



Semantic Web and Semantic Technologies to enhance Innovation and
Technology Watch processes

Alain Perez Riaño

Mondragon Goi Eskola Politeknikoa
Electronics and Computing Department

October 14, 2016



Semantic Web and Semantic Technologies to enhance Innovation and
Technology Watch processes

Alain Perez Riaño

Supervisor:

Dr. Felix Larrinaga Barrenechea
Electronics and Computing Department
Mondragon Unibertsitatea

Co-supervisor:

Dr. Juan Ignacio Igartua Lopez
Mechanics and Industrial Productions
Mondragon Unibertsitatea

*for obtaining the degree of Doctor
under the doctoral program at Mondragon Unibertsitatea:
New Information and communication technologies.*

Thesis committee:

Chairman: Dr. Edward Curry (Digital Enterprise Research Institute)
Member: Dr. Arkaitz Zubiaga Mendiakua (University of Warwick)
Member: Dr. Izaskun Fernandez Gonzalez (IK4-Tekniker)
Member: Dr. David Buján Carballal (University of Deusto)
Secretary: Dr. Leire Etxeberria Elorza (Mondragon Unibertsitatea)

October 14, 2016

A mis padres, sin vosotros nada de esto sería posible.

Originality Statement

I declare that I am the sole author of this work. This is a true copy of the final document, including any revision which may have been ordered by my examiners. I understand that my work may be available to the public, either in the school library or electronic format.

Abstract

Innovation is a key process for Small and Medium Enterprises in order to survive and evolve in a competitive environment. Ideas and idea management are considered the basis for Innovation. Gathering data on how current technologies and competitors evolve is another key factor for companies' innovation. Therefore, this thesis focuses the application of Information and Communication Technologies and more specifically Semantic Web and Semantic Technologies on Idea Management Systems and Technology Watch Systems.

Innovation and Technology Watch platform managers usually face many problems related with the data they collect and manage. Those managers have to deal with a large amount of information distributed in different platforms, not always interoperable among them. It is vital to share data between platforms so it can be converted into knowledge. Many of the tasks they perform are non productive and too much time and effort is expended on them. Moreover, Innovation process managers have difficulties in identifying why an idea contest has been successful.

Our proposal is to analyze different Information and Communication Technologies that can assist companies with their Innovation and Technology Watch processes. Thus, we studied several Semantic and Web technologies, we build some conceptual models and tested them in different case studies to see the results achieved in real scenarios.

The outcome of this thesis has been the creation of a solution architecture to enable interoperability among platforms and to ease the work of the process' managers. In this framework and to complement the architecture, two ontologies have been developed: (1) Gi2Mo Wave and (2) Mentions Ontology. On one hand, Gi2Mo Wave focused on annotating the background of idea contests, assisting on the analysis of the contests and easing its replication. On the other hand, Mentions Ontology focused on annotating the elements mentioned in plain text content, such as ideas or news items. That way, Mentions Ontology creates a way to link the related content, enabling the interoperability among content from different platforms.

In order to test the architecture, a new web Idea Management System and a Technology Watch system have been also developed. The platforms incorporate semantic ontologies and tools to enable interoperability. We also demonstrate how Semantic Technologies reduce human workload by contributing on the automatic classification of content in the Technology Watch process. Finally, conclusions have been gathered according to the results achieved testing the used technologies, identifying the ones with best results.

Laburpena

Berrikuntza prozesu oso garrantzitsu bat da Enpresa Txiki eta Ertainen lehiakor eta bizirik irauteko ingurumen lehiakor batean. Berrikuntza prozesuek ideiak eta ideien kudeaketa dituzte oinarri gisa. Teknologiek eta lehiakideek nola eboluzionatzen duten jakitzea ere garrantzitsua da enpresen berrikuntzarako, eta baita ere informazio hori kudeatzea. Beraz, Informazio eta Komunikazio sistemen aplikazioan oinarritzen da tesi hau, zehazkiago Web Semantika eta Teknologia Semantikoetan eta hauen aplikazioa Ideia Kudeaketa eta Zaintza Teknologikoko sistemetan.

Berrikuntza eta Zaintza Teknologikoko plataformen kudeatzaileek arazo larriak izaten dituzte jasotako datuekin eta haien kudeaketarekin. Kudeatzaile horiek plataforma ezberdinetan banatutako informazio kantitate handi batekin topo egiten dute eta plataforma horiek ez dira beti elkar eraginkorrak. Beraz, beharrezkoa da plataforma ezberdinetako datuak elkarren artean partekatzea gero datu horiek “ezagutza” bihurtzeko. Gainera, kudeatzaileek egiten dituzten zeregin kopuru handi bat zeregin ez emankorrak dira, denbora eta esfortzu handia suposatzen dute baliozko ezer gehitu gabe. Eta ez hori bakarrik, berrikuntza prozesuko kudeatzaileek zail izaten dute ideia lehiaketan arrakastaren arrazoiak identifikatzen.

Gure proposamena Informazio eta Komunikazio Teknologia ezberdinak frogatzea da enpresen berrikuntzako eta zaintza teknologikoko prozesuetan laguntzeko. Honela, hainbat teknologia semantiko eta web teknologia aztertu dira, modelo kontzeptual batzuk eraikitzen eta probatzen benetako erabilpen kasutan lortutako emaitzak konprobatzeko.

Tesi honen lorpena plataformen arteko elkar eraginkortasuna ahalbidetzen duen eta prozesuen kudeatzaileen lana errazten duen modelo baten sorpena izan da. Horrela eta sortutako modeloa konplimentatzeko, bi ontologia sortu dira: (1) Gi2Mo-Wave eta (2) Mentions Ontology. Alde batetik, Gi2Mo-Wave ontologia ideien eta ideia lehiaketan testuinguruaren errepresentazio semantikoan oinarritu da. Horrela testuinguruaren analisisa errazten da, ideia lehiaketa arrakastatsua errepikatzea ere errazagoa eginez. Bestalde, Mentions-Ontology ontologia eduki ezberdinen (ideiak edo berriak adibidez) testuetan aipatutako elementuen errepresentazio semantikoan oinarritu da. Horrela, Mentions Ontology ontologiak edukia elkar konektatzeko era bat sortzen du, plataforma ezberdinen edukiaren arteko elkar eraginkortasuna ahalbidetzen.

Modelo edo arkitektura hau frogatzeko, Ideia Kudeaketa Sistema eta Zaintza teknologikoko web plataforma berri batzuk garatu dira ere. Plataforma hauek tresna eta ontologia semantikoak dituzte txertatuta, beraien arteko elkar eraginkortasuna ahalbidetzeko. Gainera, teknologia semantikoen aplikazioarekin giza lan kargaren murrizketa nola gauzatu ere frogatzen dugu, Zaintza Teknologikoko edukiaren klasifikazio automatikoan ekarpenak eginez. Bukatzeko, konklusioak bildu dira erabili diren teknologien frogetatik jasotako emaitzetan oinarrituta eta emaitza onenak lortu dituztenak identifikatu dira.

Resumen

El proceso de Innovación es un proceso clave para la supervivencia y evolución de las Pequeñas y Medianas Empresas en un entorno competitivo. Las ideas y la gestión de ideas se consideran la base de la innovación. Recopilar datos sobre cómo evolucionan las actuales tecnologías y los competidores es otro factor clave para la innovación de las empresas. Por lo tanto, esta tesis se centra en la aplicación de Tecnologías de la Información y Comunicación, más concretamente la aplicación de Web Semántica y Tecnologías Semánticas en los Sistemas de Gestión de ideas y de Vigilancia Tecnológica.

Los gestores de las plataformas de innovación y de vigilancia tecnológica se enfrentan a muchos problemas relacionados con los datos que recogen y gestionan. Esos gestores se enfrentan a una gran cantidad de información distribuida en diferentes plataformas, no siempre interoperables entre ellas. Es de vital importancia que las diferentes plataformas sean capaces de compartir datos entre ellas, de modo que esos datos puedan convertirse en el conocimiento. Muchas de las tareas realizadas por estos gestores son tareas no productivas y se invierte demasiado tiempo y esfuerzo en realizarlas. Además, los responsables de los procesos de innovación tienen dificultades para identificar por qué un concurso de ideas ha sido un éxito.

Nuestra propuesta es analizar diferentes Tecnologías de Información y Comunicación que puedan ayudar a las empresas con sus procesos de Innovación y Vigilancia Tecnológica. Por ello, hemos estudiado varias tecnologías semánticas y Web, hemos desarrollado algunos modelos conceptuales y los hemos probado en diferentes casos de estudio para ver los resultados obtenidos en escenarios reales.

El resultado de este trabajo ha sido la creación de una arquitectura que permite la interoperabilidad entre plataformas y que facilita el trabajo de los responsables de los procesos. En este marco, y para complementar la arquitectura, se han desarrollado dos ontologías: (1) Gi2Mo Wave y (2) Mentions Ontology. Gi2Mo Wave se centra en la anotación del contexto de los de ideas, ayudando en el análisis de los concursos y facilitando su replicación. Por otro lado, Mentions Ontology se centra en la anotación de los elementos mencionados en el texto plano de contenidos de diferente índole, como por ejemplo ideas o noticias. Así, Mentions Ontology crea una forma de encontrar relaciones entre contenidos, lo que permite la interoperabilidad entre los contenidos de diferentes plataformas.

Con el fin de probar la arquitectura, también se han desarrollado dos plataformas: un Sistema de Gestión de Ideas y un Sistema de Vigilancia Tecnológica. Las plataformas incorporan ontologías semánticas y herramientas para permitir su interoperabilidad. Además, demostramos cómo reducir la carga de trabajo humana, mediante el uso de tecnologías semánticas para la clasificación automática del contenido del proceso de la Vigilancia Tecnológica. Por último, probando las tecnologías y herramientas se han recogido las conclusiones de acuerdo con los resultados obtenidos, identificando las que obtienen los mejores resultados.

Agradecimientos

Me gustaría expresar mi gratitud a todos aquellos que me han ayudado durante todo el desarrollo de esta tesis.

En primer lugar, estoy muy agradecido al director, Felix Larrinaga, por darme la oportunidad de participar en este trabajo de investigación y por su ayuda, experiencia y conocimientos que han sido muy útiles para el desarrollo de la tesis. Agradecer también al codirector Juan Ignacio Igartua por tomar parte en el equipo y darnos la perspectiva de Innovación del proyecto.

Además, también me gustaría agradecer a Osane Lizarralde por todos los consejos y la ayuda proporcionados durante la investigación. Gracias a Rosa Basagoiti y a Ronny Adalberto Cortez por el trabajo realizado con la última parte de la tesis.

Agradezco también a Mondragon Unibertsitatea por concederme la beca que ha hecho posible esta investigación y a Koniker por la financiación, el conjunto de datos y la experiencia proporcionada.

A todos los compañeros que están y que han pasado por el departamento por todas esas vivencias compartidas durante estos años. Espero que sigamos en contacto durante mucho tiempo.

Gracias a Nuria, por estar ahí todo este tiempo, por aguantarme y apoyarme en los buenos y en los malos momentos.

Por último, y no por ello menos importante, me gustaría dar las gracias a mis padres, que me ha apoyado en todas las decisiones que he tomado en mi vida, por sus consejos y por ayudarme de todas las maneras posibles.

Contents

Contents	xvii
List of Figures	xxi
List of Tables	xxiii
Acronyms	xxv
1 Introduction	1
1.1 Contribution area and scope	1
1.2 Problem and Motivation	2
1.3 Objectives	3
1.4 Research questions	4
1.5 Solution architecture	5
2 Foundations	7
2.1 Innovation	7
2.2 <i>Technology Watch</i> (TW)	12
2.3 Technologies to support Innovation and Technology Watch	15
2.3.1 Social Web	15
2.3.2 Semantic Technologies	18
2.3.3 <i>Semantic Web</i> (SW)	19
2.4 Applications of the Technologies in Innovation and Technology Watch con- text	28
2.4.1 Social Web applications for innovation process	28
2.4.2 Semantic Web applications for innovation process	30
2.4.3 Semantic Technologies and Semantic Web applications for Tech- nology Watch	34
2.4.4 Technology applications in the enterprise	46
2.5 Conclusions	48

3	Web Based Platform for Innovation processes	53
3.1	Introduction	53
3.2	Theory	54
	3.2.1 Innovation Process	54
	3.2.2 Technology	55
3.3	Material and Methods	57
	3.3.1 Objectives	57
	3.3.2 Metrics	58
	3.3.3 Methodology	60
	3.3.4 Platform	60
	3.3.5 Visualization	67
	3.3.6 Community modules	69
3.4	Results	70
	3.4.1 Ekiten (Mondragon Corporation)	70
3.5	Conclusions	72
4	Semantic Web to Link the Innovation Process with Internal and External Repositories	75
4.1	Introduction	75
4.2	Use Cases	77
4.3	Platform	79
4.4	Benefits	81
4.5	Conclusions	82
5	Semantic Web to enhance Innovation and Technology Watch linking	85
5.1	Introduction	86
5.2	Theory	86
	5.2.1 Innovation	87
	5.2.2 Technology Watch	88
5.3	Material and methods	89
	5.3.1 Solution architecture	89
	5.3.2 Tools	90
	5.3.3 Mentions Ontology (MO)	92
	5.3.4 Experiments and population	95
	5.3.5 Evaluation method	96
5.4	Results	98
	5.4.1 Concept definition	98
	5.4.2 Idea similarity	99
	5.4.3 News similarity	104
5.5	Conclusions	105

6	Semantic Technologies to enhance Technology Watch productivity	107
6.1	Introduction	107
6.2	Background	108
6.3	Theory	110
6.3.1	<i>Natural Language Processing</i> (NLP)	110
6.3.2	<i>Artificial Inteligence</i> (AI)	111
6.4	Material and methods	112
6.4.1	Datasets	112
6.4.2	Tools	112
6.4.3	Experiments	113
6.5	Results	116
6.6	Conclusions	119
7	Contribution	121
7.1	Cross university and company collaboration	121
7.1.1	Collaboration with <i>Digital Enterprise Research Institute</i> (DERI) of the <i>National University of Ireland</i> (NUI)	121
7.1.2	Collaboration with <i>Universidad Polit3cnica de Madrid</i> (UPM)	122
7.1.3	ISEA	122
7.1.4	Koniker	123
7.1.5	<i>Mondragon Unibertsitatea</i> (MU)	123
7.2	Projects	124
7.3	Scientific Publications	125
7.3.1	<i>A case study on the use of community platforms for inter-enterprise innovation</i> , 2011, 17th International Conference on Concurrent Enterprising (ICE 2011)	125
7.3.2	<i>INNOWEB: Gathering the context information of innovation processes with a collaborative social network platform</i> , 2013, 19th International Conference on Concurrent Enterprising (ICE 2013)	126
7.3.3	<i>The Role of Linked Data and Semantic-Technologies for Sustainability Idea Management</i> , 2013, 2nd International Symposium on Modelling and Knowledge Management for Sustainable Development (MoKMaSD 2013)	126
7.3.4	<i>Semantic Annotations to enhance Innovation and Technology Watch Interoperability</i> , pending at International Journal on Semantic Web and Information Systems (IJSWIS)	126
7.3.5	<i>A case study on the use of Machine Learning techniques for supporting Technology Watch</i> , pending at Data and Knowledge Engineering (DKE) journal	127
7.4	Summer Schools	127
8	Conclusions	129
8.1	Conclusions	129
8.2	Future Work	131

