Forestry
	Horticulture
	Insect Science
	Plant Science
	Soil Science

TABLE B.2: Subject categorization for the research area “Arts and Humanities” proposed by Scopus

Research Area	Category
Arts and Humanities	Arts and Humanities (miscellaneous)
	History
	Language and Linguistics
	Archeology (arts and humanities)
	Classics
	Conservation
	History and Philosophy of Science
	Literature and Literary Theory
	Museology
	Music
	Philosophy
	Religious Studies
Visual Arts and Performing Arts	

TABLE B.3: Subject categorization for the research area “Biochemistry, Genetics and Molecular Biology” proposed by Scopus

Research Area	Category
Biochemistry, Genetics and Molecular Biology	Biochemistry, Genetics and Molecular Biology (miscellaneous)
	Aging
	Biochemistry
	Biophysics
	Biotechnology
	Cancer Research
	Cell Biology
	Clinical Biochemistry
	Developmental Biology
	Endocrinology
	Genetics
	Molecular Biology
	Molecular Medicine
	Physiology
Structural Biology	

TABLE B.4: Subject categorization for the research area “Business, Management and Accounting” proposed by Scopus

Research Area	Category
Business, Management and Accounting	Business, Management and Accounting (miscellaneous)
	Accounting
	Business and International Management
	Management Information Systems
	Management of Technology and Innovation
	Marketing
	Organizational Behavior and Human Resource Management
	Strategy and Management
	Tourism, Leisure and Hospitality Management
Industrial Relations	

TABLE B.5: Subject categorization for the research area “Chemical Engineering” proposed by Scopus

Research Area	Category
Chemical Engineering	Chemical Engineering (miscellaneous)
	Bioengineering
	Catalysis
	Chemical Health and Safety
	Colloid and Surface Chemistry
	Filtration and Separation
	Fluid Flow and Transfer Process
	Process Chemistry and Technology

TABLE B.6: Subject categorization for the research area “Chemistry” proposed by Scopus

Research Area	Category
Chemistry	Chemistry (miscellaneous)
	Analytical Chemistry
	Electrochemistry
	Inorganic Chemistry
	Organic Chemistry
	Physical and Theoretical Chemistry
	Spectroscopy

TABLE B.7: Subject categorization for the research area “Computer Science” proposed by Scopus

Research Area	Category
Computer Science	Computer Science (miscellaneous)
	Artificial Intelligence
	Computational Theory and Mathematics
	Computer Graphics and Computer-Aided Design
	Computer Networks and Communications
	Computer Science Applications
	Computer Vision and Pattern Recognition
	Hardware and Architecture
	Human-Computer Interaction
	Information Systems
	Signal Processing
Software	

TABLE B.8: Subject categorization for the research area “Decision Sciences” proposed by Scopus

Research Area	Category
Decision Sciences	Decision Science (Miscellaneous)
	Information Systems and Management
	Management Science and Operation Research
	Statistics, Probability and Uncertainty

TABLE B.9: Subject categorization for the research area “Earth and Planetary Sciences” proposed by Scopus

Research Area	Category
Earth and Planetary Sciences	Earth and Planetary Sciences (miscellaneous)
	Atmospheric Sciences
	Computers in Earth Sciences
	Earth Surface Processes
	Economic Geology
	Geochemistry and Petrology
	Geology
	Geophysics
	Geotechnical Engineering and Engineering Geology
	Oceanography
	Paleontology
	Space and Planetary Science
Stratigraphy	

TABLE B.10: Subject categorization for the research area “Economics, Econometrics and Finance” proposed by Scopus

Research Area	Category
Economics, Econometrics and Finance	Economics, Econometrics and Finance (miscellaneous)
	Economics and Econometrics
	Finance

TABLE B.11: Subject categorization for the research area “Energy” proposed by Scopus

Research Area	Category
Energy	Energy (miscellaneous)
	Energy Engineering and Power Technology
	Fuel Technology
	Nuclear Energy and Engineering
	Renewable Energy, Sustainability and the Environment

TABLE B.12: Subject categorization for the research area “Engineering” proposed by Scopus

Research Area	Category
Engineering	Engineering (miscellaneous)
	Aerospace Engineering
	Automotive Engineering
	Biomedical Engineering
	Civil and Structural Engineering
	Computational Mechanics
	Control and Systems Engineering
	Electrical and Electronic Engineering
	Industrial and Manufacturing Engineering
	Mechanical Engineering
	Mechanics of Materials
	Ocean Engineering
	Safety, Risk, Reliability and Quality
	Media Technology
	Building and Construction
Architecture	

TABLE B.13: Subject categorization for the research area “Environmental Science” proposed by Scopus

Research Area	Category
Environmental Science	Environmental Science (miscellaneous)
	Ecological Modeling
	Ecology
	Environmental Chemistry
	Environmental Engineering
	Global and Planetary Change
	Health, Toxicology and Mutagenesis
	Management, Monitoring, Policy and Law
	Nature and Landscape Conservation
	Pollution
	Waste Management and Disposal
	Water Science and Technology

TABLE B.14: Subject categorization for the research area “Immunology and Microbiology” proposed by Scopus

Research Area	Category
Immunology and Microbiology	Immunology and Microbiology (miscellaneous)
	Applied Microbiology and Biotechnology
	Immunology
	Microbiology
	Parasitology
	Virology

TABLE B.15: Subject categorization for the research area “Materials Science” proposed by Scopus

Research Area	Category
Materials Science	Materials Science (miscellaneous)
	Biomaterials
	Ceramics and Composites
	Electronic, Optical and Magnetic Materials
	Materials Chemistry
	Metals and Alloys
	Polymers and Plastics
	Surfaces, Coatings and Films
	Nanoscience and Nanotechnology

TABLE B.16: Subject categorization for the research area “Mathematics” proposed by Scopus

Research Area	Category
Mathematics	Mathematics (miscellaneous)
	Algebra and Number Theory
	Analysis
	Applied Mathematics
	Computational Mathematics
	Control and Optimization
	Discrete Mathematics and Combinatorics
	Geometry and Topology
	Logic
	Mathematical Physics
	Modeling and Simulation
	Numerical Analysis
	Statistics and Probability
	Theoretical Computer Science

TABLE B.17: Subject categorization for the research area “Medicine” proposed by Scopus

Research Area	Category
Medicine	Medicine (miscellaneous)
	Anatomy
	Anesthesiology and Pain Medicine
	Biochemistry (medical)
	Cardiology and Cardiovascular Medicine
	Critical Care and Intensive Care Medicine
	Complementary and Alternative Medicine
	Dermatology
	Drug Guides
	Embryology
	Emergency Medicine
	Endocrinology, Diabetes and Metabolism
	Epidemiology
	Family Practice
	Gastroenterology
	Genetics (clinical)
	Geriatrics and Gerontology
	Health Informatics
	Health Policy
	Hematology
	Hepatology
	Histology
	Immunology and Allergy
	Internal Medicine
	Infectious Diseases
	Microbiology (medical)
	Nephrology
	Neurology (clinical)
	Obstetrics and Gynecology
	Oncology
	Ophthalmology
	Orthopedics and Sports Medicine
	Otorhinolaryngology
	Pathology and Forensic Medicine
	Pediatrics, Perinatology and Child Health
	Pharmacology (medical)
	Physiology (medical)
	Psychiatry and Mental Health
	Public Health, Environmental and Occupational Health
	Pulmonary and Respiratory Medicine
	Radiology, Nuclear Medicine and Imaging
Rehabilitation	
Reproductive Medicine	
Reviews and References (medical)	
Rheumatology	
Surgery	
Transplantation	
Urology	

TABLE B.18: Subject categorization for the research area “Neuroscience” proposed by Scopus

Research Area	Category
Neuroscience	Neuroscience (miscellaneous)
	Behavioral Neuroscience
	Biological Psychiatry
	Cellular and Molecular Neuroscience
	Cognitive Neuroscience
	Developmental Neuroscience
	Endocrine and Autonomic Systems
	Neurology
	Sensory Systems

TABLE B.19: Subject categorization for the research area “Nursing” proposed by Scopus

Research Area	Category
Nursing	Nursing (miscellaneous)
	Advanced and Specialized Nursing
	Assessment and Diagnosis
	Care Planning
	Community and Home Care
	Critical Care Nursing
	Emergency Nursing
	Fundamentals and Skills
	Gerontology
	Issues, Ethics and Legal Aspects
	Leadership and Management
	LPN and LVN
	Maternity and Midwifery
	Medical and Surgical Nursing
	Nurse Assisting
	Nutrition and Dietetics
	Oncology (nursing)
	Pediatrics
	Pharmacology (nursing)
	Psychiatric Mental Health
Research and Theory	
Review and Exam Preparation	

TABLE B.20: Subject categorization for the research area “Pharmacology, Toxicology and Pharmaceutics” proposed by Scopus

Research Area	Category
Pharmacology, Toxicology and Pharmaceutics	Pharmacology, Toxicology and Pharmaceutics (miscellaneous)
	Drug Discovery
	Pharmaceutical Science
	Pharmacology
	Toxicology

TABLE B.21: Subject categorization for the research area “Physics and Astronomy” proposed by Scopus

Research Area	Category
Physics and Astronomy	Physics and Astronomy (miscellaneous)
	Acoustics and Ultrasonics
	Astronomy and Astrophysics
	Condensed Matter Physics
	Instrumentation
	Nuclear and High Energy Physics
	Atomic and Molecular Physics, and Optics
	Radiation
	Statistical and Nonlinear Physics
	Surfaces and Interfaces