A Gi2MO Wave Ontology Specification	135
B Mentions Ontology Specification	147
Bibliography	161

List of Figures

1.1	Solution architecture for the thesis.	5
2.1	Innovation process baseline stages.	9
2.2	Five stages of the Technology Watch process.	14
2.3	The role of the technologies involved on the semantic web.	21
2.4	Tim Berners Lee's Semantic Web stack.	22
2.5	Linked Data Publishing Options and Workflows.	23
2.6	Linked Open Data publication classification according to Tim Berners-Lee	24
2.7	RDF graph example describing Eric Miller.	26
2.8	SPARQL as unification mechanism.	28
2.9	UML Class Diagram for the Gi2MO Ontology.	31
2.10	KISTI Reference & Academic Ontology for intellectual property and patents.	44
2.11	AtomOWL ontology for feeds (RSS).	45
3.1	Early stages of the innovation process	55
3.2	Incremental Development Cycle methodology	60
3.3	Idea generation in blog format.	61
3.4	Idea list.	61
3.5	Idea analysis implementation with idea2wiki module.	62
3.6	Wiki format for Idea enrichment.	63
3.7	Innoselect module for Idea Selection.	63
3.8	Multiple innovation processes (<i>Waves</i>) at once.	64
3.9	Real time notification selection.	65
3.10	Wave Event visualization example.	66
3.11	An example idea on RDF format.	67
3.12	Blazegraph visualization.	68
3.13	Idea Selection rating visualization example from <i>Innoselect Chart</i> module.	68
3.14	Wave Chart visualization example, showing idea related data.	69
3.15	Voting widgets in the ideas.	70
4.1	Energy Reduction and <i>Life-cycle Assessment</i> (LCA) ideas example	77
4.2	Data links example	78

4.3	Linking <i>Idea Management System</i> (IMS) with external and internal repositories.	79
4.4	Process to find related elements in Internal and External repositories.	80
4.5	Example of automatically added information from IdeaMentions module.	81
5.1	Solution architecture	90
5.2	Simplified mentioned concept identification to find content relationships.	91
5.3	Mentions Ontology's Class Diagram.	93
5.4	A graphic of a basic example of a Mentions Ontology (MO) use case.	94
5.5	Accuracy of identified concepts from the body.	98
5.6	Accuracy of identified concepts from the title.	98
5.7	Accuracy of critical concepts from the body.	99
5.8	Accuracy of critical concepts from the title.	99
5.9	Idea relationships in each iteration.	100
5.10	Idea relations by position on 1st iteration.	102
5.11	Idea relations by position on 2nd iteration.	102
5.12	Idea relations by position on 3rd iteration.	104
5.13	Idea relations by position on 4th iteration.	104
5.14	News items and ideas relationships.	105
5.15	News items and ideas relationships by position.	105
6.1	Five stages of the Technology Watch process according to SCIP.	109
6.2	Single document's hits.	114
6.3	Information flow with no interactions.	115
6.4	Information flow using automatic system.	116
6.5	Model generation.	117
6.6	Class hierarchy.	117
7.1	UML Class diagram for Gi2MO Wave Ontology	122
7.2	Current Technology Watch workflow and time spent in each step according to Koniker.	124

List of Tables

2.1	Analysis programs by Country.	36
2.2	Tools to help in the Technology Watch process stages.	40
2.3	Necessary steps to correctly implement a Text Mining project.	42
3.1	Inputs from Ekiten case study.	71
3.2	Outcomes from Ekiten case study.	71
3.3	2012 wave comparison	72
5.1	Null and alternative hypothesis for similarity experiments.	98
5.2	Statistical inference results by amount of idea-idea relationships.	104
5.3	Statistical inference results by amount of idea-idea relationships.	105
6.1	Results using “Raw text”, without and with “Semantic Annotations” (1st experiment).	118
6.3	Confusion matrix of the model with the best results (J48 algorithm - 1st hit - RAW text).	118
6.2	Results using “Experts’ information” , without and with “Semantic Annotations” (2nd experiment).	119
6.4	Superclass Confusion Matrix	119
6.5	Superclass results.	120

Acronyms

AI *Artificial Intelligence*

AJAX *Asynchronous JavaScript and XML*

API *Application Programming Interface*

CI *Competitive Intelligence*

CMS *Content Management System*

CSCW *Computer-supported cooperative work*

DERI *Digital Enterprise Research Institute*

EU *European Union*

FOAF *Friend of a Friend*

GHG *Greenhouse Gas*

GRDDL *Gleaning Resource Descriptions from Dialects of Languages*

HTML *HyperText Markup Language*

HTTP *Hypertext Transfer Protocol*

ICT *Information and Communication Technology*

IMS *Idea Management System*

LCA *Life-cycle Assessment*

LD *Linked Data*

LOD *Linked Open Data*

MO *Mentions Ontology*

MU *Mondragon Unibertsitatea*

MUTO *Modular Unified Tagging Ontology*

NLP *Natural Language Processing*
NUI *National University of Ireland*
OSS *Open Source Software*
OWL *Web Ontology Language*
POWDER *Protocol for Web Description Resources*
RDF *Resource Description Framework*
RDFa *Resource Description Framework in Attributes*
RDFs *Resource Description Framework Schema*
R+D+I *Research, Development and Innovation*
RSS *Rich Site Summary*
RTW *Real-Time Web*
SCIP *Society for Competitive Intelligence Professionals*
SKOS *Simple Knowledge Organization System*
SME *Small and Medium Enterprise*
SPARQL *SPARQL Protocol and RDF Query Language*
SQL *Structured Query Language*
SVM *Support Vector Machine*
SW *Semantic Web*
TM *Text Mining*
TW *Technology Watch*
UPM *Universidad Politécnica de Madrid*
URI *Uniform Resource Identifier*
W3C *World Wide Web Consortium*
XBRL *eXtensible Business Reporting Language*
XHTML *eXtensible HyperText Markup Language*
XML *eXtensible Markup Language*
XSLT *eXtensible Stylesheet Language Transformations*

Chapter 1

Introduction

Innovation and *Technology Watch* (TW) processes are key for *Small and Medium Enterprises* (SMEs) survival and evolution. Those processes depend on valid information gathering and a correct management. *Information and Communication Technologies* (ICTs) have been proven as technologies that can highly improve the management of information. This thesis focuses on the study of different ICTs that can improve those processes and their interoperability.

In this chapter, the thesis contribution area, motivation, objectives, research questions and solution architectures are outlined. We summarize the problems of current Innovation and TW platforms, we show the motivation of the team and the objectives defined due to the motivation. Related to the objectives, the research questions formulated for this thesis are set out to answer, use to focus down the research and later evaluated using clear research methodologies. Finally, we summarize the solution architecture that is proposed by the thesis to answer those questions.

1.1 Contribution area and scope

The contribution area of this thesis is in the field of Innovation and TW processes, their interaction and reduction of human workload achieved with content management.

For the Innovation process we consider the early stages of idea management. This is the idea life-cycle from its definition to its implementation. This process is supported by the knowledge provided by participants, both internal and external to the company, that contribute with new ideas or content that complement existent ideas. IMSs are essential tools to this process. IMSs are software platforms that enable the collection and management of ideas.

Knowing how the competitors, the market and technologies evolve is essential for companies in order to survive in a competitive environment. TW is a process that enables the capturing of information relevant to a company's business from outside and inside the organization. The information collected refers to competitors, technologies, news, patents, etc. and can be stored and managed using software platforms.

Both processes rely on ICT platforms to store and manage content related to the

company. Therefore, this thesis focuses on the identification of different ICTs that can provide additional management support to both processes and improve them, in areas like interoperability and task automation. Moreover we focus on testing those technologies in real use cases to prove their functionality in real environments.

1.2 Problem and Motivation

Innovation is key for the survival and evolution of the enterprises. As will be shown in Chapter 2 many information and communication tools and technologies are used in the context of Innovation, providing support and improving competitiveness. Among those tools *Idea Management Systems* (IMSs) are essential. IMSs are platforms that help generating ideas and enabling cooperation among different innovation agents. TW platforms are also key for innovation, gathering data about a companies business (competitors, technologies, markets ...) is important not to fall behind. IMS and TW platforms usually work independently and do not exchange information.

SMEs do not always have large teams involved in these processes and they have trouble managing all gathered information. This may lead into companies giving up upon Innovation and TW processes losing important input for the knowledge of the company. The main issues related to data and ICTs in the Innovation and TW processes encountered are:

- **(1) Data-overflow.** One of the biggest problems TW and Innovation processes face is the amount of information they gather. These processes can gather hundreds of ideas, news about technologies, markets or competitors, patents, articles ... It is not easy to manage such large amount of information. Experts need to read and rate large volumes of information. Much of the content may not be relevant to the enterprises (noisy content). Many times content is duplicated or closely related to already processed content. Human management of noisy or repeated content demands resources without adding value to the process.
- **(2) Issues on successful idea contest identification and replication.** SMEs usually launch innovation contests in order to generate new ideas for specific topics. It is not always easy to identify why some contest are successful and others not. Most IMSs are idea centered and do not gather information about the background where the ideas have been created. Therefore, it is important to identify and annotate the factors and events that make a contest successful. That way, those factors and events can be replicated in the future.
- **(3) Lack of interoperability among platforms.** Many different platforms are related to the innovation process. Usually those platforms do not interoperate. Making those platforms interoperable may assist on generating better ideas automatically adding related information to the idea generation process.
- **(4) Waste of time of experts in non-productive tasks.** Gathering as much data as possible is key for innovation, but it is usually complicated to transform that

data on relevant information. Experts expend much time reading and classifying gathered data in order to make it useful. Thus, ICTs could be used in order to assist those experts, automating tasks and reducing their workload, providing them with more time to spend on more productive tasks.

Our motivation of the thesis was to improve the current situation of innovation software solutions according to the problems mentioned above. In particular, we focus on the fact that many improvements can be made on the interoperability and automation of the processes.

1.3 Objectives

The primary objective of this thesis is to deliver ICT based solutions to improve the Innovation and TW processes of SME. Considering the problems and motivation identified in the previous section, the following specific objectives have been marked as the outcome for the thesis:

- **(1) Propose a conceptual model for the identification of successful idea contests and their replication.** In order to help on successful Idea contest replication, our objective is to build a model that gathers context information for the IMS. The model architecture will be implemented in IMS platforms to support on the replication of successful idea contests. These platforms will ease the work of IMS managers providing the knowledge about successful campaigns and enabling their replication. This objective is addressed in chapter 3, where an ontology to gather context information and an IMS platform that accommodates that ontology and other semantic tools are proposed.
- **(2) Identify semantic web methods that provide additional content to IMSs from repositories inside and outside the company.** Semantic repositories inside and outside the companies have information that can be exploited in the Innovation process. Therefore, identifying semantic web methods to make use of that information is one of the objectives of this thesis. This objective is addressed in chapter 4, where several technologies to link ideas with real internal and external repositories to the company are tested for a specific domain (sustainability).
- **(3) Propose a concept model to enable the interoperability among platforms linking content.** One of the objectives of this thesis is to create an architecture that enables interoperability among IMS and TW platforms. Re-using information from other platforms could be essential for the competitiveness and survival of SMEs. If data from different platforms of the company cannot be used in the Innovation process, that information may not be useful. Therefore, enabling data interoperability among platforms may be the best way to exploit this data and transform it into knowledge. This objective is addressed in chapter 5, where a model to enable interoperability among Innovation and TW platforms is presented and added to the platforms used in the thesis.

- **(4) Reduce the workload of the experts in the TW process.** Our objective is to use different semantic technologies in order to automate non-productive classification tasks and reduce human workload in the TW process. Due to the data-overflow, if too much time and effort is needed to process the data, many useful information gets lost among non-useful information. This objective is addressed in chapter 6, building some tools to automatically classify documents. This way the amount of readings performed by human agents are reduced.

1.4 Research questions

In addition to identifying the motivation, problems and objectives of the thesis, we have also formulated some research questions. The research questions specify the main issues that we intend to solve based on the stated motivation, problems and objectives. This way, we could later on validate the contributions of the thesis. The questions defined are stated below:

- **(1) Can a conceptual model help on replicate successful Idea Contests?** Our hypothesis is that conceptual model implemented in a web IMS can help on replicating successful idea contests in real use cases, gathering background or context metrics. Later on, the IMS managers could analyze gathered metrics and determine how their changes impact on the idea contest. This could help on identifying the elements that make the context successful or make it fail. This question is addressed in chapter 3, building an ontology to gather background information for ideas and idea contests.
- **(2) Can semantic methods enable content linking among IMS platforms and semantic repositories internal and external to the companies?** Our hypothesis is that Semantic Web can help Innovation platforms in real scenarios linking ideas from IMS platforms to content inside semantic repositories. Therefore, we aim to test semantic web solutions with real data, to identify specifically how they can aid Innovation processes. This question is addressed in chapter 4, using different methods to use data from real sustainability repositories in an IMS.
- **(3) Can interoperability among content from Innovation and TW platforms be generalized in a single model?** We consider that interoperability is one of the biggest issues for Innovation and TW processes. Therefore, our hypothesis is that we can build a linking architecture on those processes that can be applicable to any text based content. This could ease the re-usability of the data and its transformation into knowledge. This objective is addressed in chapter 5, building an architecture and a new ontology. This way, interoperability among ideas and news items from Innovation and TW platforms has been enabled.
- **(4) Which are the best semantic technologies that help reducing the time spent on non-productive tasks inside the TW process?** Being the human workload one of the largest costs of these processes, reducing the time spent on

non-productive tasks is a very important matter. Therefore, our hypothesis is that we can use existing semantic technologies to reduce document classification tasks on real TW scenarios. This question is addressed in chapter 6, testing different semantic technologies in the same real use case to find the one retrieving the better results.

1.5 Solution architecture

In order to fulfil the objectives and research question of the thesis, we propose different solutions gathered in a single architecture to make it easier to understand. This architecture is a layered solution that presents in a single view the technologies, the tools and content considered in this thesis. The solution architecture can be seen in figure 1.1. Beginning from the bottom layer that architecture is described below:

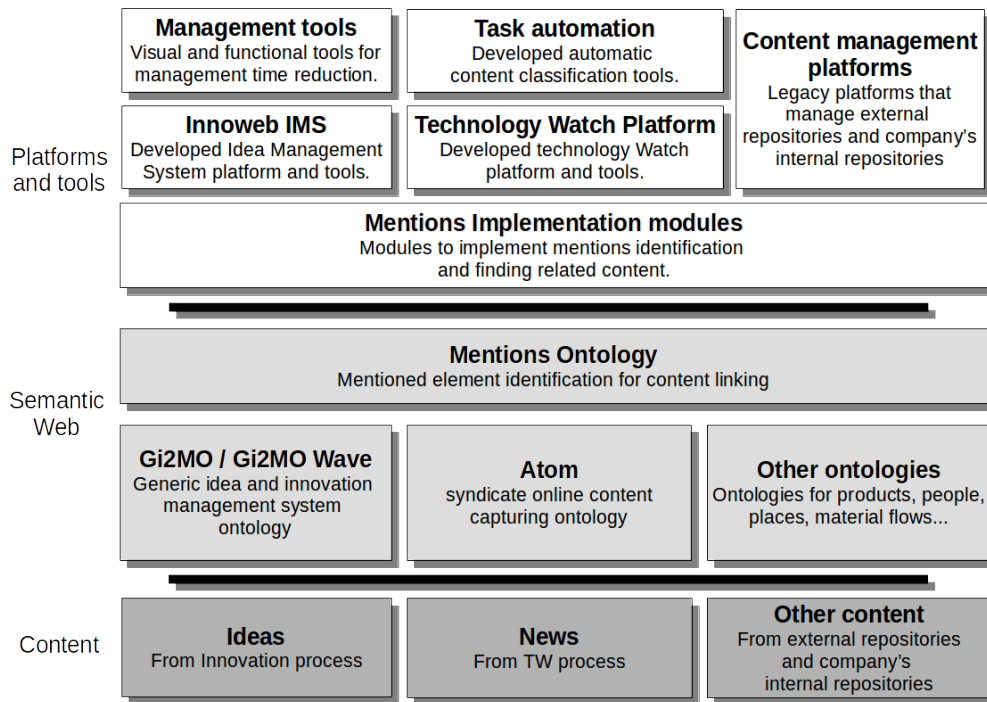


Figure 1.1: Solution architecture for the thesis.

- **(1) Content layer:** This layer represents the content managed in this thesis. First we have the ideas gathered from innovation processes. Secondly we have news items gathered from TW platforms. Finally, we have different content from repositories external and internal to the company, such as people, products, rooms, energy consumption...

- **(2) Semantic Web layer:** This layer contains the semantic representation of the content layer. Each type of content uses an ontology to represent each domain semantically. Moreover, there are two ontologies developed in this thesis: (1) Gi2Mo Wave ontology that represents the background of the idea contests semantically and (2) Mentions Ontology that represents the relationships among content from same and different domains.
- **(3) Platform and tools layer:** This layer represents the platforms that manage the content and its semantic representation (from previously described layers). Furthermore, it represents the modules or tools to implement the "mentions" functionality that links the content of the 3 columns. Finally it also represents the tools that automate tasks and help content managers in reducing non-productive time.

The research on all proposed solutions has been based on the observation of different case studies, most of them real scenarios for companies. **Chapter 2** describes in detail the foundations of this thesis. **Chapter 3** faces the 1st research question, building the left column of the solution architecture (figure 1.1), based on the use of IMSs on real use cases from different companies. **Chapter 4** faces the 2nd research question of the thesis, testing how semantic web methods can be used to enhance Innovation processes with semantic repositories. In order to perform this tests, we used real repositories from DERI. This is represented in the right column of the solution architecture. **Chapter 5** faces the 3rd research question of the thesis, finding a way to link content in order to share data among different platforms. This is represented by the 2 horizontal blocks of the solution architecture (*Mentions ontology* and *Mentions Implementation modules*). **Chapter 6** face the 4th and last research question of the thesis, testing which ICTs are best for human workload reduction on non-productive tasks, more specifically it focusses on classification techniques applied to the TW process. To perform this test, a real multi-class classification case study has been used, based on data provided by Koniker. This is represented by the central column of the solution architecture. **Chapter 7** gathers the contributions made by this thesis to the community, describing the cross university and company collaboration, the projects where we have participated, the scientific publications and the peer reviewed summer schools where we have participated. Finally, **Chapter 8** gathers the main conclusions of the thesis and the possible future work.

Chapter 2

Foundations

This chapter focuses on the study of Innovation and TW processes. The general concepts related to them are outlined first. The first section presents an historical approach to the different perspectives of the innovation concept. The existing types of innovations, the stages that compose the innovation process and the issues to be solved are also presented. The next section focuses on the Technology Watch concept, the stages that compose it and the issues to be addressed.

The chapter continues outlining the technologies selected to address the identified issues of both processes. This section is divided in three parts. First, the Social Web and its benefits for collaborative processes is presented. Next, the Semantic Technologies capable of supporting task automation are described. Finally, the Semantic Web and its ability of interlinking different systems is outlined.

The following section presents the state of the art on the application of the technologies outlined in the previous section on the Innovation and TW processes. The first two subsections gather the application of Social Web and Semantic Web in the Innovation process. Next, Semantic Technologies and Semantic Web applications for TW processes are described.

Finally, taking into account all the previous topics, the conclusions of the study are outlined.

2.1 Innovation

In the twentieth century (1934), Schumpeter [115] made one of the first definitions on innovation. From its traditional definition, innovation encompass the following five cases:

1. Market introduction of a new good.
2. A new method of production.
3. The opening of a new market in a country.
4. The conquest of a new source of supply of semi-finished products or raw materials.

5. The implementation of a new structure in a market.

Half a century later, Padmore, Schuetze and Gibson [98] summarized Schumpeter's definition by saying that innovation is any change in inputs, methods, or outputs that manages to improve the trading position of a company and it's new to the its current market.

Gee [55] and Pavon and Goodman [101], incorporate the concept of *process* to these definitions, where from an idea, invention, or recognition of a need, a useful product, technique or service is developed in order to make it commercially accepted. They even consider innovation the improvement of a product to meet market needs. In the same line, and with reference to technological innovation, Cantisani [23] defines innovation as "the sequence of activities to generate new techniques with the help of science and its method". Amabile [3] incorporates the nuances of creative ideas as a source of organizational innovation, while Porter [105] identifies innovation as "a new way of doing things, which is marketed".

Galanakis [54] defines these issues as the use of "new or existing scientific or technological knowledge... to generate ideas that give rise to innovation (something new) ." Innovation is understood as an idea, process, system, method, service, product, policy, etc. characterized as new or improved and commercially accepted. There is a change from a linear view, where activities take place in a sequential manner, to a new one, where activities overlap and have multiple feedback loops.

Both concepts, *innovation* and *innovation process*, have incorporated these latter aspects, the partition of different actors throughout the process and decision-making aspects, in order to reduce development time of the innovation process, and key aspects such as the incorporation of new technologies and train network to accelerate the process of knowledge capture and transfer of learning in a collaborative environment with other agents that ensures mutual benefit [45].

The first reference to the models of innovation is the one that today is known as the linear model of innovation. People like Godin [57] studied this concept and made a review of its origin and historical evolution. Although used, criticized and improved by several authors, this model has rarely been cited as an original source. Some authors place that source in Bush [22], but others disagree [57].

The mentioned linear model only provides a new product market perspective, but it is not the only one. Although it is still in use, other models have appeared throughout the twentieth century. Rothwell [112] mentioned four generations of models on the evolution of innovation and predicted a fifth where networking becomes important [113]. Below those generations of models on the evolution of innovation are listed:

1. Technology push.
2. Market pull.
3. Coupling models.
4. Integrated Model.

5. System integration and networking Model.

Hobday [64] took Rothwell's approach and presented the five generations of innovation models (adding the *Systems integration and networking model*) also indicating their advantages and weaknesses. More recently, Cantisani [23] analysed the different generations, focusing on the first three, and collecting contributions of various authors like Bush [22] or Stokes [120].

After 30 years of research in relation to innovation models (1977-2006), Errasti, Oyarbide & Zabaleta [47] conclude that most of them agree on a common baseline, and that the main difference is the adaptation of each to a particular case. The common baseline process stages can be found on figure 2.1 and each stage is described below:



Figure 2.1: Innovation process baseline stages.

1. *Idea generation*: this stage gathers new ideas. This stage can also gather comments or ratings of involved agents.
2. *Idea analysis*: this stage is the first filter where the ideas are analysed by experts and some of them are set aside.
3. *Idea enrichment*: ideas that pass the previous stage are enriched by experts so they can make a deeper study and add valuable information into them.
4. *Idea selection*: after enriching the idea, experts analyse them again and rate the ideas according to some criteria. This way, the second filter is performed and only feasible selected ideas are taken into account, becoming projects.
5. *Idea development*: idea developing planning is approached in this stage. Studies are conducted on many issues; market, technology, business plans, risks, possible collaborations, competitors, prototypes...
6. *Idea implementation*: the last stage is concerned with implementing the idea and bringing it to the market.

The economic challenges that arise require new models and concepts around innovation. Among the new models, *Open Innovation* concept is more emphasized, coined by Chesbrough in 2003 [27] and later studied by Christensen et al. [29]; Chesbrough & Crowther [26]; Almirall [2]; Dodgson et al. [42]; Enkel & Gassmann [44]; Fredberg et al. [53] and the European Commission [30]. The Basque Government has also internalized this new concept of open innovation. This is reflected in the *Science and Technology Plan* underway since 2007 [124], [125], [126].

Open innovation is based on the following principles [27]:

- Not all the smart people work for us - we need to leverage external knowledge.
- The R&D outsourcing can generate significant value for us.
- The research does not have to originate from our own work to make it profitable for us.
- A robust business model is more important than being first to market.
- Both internal and external ideas are essential to winning.
- We can capitalize on our intellectual property and we should buy others when we need it.

There are several ways to practice open innovation. Enkel and Gassmann [44] suggest some examples:

- Integration of customers and suppliers.
- Through listening, as innovation clusters.
- Applying innovation across industries.
- Buying intellectual property.
- Investing in the creation of global knowledge.

The first open innovation models have been studied in the Open Source Software (OSS) development industry and later were transferred to more general practices of open innovation. West and Gallager [130] identify three main threats (motivation, integration and innovation exploitation) and define four generic strategies for open innovation:

- Combined R&D - shared R&D (requires a change in culture).
- Spinouts - an escape from the bureaucracies of big business.
- Sale of accessories - accept the commercialization or development of differentiated products based on commodities.
- Accessories donated - general purpose technologies that are sold so that users can develop differentiated products (eg, user folders).

Open innovation is assuming several changes, one is the ability to be able to collaborate with many people. Surowiecki [121] called it *The wisdom of crowds*, assuming that the collective intelligence exceeds that of a few people, both in terms of ideas and knowledge. The use of the power of crowds to increase the capacity for innovation is closely linked to community-based innovation. In addition, there is a general assumption that even though open innovation increases the potential creativity in the innovation process, also increases the complexity involved in managing the process.

Another tendency linked to Open Innovation is Collaborative Innovation. It involves the participation of multiple actors in the ecosystem of organization, ranging from employees to the competitor. Different authors see collaborative innovation from two perspectives:

1. *Firm-centric innovation*: Innovation within the organization.
2. *Network-centric innovation*: Innovation included in an extended organization concept.

Firm-centric innovation focuses on both internal resources as a source of acceleration of innovation processes, while the network-centric innovation, the process extends beyond the boundaries of the organization. In reference to these different levels of openness, there are three sub-models:

1. *Ecosystem innovation*. This concept addresses the classical extension of the value chain of an organization. Usually, some stakeholders have more knowledge in certain areas of the value chain than others. Within this process of opening innovation, both the organization and the different stakeholders benefit from sharing knowledge. Google is one of the best examples of this innovation model [69].
2. *User innovation*. It is a concept introduced and developed by Prof. Eric Von Hippel [127], [128]. User innovation concerns the innovations achieved by end users and producers. Perhaps the most documented case study is the Lego [76]. Lego engineers had been working for seven years in the development of Lego Mindstorms robotic game and only three weeks after its release there were thousands of hackers working on new developments of robots. From there, Lego knew organize various initiatives whose main base line has been the optimization of the product. Currently, there are more than 20,000 Lego fans who organized an online community of innovation.
3. *Crowdsourcing*. This term was first introduced by Jeff Howe [65] and is defined as the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call'. Procter & Gamble and the *Connect and Develop* platform is one of the best examples of crowdsourcing [66].

When talking about innovation processes, it is associated with the research and organization development (R&D) and numerous studies have focused on analysing different aspects of this activity in the organization, the R&D effort, how to organize this activity and the innovation results achieved [45], [10], [25], [85], [60], [38], [94], [122], [6].

In reference to the operational part of the innovation process, innovation requires a flow of ideas, obtained through formal and informal processes [75]. This process is more effective in organizations that combine these two features: first, the control directs initiatives at the different layers of the organization and second, the great commitment

of the participants of the organization with respect to this process [95]. In addition, staff of organizations tend to be highly trained in innovation and knowledge commonly shared with scientists from different disciplines and departments, ensuring a direct connection between business opportunities and organizational capacity and production.

In conclusion, the most critical phase of the innovation process is *the management of ideas*. Once the idea is selected, the next stage is to manage its implementation. Thus the main difficulties are in the early stages of the innovation process: those ranging from the creation of the idea, its analysis, the enrichment of this and finally the selection.

As it will be shown in section 2.3.3, ICTs can be used to create links between the data and generating an interoperable data space where related information can be found and exploited. Igartua, J.I. [68] confirmed in 2010 the relationship between the use of Innovation Management Tools and innovation activity, showing the need of ICT tools in the process. Also in 2010, Westerski et al. [133] wrote about innovation, interoperability and linking. They modelled the data of IMSs using Semantic Web principles as a first step to achieve their goals: “*knowledge management based on interlinking of enterprise systems and web assets to increase information awareness and help innovation assessment*”.

Another problem is the large amount of data. If a company has a large idea flow, they could spend too much time managing. Using semantic technologies (see section 2.3.2) to automate this work could reduce this non-productive time.

However, there are many problems when it comes to face the innovation process. One of the biggest problems is *interoperability*. Heterogeneous systems contain relevant information for innovation, but it is not easy to reuse or share among those independent systems. Most of the time, users have to jump from a system to another and lose time in order to get the required information. Information interoperability among different systems could provide unified scenario where users could consume data from other systems. For example a company could use the information gathered in the TW process (see section 2.2) and offer that data to other systems available in the innovation process enabling the generation of richer ideas. But the interoperability problem does not affect only different information systems, there are problems to link related ideas in the innovation process itself.

As previously mentioned, R&D is closely associated with Innovation process. Therefore, Technology Watch process and its implication within the innovation has been studied. Next section (2.2) presents the state of the art for that process.

2.2 *Technology Watch (TW)*

As reflected in the standard UNE 166006:2011 (*R&D&I management: Technological watch and competitive intelligence system*), TW is an *organized, selective and permanent process*, to capture information from outside and inside the organization about science and technology. All of this in order to select, analyse, disseminate and communicate the information, turning it into knowledge, making decisions with less risk and anticipating change. This way, TW represents a key tool in the *Research, Development and Innovation (R+D+I)* process.