TABLE B.22: Subject categorization for the research area “Psychology” proposed by Scopus

Research Area	Category
Psychology	Psychology (miscellaneous)
	Applied Psychology
	Clinical Psychology
	Developmental and Educational Psychology
	Experimental and Cognitive Psychology
	Neuropsychology and Physiological Psychology
	Social Psychology

TABLE B.23: Subject categorization for the research area “Social Sciences” proposed by Scopus

Research Area	Category
Social Sciences	Social Sciences (miscellaneous)
	Archeology
	Development
	Education
	Geography, Planning and Development
	Health (social science)
	Human Factors and Ergonomics
	Law
	Library and Information Sciences
	Linguistics and Language
	Safety Research
	Sociology and Political Science
	Transportation
	Anthropology
	Communication
	Cultural Studies
	Demography
	Gender Studies
	Life-span and Life-course Studies
	Political Science and International Relations
Public Administration	
Urban Studies	
Social Work	

TABLE B.24: Subject categorization for the research area “Veterinary” proposed by Scopus

Research Area	Category
Veterinary	Veterinary (miscellaneous)
	Equine
	Food Animals
	Small Animals

TABLE B.25: Subject categorization for the research area “Dentistry” proposed by Scopus

Research Area	Category
Dentistry	Dentistry (miscellaneous)
	Dental Assisting
	Dental Hygiene
	Oral Surgery
	Orthodontics
	Periodontics

TABLE B.26: Subject categorization for the research area “Health Professions” proposed by Scopus

Research Area	Category
Health Professions	Respiratory Care
	Sports Science
	Health Professions (miscellaneous)
	Chiropractics
	Complementary and Manual Therapy
	Emergency Medical Services
	Health Information Management
	Medical Assisting and Transcription
	Medical Laboratory Technology
	Medical Terminology
	Occupational Therapy
	Optometry
	Pharmacy
	Physical Therapy, Sports Therapy and Rehabilitation
	Podiatry
	Radiological and Ultrasound Technology
Speech and Hearing	

B.2 Web of Science

The Web of Science's categories are presented according to the research areas below.

TABLE B.27: Subject categorization for the research area "Arts & Humanities" proposed by Web of Science

Research Area	Category
Arts & Humanities	Architecture
	Art
	Arts & Humanities Other Topics
	Asian Studies
	Classics
	Dance
	Film, Radio & Television
	History
	History & Philosophy of Science
	Literature
	Music
	Philosophy
	Religion
	Theater

TABLE B.28: Subject categorization for the research area “Life Sciences & Biomedicine” proposed by Web of Science (Part I)

Research Area	Category
Life Sciences & Biomedicine	Agriculture
	Allergy
	Anatomy & Morphology
	Anesthesiology
	Anthropology
	Behavioral Sciences
	Biochemistry & Molecular Biology
	Biodiversity & Conservation
	Biophysics
	Biotechnology & Applied Microbiology
	Cardiovascular System & Cardiology
	Cell Biology
	Critical Care Medicine
	Dentistry, Oral Surgery & Medicine
	Dermatology
	Developmental Biology
	Emergency Medicine
	Endocrinology & Metabolism
	Entomology
	Environmental Sciences & Ecology
	Evolutionary Biology
	Fisheries
	Food Science & Technology
	Forestry
	Gastroenterology & Hepatology
	General & Internal Medicine
	Genetics & Heredity
	Geriatrics & Gerontology
	Health Care Sciences & Services
	Hematology
	Immunology
	Infectious Diseases
	Integrative & Complementary Medicine
	Legal Medicine
	Life Sciences Biomedicine Other Topics
	Marine & Freshwater Biology
	Mathematical & Computational Biology
	Medical Ethics
	Medical Informatics
	Medical Laboratory Technology
	Microbiology
Mycology	
Neurosciences & Neurology	
Nursing	
Nutrition & Dietetics	

TABLE B.29: Subject categorization for the research area “Life Sciences & Biomedicine” proposed by Web of Science (Part II)

Research Area	Category
Life Sciences & Biomedicine	Obstetrics & Gynecology
	Oncology
	Ophthalmology
	Orthopedics
	Otorhinolaryngology
	Paleontology
	Parasitology
	Pathology
	Pediatrics
	Pharmacology & Pharmacy
	Physiology
	Plant Sciences
	Psychiatry
	Public, Environmental & Occupational Health
	Radiology, Nuclear Medicine & Medical Imaging
	Rehabilitation
	Reproductive Biology
	Research & Experimental Medicine
	Respiratory System
	Rheumatology
	Sport Sciences
	Substance Abuse
	Surgery
	Toxicology
	Transplantation
	Tropical Medicine
	Urology & Nephrology
	Veterinary Sciences
	Virology
	Zoology

TABLE B.30: Subject categorization for the research area “Physical Sciences” proposed by Web of Science

Research Area	Category
Physical Sciences	Astronomy & Astrophysics
	Chemistry
	Crystallography
	Electrochemistry
	Geochemistry & Geophysics
	Geology
	Mathematics
	Meteorology & Atmospheric Sciences
	Mineralogy
	Mining & Mineral Processing
	Oceanography
	Optics
	Physical Geography
	Physics
	Polymer Science
	Thermodynamics
Water Resources	

TABLE B.31: Subject categorization for the research area “Social Sciences” proposed by Web of Science

Research Area	Category
Social Sciences	Archaeology
	Area Studies
	Biomedical Social Sciences
	Business & Economics
	Communication
	Criminology & Penology
	Cultural Studies
	Demography
	Education & Educational Research
	Ethnic Studies
	Family Studies
	Geography
	Government & Law
	International Relations
	Linguistics
	Mathematical Methods In Social Sciences
	Psychology
	Public Administration
	Social Issues
	Social Sciences Other Topics
	Social Work
	Sociology
Urban Studies	
Women’s Studies	

TABLE B.32: Subject categorization for the research area “Technology” proposed by Web of Science

Research Area	Category
Technology	Acoustics
	Automation & Control Systems
	Computer Science
	Construction & Building Technology
	Energy & Fuels
	Engineering
	Imaging Science & Photographic Technology
	Information Science & Library Science
	Instruments & Instrumentation
	Materials Science
	Mechanics
	Metallurgy & Metallurgical Engineering
	Microscopy
	Nuclear Science & Technology
	Operations Research & Management Science
	Remote Sensing
	Robotics
	Science & Technology Other Topics
	Spectroscopy
	Telecommunications
Transportation	

B.3 The Unified Publications' Areas and Categories Dataset

The unified categories are presented according to the research areas below.

TABLE B.33: Subject categorization for the research area "Agricultural and Biological Sciences" proposed for C-HyRA

Category1	Category2	Category3	Category4
Agricultural and Biological Sciences (miscellaneous)	Life Sciences and Biomedicine	Agriculture	
Agronomy and Crop Science			
Animal Science and Zoology	Life Sciences and Biomedicine	Zoology	
Aquatic Science	Life Sciences and Biomedicine	Fisheries	Marine and Freshwater Biology
Ecology, Evolution, Behavior and Systematics	Life Sciences and Biomedicine	Mycology	
Food Science	Life Sciences and Biomedicine	Food Science and Technology	
Forestry	Life Sciences and Biomedicine		
Horticulture			
Insect Science	Life Sciences and Biomedicine	Entomology	
Plant Science	Life Sciences and Biomedicine	Plant Sciences	Mycology
Soil Science	Technology	Remote Sensing	

TABLE B.34: Subject categorization for the research area “Arts and Humanities” proposed for C-HyRA

Category1	Category2	Category3	Category4
Arts and Humanities (miscellaneous)	Arts and Humanities Other Topics	Life Sciences and Biomedicine	Psychiatry
History			
Language and Linguistics			
Archeology (arts and humanities)			
Classics			
Conservation			
History and Philosophy of Science			
Literature and Literary Theory	Literature		
Museology			
Music			
Philosophy			
Religious Studies	Religion		
Visual Arts and Performing Arts	Art		
Theater			
Asian Studies			
Dance			
Film, Radio and Television			
Architecture			

TABLE B.35: Subject categorization for the research area “Biochemistry, Genetics and Molecular Biology” proposed for C-HyRA

Category1	Category2	Category3	Category4	Category5	Category6	Category7	Category8
Biochemistry, Genetics and Molecular Biology (miscellaneous)	Life Sciences and Biomedicine	Biochemistry and Molecular Biology	Life Sciences Biomedicine Other Topics	Evolutionary Biology	Mathematical and Computational Biology	Physical Sciences	Crystallography
Aging							
Biochemistry	Life Sciences and Biomedicine	Biochemistry and Molecular Biology					
Biophysics	Life Sciences and Biomedicine						
Biotechnology	Technology	Science and Technology Other Topics					
Cancer Research	Life Sciences and Biomedicine	Oncology					
Cell Biology	Life Sciences and Biomedicine						
Clinical Biochemistry							
Developmental Biology	Life Sciences and Biomedicine	Evolutionary Biology					
Endocrinology	Life Sciences and Biomedicine	Reproductive Biology					
Genetics	Life Sciences and Biomedicine	Genetics and Heredity					
Molecular Biology	Life Sciences and Biomedicine	Biochemistry and Molecular Biology					
Molecular Medicine							
Physiology	Life Sciences and Biomedicine						
Structural Biology							

TABLE B.36: Subject categorization for the research area “Business, Management and Accounting” proposed for C-HyRA

Category1	Category2	Category3
Business, Management and Accounting (miscellaneous)	Social Sciences	Business and Economics
Accounting		
Business and International Management		
Management Information Systems		
Management of Technology and Innovation		
Marketing		
Organizational Behavior and Human Resource Management		
Strategy and Management		
Tourism, Leisure and Hospitality Management		
Industrial Relations		

TABLE B.37: Subject categorization for the research area “Chemical Engineering” proposed for C-HyRA

Category1
Chemical Engineering (miscellaneous)
Bioengineering
Catalysis
Chemical Health and Safety
Colloid and Surface Chemistry
Filtration and Separation
Fluid Flow and Transfer Process
Process Chemistry and Technology

TABLE B.38: Subject categorization for the research area “Chemistry” proposed for C-HyRA

Category1	Category2	Category3	Category4	Category5	Category6
Chemistry (miscellaneous)	Physical Sciences	Crystallography	Mining and Mineral Processing	Technology	Imaging Science and Photographic Technology
Analytical Chemistry					
Electrochemistry	Physical Sciences				
Inorganic Chemistry					
Organic Chemistry					
Physical and Theoretical Chemistry	Physical Sciences	Thermodynamics			
Spectroscopy	Technology				

TABLE B.39: Subject categorization for the research area “Computer Science” proposed for C-HyRA

Category1	Category2	Category3	Category4	Category5	Category6
Computer Science (miscellaneous)	Technology	Life Sciences and Biomedicine	Mathematical and Computational Biology	Telecommunications	
Artificial Intelligence	Technology	Robotics			
Computational Theory and Mathematics					
Computer Graphics and Computer-Aided Design					
Computer Networks and Communications	Technology	Science and Technology Other Topics	Telecommunications		
Computer Science Applications	Technology	Automation and Control Systems	Imaging Science and Photographic Technology	Robotics	Transportation
Computer Vision and Pattern Recognition	Technology	Imaging Science and Photographic Technology			
Hardware and Architecture	Technology	Robotics			
Human-Computer Interaction	Technology	Robotics			
Information Systems	Technology	Science and Technology Other Topics			
Signal Processing					
Software	Technology	Robotics			

TABLE B.40: Subject categorization for the research area “Decision Sciences” proposed for C-HyRA

Category1	Category2	Category3	Category4
Decision Science (Miscellaneous)			
Information Systems and Management	Technology	Science and Technology Other Topics	
Management Science and Operation Research	Technology	Operations Research and Management Science	Transportation
Statistics, Probability and Uncertainty			

TABLE B.41: Subject categorization for the research area “Earth and Planetary Sciences” proposed for C-HyRA

Category1	Category2	Category3	Category4	Category5
Earth and Planetary Sciences (miscellaneous)	Technology	Remote Sensing	Physical Sciences	Physical Geography
Atmospheric Sciences	Physical Sciences	Meteorology and Atmospheric Sciences		
Computers in Earth Sciences	Technology	Remote Sensing		
Earth Surface Processes				
Economic Geology	Physical Sciences	Geology		
Geochemistry and Petrology	Physical Sciences	Geochemistry and Geophysics	Mineralogy	Mining and Mineral Processing
Geology	Physical Sciences	Technology	Remote Sensing	Physical Geography
Geophysics	Physical Sciences	Geochemistry and Geophysics	Mineralogy	
Geotechnical Engineering and Engineering Geology	Physical Sciences	Geology	Mining and Mineral Processing	
Oceanography	Physical Sciences			
Paleontology	Life Sciences and Biomedicine			
Space and Planetary Science				
Stratigraphy	Physical Sciences	Geology		

TABLE B.42: Subject categorization for the research area “Economics, Econometrics and Finance” proposed for C-HyRA

Category1	Category2	Category3
Economics, Econometrics and Finance (miscellaneous)	Social Sciences	Business and Economics
Economics and Econometrics	Social Sciences	Business and Economics
Finance		

TABLE B.43: Subject categorization for the research area “Energy” proposed for C-HyRA

Category1	Category2	Category3
Energy (miscellaneous)	Technology	Energy and Fuels
Energy Engineering and Power Technology		
Fuel Technology		
Nuclear Energy and Engineering	Technology	Nuclear Science and Technology
Renewable Energy, Sustainability and the Environment		

TABLE B.44: Subject categorization for the research area “Engineering” proposed for C-HyRA

Category1	Category2	Category3	Category4	Category5	Category6
Engineering (miscellaneous)	Technology	Science and Technology Other Topics	Telecommunications	Metallurgy and Metallurgical Engineering	
Aerospace Engineering					
Automotive Engineering	Technology	Transportation			
Biomedical Engineering					
Civil and Structural Engineering					
Computational Mechanics					
Control and Systems Engineering	Technology	Automation and Control Systems	Robotics		
Electrical and Electronic Engineering	Technology	Automation and Control Systems	Remote Sensing	Robotics	Telecommunications
Industrial and Manufacturing Engineering					
Mechanical Engineering	Technology	Robotics	Mechanics	Transportation	
Mechanics of Materials	Technology	Mechanics	Physical Sciences	Thermodynamics	Metallurgy and Metallurgical Engineering
Ocean Engineering					
Safety, Risk, Reliability and Quality					
Media Technology	Technology	Imaging Science and Photographic Technology			
Building and Construction	Technology	Construction and Building Technology			
Architecture					

TABLE B.45: Subject categorization for the research area “Environmental Science” proposed for C-HyRA

Category1	Category2	Category3	Category4	Category5
Environmental Science (miscellaneous)				
Ecological Modeling				
Ecology	Life Sciences and Biomedicine	Environmental Sciences and Ecology		
Environmental Chemistry				
Environmental Engineering	Technology	Science and Technology Other Topics		
Global and Planetary Change				
Health, Toxicology and Mutagenesis	Life Sciences and Biomedicine	Toxicology		
Management, Monitoring, Policy and Law				
Nature and Landscape Conservation	Life Sciences and Biomedicine	Biodiversity and Conservation		
Pollution				
Waste Management and Disposal				
Water Science and Technology	Physical Sciences	Water Resources	Technology	Science and Technology Other Topics

TABLE B.46: Subject categorization for the research area “Immunology and Microbiology” proposed for C-HyRA

Category1	Category2	Category3
Immunology and Microbiology (miscellaneous)		
Applied Microbiology and Biotechnology	Life Sciences and Biomedicine	Biotechnology and Applied Microbiology
Immunology	Life Sciences and Biomedicine	
Microbiology	Life Sciences and Biomedicine	
Parasitology	Life Sciences and Biomedicine	
Virology	Life Sciences and Biomedicine	

TABLE B.47: Subject categorization for the research area “Materials Science” proposed for C-HyRA

Category1	Category2	Category3	Category4	Category5
Materials Science (miscellaneous)	Technology	Physical Sciences	Crystallography	Thermodynamics
Biomaterials				
Ceramics and Composites				
Electronic, Optical and Magnetic Materials	Technology	Imaging Science and Photographic Technology	Telecommunications	
Materials Chemistry	Physical Sciences	Mining and Mineral Processing		
Metals and Alloys	Technology	Metallurgy and Metallurgical Engineering	Physical Sciences	Mining and Mineral Processing
Polymers and Plastics	Physical Sciences	Polymer Science		
Surfaces, Coatings and Films				
Nanoscience and Nanotechnology				

TABLE B.48: Subject categorization for the research area “Mathematics” proposed for C-HyRA

Category1	Category2	Category3	Category4
Mathematics (miscellaneous)	Physical Sciences	Life Sciences and Biomedicine	Mathematical and Computational Biology
Algebra and Number Theory			
Analysis			
Applied Mathematics	Technology	Robotics	
Computational Mathematics			
Control and Optimization			
Discrete Mathematics and Combinatorics			
Geometry and Topology			
Logic			
Mathematical Physics			
Modeling and Simulation	Technology	Robotics	
Numerical Analysis			
Statistics and Probability			
Theoretical Computer Science			

TABLE B.49: Subject categorization for the research area “Medicine” proposed for C-HyRA (Part I)

Category1	Category2	Category3	Category4	Category5	Category6	Category7	Category8
Medicine (miscellaneous)	Life Sciences and Biomedicine	Life Sciences Biomedicine Other Topics	Research and Experimental Medicine	Tropical Medicine	Technology	Microscopy	Respiratory System
Anatomy	Life Sciences and Biomedicine	Anatomy and Morphology	Technology	Microscopy			
Anesthesiology and Pain Medicine	Life Sciences and Biomedicine	Anesthesiology					
Biochemistry (medical)							
Cardiology and Cardiovascular Medicine	Life Sciences and Biomedicine	Cardiovascular System and Cardiology					
Critical Care and Intensive Care Medicine	Life Sciences and Biomedicine	Critical Care Medicine	Respiratory System				
Complementary and Alternative Medicine	Life Sciences and Biomedicine	Integrative and Complementary Medicine					
Dermatology	Life Sciences and Biomedicine						
Drug Guides							
Embryology							
Emergency Medicine	Life Sciences and Biomedicine						
Endocrinology, Diabetes and Metabolism	Life Sciences and Biomedicine	Endocrinology and Metabolism					
Epidemiology							
Family Practice							
Gastroenterology	Life Sciences and Biomedicine	Gastroenterology and Hepatology					

TABLE B.50: Subject categorization for the research area “Medicine” proposed for C-HyRA (Part II)