The information that must be watched comes from different domains and formats:

- Patents, utility models, industrial designs (national, European or global). Often the time of the presentation is important, other times, the expiring time.
- Legislation and Regulations that may affect the activity of the company, its customers or suppliers.
- Socio-economic situation in the home or target countries of the company.
- Scientific and technical news on specialist journals, symposia, conferences and other scientific events.
- Doctoral theses and scientific and technical publications from universities, research centres and agencies.
- Sector news (without neglecting other sectors that can have positive or negative interference with the business of the company).
- Information on grants and subsidies.
- Products, prices, quality and sale conditions of competitors.
- Trade Shows: emerging industries, new competitors, distribution strategies, new products, etc.
- Direct personal contacts with competitors, suppliers, research centres, universities, etc.

The whole process of well analysed information capturing that becomes into knowledge for the company and its use within the organization, is a practice known as *Competitive Intelligence* (CI). CI analyses the factors influencing the competitiveness of the company in order to generate competitive strategies and to act successfully in the generation of innovation in the global environment of Business Intelligence.

There is almost universal agreement in relation to Technology watch and CI structure. In its simplest form, it is a process of adding value to information, analysing and producing knowledge in an intelligent way [73]. *Society for Competitive Intelligence Professionals* (SCIP), the most representative association worldwide in CI field, identifies five steps in the process. Those stages can be seen on figure 2.2 and are described below:

- Planning: work with decision making agents to discover and define the TW requirements.
- Information gathering: includes source identification; keyword or taxonomy definition, search, data access and data extraction.
- Data analysis: data interpretation and compilation of recommended actions.
- Dissemination: deliver the findings to decision making responsible.



Figure 2.2: Five stages of the Technology Watch process.

- Feedback or evaluation process: taken into account the response of decision makers and gather the new requirements to continue with the process.

The term “watch” was first applied to technology and was part of the management models of innovation and technology [92], [93]. TW was understood as a function for analysing the innovative behaviour of direct and indirect competitors. This can be done exploring all sources of information (books, grey literature, patent offices, etc..), examining the products on the market (built technology analysis), attending trade shows and conferences to position over competitors and gathering knowledge and technologies that will dominate the future [91].

Rouach [114], leaving aside the technological field, writes about the function of watching, in general, and describes it as the art of discovering, collecting, processing, storing information and relevant signals (weak and strong) that will guide the future and protect us from competitors’ present and future attacks. Technology watch is all about external information transfer to the inside of the company, trying to gather relevant information and send it to the right people at the right time.

Palop and Vicente [99] clarify that companies must make a systematic and organized effort to enable them to observe, capture, analyse and disseminate information from the economic, technological, social or business environment in order to make appropriate decisions with minimal risk.

In 2006, Durán et al. [43] agree with the process for TW, and claim its benefits. They identified the importance of the classification of the information and proposed to use list of controlled terms, classified and grouped according to different points of view. In 2009, Fernandez et al. [50] also proposed to undertake technology watch as a *process*, indicated that it is the starting point of the innovation process and identified the need of establishing key words in order to tag the information gathered.

One of the most important issues about TW is the amount of time spent by experts on some of the stages of the process (see section 7.1.4 and figure 7.2). This happens usually

because of the great amount of data collected in the process. Semantic Technologies can be used in order to automatize as many tasks as possible and making the process faster (see section 2.3.2).

Interoperability is also an issue that opens research opportunities in this process. Much information can be gathered by Technology Watch processes, but if it is not exploitable by other systems (such as the previously mentioned Innovation process), the value of that information decreases. Thus, Semantic Web can be used in order to store that data in an interoperable format and link it with other systems (see section 2.3.3) for its exploitation.

Next section (2.3) describes the selected technologies to solve these issues and the ones that were identified for the Innovation process (see section 2.1).

2.3 Technologies to support Innovation and Technology Watch

Emerging ICT have revolutionized the TW field by offering new options when seeking and gathering information. Internet has been a huge source of information where it is not difficult to get large amount of data but it is hard to obtain relevant information. Collaboration has been defined as key for the Innovation process. The Social Web enables the simultaneous interaction and participation of multiple agents over the same process. Both, TW and innovation processes can benefit from the sharing of related content. They need technologies that link that content and enable interoperability. Thus, this section also analyses Semantic Web Technologies and its interoperability capabilities.

In this section, the three key technologies selected for innovation and technology watch will be presented: Social Web to manage the collective knowledge of people; Semantic Technologies to support task automation for the reduction of the amount of non-productive time; and Semantic Web to manage global knowledge with data extracted from the network and link different type of content.

2.3.1 Social Web

This subsection describes the Social Web, some tools and advantages it can bring for the Innovation and Technology Watch processes. Social Web is the evolution of the first Web, transforming the plain information web into a collaborative web.

The term Web 2.0, that some use to refer to the Social Web, is attributed to Tim O'Reilly [96]. It refers to a second generation in the history of the development of communications in the Web environment. Using the Internet as a platform, the information is always available and it can be collected at any time. In the beginning of the Web, the information was provided in a static way. The second generation focuses on providing tools to produce dynamic web pages where people can add information to the web.

With those dynamic capabilities what prevails in Web 2.0 are the users, the clients. Hence the concept of *Collective Intelligence* or *Wisdom of crowds* [121]. The sum of intelligences is superior to each individual product. Exchanging ideas generates more

ideas; share, contribute, have attitude to create content is the philosophy of Web 2.0. Therefore, Innovation processes can benefit from this collaboration among agents, giving them the chance to work together.

In order to achieve that, Social Web is composed of 5 tool types in order to manage on-line communication. Below, those types are mentioned with some examples:

- Content Publishing Tools (Drupal¹, WordPress², Sharepoint³ ...).
- Tools that manage shared content (Wikipedia⁴, MusicBrainz⁵ ...).
- Tools to share multimedia, such as videos, photos, etc.. (YouTube⁶, Vimeo⁷, Flickr⁸ ...).
- Social networking tools (Facebook⁹, Twitter¹⁰, Google+¹¹ ...).
- Virtual worlds (Second Life¹², PlayStation Home¹³ ...).

Publishing tools are used to provide a common place to gather the knowledge of innovation agents. Those tools can be used as a base to generate platforms to gather ideas, and also comments or ratings about those ideas.

Content and multimedia sharing tools can be used by Technology Watch and Innovation processes. They can have relevant information related to those processes, and enabling the interoperability between those contents, sharing tools can add valuable information to the processes.

Finally, Social networking tools and Virtual Worlds can be used in order to find new agents that can be interested on the Innovation processes, attracting them to participate. This way, Social Web enables Open Innovation, described on section 2.1.

There is also another Social Web term, Enterprise Social Software or Enterprise 2.0, that refers to the application of Social Web applications to enterprise environments. Most demanded functionalities for these applications are similar to those of Facebook or LinkedIn but with more control and governance.

As social networking started to grow in popularity a new breed of Web applications took on the market among enterprises; community platforms. Among the core features, community platforms offer all the functionalities inherited from social Web technologies like blogging, wikis or social networking [100].

¹<https://drupal.org/>

²<https://wordpress.com/>

³<https://products.office.com/es-es/sharepoint/collaboration>

⁴<https://wikipedia.org/>

⁵<https://musicbrainz.org/>

⁶<https://youtube.com>

⁷<https://vimeo.com/>

⁸<https://flickr.com/>

⁹<https://facebook.com/>

¹⁰<https://twitter.com/>

¹¹<https://plus.google.com/>

¹²<https://secondlife.com/>

¹³Closed on March 31, 2015

A issue of the series of McKinsey reports on Web 2.0 adoption shows very positive results on the use of social technologies and a majority of respondents say their companies enjoy measurable business benefits from using Web 2.0 [20]. The use of social webs in the context of enterprise is incipient.

Many companies started to use Social Web for internal communication, and coordination, and the adoption of such tools has grown rapidly. As a consequence a number of commercial tools have emerged to provide Social Web infrastructure, such as: Social-Text¹⁴, Sharepoint¹⁵, CONFLUENCE¹⁶, and Yammer¹⁷.

Apart from the promises of making the internal communication more efficient, and raising the awareness of the ongoing work in the company, the Social Web tools also help constitute a sort of memory of the companies life, and produces other beneficial side effects. For example, some companies have used the Social Web tools to leverage their social capital and find experts for certain projects inside the company. In particular Kolari [77], proposes to use corporate blogs for this task. Social bookmarking has also been suggested as a useful application of the Social Web in the enterprise[89].

Some Enterprise 2.0 tools related to the Innovation process are the *Idea Management Systems* (IMSs). They support innovation by providing the tools necessary to manage ideas created in the innovation process, from the generation of an idea to its selection and launch. Such systems provide support for the idea funnel process¹⁸ “where ideas are systematically filtered and assessed against criteria and only the most valuable ones are implemented and put into practice”¹⁹. Even if idea management systems have been in the market for a while there has been a lately increase in their popularity particularly with the advent of the social Web and the rise of social business software or community platforms. Idea management systems are sometimes seen as subset of such platforms.

Besides, the increasing interest in IMS [31] has pushed community platform vendors to integrate idea management and community software into a single type of platform that includes idea management functionalities and social technologies. Although many barriers have been detected, such as no implication of managers, the need of a cultural change or hierarchical structure has been identified as the most important one. Initial data show very quantifiable benefits and there is no doubt about the upward trend of adoption of these technologies. A high percentage of companies have planned to increase the investment in 2.0 technologies.

In conclusion, Social Web enables gathering information from many agents and even finding new agents to help on Innovation and Technology Watch processes. One of the main problems is the large amount of information available and its management. Some kind of automation is necessary to reduce the burden of treating all that information. Next subsection (2.3.2) describes Semantic Technologies and how they can help in order to solve this type of problems.

¹⁴<http://socialtext.com>

¹⁵<https://products.office.com/es-es/sharepoint/collaboration>

¹⁶<http://www.atlassian.com/software/confluence>

¹⁷<https://www.yammer.com/>

¹⁸<http://www.ifm.eng.cam.ac.uk/dstools/paradigm/innova.html>

¹⁹<http://www.think-differently.org/2007/06/what-is-idea-management-system.html>

2.3.2 Semantic Technologies

Semantic Technologies help to derive meaning from information. Their main goal is understanding large or complex sets of data, without having any previous knowledge about it. Semantic Technologies are in many cases complements or supporting tools for the Semantic Web application.

This subsection describes Semantic Technologies that can help in the Innovation and TW processes. NLP is proposed for machine understanding of human language and AI is adopted for task automation.

Natural Language Processing (NLP)

Natural Language Processing (NLP)'s objective is to enable computers to make sense of human language.

In 2003, Chowdhury [28] described NLP as "an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things". The human language has a structure, called grammar, and understanding that structure is one of the biggest efforts in NLP. This is a difficult task, even humans have difficulties in some contexts. For example, trying to understand a short sentence as *Do you see the man with the telescope?*, could mean two different things:

1. You are using the telescope, and if you can see the man.
2. If you can see a man that is using a telescope.

Depending on the context of the sentence both can be right. So, in order for human language to make sense for computers, it is not enough to understand the grammar but also understanding the whole context.

NLP can be used for different purposes:

- **Entity extraction:** identifies proper nouns and other specific information from plain text. Tries to map terms to concepts. For example, the text "Resource Description Framework" should map to the same concept as RDF.
- **Auto-categorization:** classify different documents by grouping them based on certain criteria.
- **Question Answering:** provide answers to questions applying previously learned knowledge. For example, an IMB tool called *Watson*²⁰ [51] can learn through iterations and deliver evidence based responses. It generates hypotheses, recognizing that there are different probabilities of various outcomes. Watson "learns" tracking feedback, learning from success and failure, to improve future responses.

²⁰<http://www.ibm.com/watson>

Several studies analyse the use of NLP as the supporting technique for the Semantic Web, being *entity extraction* one of the main areas of interest [110], [83]. There are also some works that aim to extract text and map it to an ontology or semantic element [134], [62].

On 2013, Convertino et al. also analyzed some NLP tools in order to extract information from ideas and their comments [32]. They used those NLP tools to identify the “core”

In conclusion, NLP can be used to identify concepts from a text, enabling the identification of the elements appearing on it. That way, *entities* can be identified and linked between them, relating text with those entities. This enables interoperability without explicitly defining those entities. Moreover, NLP can be used to automate the content highlighting tasks, to help users better understand the content generated in different platforms.

Artificial Intelligence (AI)

Artificial Intelligence (AI) or Computational Intelligence “is the study of the design of intelligent agents” [104]. An agent is something that acts in an environment and an intelligent agent is an agent that acts intelligently, doing something appropriate for its circumstances and its goals. It is not necessarily something intelligent, but something that follows a logic. AI can be used for different purposes, below, three examples that can help Innovation and Technology Watch processes on task automation are mentioned:

- **Classification:** different classification algorithms can be trained giving them some examples as training set. It can learn from some of the features of a document and its classification. Then if a new document is analysed, the algorithms can predict its classification. This way, some classification tasks that are made in the TW process can be automatized, reducing the amount of time spent on them. There are several algorithms based on different approaches, that can perform differently depending on the classification context, amount of input features, size of the training set... Baharudin et. al. [9] propose several AI learning algorithms, 3 examples could be *Naive Bayes*, *Support Vector Machine* (SVM) or *Decision Trees*.
- **Clustering:** giving some objects to a clustering algorithm, it finds similarities among the objects and groups them in different *clusters*. In comparison with classification, the clustering algorithms do not need a previously classified dataset, and it can find relationships among the objects automatically.
- **Pattern recognition:** giving some inputs to an *Artificial Neural Network*, different patterns can be found. It is inspired on the brain and how it works. It can be used for clustering, filtering or even classification.

2.3.3 Semantic Web (SW)

This subsection describes the *Semantic Web* (SW), and its ability for creating interoperable data spaces. Next it shows the SW technologies and the standards that can be

useful for the Innovation and Technology Watch processes. Finally, it presents the *Open Data* and *Linked Data* concepts and their principles.

SW (also known as *Web 3.0*, *the Web of Data* or *the Linked Data Web*) supposes the next major evolution in connecting information. SW proposes the principles and standards for data interlinking with the objective of making data understandable by computers. This enables automation of sophisticated tasks. While semantic technologies are algorithms and solutions that bring structure and meaning to information, SW technologies are W3C technology standards that aim to bring semantics to the data in the web, and making easier to link different kind of data.

Berners-Lee's description [13] says that "The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation". As the Web was the way to link document with each other, SW tries to do the same with data. Christian Bizer [16] describes one of the problems that SW tries to solve: "Traditionally, data published on the Web has been made available as raw dumps in formats such as CSV or *eXtensible Markup Language* (XML), or marked up as *HyperText Markup Language* (HTML) tables, sacrificing much of its structure and semantics.". With the Semantic Web not only the documents are linked, giving semantics and structure to the data, they can be interlinked too.