Category1	Category2	Category3	Category4	Category5	Category6	Category7	Category8
Genetics (clinical)							
Geriatrics and Gerontology	Life Sciences and Biomedicine						
Health Informatics	Life Sciences and Biomedicine	Medical Informatics					
Health Policy							
Hematology	Life Sciences and Biomedicine						
Hepatology							
Histology	Technology	Microscopy					
Immunology and Allergy	Life Sciences and Biomedicine	Allergy					
Internal Medicine	Life Sciences and Biomedicine	General and Internal Medicine					
Infectious Diseases	Life Sciences and Biomedicine						
Microbiology (medical)							
Nephrology	Life Sciences and Biomedicine	Urology and Nephrology					
Neurology (clinical)							
Obstetrics and Gynecology	Life Sciences and Biomedicine						
Oncology	Life Sciences and Biomedicine						
Ophthalmology	Life Sciences and Biomedicine						
Orthopedics and Sports Medicine	Life Sciences and Biomedicine	Orthopedics					
Otorhinolaryngology	Life Sciences and Biomedicine						

TABLE B.51: Subject categorization for the research area “Medicine” proposed for C-HyRA (Part III)

Category1	Category2	Category3	Category4	Category5	Category6	Category7	Category8
Pathology and Forensic Medicine	Life Sciences and Biomedicine	Pathology	Legal Medicine	Technology	Microscopy		
Pediatrics, Perinatology and Child Health	Life Sciences and Biomedicine	Pediatrics					
Pharmacology (medical)	Life Sciences and Biomedicine	Pharmacology and Pharmacy					
Physiology (medical)	Life Sciences and Biomedicine	Physiology					
Psychiatry and Mental Health	Life Sciences and Biomedicine	Psychiatry	Substance Abuse				
Public Health, Environmental and Occupational Health	Life Sciences and Biomedicine	Public, Environmental and Occupational Health	Substance Abuse				
Pulmonary and Respiratory Medicine	Life Sciences and Biomedicine	Respiratory System					
Radiology, Nuclear Medicine and Imaging	Life Sciences and Biomedicine	Radiology, Nuclear Medicine and Medical Imaging					
Rehabilitation	Life Sciences and Biomedicine						
Reproductive Medicine							
Reviews and References (medical)							
Rheumatology	Life Sciences and Biomedicine						
Surgery	Life Sciences and Biomedicine	Technology	Robotics				
Transplantation	Life Sciences and Biomedicine						
Urology	Life Sciences and Biomedicine	Urology and Nephrology					

TABLE B.52: Subject categorization for the research area “Neuroscience” proposed for C-HyRA

Category1	Category2	Category3	Category4
Neuroscience (miscellaneous)	Life Sciences and Biomedicine	Behavioral Sciences	Substance Abuse
Behavioral Neuroscience	Life Sciences and Biomedicine	Behavioral Sciences	
Biological Psychiatry	Life Sciences and Biomedicine	Psychiatry	
Cellular and Molecular Neuroscience			
Cognitive Neuroscience			
Developmental Neuroscience			
Endocrine and Autonomic Systems			
Neurology	Life Sciences and Biomedicine	Neurosciences and Neurology	
Sensory Systems			

TABLE B.53: Subject categorization for the research area “Nursing” proposed for C-HyRA

Category1	Category2	Category3
Nursing (miscellaneous)	Life Sciences and Biomedicine	
Advanced and Specialized Nursing		
Assessment and Diagnosis		
Care Planning		
Community and Home Care		
Critical Care Nursing		
Emergency Nursing		
Fundamentals and Skills		
Gerontology		
Issues, Ethics and Legal Aspects	Life Sciences and Biomedicine	Medical Ethics
Leadership and Management		
LPN and LVN		
Maternity and Midwifery		
Medical and Surgical Nursing		
Nurse Assisting		
Nutrition and Dietetics	Life Sciences and Biomedicine	
Oncology (nursing)	Life Sciences and Biomedicine	Oncology
Pediatrics	Life Sciences and Biomedicine	
Pharmacology (nursing)	Life Sciences and Biomedicine	Pharmacology and Pharmacy
Psychiatric Mental Health		
Research and Theory		
Review and Exam Preparation		

TABLE B.54: Subject categorization for the research area “Pharmacology, Toxicology and Pharmaceutics” proposed for C-HyRA

Category1	Category2	Category3
Pharmacology, Toxicology and Pharmaceutics (miscellaneous)		
Drug Discovery		
Pharmaceutical Science	Life Sciences and Biomedicine	Pharmacology and Pharmacy
Pharmacology	Life Sciences and Biomedicine	Pharmacology and Pharmacy
Toxicology	Life Sciences and Biomedicine	

TABLE B.55: Subject categorization for the research area “Physics and Astronomy” proposed for C-HyRA

Category1	Category2	Category3	Category4	Category5	Category6
Physics and Astronomy (miscellaneous)	Physical Sciences	Physics	Thermodynamics		
Acoustics and Ultrasonics	Technology	Acoustics			
Astronomy and Astrophysics	Physical Sciences				
Condensed Matter Physics	Physical Sciences	Crystallography	Technology	Metallurgy and Metallurgical Engineering	
Instrumentation	Technology	Instruments and Instrumentation	Microscopy	Remote Sensing	
Nuclear and High Energy Physics	Technology	Nuclear Science and Technology			
Atomic and Molecular Physics, and Optics	Physical Sciences	Optics	Thermodynamics	Technology	Imaging Science and Photographic Technology
Radiation					
Statistical and Nonlinear Physics					
Surfaces and Interfaces					

TABLE B.56: Subject categorization for the research area “Psychology” proposed for C-HyRA

Category1	Category2	Category3	Category4
Psychology (miscellaneous)	Social Sciences	Life Sciences and Biomedicine	Behavioral Sciences
Applied Psychology			
Clinical Psychology			
Developmental and Educational Psychology			
Experimental and Cognitive Psychology			
Neuropsychology and Physiological Psychology			
Social Psychology			

TABLE B.57: Subject categorization for the research area “Social Sciences” proposed for C-HyRA

Category1	Category2	Category3	Category4
Social Sciences (miscellaneous)	Social Sciences Other Topics	Life Sciences and Biomedicine	Behavioral Sciences
Archeology	Archaeology		
Development			
Education	Education and Educational Research		
Geography, Planning and Development	Geography	Physical Sciences	Physical Geography
Health (social science)			
Human Factors and Ergonomics			
Law	Government and Law		
Library and Information Sciences	Technology	Information Science and Library Science	Science and Technology Other Topics
Linguistics and Language	Linguistics		
Safety Research			
Sociology and Political Science	Sociology		
Transportation	Technology		
Anthropology	Life Sciences and Biomedicine	Behavioral Sciences	
Communication			
Cultural Studies			
Demography			
Gender Studies			
Life-span and Life-course Studies			
Political Science and International Relations	International Relations		
Public Administration			
Urban Studies			
Social Work			
Biomedical Social Sciences			
Criminology and Penology			
Ethnic Studies			
Family Studies			
Mathematical Methods In Social Sciences			
Social Issues			
Area Studies			
Women’s Studies			

TABLE B.58: Subject categorization for the research area “Veterinary” proposed for C-HyRA

Category1	Category2	Category3
Veterinary (miscellaneous)	Life Sciences and Biomedicine	Veterinary Sciences
Equine		
Food Animals		
Small Animals		

TABLE B.59: Subject categorization for the research area “Dentistry” proposed for C-HyRA

Category1	Category2	Category3
Dentistry (miscellaneous)	Life Sciences and Biomedicine	Dentistry, Oral Surgery and Medicine
Dental Assisting		
Dental Hygiene		
Oral Surgery		
Orthodontics		
Periodontics		

TABLE B.60: Subject categorization for the research area “Health Professions” proposed for C-HyRA

Category1	Category2	Category3	Category4
Respiratory Care	Life Sciences and Biomedicine	Respiratory System	
Sports Science	Life Sciences and Biomedicine	Sport Sciences	Rehabilitation
Health Professions (miscellaneous)	Life Sciences and Biomedicine	Health Care Sciences and Services	
Chiropractics			
Complementary and Manual Therapy			
Emergency Medical Services			
Health Information Management			
Medical Assisting and Transcription			
Medical Laboratory Technology	Life Sciences and Biomedicine	Technology	Microscopy
Medical Terminology			
Occupational Therapy			
Optometry			
Pharmacy			
Physical Therapy, Sports Therapy and Rehabilitation	Life Sciences and Biomedicine	Rehabilitation	
Podiatry			
Radiological and Ultrasound Technology			
Speech and Hearing			

Bibliography

- [1] J. Alvarado-Uribe, A. Becerril García, M. Gonzalez-Mendoza, R. Lozano Espinosa, J. M. Molina Espinosa, Semantic approach for discovery and visualization of academic information structured with oai-pmh, *Acta Polytechnica Hungarica* 14 (3) (2017) 129–148. [doi:10.12700/APH.14.3.2017.3.8](https://doi.org/10.12700/APH.14.3.2017.3.8).
- [2] J. Alvarado-Uribe, A. Gómez-Oliva, A. Y. Barrera-Animas, G. Molina, M. Gonzalez-Mendoza, M. C. Parra-Meroño, A. J. Jara, Hyra: A hybrid recommendation algorithm focused on smart poi. ceutí as a study scenario, *Sensors* 18 (3) (2018) 890.
- [3] J. Alvarado-Uribe, M. González-Mendoza, N. Hernández-Gress, C. E. Escobar-Ruiz, M. U. Hernández-Camacho, Una herramienta visual para la búsqueda semántica rdf, *Research in Computing Science* 95 (2015) 9–22.
- [4] J. Alvarado-Uribe, A. Gómez-Oliva, G. Molina, M. Gonzalez-Mendoza, M. C. Parra-Meroño, A. J. Jara, Towards the Development of a Smart Tourism Application Based on Smart POI and Recommendation Algorithms: Ceutí as a Study Case, Vol. 612, Springer, Cham, 2018, pp. 904–916.
- [5] K. Schlegel, J. Hare, Hype cycle for analytics and business intelligence, 2017, Tech. rep., Gartner (July 2017).
- [6] V. L. Sauter, *Decision Support Systems for Business Intelligence*, 2nd Edition, John Wiley & Sons, 2010.
- [7] G. Phifer, Hype cycle for web computing, 2014, Tech. rep., Gartner (July 2014).
- [8] G. Phifer, Hype cycle for web computing, 2015, Tech. rep., Gartner (July 2015).
- [9] M. Revang, Hype cycle for web computing, 2016, Tech. rep., Gartner (August 2016).

- [10] M. Chen, S. Mao, Y. Liu, Big data: A survey, *Mobile Networks and Applications* 19 (02) (2014) 171–209.
- [11] B. Khaleghi, A. Khamis, F. O. Karray, S. N. Razavi, Multisensor data fusion: A review of the state-of-the-art, *Information Fusion* 14 (1) (2013) 28–44.
- [12] C. Thomas, N. Balakrishnan, Modified evidence theory for performance enhancement of intrusion detection systems, in: *Information Fusion, 2008 11th International Conference on*, IEEE, Cologne, 2008, pp. 1–8.
- [13] D. L. Hall, J. Llinas, An introduction to multisensor data fusion, *Proceedings of the IEEE* 85 (1) (1997) 6–23.
- [14] A. Margara, J. Urbani, F. van Harmelen, H. Bal, Streaming the web: Reasoning over dynamic data, *Web Semantics: Science, Services and Agents on the World Wide Web* 25 (2014) 24–44.
- [15] D. J. Hand, [Principles of data mining](#), *Drug Safety* 30 (7) (2007) 621–622. doi:10.2165/00002018-200730070-00010.
URL <https://doi.org/10.2165/00002018-200730070-00010>
- [16] G. Wienold, Some basic aspects of text processing, *Poetics Today* 2 (4) (1981) 97–109.
- [17] E. Turban, R. Sharda, D. Delen, *Decision Support and Business Intelligence Systems*, 9th Edition, Pearson Education, Inc., 2011.
- [18] N. Shadbolt, W. Hall, T. Berners-Lee, The semantic web revisited, *IEEE Intelligent Systems* 21 (3) (2006) 96–101.
- [19] P. F. Patel-Schneider, I. Horrocks, Position paper: A comparison of two modelling paradigms in the semantic web, in: *Proceedings of the 15th international conference on World Wide Web*, ACM, 2006, pp. 3–12.
- [20] W. W. W. Consortium, [Inference](#), Official website (2015).
URL <http://www.w3.org/standards/semanticweb/inference#specifications>
- [21] I. Subirats-Coll, Seven things you should know about linked data, *COAR Repository Observatory* (2).

- [22] Springer, [Online tools & social media for authors and editors](http://www.springer.com/gp/authors-editors/book-authors-editors/book-authors-helpdesk/online-tools-social-media-for-authors/3340), Official Website (2015).
URL <http://www.springer.com/gp/authors-editors/book-authors-editors/book-authors-helpdesk/online-tools-social-media-for-authors/3340>
- [23] AMIPCI, PULPO, ELOGIA, [11° estudio sobre los hábitos de los usuarios de internet en México 2015](https://amipci.org.mx/images/AMIPCI_HABITOS_DEL_INTERNAUTA_MEXICANO_2015.pdf), Official website of AMIPCI (2015).
URL https://amipci.org.mx/images/AMIPCI_HABITOS_DEL_INTERNAUTA_MEXICANO_2015.pdf
- [24] M. d. L. Martínez-Villaseñor, M. González-Mendoza, N. Hernández-Gress, Towards a ubiquitous user model for profile sharing and reuse, *Sensors* 12 (10) (2012) 13249–13283.
- [25] Y. Yu, X. Chen, A survey of point-of-interest recommendation in location-based social networks, in: *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, 2015, pp. 53–60.
- [26] M. Rowe, Interlinking distributed social graphs, in: *Linked Data on the Web (LDOW2009)*, Madrid, Spain, 2009.
- [27] S. Brin, L. Page, Reprint of: The anatomy of a large-scale hypertextual web search engine, *Computer networks* 56 (18) (2012) 3825–3833.
- [28] A. Gelbukh, G. Sidorov, *Procesamiento automático del español con enfoque en recursos léxicos grandes*, Instituto Politécnico Nacional, 2010.
- [29] F. E. White, Data fusion lexicon, Tech. rep., Data Fusion Panel of the Joint Directors of Laboratories, Technical Panel for C3, San Diego, CA (October 1991).
- [30] D. F. Barbieri, D. Braga, S. Ceri, E. D. Valle, M. Grossniklaus, C-sparql: Sparql for continuous querying, in: *Proceedings of the 18th international conference on World wide web*, ACM, Madrid, Spain, 2009, pp. 1061–1062.
- [31] D. Anicic, P. Fodor, S. Rudolph, N. Stojanovic, Ep-sparql: A unified language for event processing and stream reasoning, in: *Proceedings of the 20th international conference on World wide web*, WWW '11, ACM, Hyderabad, India, 2011, pp. 635–644.
- [32] M. Gebser, T. Grote, R. Kaminski, P. Obermeier, O. Sabuncu, T. Schaub, Stream reasoning with answer set programming: Preliminary report, *KR* 12 (2012) 613–617.

- [33] M. Gebser, T. Grote, R. Kaminski, P. Obermeier, O. Sabuncu, T. Schaub, [Stream reasoning with answer set programming: Extended version](#), <http://www.cs.uni-potsdam.de/wv/pdfformat/gegrkaobsasc12a.pdf> (accessed on 15 November 2016), unpublished draft. Available at (oclingo) (2012).
URL <http://www.cs.uni-potsdam.de/wv/pdfformat/gegrkaobsasc12a.pdf>
- [34] G. Sidorov, Construcción no lineal de n-gramas en la lingüística computacional, Sociedad Mexicana de Inteligencia Artificial, 2013.
- [35] V. Lopez, M. Pasin, E. Motta, [The Semantic Web: Research and Applications: Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29–June 1, 2005. Proceedings](#), Vol. 3532 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, Ch. AquaLog: An Ontology-Portable Question Answering System for the Semantic Web, pp. 546–562. doi:10.1007/11431053_37.
URL http://dx.doi.org/10.1007/11431053_37
- [36] A. A. A. Youssif, G. A. Z., A. E. A., Hsws: Enhancing efficiency of web search engine via semantic web, in: Proceedings of the International Conference on Management of Emergent Digital EcoSystems, ACM, 2011, pp. 212–219.
- [37] A. Mehta, P. Makkar, S. Palande, S. B. Wankhede, Semantic web search engine, International Journal of Engineering Research and Technology 4 (04) (2015) 687–691.
- [38] M. Amoretti, L. Belli, F. Zanichelli, Utravel: Smart mobility with a novel user profiling and recommendation approach, Pervasive and Mobile Computing 38 (2) (2017) 474–489.
- [39] B. Xie, X. Tang, F. Tang, Hybrid recommendation base on learning to rank, in: Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2015 9th International Conference on, IEEE, 2015, pp. 53 – 57.
- [40] L. Guo, H. Jiang, X. Wang, F. Liu, Learning to recommend point-of-interest with the weighted bayesian personalized ranking method in lbsns, Information 8 (1) (2017) 20.
- [41] Q. Yuan, G. Cong, A. Sun, Graph-based point-of-interest recommendation with geographical and temporal influences, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, Shanghai, China, 2014, pp. 659–668.

- [42] B. Liu, Y. Fu, Z. Yao, H. Xiong, Learning geographical preferences for point-of-interest recommendation, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Chicago, Illinois, USA, 2013, pp. 1043 – 1051.
- [43] E.-y. Kang, H. Kim, J. Cho, [Personalization Method for Tourist Point of Interest \(POI\) Recommendation](#), Vol. 4251, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, Ch. Knowledge-Based Intelligent Information and Engineering Systems: 10th International Conference, KES 2006, Bournemouth, UK, October 9-11, 2006. Proceedings, Part I, pp. 392–400. doi:10.1007/11892960_48.
URL https://doi.org/10.1007/11892960_48
- [44] M. Ye, P. Yin, W.-C. Lee, D.-L. Lee, Exploiting geographical influence for collaborative point-of-interest recommendation, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, Beijing, China, 2011, pp. 325 – 334.
- [45] J. J.-C. Ying, E. H.-C. Lu, W.-N. Kuo, V. S. Tseng, [Urban point-of-interest recommendation by mining user check-in behaviors](#), in: Proceedings of the ACM SIGKDD International Workshop on Urban Computing, UrbComp '12, ACM, Beijing, China, 2012, pp. 63–70. doi:10.1145/2346496.2346507.
URL <http://doi.acm.org/10.1145/2346496.2346507>
- [46] N. Zheng, X. Jin, L. Li, Cross-region collaborative filtering for new point-of-interest recommendation, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, Rio de Janeiro, Brazil, 2013, pp. 45 – 46.
- [47] X. Liu, Y. Liu, K. Aberer, C. Miao, [Personalized point-of-interest recommendation by mining users' preference transition](#), in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM '13, ACM, San Francisco, CA, USA, 2013, pp. 733–738. doi:10.1145/2505515.2505639.
URL <http://doi.acm.org/10.1145/2505515.2505639>
- [48] K. Meehan, T. Lunney, K. Curran, A. McCaughey, Context-aware intelligent recommendation system for tourism, in: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on, IEEE, San Diego, 2013, pp. 328–331.