Two are the main technologies for machines to comprehend linked data on the web: *Resource Description Framework* (RDF) and *Web Ontology Language* (OWL). Those technologies combined try to make explicit descriptions of content on the web, whether catalogues, forms, maps or any type of documentary objects. This way, the content of the web is structured and linked as in any database, creating the *Linked Data* (LD). This information semantically structured allows content managers, and therefore machines, to interpret digital documents and perform intelligent search, capture and processing of information. Thus, Semantic Web is all about having linked and well defined data on the Web so they can not just display the data, but also: automate tasks, integrate and reuse data between applications.

The role of the technologies involved on the semantic web can be graphically seen on Figure 2.3. On the lower level we can see *the information* itself. It is a structured or semi-structured *dataset* exposed on the Internet. The second level contains the abstract formalization of data or information. It is built by triplets that formalize relationships between data, which can be represented by directed graphs, and by integrating additional meta-cognition from vocabularies, ontologies, classes, etc. The third level contains the applications or services, which leverage data sets formalized in the previous level to get the desired results. These applications may be based on the *query* of formalized datasets, from the previous level, in the same way that *Structured Query Language* (SQL) queries a database. They can also be based on inference or logical calculations, but the computational complexity of this task may represent a practical limitation.

Most of the languages or technologies involved on the Semantic Web are being standardised and reviewed by the *World Wide Web Consortium* (W3C), an essential task if it aims to be a common mechanism of representation, exhibition and exploitation of the

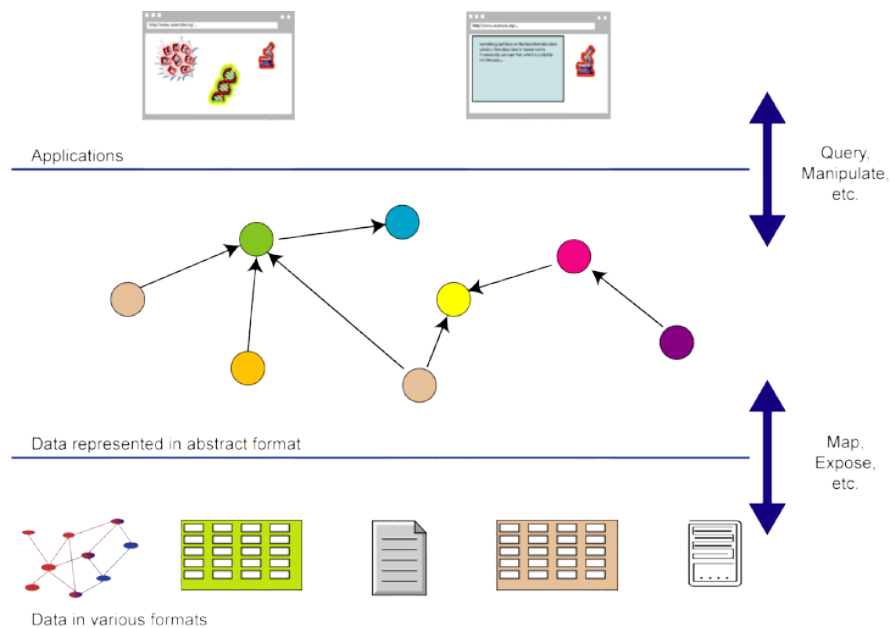


Figure 2.3: The role of the technologies involved on the semantic web (by Herman, I.[63], page 74)

information.

Besides of standardising technologies, the crux of the matter from the second strategic response produced lately in the Semantic Web is *linking the data*. This means that the goal is not merely the annotation or semantic mark-up of the information in isolated repositories, but is the establishment of the greatest possible number of interconnections between different repositories or datasets.

The Linked Open Data (*Linked Open Data* (LOD)) project²¹ is the leading exponent of this response. It has been named as *the seed* that will bring the authentic Web of Data. The *LOD cloud* gives support to some of the most attractive semantic web applications to date, such as DBpedia²² or the BBC's music portal ²³.

Bellow, some Semantic Web technologies and concepts will be explained.

Semantic Web Technologies and Concepts

Semantic Web technologies are a family of standard technologies from the World Wide Web Consortium (W3C) designed to relate and describe data from the Web and the enterprises. On 2000, Berners Lee [11] presented those technologies in different layers (see figure 2.4). The lower levels are related with the most *technical* aspects of the

²¹<http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

²²<http://dbpedia.org>

²³<http://www.bbc.co.uk/music>

data (how to represent an entity identifier, how to describe the data...), while the upper ones describe the *logic* of that data (if a relation can exist, rules to infer information automatically...).

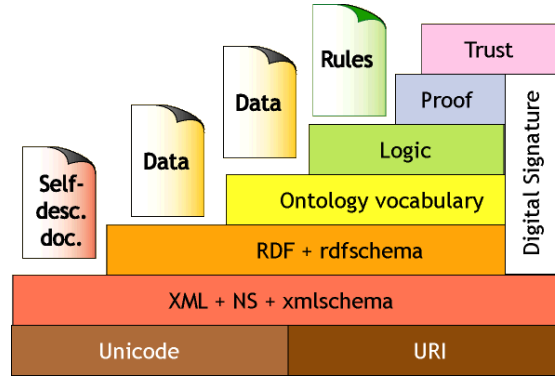


Figure 2.4: Tim Berners Lee's Semantic Web stack.

Next, previously mentioned LD's principles along with Semantic Web basic concepts and some of standards technologies are presented.

- Principles of the Linked Data and its Publication

Since the announcement by Sir Tim Berners-Lee and its adoption as an official project of the W3C, the publication of new data sets has been massive. Thus, the acceptance of Linked Data principles [17] and the publication of new data sets has brought the Web to a global space that allows the connection of different data sources, leading to the first initiatives for search engines and indexers creation in order to exploit these data [12] [61] [97].

The four principles where the Web of Data should be supported are enumerated next. They were proposed on 2006 by Berners-Lee²⁴ and they are also known as Linked Data principles:

1. Use *Uniform Resource Identifiers* (URIs) as names for things.
2. Use *Hypertext Transfer Protocol* (HTTP) URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information using the standards.
4. Including links to other URIs so that people can discover more things.

The process of Linked Data publishing requires the adoption of those basic principles. The use of Semantic Web standards along with those principles make interoperability and reuse of data more efficient. However, this does not mean that current data management

²⁴<http://www.w3.org/DesignIssues/LinkedData.html>

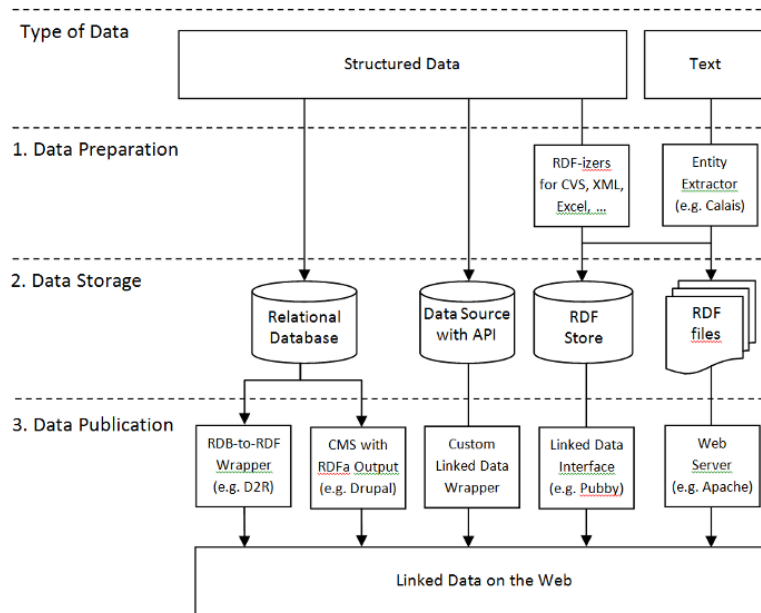


Figure 2.5: Linked Data Publishing Options and Workflows.[15]

systems such as relational databases must disappear. Adding a technology layer to interconnect such systems with the Web of data can be beneficial. There are various mechanisms for accomplishing this and they are summarized in figure 2.5.

Berners-Lee²⁵ presented a classification that measures the level of commitment of a publication with the LD. Those levels are presented next and in figure 2.6:

- ★ : Available on the web (whatever format) but with an open licence, to be Open Data.
- ★★ : Available as machine-readable structured data (e.g. excel instead of image scan of a table).
- ★★★ : As the previous one plus non-proprietary format (e.g. CSV instead of excel).
- ★★★★ : All the above plus, Use open standards from W3C (RDF and *SPARQL Protocol and RDF Query Language* (SPARQL)) to identify things, so that people can point at your stuff.
- ★★★★★ : All the above, plus: Link your data to other people's data to provide context.

For example, Open Data Euskadi²⁶ could get between 3 and 4 stars on this classification. Some contents are in RDF format, but they are not linked between them.

²⁵<http://www.w3.org/DesignIssues/LinkedData.html>

²⁶<http://opendata.euskadi.net/>



Figure 2.6: Linked Open Data publication classification according to Tim Berners-Lee

-Vocabularies and *Simple Knowledge Organization System (SKOS)*

On the Semantic Web, vocabularies define the concepts and relationships (also referred to as *terms*) used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms. In practice, vocabularies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only)²⁷.

*Simple Knowledge Organization System (SKOS)*²⁸ provides a model for expressing the basic structure and content of conceptual schemes that are also called vocabularies: thesauri, classification schemes, taxonomies, hierarchies, lists of keywords, tags, folksonomies ... and any other similar purpose scheme. In this sense, by its profoundly theoretical-conceptual nature, SKOS probably represents a first approach to the descriptive characteristics of the ontologies.

According to W3C in basic SKOS, conceptual resources (concepts) are identified with URIs, labeled with strings in one or more natural languages, documented with various types of note, semantically related to each other in informal hierarchies and association networks, and aggregated into concept schemes.

In advanced SKOS, conceptual resources can be mapped across concept schemes and grouped into labelled or ordered collections. Relationships can be specified between concept labels. Finally, the SKOS vocabulary itself can be extended to suit the needs of particular communities of practice or combined with other modelling vocabularies.

There is an interest in providing SKOS descriptions for all conceptual schemes, not only new ones but also to those already known for decades. The interest lies not only in its own formal description and/or publication, but in the linking that enables the web, and in particular the semantic web, that could help improving the interoperability

²⁷<http://www.w3.org/standards/semanticweb/ontology>

²⁸<http://www.w3.org/TR/skos-primer/>

between different conceptual schemas. This problem is known from the very existence of the classifications, but it sharpens precisely in the digital environment.

-Ontologies, OWL and OWL2

Actually, there is not a very clear difference between *vocabularies* and *ontologies*, but the second term is the one that it is more associated with the idea of the Semantic Web. According to W3C²⁹, the trend is to use the word *ontology* for more complex, and possibly quite formal collection of terms, whereas *vocabulary* is used when such strict formalism is not necessarily used or only in a very loose sense.

OWL³⁰ is the language used to describe the ontologies and is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine *interoperability* of Web content than that supported by XML, RDF, and *Resource Description Framework Schema* (RDFs) by providing additional vocabulary along with a formal semantics.

There is a second version of OWL³¹ and many of its modifications are relatively simple (many of them syntactic helps), but there is one that is crucial: the existence of new 3 profiles³²:

- OWL 2 EL is appropriate for applications where large ontologies are needed and where there are no high level response requirements.
- OWL 2 QL allows using the standard database query technology (SQL), is therefore suitable for relatively simple ontologies but which possibly apply to a very large set of data.
- OWL 2 RL is suitable for operate with relatively simple ontologies, but directly in the RDF triplets.

The purpose of these profiles is to enable better ontology based practical applications, one of the less developed aspects in the Semantic Web by now.

-RDF

Resource Description Framework (RDF)³³ is a language to represent relation between resources on the Web. Conceptually, this relations can be represented using a *graph*, therefore sometimes it is said that RDF is a triplets expressing language. Each of the triplets links two pieces of data using a relation (using a graph terminology, a labelled connection between two nodes).

The specific syntax to write the triplets may vary, and there are mainly two varieties:

- RDF/XML³⁴, that uses XML to represent the RDF relations.

²⁹<https://www.w3.org/standards/semanticweb/ontology>

³⁰<http://www.w3.org/TR/owl-features/>

³¹<http://www.w3.org/TR/owl2-overview/>

³²<http://www.w3.org/TR/owl2-profiles/>

³³<http://www.w3.org/TR/rdf-concepts/>

³⁴<http://www.w3.org/TR/REC-rdf-syntax/>



Figure 2.7: RDF graph example describing Eric Miller.³⁶

- Turtle³⁵, a more human readable language.

An RDF graph example can be seen on figure 2.7. The nodes are both green and orange. The difference is that the first ones are URIs, or internet resources, and the last ones *literal* values. Two nodes are linked with an arrow, and its label is the name of the relation. The generalization of this concepts is made this way:

- The origin node is the subject.
- The destiny node is the object.
- The labelled arrow is the predicate.

So an RDF triplet has this form:

Subject -*Predicate*→ **Object**

Finally, the triplet's three components are URIs, or syntactically valid URIs. The unique exception can be a literal object. It is not mandatory for the URIs to be an existing one in the *real* Web or on the Internet, they are used as identifiers. This is because the machines need a formal artificial syntax to manage the resources, so the URI syntax has been used, without having any capacity loss.

-Microformats and *Resource Description Framework in Attributes* (RDFa)

Microformats and *Resource Description Framework in Attributes* (RDFa) are two tools with a similar purpose to RDF: to label semantically content of the Web. The main difference is that RDF describes the data and its relations in its own language and dataspace, while Microformats and RDFa describe that data inside of the traditional web content languages, such as HTML or *eXtensible HyperText Markup Language* (XHTML). This enables semantic annotation without any additional infrastructure to it.

³⁵<http://www.w3.org/TeamSubmission/turtle/>

³⁶<http://www.w3.org/TR/rdf-primer/#intro>

While Microformats define both the syntax for semantic labelling in XHTML and the vocabulary, RDFa³⁷ only defines the integration syntax. This makes RDFa vocabulary independent, making it possible to use any vocabulary, such as Dublin Core, BibTeX, *Friend of a Friend* (FOAF), etc. Therefore, RDFa allows more freedom of labelling. On the contrary, microformats, based on a well-defined vocabulary, allow applications to use the tagged information more easily. This makes RDFa more flexible and Microformats easier to use.

The RDFa standard also describes the way to process the RDFa+XHTML pages in order to extract the corresponding RDF triplets, making it possible to use Web pages as RDF repositories.

-GRDDL *Gleaning Resource Descriptions from Dialects of Languages* (GRDDL)³⁸ is a mechanism to obtain RDF triplets from documents based on XML and in particular XHTML pages. Authors may explicitly associate documents with transformation algorithms, typically represented in *eXtensible Stylesheet Language Transformations* (XSLT), using a link element in the head of the document.