- [49] B. Liu, H. Xiong, S. Papadimitriou, Y. Fu, Z. Yao, A general geographical probabilistic factor model for point of interest recommendation, *IEEE Transactions on Knowledge and Data Engineering* 27 (5) (2015) 1167–1179.
- [50] W. Zhang, J. Wang, Location and time aware social collaborative retrieval for new successive point-of-interest recommendation, in: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ACM, Melbourne, Australia, 2015, pp. 1221 – 1230.
- [51] Z. Yu, H. Xu, Z. Yang, B. Guo, Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints, *IEEE Transactions on Human-Machine Systems* 46 (1) (2016) 151 – 158, iEEE.
- [52] M. A. Alles, *Selección por competencias*, Ediciones Granica S.A., 2006.
- [53] M. Golemati, A. Katifori, C. Vassilakis, G. Lepouras, C. Halatsis, Creating an ontology for the user profile: Method and applications, in: *Proceedings of the first RCIS conference*, 2007, pp. 407–412.
- [54] M. Sutterer, O. Droegehorn, K. David, Upos: User profile ontology with situation-dependent preferences support, in: *Advances in Computer-Human Interaction, 2008 First International Conference on*, IEEE, Sainte Luce, 2008, pp. 230 – 235.
- [55] A. Katifori, C. Vassilakis, I. Daradimos, G. Lepouras, Y. Ioannidis, A. Dix, A. Poggi, T. Catarci, Personal ontology creation and visualization for a personal interaction management system, in: *Proceedings of PIM Workshop, CHI 2008, Vol. 15*, ACM, Florence, Italy, 2008.
- [56] J. Stan, E. Egyed-Zsigmond, A. Joly, P. Maret, A user profile ontology for situation-aware social networking, in: *3rd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI2008)*, Patras, Greece, 2008.
- [57] W. T. Niu, J. Kay, [Personaf: framework for personalised ontological reasoning in pervasive computing](#), *User Modeling and User-Adapted Interaction* 20 (1) (2010) 1–40, springer Netherlands. doi:10.1007/s11257-009-9068-2.
URL <http://dx.doi.org/10.1007/s11257-009-9068-2>

- [58] A. Dix, A. Katifori, G. Lepouras, C. Vassilakis, N. Shabir, Spreading activation over ontology-based resources: from personal context to web scale reasoning, *International Journal of Semantic Computing* 4 (1) (2010) 59–102, world Scientific.
- [59] K.-L. Skillen, L. Chen, C. D. Nugent, M. P. Donnelly, W. Burns, I. Solheim, [Ubiquitous Computing and Ambient Intelligence: 6th International Conference, UCAmI 2012, Vitoria-Gasteiz, Spain, December 3-5, 2012. Proceedings](#), Vol. 7656 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, Ch. Ontological User Profile Modeling for Context-Aware Application Personalization, pp. 261–268. doi:10.1007/978-3-642-35377-2_36.
URL http://dx.doi.org/10.1007/978-3-642-35377-2_36
- [60] L. Yao, J. Tang, J. Li, A unified approach to researcher profiling, in: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI '07*, IEEE Computer Society, Washington, DC, USA, 2007, pp. 359–366.
- [61] A. Katifori, C. Vassilakis, A. Dix, Ontologies and the brain: Using spreading activation through ontologies to support personal interaction, *Cognitive Systems Research* 11 (1) (2010) 25–41, elsevier B. V.
- [62] G. A. Becerril, E. R. Lozano, E. J. M. Molina, Enfoque semántico para el descubrimiento de recursos sensible al contexto sobre contenidos académicos estructurados con oai-pmh, *Computación y Sistemas* 20 (1) (2016) 127–142.
- [63] G. Cong, W. Fan, F. Geerts, X. Jia, S. Ma, Improving data quality: Consistency and accuracy, in: *Proceedings of the 33rd international conference on Very large data bases, VLDB Endowment*, ACM, Vienna, Austria, 2007, pp. 315–326.
- [64] Instituto Tecnológico y de Estudios Superiores de Monterrey, [Siip - sistema de información de investigación y posgrado](#), <https://prod127ws.itesm.mx:4443/siip/contenido/welcome.faces?vista=A&cveCampus=Q&cvePrograma=MSM> (accessed on 30 March 2018) (2018).
URL <https://prod127ws.itesm.mx:4443/siip/contenido/welcome.faces?vista=A&cveCampus=Q&cvePrograma=MSM>
- [65] Clarivate Analytics, [Data integration](#), <https://clarivate.com/products/data-integration/> (accessed on 11 October 2018) (2018).
URL <https://clarivate.com/products/data-integration/>

- [66] LinkedIn, [Getting started with the rest api](https://developer.linkedin.com/docs/rest-api), <https://developer.linkedin.com/docs/rest-api> (accessed on 11 April 2018) (2018).
URL <https://developer.linkedin.com/docs/rest-api>
- [67] dblp, [Welcome to dblp](https://dblp.uni-trier.de/), <https://dblp.uni-trier.de/> (accessed on 12 April 2018) (2018).
URL <https://dblp.uni-trier.de/>
- [68] Tecnológico de Monterrey, [Tecnológico de monterrey](http://semtech.mty.itesm.mx:8080/vivo2/), <http://semtech.mty.itesm.mx:8080/vivo2/> (accessed on 02 May 2018) (2018).
URL <http://semtech.mty.itesm.mx:8080/vivo2/>
- [69] F. J. Cantú, H. G. Ceballos, A multiagent knowledge and information network approach for managing research assets, *Expert Systems with Applications* 37 (7) (2010) 5272–5284, elsevier.
- [70] M. Ley, [Dblp: Some lessons learned](https://doi.org/10.14778/1687553.1687577), *Proceedings of the VLDB Endowment* 2 (2) (2009) 1493–1500. doi:10.14778/1687553.1687577.
URL <https://doi.org/10.14778/1687553.1687577>
- [71] LinkedIn, [Linkedin](https://www.linkedin.com/), <https://www.linkedin.com/> (accessed on 11 October 2018) (2018).
URL <https://www.linkedin.com/>
- [72] L. Haak, J. Brown, M. Buys, M. Calvo, A. Cardoso, D.-P. Ade, P. Demain, T. Demeranville, M. Duine, P. Gopinath, S. Harley, A. Heredia, S. Hershberger, J. Hu, L. Krzmarich, A. Meadows, G. Mejias, N. Miyairi, A. Montenegro, N. George, E. Olson, L. Paglione, J. Perez, R. Peters, W. Simpson, S. Joseph, L. J. Wilkinson, C. Wilmers, D. Wright, A. T. Wrigley, A. V. Wynne, [Orcid public data file 2017](https://figshare.com/articles/ORCID_Public_Data_File_2017/5479792), https://figshare.com/articles/ORCID_Public_Data_File_2017/5479792 (accessed on 11 October 2018) (2017). doi:10.6084/m9.figshare.5479792.v1.
URL https://figshare.com/articles/ORCID_Public_Data_File_2017/5479792
- [73] E. Sutanta, R. Wardoyo, K. Mustofa, E. Winarko, Survey: Models and prototypes of schema matching, *International Journal of Electrical and Computer Engineering* 6 (3) (2016) 1011–1022.
- [74] DuraSpace Wiki, [Vivo](https://wiki.duraspace.org/display/VIVO), <https://wiki.duraspace.org/display/VIVO> (accessed on 13 March 2018) (2018).
URL <https://wiki.duraspace.org/display/VIVO>

- [75] K. Geis, [Source ontologies for vivo](https://wiki.duraspace.org/display/VIVODOC19x/Source+ontologies+for+VIVO), <https://wiki.duraspace.org/display/VIVODOC19x/Source+ontologies+for+VIVO> (accessed on 22 October 2018) (April 2018).
URL <https://wiki.duraspace.org/display/VIVODOC19x/Source+ontologies+for+VIVO>
- [76] D. Brickley, L. Miller, [Foaf vocabulary specification 0.99](http://xmlns.com/foaf/spec/), <http://xmlns.com/foaf/spec/> (accessed on 02 November 2016) (January 2014).
URL <http://xmlns.com/foaf/spec/>
- [77] S. Perreault, [vcard format specification](https://tools.ietf.org/html/rfc6350), <https://tools.ietf.org/html/rfc6350> (accessed on 02 November 2016) (August 2011).
URL <https://tools.ietf.org/html/rfc6350>
- [78] R. Iannella, J. McKinney, [vcard ontology - for describing people and organizations](https://www.w3.org/TR/vcard-rdf/#RFC6350), <https://www.w3.org/TR/vcard-rdf/#RFC6350> (accessed on 02 November 2016) (May 2014).
URL <https://www.w3.org/TR/vcard-rdf/#RFC6350>
- [79] DCMI, [Dcmi metadata terms](http://dublincore.org/documents/dcmi-terms/), <http://dublincore.org/documents/dcmi-terms/> (accessed on 03 November 2016) (June 2012).
URL <http://dublincore.org/documents/dcmi-terms/>
- [80] DCMI, [Dublin core metadata element set, version 1.1](http://dublincore.org/documents/dces/), <http://dublincore.org/documents/dces/> (accessed on 02 November 2016) (June 2012).
URL <http://dublincore.org/documents/dces/>
- [81] A. Miles, S. Bechhofer, [Skos simple knowledge organization system rdf schema](https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html), <https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html> (accessed on 03 November 2016) (August 2008).
URL <https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html>
- [82] C. Hauschke, [Vivo ontology domain definition](https://wiki.duraspace.org/display/VIVODOC110x/VIVO+Ontology+Domain+Definition), <https://wiki.duraspace.org/display/VIVODOC110x/VIVO+Ontology+Domain+Definition> (accessed on 22 October 2018) (April 2018).
URL <https://wiki.duraspace.org/display/VIVODOC110x/VIVO+Ontology+Domain+Definition>

- [83] M. Ley, *Dblp xml requests*, <http://dblp.uni-trier.de/xml/docu/dblpxmlreq.pdf> (accessed on 12 April 2018) (2009).
URL <http://dblp.uni-trier.de/xml/docu/dblpxmlreq.pdf>
- [84] Elsevier, *Scopus*, <https://www.elsevier.com/solutions/scopus> (accessed on 14 April 2018) (2018).
URL <https://www.elsevier.com/solutions/scopus>
- [85] Elsevier, *Elsevier developers*, <https://dev.elsevier.com/> (accessed on 14 April 2018) (2018).
URL <https://dev.elsevier.com/>
- [86] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, The Morgan Kaufmann Series in Data Management Systems, 2011.
- [87] starr.net, *Html codes*, <http://www.starr.net/is/type/htmlcodes.html> (accessed on 17 April 2018) (2018).
URL <http://www.starr.net/is/type/htmlcodes.html>
- [88] LuminosoInsight, *python-ffty*, <https://github.com/LuminosoInsight/python-ffty> (accessed on 13 April 2018) (2018).
URL <https://github.com/LuminosoInsight/python-ffty>
- [89] R. Speer, *ffty*, <http://ffty.readthedocs.io/en/latest/> (accessed on 13 April 2018) (2017).
URL <http://ffty.readthedocs.io/en/latest/>
- [90] M. Caraciolo, Collaborative filtering: Implementation with python!, <http://aimotion.blogspot.com.es/2009/11/collaborative-filtering-implementation.html> (accessed on 27 April 2017).
- [91] NumPy developers, *Numpy*, <http://www.numpy.org/> (accessed on 27 April 2017).
- [92] SciPy developers, *Scipy.org*, <https://scipy.org/> (accessed on 27 April 2017).
- [93] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* 12 (2011) 2825–2830.

- [94] DuraSpace, [Vivo](http://vivoweb.org/), <http://vivoweb.org/> (accessed on 13 March 2018) (2018).
URL <http://vivoweb.org/>
- [95] H. G. Ceballos Cancino, [Investigación: Investigadores](http://bibliotecatec21.mty.itesm.mx/c.php?g=119173&p=1016171), <http://bibliotecatec21.mty.itesm.mx/c.php?g=119173&p=1016171> (accessed on 22 March 2017) (March 2017).
URL <http://bibliotecatec21.mty.itesm.mx/c.php?g=119173&p=1016171>
- [96] Physical Web, Walk up and use anything, <https://google.github.io/physical-web/> (accessed on 13 March 2017).
- [97] K. Meehan, T. Lunney, K. Curran, A. McCaughey, Aggregating social media data with temporal and environmental context for recommendation in a mobile tour guide system, *Journal of Hospitality and Tourism Technology* 7 (3) (2016) 281–299.
- [98] SmartSDK, Smart pois: a fiware-based technology to engage users and make cities more sustainable, <https://www.smartsdk.eu/2017/02/16/smartpoi/> (accessed on 13 March 2017).
- [99] HOP Ubiquitous, Smart poi, https://storage.googleapis.com/smartcity/SmartPOI_A4_lr.pdf (accessed on 23 March 2017).
- [100] Yahoo Answers, All categories, <https://answers.yahoo.com/dir/index> (accessed on 11 November 2017).
- [101] S. Dooms, T. De Pessemier, L. Martens, An online evaluation of explicit feedback mechanisms for recommender systems, in: 7th International Conference on Web Information Systems and Technologies (WEBIST-2011), Ghent University, Department of Information technology, 2011, pp. 391–394.
- [102] B. Kitchenham, S. L. Pfleeger, Principles of survey research: part 5: populations and samples, *ACM SIGSOFT Software Engineering Notes* 27 (5) (2002) 17–20.
- [103] A. Gómez Oliva, M. Server Gómez, A. J. Jara, M. C. Parra-Meroño, Turismo inteligente y patrimonio cultural: Un sector a explorar en el desarrollo de las smart cities, *Internacional Journal of Scientific Management and Tourism* 3 (2017) 389–411.
- [104] Web of Science, [Research areas \(categories / classification\)](https://apps.webofknowledge.com), <https://apps.webofknowledge.com> (accessed on 26 January 2018), last modified:

- 11/30/2017 (November 2017).
URL <https://apps.webofknowledge.com>
- [105] Tecnológico de Monterrey, [Campus](#), Official Website (2017).
URL <https://tec.mx/es>
- [106] DuraSpace Wiki, [Vivo 1.8.x documentation: A simple installation](#), <https://wiki.duraspace.org/display/VIVODOC18x/A+simple+installation> (accessed on 13 March 2018) (2018).
URL <https://wiki.duraspace.org/display/VIVODOC18x/A+simple+installation>
- [107] Oracle, [Integrated cloud applications & platform services](#), <https://www.oracle.com/index.html> (accessed on 13 March 2018) (2018).
URL <https://www.oracle.com/index.html>
- [108] The Apache Software Foundation, [Apache tomcat](#), <http://tomcat.apache.org/> (accessed on 13 March 2018) (2018).
URL <http://tomcat.apache.org/>
- [109] C. Hock-Chuan, [How to install apache tomcat 9 \(on windows, mac os x, ubuntu\) and get started with java servlet programming](#), https://www.ntu.edu.sg/home/ehchua/programming/howto/Tomcat_HowTo.html (accessed on 13 March 2018), last modified: February, 2018 (2018).
URL https://www.ntu.edu.sg/home/ehchua/programming/howto/Tomcat_HowTo.html
- [110] The Apache Software Foundation, [Apache ant](#), <http://ant.apache.org/> (accessed on 13 March 2018) (2018).
URL <http://ant.apache.org/>
- [111] Mkyong.com, [How to install apache ant on windows](#), <https://www.mkyong.com/ant/how-to-install-apache-ant-on-windows/> (accessed on 13 March 2018) (2018).
URL <https://www.mkyong.com/ant/how-to-install-apache-ant-on-windows/>
- [112] Oracle, [Mysql](#), <https://dev.mysql.com> (accessed on 13 March 2018) (2018).
URL <https://dev.mysql.com>

- [113] D. S. Alves Pérez, *Mysql como comando en el cmd*, <http://solucionesia.blogspot.mx/2011/11/mysql-como-comando-en-el-cmd.html> (accessed on 13 March 2018) (2018).
URL <http://solucionesia.blogspot.mx/2011/11/mysql-como-comando-en-el-cmd.html>
- [114] DuraSpace, *Vivo v1.8.1*, <https://github.com/vivo-project/VIVO/releases/tag/rel-1.8.1> (accessed on 13 March 2018) (2018).
URL <https://github.com/vivo-project/VIVO/releases/tag/rel-1.8.1>
- [115] FIWARE, *Developers: Start using fiware right now*, <https://www.fiware.org/developers-entrepreneurs/> (accessed on 16 November 2016) (2016).
URL <https://www.fiware.org/developers-entrepreneurs/>
- [116] FIWARE, *Handling authorization and access control to apis*, <http://fiwaretourguide.readthedocs.io/en/latest/handling-authorization-and-access-control-to-apis/introduction/> (accessed on 16 November 2016) (2016).
URL <http://fiwaretourguide.readthedocs.io/en/latest/handling-authorization-and-access-control-to-apis/introduction/>
- [117] Gartner, *It glossary: Extensible markup language (xml)*, <https://www.gartner.com/it-glossary/xml-extensible-markup-language> (accessed on 12 April 2018) (2018).
URL <https://www.gartner.com/it-glossary/xml-extensible-markup-language>
- [118] Gartner, *It glossary: Sql (structured query language)*, <https://www.gartner.com/it-glossary/sql-structured-query-language/> (accessed on 12 April 2018) (2018).
URL <https://www.gartner.com/it-glossary/sql-structured-query-language/>
- [119] Y. N. Silva, I. Almeida, M. Queiroz, *Sql: From traditional databases to big data*, in: *Proceedings of the 47th ACM Technical Symposium on Computing Science Education, SIGCSE '16*, ACM, New York, NY, USA, 2016, pp. 413–418.
- [120] S.-H. Cha, *Comprehensive survey on distance/similarity measures between probability density functions*, *INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES* 1 (4) (2007) 300–307.
- [121] Laerd Statistics, *Pearson product-moment correlation*, <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php> (accessed on 29 April 2017).

- [122] B. McCune, J. B. Grace, D. L. Urban, *Analysis of Ecological Communities*, 2nd Edition, MjM Software Design, USA, 2002.