For example, if a web page has been partially labelled using the *hCalendar* microformat, the existing transformation can be indicated as follows:

```
<link
  rel="transformation"
  href="http://www.w3.org/2002/12/cal/glean-hcal"
/>
```

This way, *glean-hcal.xsl* style-sheet will apply to the web page in order to obtain the corresponding RDF triplets.

-SPARQL

SPARQL Protocol and RDF Query Language (SPARQL)³⁹ is the query language for RDF triplets. It is also a protocol⁴⁰ and it works as a web service: using this protocol, a SPARQL query is converted to a web service request, where a SPARQL processor makes the query and returns the data in XML format.

SPARQL is a standard mechanism for semantic information exploitation. It has become into the unification point for applications' semantic information gathering, as indicated in Figure 2.8.

Thus, by offering a *SPARQL endpoint*, major data repositories can expose their data in a common way, either RDF native or databases with other management systems. This way, *wrappers* such as GRDDL or RDFa can be used as data spaces, and in conjunction with SPARQL, they allow accessing to a big amount of data from the semantic web to any application.

³⁷<http://www.w3.org/TR/xhtml-rdfa-primer/>

³⁸<http://www.w3.org/TR/grddl-primer/>

³⁹<http://www.w3.org/TR/rdf-sparql-query/>

⁴⁰<http://www.w3.org/TR/rdf-sparql-protocol/>

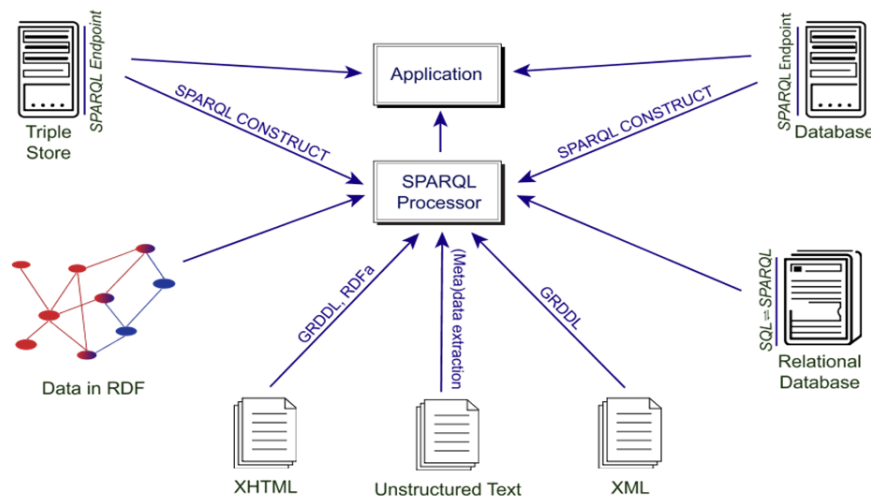


Figure 2.8: SPARQL as unification mechanism (by Herman, I. [63], page 140)

2.4 Applications of the Technologies in Innovation and Technology Watch context

This section presents the state of the art on the application of the previously outlined technologies in the Innovation and Technology Watch context, outlining some existing tools. The study begins analysing the application of Social Web in the Innovation process and continues outlining the application of the Semantic Web in that same context. Next the application of Semantic technologies and Semantic Web in the different stages of the Technology Watch process is presented. Finally, different applications of those technologies in the enterprise are described.

2.4.1 Social Web applications for innovation process

This subsection describes the applications of Social Web technologies in the Innovation process. It shows the current importance of Social Web platforms in the enterprise, some market solutions and efforts made by the community.

In 2010 several architectures of participation were analysed by Errasti et al. [46] including Facebook, Digg, Wikipedia, IBM Idea Factory, IdeaScale Innocentive, SalesFoce Idea Management, Hominex, Mindmeister, LaboraNova, IBM Lotus Connections 2.5, Microsoft Office Sharepoint, Brightidea, Imaginatik Idea Central, Jive SBS, Elgg, Drupal, Liferay, Joomla and Plone. Over 20 factors were considered as the comparison criteria for the different platforms. The most relevant criteria was; context gathering, type of license, ease of use, developing language, operative system, integration with social networks, integration with real time web, semantic web, blogs, wikis, RSS, email, etc. The main conclusions extracted from the analysis were:

- All analysed IMSs are idea centred, gathering little context information in the best cases.
- There are open source platforms with similar characteristics to proprietary software. The inclination to select open source is reinforced.
- Any existent or new innovation support platform must consider the integration of its mechanism with most popular social web platforms, such as Facebook or Twitter in order to be successful by means of participation. Share ideas among those platforms guarantees reaching collaborators and in some cases open the process to new participants.
- Drupal and Liferay obtained the highest score among the tested tools, but the dimension of community users and efforts to integrate semantic web technology currently favours Drupal.

Since the work presented by Errasti et al., the IMS market has presented a whole bunch of solutions. The number keeps changing with new introductions, failures, mergers, etc.

More recently, a list of innovation and idea management software has been compiled by Ron Shulkin⁴¹ and Lauchlan Mackinnon⁴² in their respective blogs with over 40 references. The fact is that the market is so huge every vendor in the space can make enough sales to sustain themselves. The market for IMSs tools has grown into a noteworthy niche solution market. As is normal with any emerging technology market, even if the solutions are powerful enough, lack of user awareness creates a barrier to market growth and provider success.

From the *European Union* (EU) research perspective, there have been a few projects dealing with idea management in the last years:

- Disrupt-IT (IST-FP5-33372) is a project finished in 2004 that aimed at the specification of a dynamic management methodology to foster disruptive innovation in smart organisations. One of the outcomes of the project was an idea generation tool that provided a very basic interface for idea posting and evaluation. This initial platform developed by Atos has evolved into a new web site for idea collection, management and evaluation based on social Web technologies⁴³.
- Laboranova⁴⁴ (IST-FP6-035262) is a project that finished in 2010 aimed at supporting innovators, teams and companies within the development and management of innovative ideas and concepts in the early stages of the innovation process. The Laboranova consortium has developed and combined models and tools in three

⁴¹<http://www.examiner.com/article/a-list-of-every-innovation-collaboration-cms-ideamanagement-tool>

⁴²<http://www.ideamanagementsystems.com/2010/10/44-idea-management-software-solutions.html>

⁴³<http://pgi2-en.atosorigin.es/node/236>

⁴⁴<http://www.laboranova.com>

specific areas: ideation, connection and evaluation of ideas. There are a total number of 13 tools as a result of the project that provide support to different types of innovation and different stages of the innovation process. There are tools that support creativity, ideation, prediction markets or innovation jams. Most of the tools are based on Service-Oriented Architectures and base their data models in existing ontologies but no native support for semantic technologies is provided.

There are more European projects about innovation platforms, but they focus on specific application domains, such as:

- INNOVATION PLATFORM (2010-13): Innovation Management Platform for Aeronautics⁴⁵.
- BIO TIC (2012-2014): the Industrial Biotech Research and Innovation Platforms Centre towards Technological Innovation and solid foundations for a growing industrial Biotech sector in Europe⁴⁶.

2.4.2 Semantic Web applications for innovation process

The Semantic Web is a vision of the future where all information is machine processable and computers may interpret available content and its relationships to help humans in accessing, browsing and searching information.

This subsection describes the applications of Semantic Web technologies in the Innovation process. It will show some ontologies that aim to describe the process and the different efforts made in order to support the process.

In the Innovation field and more specifically in the idea generation stage the research has focused in the modelling and application of meta-data for interlinking on Idea Management Systems (IMS). Two are the relevant innovation ontologies encountered in the literature; the innovation management ontology presented by Christopher Riedl [109] and the GI2MO ontology presented by Adam Westerski [131].

Riedl presents an ontology (Idea Ontology) that applies an approach where the effort concentrates on the integration of idea repositories and little impact is put on interlinking (i.e. relationships and dependencies between concepts).

Westerski [131] cover this aspect by proposing an ontology for IMS where the whole innovation process and its life cycle is taken into account. He presents a formalization of meta-data that can be used to describe ideas and associated information throughout the innovation process. As a result, concepts such as the idea meta-data changes in time and the role of various actors in the Innovation process influence the model in significantly bigger degree than in the case of Idea Ontology. The ontology is proposed as a universal meta-data schema to be applied in any sort of IMS (see figure 2.9).

Aside those ontologies, there are other attempts to construct models for concepts related to different aspects of the innovation process. Bullinger[21] proposes the concept

⁴⁵http://cordis.europa.eu/projects/rcn/97671_en.html

⁴⁶http://cordis.europa.eu/projects/rcn/104298_en.html

⁴⁷<http://www.gi2mo.org/ontology/>

of OntoGate for idea assessment through usage of ontologies that model domain specific knowledge (e.g. product structure, market description, organization strategy etc.). The proposal of Bullinger complements GI2MO Ontology as a tool that can be connected with existing Idea Management System meta-data to provide a new solution for idea assessment.

Stankovic et al. [119] propose an ontology related to innovation modelling that covers serialization of information system meta-data for integration mainly targeting Idea Marketplaces and as a result only focuses on modelling aspects related to challenges and competitions that are central for this group of systems.

Lorenzo et al. [82] propose an ontology for brainstorming systems that covers a large number of concepts related to idea modelling and communities. They focus modelling community collaborative processes and pay less attention to the management, assessment and measurement aspects that are central for Idea Management Systems.

GI2MO ontology has been adopted in this work as the most representative model for the innovation process and consequently it will be the starting point for the research to be carried out. Along with the ontology Adam Westerski has contributed in his thesis work with additional semantic solutions and experiments in the innovation context. The overall objective of the thesis is to obtain more meta-data that would allow better comparison and assessment of ideas. The contribution areas are summarised next:

- **Community Opinion Model for Idea Management Systems** - The objective at this stage is to use opinion mining techniques on community generated content in the form of idea comments to analyse this data and generate additional metrics for idea assessment. The results show that the utilized technology and solutions build on top of the IMS delivering new information for decision makers that have impact on the acceptance or rejection of ideas.
- **Idea Characteristics Model for Idea Management Systems** - In this field the thesis identifies the characteristics specific for ideas in the Idea Management Systems independently of the domain or market segment in which the system is deployed, proposing a new taxonomy. It also proposes a number of experiments where those characteristics are applied both manually and in an automatic manner using semantic technologies (machine learning approach). Finally, the thesis delivers a study of transforming annotations into metrics that identify information stored in the IMS. Although metrics derived from innovation models are equally relevant to identification of winning ideas as any other currently used community metrics (up/down ratings, comment count etc.), the proposed metrics allow to identify community behaviour to verify if submitted ideas follow the initially set goals of organizers.
- **Idea Relationships Model for Idea Management Systems** - The final contribution explores further the relationship types as well as identifies if the new meta-data obtained via previous contributions can be used to facilitate relationship identification (idea dependencies). The contributions in this area are: 1) a classification of the systems related to IMS; 2) a methodology for interlinking those systems by

means of extending GI2MO ontology; 3) hierarchy of relationships between ideas and its use for clustering of ideas.

Among the many lines of future research identified in Westerski's thesis, it is worth mentioning the following because of their relation with the proposal outlined in this document:

- Further research on automatic idea annotation. The thesis experimented with automatic idea annotation using a single method: supervised machine learning approach based on k-NN algorithm with nearest neighbours detected using keyword similarity. Westerski proposes the evaluation of other methods and recommends automatic annotation for the full taxonomy.
- Clustering based on idea characteristics. The thesis delivered a number of studies on idea characteristics and quantitative analysis of those characteristics. The author proposes to test idea similarity considering more than a single characteristic. A potential direction for this kind of research could be the use of clustering algorithms and the treatment of idea characteristics as feature vectors. It remains to be seen if such methods could deliver distinctive clusters of ideas that could have a meaningful impact on analysis of idea datasets.
- Idea annotation with domain ontologies. The thesis studied automatic annotation in the context of domain independent taxonomy in order to deliver a tool for comparison of different IMS deployment. As an extension of this work, the thesis also investigated impact of those annotations on idea similarity. In terms of future work, the author proposes to evaluate the use of domain ontologies in the same way, i.e. automatic annotation of ideas with concepts related to that domain, development of metrics based on those annotations, and computation of idea similarity based on domain related annotations.
- Improved Enterprise Linked Data evaluation. Westerski's work had a major problem for the evaluation of idea relationships. The lack of sufficient data and large enterprise partners that would share their information for the needs of experiments. In terms of future work, he proposes the evaluation of the proposed solution in the environment of a large enterprise.
- Usage of the newly discovered idea relationships. The thesis presented results of clustering based on new relationships. The author points to the use of new relationships in other ways like idea recommendation or ranking. The future work in this area should verify if such a use would aid idea assessment and if ranking generated based on relationships would have an impact on ideas implemented.
- Automatic idea mash-ups. The research has shown that ideas are not only duplicates but are connected to each other in a variety of ways. The thesis experimented with downsizing the idea dataset via clustering but perhaps a viable future line of research could be allowing users to mash-up ideas together from the existing idea

database. The room for novelties is quite broad there and could include research on: idea mash-up operators, idea similarity for automatic mash-up suggestions, metrics for ranking the mashed ideas vs. regular ones, and finally research on incentives and community take-up with relation to reusing ideas of other people.

As it will be shown later the research proposal of this work will be based on this contribution, exploring the relation among IMS and TW platforms or repositories.

It is expected that by applying Semantic Web features into the process, performance will improve. For example avoiding data duplicity, applications will be able to offer or suggest similar ideas. Moreover, semantic data will enable the possibility of grouping similar ideas and make it easier to users to handle all the information. By adding semantic information to data, users will retrieve better solutions and more related information from all the content.

2.4.3 Semantic Technologies and Semantic Web applications for Technology Watch

This subsection describes the applications of Semantic Web and Semantic technologies in the TW process. First, it introduces the functions where ICT can help TW process and a classification of the technologies. Next, the section presents information sources, the formats used to publish that information and the tools employed to exploit data. After that, existing Semantic Technology tools that can help on that data exploitation will be presented. Finally, Semantic Web applications that assist TW will be described.

In relation to the role of the techniques and tools of TW and *Competitive Intelligence* (CI), many of them have found great support in ICT through different software tools, basically, because the treatment of a large amount of information without any computer aid could be very time-consuming. As well as collaborating with software tools in order to monitor both competitive and technological environment, ICT can perform an extraordinary work in the development of CI systems. The following functions can be performed through software tools:

- Identify information sources.
- Capture and organize information according to the needs of the organization.
- Support the work of the Competitive Intelligence Unit by assigning tasks and monitoring the treatment that has been given to a particular document.
- To facilitate the analysis task, distribution of information and the results obtained from the analysis.

ICT tools optimize each stage of the CI process. Thanks to a higher level of automation, systematization and personalization, they enable greater investment in higher value-added tasks from the professional CI, and also to cover more tasks in less time. Moreover, ICT tools include great advantages to users by providing key relevant information, easier to analyze and therefore easier to consume.

These ICTs have been very useful for the TW process. Table 2.1 lists the countries and their most recent and important developed TW tools.

It should be noticed that the existing information sources are many and each one focuses on specific geographical areas. This makes the TW process complex and the need to look something up in numerous sites. Some of this information sources and patent databases that provide the data are listed below:

- OEPM (Spanish Patent and Brand Office) under the Ministry of Industry, Tourism and Trade, offers through its website access to databases of different kinds, inventions and designs in Spanish (Invenes), inventions in Spain and Latin America (Latipat), brands location, international classification of products, services and patents, records condition, issues of law and access to other databases in other languages. Besides, it also offers other related services to technology watch such as creating Patent Technology reports, Technology Watch reports, Retrospective searches and Technology Watch bulletins.
- European Patent Office: it belongs to The European Patent Organisation, an intergovernmental european organization founded in 1977. It allows searching for patents in different languages (German, English and French) on its database (Espacenet⁴⁸). It also offers other services such as searching of publications, alerts about changes in documents, patents reports and RSS service (Open Patent Services). Finally, it allows worldwide access to other data sources.
- United States Patent and Trademark Office: It is under the Department of Commerce of the United States and provides numerous utilities for patent and trademark search. Not only offers tools and services similar to those of the other offices but provides an extensive legal service in relation to patents.

Many countries also offer free access to their patent collections. Below, some of those country databases are listed:

- Japan Patent Office (JPO)⁴⁹. This site also offers access to automatic translations of Japanese patents .
- World Intellectual Property Organization (WIPO) provides the search service PATENTSCOPE[®]⁵⁰, which contains a search engine of international published patents and automatic translations of some documents as a list of worldwide patent databases.
- Korean Intellectual Property Rights Information Service (KIPRIS)⁵¹.

⁴⁸<http://es.espacenet.com/>

⁴⁹<http://www.jpo.go.jp/>

⁵⁰<http://patentscope.wipo.int>

⁵¹<http://eng.kipris.or.kr/>

SOFTWARE	CONTRIBUTION
FRANCE	
Matheo Patent	French Software specialized in the scientometric treatment of information, particularly aspects relating to Intellectual and Industrial Property.
Tetralogie	Complete solution for analysing large volumes of scientific information and patents
Dataview	Software tool for treatment of scientific and technological information experts.
JAPAN	
PAT-LIST	It offers two products for Industrial Property Management: - PAT-LIST-WPI ver. 3.0E - PAT-LIST-CLM (IFI) ver. 3.0E
UNITED STATES	
PatentLab II	Analysis software capable of visualizing the relationships between a wide range of patent data in graphs, tables and reports. Analyses data for the company to make important decisions.
Derwent Analytics	Data mining software and visualization of large volumes of patent information obtained from the databases of Thomson Derwent. The aim is to extract critical approaches to make business decisions. Its use is easy and intuitive.
VantagePoint	It enables to quickly analyse the search results of bibliographic databases and literature R&D. It is specifically designed in order to interpret search results from science and technology databases.
ClearForest Analytics	Analytical system specifically designed for text analysis, which adds value to the existing tools for Business Intelligence.
Aurigin-Aureka	It enables to annotate and organize data, to use advanced analysis tools, data mining and text, or to disseminate its results and provide visualizations.
Anacubis Desktop	New and sophisticated tool of analysis of relationships between entities. It's used to analyse data obtained from a wide range of sources and file types, both textual and quantitative.
SWEDEN	
Bibexcel	Program developed specifically for the handling and transformation of bibliographic records.

Table 2.1: Analysis programs by Country.

- Other International Intellectual Property Offices that offer search engines in databases are: Australia⁵², Canada⁵³, Denmark⁵⁴, Finland⁵⁵, France⁵⁶, Great Britain⁵⁷, Germany⁵⁸, India⁵⁹, Israel⁶⁰, Sweden⁶¹, Norway⁶², Switzerland⁶³ and Taiwan⁶⁴.

Not all the sources give their information in the same format. The formats used in blogs, forums, websites, newspaper articles, etc., usually are those of the web (HTML, XML or similar). For publications, bulletins, patents... besides the above web formats, other ones are often used, especially PDF or Postscript. In many cases this type of information is stored in local databases which do not always allow open access. Many organizations, especially the administration, put general data to everyone's disposal on what is commonly known as Open Data. It is becoming more likely that significant data about technology watch is offered to companies as Open Data. The formats used in this philosophy are varied: DOC, XSL, ODF, PDF, CSV, ZIP or RDF-XML and they can also be offered through *Application Programming Interfaces* (APIs) or Web Services.

There are many ways to exploit these digital sources, using tools to gather all that data. Duran et al. [43] described some tools to obtain information from different sources. Fernandez et al. [50] also mentioned some tools to help in the stages of TW process. Table 2.2 shows those tools and some recently produced applications, that can be useful for TW.

NAME	DESCRIPTION
Monitoring media	
Iconoce ⁶⁵	Scans about 500 media, collecting 1,500 news/day. Allows full text search. Alert Service.
iMente ⁶⁶	Collects news headlines from 13,000 sources in 13 languages, 55,000 news/day. Allows search in title and text. Alert Service.
Continued on next page	

⁵²<http://www.ipaustralia.gov.au/auspat/>

⁵³<http://brevets-patents.ic.gc.ca>

⁵⁴<http://www.dkpto.org/>

⁵⁵<http://patent.prh.fi/>

⁵⁶<http://www.inpi.fr/fr/services-et-prestations/bases-de-donnees-gratuites/>

[base-statut-des-brevets.html](http://www.inpi.fr/fr/services-et-prestations/bases-de-donnees-gratuites/base-statut-des-brevets.html)

⁵⁷<http://www.ipo.gov.uk/>

⁵⁸<http://www.dpma.de/>

⁵⁹<http://ipindiaservices.gov.in/patentsearch/>

⁶⁰<http://www.justice.gov.il/MOJEng/RashamHaptentim/>

⁶¹<http://www.prv.se/en/>

⁶²<http://www.patentstyret.no/en/>

⁶³<http://was.prv.se/spd/search?lang=en>

⁶⁴<http://twpat.tipo.gov.tw/tipotwoc/tipotwekm>

⁶⁵<http://www.iconoce.com>

⁶⁶<http://www.imente.com>

Acceso ⁶⁷	Makes selective gathering from different type of: Press releases. Events or courses organized by a company. Financial Releases. Newswire.
My News ⁶⁸	Full text of more than 100 newspapers, mostly Spanish. Alert Service. Check online archive.
Spanish and international company watch	
Ardan ⁶⁹	Business reports. Analysis on competition.
Informa ⁷⁰	Business data and reports. Balance analysis.
e-Infoma ⁷¹	Business data, business reports and grants.
Dun&Bradstreet ⁷²	Business reports. Credit reports. Incidents notifications.
Dialog Company Profiles ⁷³	Gathering information about companies from other databases.
Webpage watch	
Karnak ⁷⁴	Sends an alert with the number of new results on a topic of interest.
Tracerlock ⁷⁵	Sends an alert with the new results on a topic of interest or the changes of a web page.
Northernlight ⁷⁶	Sends an alert about new results on a topic of interest.
Web monitoring with crawlers agents	
WebsiteWatcher ⁷⁷	Enables: Checking an unlimited number of web sites for changes. Record macros to reach websites whose address is ignored. Use regular expressions to define the filtering mechanisms. Check for updates automatically at a speed of over 100 web addresses per minute.
Continued on next page	

⁶⁷<http://www.acceso.com>

⁶⁸<http://www.mynews.es/>

⁶⁹<http://www.ardan.es>

⁷⁰<http://www.informa.es>

⁷¹<http://www.e-informa.com>

⁷²<http://www.dnb.com>

⁷³<http://login.profiles.dialog.com/dialog>

⁷⁴<http://www.karnak.com>

⁷⁵<http://www.tracerlock.com>

⁷⁶<http://www.northernlight.com>

⁷⁷<http://www.aignes.com>

Copernic AgentPro ⁷⁸	Meta search engine able to filter results with keywords, sort by relevance, etc. You can send email with new search results or changes in a page. It integrates as IExplorer toolbar.
Web browser pluggins (IExplorer.)	
Vivisimo tool bar ⁷⁹	It is a meta engine that analyzes the text of the results, create categories according to most representative terms and groups the results in those categories.
Copernic-Meta ⁸⁰	Integrates the search window on the taskbar. It can be customized in order to include any search engine.
Scirus tool bar ⁸¹	Permite lanzar búsquedas en varias secciones de Scirus o en el buscador Alltheweb. Allows you search in several sections in Scirus or Alltheweb search engines.
Agregators	
Bloglines ⁸²	It is a web based RSS feed agregator. Once registered in the web application, a user can subscribe to different feeds and recive the updates automatically.
Feedly ⁸³	It is a RSS feed agregator, similar to Bloglines. It has been one of the most benefited applications of the Google Reader closure.
Google Reader ⁸⁴	It was a RSS feed agregator similar to Bloglines.
Google Currents	It is a publication agregator. The main difference with the previous ones is that the user does not enter the feeds that they want to subscribe to, they just agregate some interesting topics and they recive related publication. This makes easier for the users to search an specific subject, but they lose control over the information sources.
Flipboard ⁸⁵	It is another publication agregator.
Del.ici.ous ⁸⁶	It is a social bookmark manager. It is not exactly an agregator, but any user can add gather bookmarks and group them using folcsonomies (tags). The system can also thell the users how many people has the same bookmark and suggest tags for it.
Continued on next page	

⁷⁸<http://www.copernic.com>

⁷⁹<http://vivisimo.com/toolbar/toolbar-download.html>

⁸⁰<http://www.copernic.com/en/products/meta>

⁸¹<http://www.scirus.com/srsapp/toolbar/>

⁸²<http://www.bloglines.com>

⁸³<http://www.feedly.com>

⁸⁴Discontinued service since July 2013

⁸⁵<http://flipboard.com/>

⁸⁶<http://delicious.com/>

Data collector tools	
Zotero	It is a Firefox web browser plugin. It creates a database with user's bookmarks and enables exporting that database in different formats (reports, bibliography...).
Connotea ⁸⁷	It was a free online reference management and sharing tool, that could gather bibliography, keywords and tags, and share them.
Bibliography reference management tools	
Endnote	It is a web based bibliography management tool.
JabRef ⁸⁸	It is a bibliography management tool that works locally. Stores the bibliography in a <i>.bib</i> data base and works natively with writing tools like <i>LaTeX</i> .
RefWorks ⁸⁹	It is a web bibliography management tool. Stores many kind of bibliography formats and can export or show the bibliography in many styles.

Table 2.2: Tools to help in the Technology Watch process stages.

The data gathered using these tools is usually extensive and hard to exploit, therefore, Semantic Technologies are used to automatize tasks and classify data. The use of those technologies in the field of knowledge acquisition is extensive. Many are the publications that employ NLP, AI or Semantic Search to withdrawn additional information from databases, web sites, documents or similar. The objective at this stage is to focus the study on the Technology Watch area. The scope of application of the technology is diverse but can concentrate in information search, information identification and extraction, and classification or categorization.

One of the techniques that has attracted much of the scientific interest is the application of NLP with the goal to turn text into data for analysis, document classification or domain identification. This technique is called *Text Mining* (TM) or text data mining and refers to the process of deriving high-quality information from text. Text Mining is now a wide area of research that provides useful techniques that can be used in the context of technology watch.

According to Jacquenet and LARGERON [71], the term appears for the first time in 1995 (Feldman[49]) and was defined by Sebastiani [116] as the set of tasks designed to extract the potentially useful information, by analysis of large quantities of texts and detection of frequent patterns. Losiewicz et al. [84] for example show that clustering techniques, automatic summaries, information extraction can be of great help for business

⁸⁷Discontinued service since March 2013

⁸⁸<http://jabref.sourceforge.net/>

⁸⁹<http://www.refworks.com/>

leaders. Zhu and Porter [135] show how bibliometrics can be used to detect technology opportunities from competitors information found in electronic documents. Another use of text mining techniques for technology watch has been proposed in Lent et al. [80] where the authors tried to find new trends from an IBM patent database using sequential pattern mining algorithms.

In all these works, text mining techniques have been mainly used to help managers dealing with large amount of data in order to find out frequent useful information or discover related work linked to their main concerns.

Bolasco et al. [18] presents a study on the application of Text Mining in different scopes. They claim that there is not a "ready for use" instrument available for users and usable to handle an entire TW process. Instead they identify concrete cases of application. More specifically in the analysis of patents, they identify the use of TM techniques to retrieve textual information in the short description of the patent, to complete the picture offered by the codified information. To achieve that purpose they propose the following TM techniques; text indexing and extraction, concept clustering and graphical display and navigation. They also identified the use in the KM field (document classification) and in Customer Resource Management (CRM) to collect or induce customer opinion (Opinion Mining). The article also describes the necessary steps to correctly implement a TM project. These are presented in table 2.3.

Finally they identify the following issues that need to be addressed when boarding a TM project:

- The intervention of experts at the time of annotation or term list (taxonomies) definition is a must.
- Relevant sources and documents must be identified and metrics to measure the impact set.
- The importance of modelling or structuring information based on a specific domain is crucial.
- Content enrichment is limited by the definition of the domain. The established scope for the domain limits the learning capabilities.

Text Mining techniques have been employed not only to detect frequent useful information. One important goal of technology watch and more generally business intelligence is to detect new, rare, unexpected and hence generally infrequent information. In Jacquenet and Largeron [71], text mining techniques are used to discover unexpected documents in large corpora of documents (patents, scientific papers, data-sheets...).

Apart from Text Mining, other Semantic Technologies have been also employed in fields that can relate to TW. For example in the field of Topic Detection Tracking, Rajaraman and Tan [108] proposed to discover trends from a stream of text documents using neural networks.

Ibekwe-SanJuan [67] propose Semantic Technologies to find relations among text for thesaurus construction and maintenance. In their methodology, they combine NLP

DOCUMENT PRE-PROCESSING

- | | |
|---|---|
| 1 | Definition of rules for extraction/collection of text (data selection and filtering) |
| 2 | Definition and identification of document format. |
| 3 | Text normalisation (cleaning, recognition of dates and of currencies, ...). |
| 4 | Reduction and transformation of text (elimination of stop words, identification of proper names). |

LEXICAL PROCESSING

- | | |
|---|---|
| 5 | Choice of unit analysis: words (tokens or lemmas) and multiword expressions, terms. |
| 6 | Definition of grammatical rules to solve text ambiguity. |
| 7 | Linguistic and lexical analysis (lemmatisation, key words detection, other tagging). |
| 8 | Definition of semantic categories to be searched for in the text (marking of terminology of interest), extraction of key words. |
| 9 | Classification according to concepts and/or other meta-data, information extraction. |

TM PROCESSING

- | | |
|----|--|
| 10 | Classification of texts. |
| 11 | Clusterisation of texts and summarisation. |
| 12 | Knowledge extraction (in some cases integrated with the aid of experts). |
| 13 | Visualisation techniques. |
| 14 | integration of TM results with data mining processes. |

Table 2.3: Necessary steps to correctly implement a Text Mining project.

with a clustering algorithm and an information visualization interface. The resulting system called TermWatch, extracts terms from a text collection, mines semantic relations between them using complementary linguistic approaches and clusters terms using these semantic relations. They point to the possible advantages of using external semantic resources to complement the other two types of resources they use; internal evidence (structure of the terms themselves) and contextual relation markers (conceptually related terms). This can be done via a domain thesaurus, ontology or a general language resource.