Published Papers

Semantic Approach for Discovery and Visualization of Academic Information Structured with OAI-PMH

Joanna Alvarado-Uribe¹, Arianna Becerril Garcia², Miguel Gonzalez-Mendoza¹, Rafael Lozano Espinosa¹, José Martín Molina Espinosa¹

¹ Tecnológico de Monterrey, School of Engineering and Sciences, Av. Eugenio Garza Sada No. 2501 Sur, Col. Tecnológico, 64849, Monterrey, N.L., México, {A00987514, mgonza, ralozano, jose.molina}@itesm.mx

² Universidad Autónoma del Estado de México, Instituto Literario Ote. No. 100, Col. Centro, 50000, Toluca, Estado de México, México, abecerrilg@uaemex.mx






Abstract: There are different channels to communicate the results of a scientific research; however, several research communities state that the Open Access (OA) is the future of academic publishing. These Open Access Platforms have adopted OAI-PMH (Open Archives Initiative - the Protocol for Metadata Harvesting) as a standard for communication and interoperability. Nevertheless, it is significant to highlight that the open source knowledge discovery services based on an index of OA have not been developed. Therefore, it is necessary to address Knowledge Discovery (KD) within these platforms aiming at students, teachers and/or researchers, to recover both, the resources requested and the resources that are not explicitly requested – which are also appropriate. This objective represents an important issue for structured resources under OAI-PMH. This fact is caused because interoperability with other developments carried out outside their implementation environment is generally not a priority (Level 1 "Shared term definitions"). It is here, where the Semantic Web (SW) becomes a cornerstone of this work. Consequently, we propose OntoOAI, a semantic approach for the selective knowledge discovery and visualization into structured information with OAI-PMH, focused on supporting the activities of scientific or academic research for a specific user. Because of the academic nature of the structured resources with OAI-PMH, the field of application chosen is the context information of a student. Finally, in order to validate the proposed approach, we use the RUDAR (Roskilde University Digital Archive) and REDALYC (Red de Revistas Científicas de América Latina y el Caribe, España y Portugal) repositories, which implement the OAI-PMH protocol, as well as one student profile for carrying out KD.

Keywords: the Semantic web; knowledge discovery; user profile ontology; ontology merging; OAI-PMH; visualization



Article

HyRA: A Hybrid Recommendation Algorithm Focused on Smart POI. Ceutí as a Study Scenario

Joanna Alvarado-Uribe ^{1,*} , Andrea Gómez-Oliva ^{2,3}, Ari Yair Barrera-Animas ¹ , Germán Molina ², Miguel Gonzalez-Mendoza ¹ , María Concepción Parra-Meroño ³  and Antonio J. Jara ⁴ 

¹ Computer Science Department, Tecnológico de Monterrey, School of Engineering and Sciences, Carretera Lago de Guadalupe Km. 3.5, Col. Margarita Maza de Juárez, Atizapán de Zaragoza 52926, Estado de Mexico, Mexico; ybarrera@itesm.mx (A.Y.B.-A.); mgorza@itesm.mx (M.G.-M.)

² HOP Ubiquitous S.L., Calle Luis Buñuel No. 6, 30562 Ceutí, Murcia, Spain; andrea@hopu.eu (A.G.-O.); german@hopu.eu (G.M.)

³ Social Sciences, Law and Business Department, Universidad Católica de Murcia (UCAM), Business Administration, Marketing and Economics, Campus de los Jerónimos, Guadalupe, 30107 Murcia, Spain; agomez14@alu.ucam.edu (A.G.-O.); mcparra@ucam.edu (M.C.P.-M.)

⁴ Institute of Information Systems, University of Applied Sciences Western Switzerland, ConEx Lab, 3960 Sierre, Switzerland; jara@ieee.org

* Correspondence: joanna.alvarado@itesm.mx; Tel.: +52-773-734-3235

Received: 7 January 2018; Accepted: 7 March 2018; Published: 17 March 2018

Abstract: Nowadays, Physical Web together with the increase in the use of mobile devices, Global Positioning System (GPS), and Social Networking Sites (SNS) have caused users to share enriched information on the Web such as their tourist experiences. Therefore, an area that has been significantly improved by using the contextual information provided by these technologies is tourism. In this way, the main goals of this work are to propose and develop an algorithm focused on the recommendation of Smart Point of Interaction (Smart POI) for a specific user according to his/her preferences and the Smart POIs' context. Hence, a novel Hybrid Recommendation Algorithm (HyRA) is presented by incorporating an aggregation operator into the user-based Collaborative Filtering (CF) algorithm as well as including the Smart POIs' categories and geographical information. For the experimental phase, two real-world datasets have been collected and preprocessed. In addition, one Smart POIs' categories dataset was built. As a result, a dataset composed of 16 Smart POIs, another constituted by the explicit preferences of 200 respondents, and the last dataset integrated by 13 Smart POIs' categories are provided. The experimental results show that the recommendations suggested by HyRA are promising.

Keywords: recommendation algorithm; point-of-interest; similarity and distance measures; aggregation operator; POI category; geographical influence; tourism

1. Introduction

Nowadays, Physical Web [1] together with the increase in the use of mobile devices, Global Positioning System (GPS), and Social Networking Sites (SNS) have caused users to share enriched information on the Web such as their tourist experiences [2]. Nevertheless, generally, tourist guide applications are based on information heavily related to the location, disregarding other types of context information, which can be provided by the user. Consequently, all information is available to all users in touch, leading to the issue known as "information overload" [3] as well as problems of inappropriate suggestions [4]. Such facts entail the need to enhance the user's individual tourist experience according to his/her preferences and context information.

Una herramienta visual para la búsqueda semántica RDF

Joanna Alvarado-Uribe¹, Miguel González-Mendoza¹, Neil Hernández-Gress¹,
Carlos Eli Escobar-Ruiz² y Marcos Uriel Hernández-Camacho²

¹Tecnológico de Monterrey, Campus Estado de México,
México

²Universidad Politécnica de Chiapas, Chiapas,
México

joanna.1890@gmail.com; {mgonza, ngress}@itesm.mx; carlosescobar@
portaltuxtla.com; uriel.hdzc@gmail.com
<http://www.itesm.mx>
<http://www.upchiapas.edu.mx>

Resumen. La cantidad de información que uno o más usuarios de Internet generan para la Web Semántica está incrementando diariamente. Por esto, es necesario desarrollar herramientas que nos permitan mostrar esta información de una manera rápida, simple y fácil de entender. De acuerdo con esta premisa, hemos desarrollado una herramienta de visualización de datos semánticos, denominada DBPedia Search, capaz de: 1) consultar cualquier base de datos de tripletas que cuente con un *endpoint* de SPARQL y; 2) generar gráficos, mapas de calor y mapas de geolocalización de manera automática, con base en la información obtenida de la búsqueda realizada por el usuario. El objetivo principal es realizar una búsqueda y un análisis simplificados de los datos semánticos y presentarlos gráficamente.

Palabras clave: DBPedia search, visualización, *Endpoint* de SPARQL, tripletas.

1. Introducción

La Web Semántica es percibida como un área de investigación multidisciplinaria que combina campos científicos como la Inteligencia Artificial, Ciencias de la Información, Teoría de Algoritmo y de la Complejidad, Teoría de Base de datos, Redes de Computadoras, entre otros [1].

La Web Semántica se basa en la idea de agregar más semántica legible por la computadora a la información web a través de anotaciones escritas en *Resource Description Framework* (RDF) [2]. El modelo RDF se introdujo en 1999 como una recomendación del *World Wide Web Consortium* (W3C). Debido a esto,

Towards the Development of a Smart Tourism Application Based on Smart POI and Recommendation Algorithms: Ceutí as a Study Case

Joanna Alvarado-Uribe¹(✉), Andrea Gómez-Oliva^{2,3}, Germán Molina³, Miguel Gonzalez-Mendoza¹, María Concepción Parra-Meroño², and Antonio J. Jara⁴

¹ School of Engineering and Sciences, Tecnológico de Monterrey, Carretera Lago de Guadalupe Km. 3.5, Col. Margarita Maza de Juárez, 52926 Atizapán de Zaragoza, Estado de México, Mexico
{A00987514,mgonza}@itesm.mx

² Universidad Católica de Murcia (UCAM), Campus de los Jerónimos, Guadalupe, 30107 Murcia, Spain
mcparra@ucam.edu

³ HOP Ubiquitous S.L., Calle Luis Buñuel No. 6, Ceutí, 30562 Murcia, Spain
{andrea,german}@hopu.eu

⁴ University of Applied Sciences Western Switzerland, 3960 Sierre, Switzerland
jara@ieee.org

Abstract. Nowadays, major industry, government, and citizen initiatives are boosting the development of smart applications and services that improve the quality of life of people in domains such as mobility, security, health, and tourism, using both emerging and existing technologies. In particular, a smart tourist destination aims to improve both the citizen's quality of life and the tourist experience making use of innovation and technology. In this way, the main idea of this work is to develop a smart application focused on improving the tourist experience. The application will be based on a new concept called Smart Point of Interaction (Smart POI), the user experience research in this area, as well as a Smart POI recommendation algorithm capable of considering both user preferences and geographical influence when calculating new suggestions for users. For the experimental phase, two scenarios are considered: a simulated story and a real-world environment. In the real-world scenario, the town of Ceutí will be the first scope while for the simulated scenario, a database will be generated through surveys. As a first result, the points of interest, the target audience, and the features that will constitute the database representative of the user profile have been defined according to the real-world scenario in Ceutí. Moreover, the incorporation of an explicit feedback mechanism for the Smart POIs has been proposed as an initial approach to address user preferences.

This doctoral dissertation was typed in using LATEX2¹ by Joanna ALVARADO URIBE

¹The style used to set up this dissertation was prepared by the Group of Intelligent Systems of the Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Estado de Mexico.