The use of the Semantic Web in the Technology Watch field has also been the target for this study. Although we have not found an ontology that represents the Technology Watch process as was the case with the Innovation process the work developed in the field is prolix. Mainly the ontologies encountered deal with the representation or modelling of information extracted during the Technology Watch process but do not address the process itself or the interoperability issues of this process with other enterprises systems or platforms. The classification identified in this area, groups proposals where general ontologies are presented and proposals where specific ontologies are developed to solve Technology Watch challenges.

General ontologies model how the information of publications, patents, feeds or similar content should be structured semantically. With relation to publications and patents, the Korea Institute of Science and Technology Information (KISTI) provides a platform called OntoFrame⁹⁰. OntoFrame is an information service platform that uses Semantic Web technologies. It includes OntoURI a semantic knowledge management tool that creates ontology schema and instances and identifies co-references between ontology individuals; OntoReasoner an inference engine that stores and infers ontology-based RDF triples and answers SPARQL queries; and Mariner, that provides search functionality. The ontology models research entities (e.g., persons and institutions), their accomplishments (e.g., articles), publications which indicate specific journal issues or proceedings, locations and topics.

The ontology schema model and the ontology instance model are subject to license although a non-commercial permit could be acquired.

Another general ontology for the modelling of information received by means of syndication is AtomOWL. AtomOWL is an ontology whose aim is to capture the semantics of rfc4287. RFC4287 is a format to syndicate online content, such as weblogs, podcasts, video-casts, etc. Syndication is a helpful way to alert interested readers about changes in a web site (new content or changed content). *Rich Site Summary* (RSS) is commonly used once sources of information are identified in the Technology Watch process.

As AtomOwl captures the semantics of rfc4287 it is easy to convert rfc4287 feeds to AtomOwl statements (see figure 2.11) and thus add them to a triple database, which can then be queried using a SPARQL endpoint. This should help looking for updated content by making powerful queries. AtomOWL being built on RDF is very easily extensible and it meshes with other Ontologies such as FOAF or SIOC.

Other research points to the importance of developing specific domain relevant ontologies. That is, they propose the use of ontologies that model the Technology Watch

⁹⁰ <http://www.w3.org/2001/sw/sweo/public/UseCases/OntoFrame/>

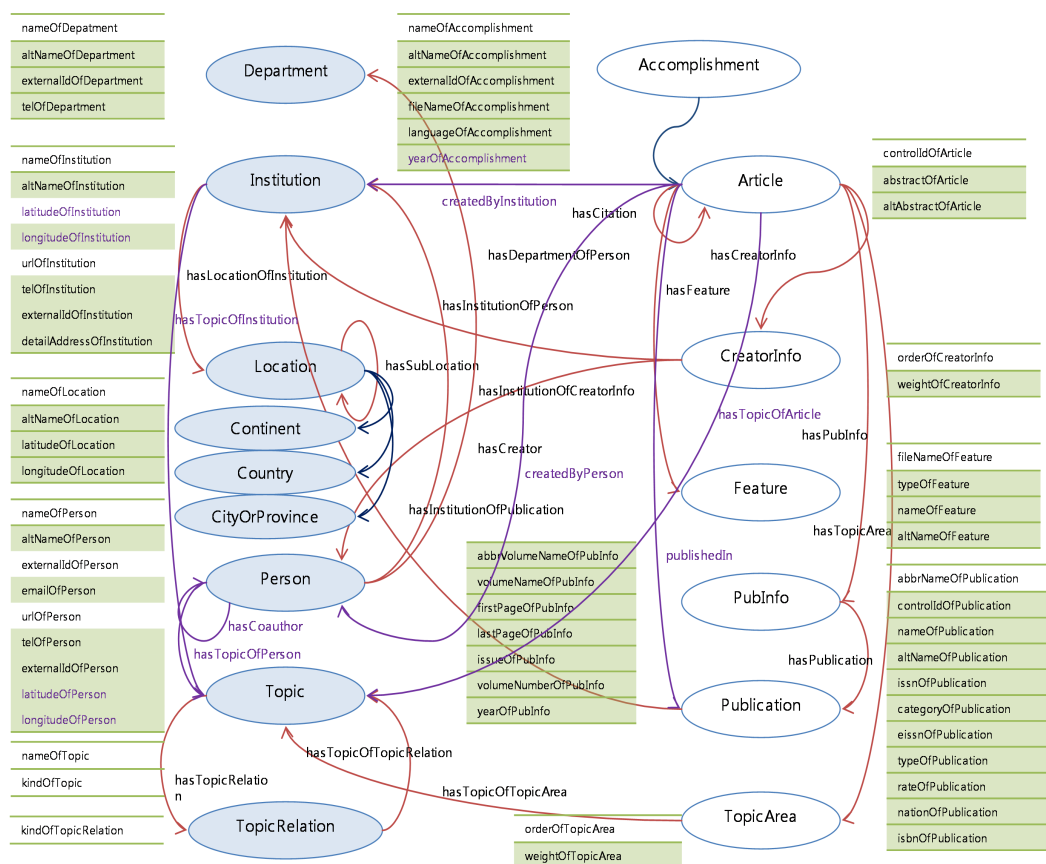


Figure 2.10: KISTI Reference & Academic Ontology for intellectual property and patents.

target domain.

Thus, Maynard et al. [87] present a knowledge Management platform that integrates a variety of technologies to observe resources in the internet. The method for Information Extraction (IE) and annotation proposes the use of specific ontologies (Employment and Chemist Industry domains).

They claim that “one of the important differences between traditional IE and Ontology-Based IE (OBIE) is the use of a formal ontology rather than a flat lexicon or gazetteer structure. The advantage of OBIE over traditional IE is that the output (semantic meta-data about the text) is linked to an ontology, so this enables us to extract much more meaningful information about the text, for example making use of relational information or performing reasoning. Another difference is that OBIE not only finds the (most specific) type of the extracted entity, but it also identifies it, by linking it to its semantic description in the ontology. This allows entities to be traced across documents and their descriptions to be enriched through the IE process". Thus the application of the platform (including the domain ontology) can be used for analysis and enhancing discovered information (advantages of identifying instances from the ontology in the text).

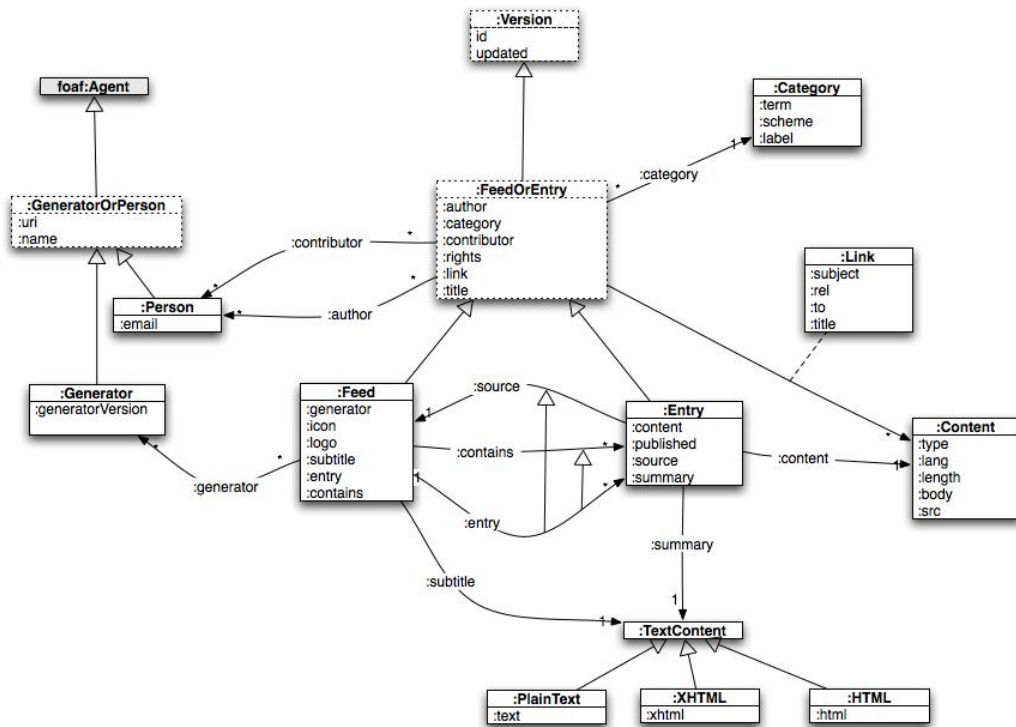


Figure 2.11: AtomOWL ontology for feeds (RSS).

The tools proposed are:

- A tool/model for the development of ontologies, which can be used to describe concepts and trends in the user's domain of interest;
- A tool/model for the development of generic and targeted search agents which can use these ontologies to search for business intelligence from diverse web-based sources;
- A platform for integrating information from various sources and consolidating, analysing and publishing this information.

They also outline the 5 basic stages of their method:

- Web mining application to find relevant documents (or manual input of relevant documents).
- Selection of concepts in which the user is interested;
- Information Extraction (IE) and annotation (identifying instances from the ontology in the text).

- Visual presentation of results (annotation of instances) and statistical analysis.
- Ontology modification (an ontology editor is used to enrich the existing ontology from the results of the analysis).

The main inputs for the platform are the feeds gathered from Internet and the domain ontology. They also identify the need of domain specific lists to support annotation with the ontology. Here there is room for domain experts allowing manually annotation at the beginning to follow with automatic annotation once the taxonomy and the ontology are well defined.

As drawbacks they identify the following:

- Annotation complexity depends on the domain's complexity. The identification is easier in rule-based systems than in learning-based systems where training data is required. Part of the annotation is manual.
- Adaptation of the platform/method to new domains is not easy task for non domain/IE experts.

2.4.4 Technology applications in the enterprise

This section shows how Semantic Web and Semantic Technologies currently help the enterprises in order to integrate data from different sources, briefly describing some existing tools.

In order to address data integration issues in the Enterprise, Semantic Web provides an adapted solution. Since it defines languages and design principles for data interoperability, it can be efficiently used to integrate data from several sources. Among the seminal work in the area, following the long tradition of middle-ware systems, Maedche et al. [86] defines enterprise knowledge management systems using ontologies through OMKS - Ontology-based Knowledge Management System. This focuses on integrating and aligning several internal data sources (databases, directories ...) thanks to a central middle-ware system. Recently, several use-cases and case-studies about Semantic Web technologies in the enterprise gathered by the W3C⁹¹ also showcase the use of Semantic Web for make enterprise data more interoperable, such as biomedical information management at Eli Lilly⁹². A relevant approach in these use-cases is the work done by NASA for expert finding⁹³.

Other approaches for semantic web in the enterprise focus on direct data exchanges between applications in the Enterprise Architecture Integration (EAI) realm, as proposed by Anicic et al. [5], using OWL ontologies and dedicated scripts to align XML inputs and outputs of several applications. However, these approaches generally do not take into account users and their social interactions, as they mainly focus on knowledge extracted

⁹¹<http://www.w3.org/2001/sw/sweo/>

⁹²<http://www.w3.org/2001/sw/sweo/public/UseCases/Lilly/>

⁹³<http://www.w3.org/2001/sw/sweo/public/UseCases/Nasa/>

from static knowledge-bases and do not convey a collaborative aspect, neither exploit fully knowledge from the Web to augment their internal information system.

Recently, similar efforts have also been tackled by various projects, including:

- LOD2⁹⁴, that provide in particular “Supporting both institutions as well as individuals with tutorials and best practices concerning Linked Data publication and consumption” which could be used for enterprises that wish to integrate data using Semantics.
- The Corporate Semantic Web project⁹⁵ that focuses on various facts of improving enterprise work within the enterprise.

In addition to the previous work on integrating legacy data with Semantic Web technologies, some research has extensively done in order to combine Social Web applications and Semantic Web technologies[19] in the enterprise. Applications such as OpenLink DataSpace offers a complete Web 2.0 suite including blogs, wikis and bookmarking systems, all relying on Semantic Web technologies. They are based on ontologies such as SIOC⁹⁶ or MOAT⁹⁷ providing an integrated architecture. Another related system is Talis Engage⁹⁸, a collaborative platform based on different ontologies also including SIOC. In terms of project, relevant work include:

- European Organik project⁹⁹ also aims at extending the original Enterprise 2.0 vision[14].
- The KiWi project¹⁰⁰ also focuses on enterprise social networking using semantics, and provides use-case in the enterprise for semantic publishing and search.
- Not directly focused to the enterprise the IP FP7 project IKS¹⁰¹ aims an extending CMS with semantic Web technologies. By working with Drupal, a well-known CMS, and extending it with Semantic Web features, the IdeaWatch project could liaise with IKS to augment the EU-research on Semantic CMS.
- The IP ECOSPACE project also focused on integration of *Computer-supported cooperative work* (CSCW) components using Semantic Web technologies, and used notably SIOC, that will serve as a base model in our context. That way, we can build upon the results of EcoSpace in IdeaWatch.

⁹⁴<http://lod2-project.eu/>

⁹⁵<http://www.corporate-semantic-web.de>

⁹⁶<http://sioc-project.org>

⁹⁷<http://moat-project.org>

⁹⁸<http://talis.com/engage>

⁹⁹<http://www.organik-project.eu/>

¹⁰⁰<http://kiwi-project.eu>

¹⁰¹<http://www.iks-project.eu/>

2.5 Conclusions

This section summarizes all the conclusions drawn from the state of the art and set the antecedents for this proposal. First a brief description on general issues related to the processes analysed in this work, Innovation and Technology Watch process, is given. Next, the importance of interoperability is addressed. The work continues with the identification of the ICTs applied in each of the stages of both processes. The level of implementation or usage of these technologies in industry solutions today is outlined next. Finally, areas of further research on the application of the Social Web, Semantic technologies and Semantic Web are outlined.

Innovation is essential for business survival. For innovation to be successful, collaboration among different agents and systems is essential. Among those systems a well planned Technology Watch process is vital. Sometimes TW acts as the trigger of an innovation process. That is, as the result of a TW process new ideas arise and the initial phase of the innovation process starts. In other situations, idea validity needs to be assured by contrasting technology, market viability, possible anticipation of the competence (patents), risks and other Technology Watch issues. This synergy between both processes occurs more than once during the different innovation process stages. TW makes possible anticipating, reducing risks, progressing or cooperating. Both Innovation and Technology Watch processes are iterative.

Sharing data between processes and within the stages of a process is another important issue. The actors for the different stages rely on the quality of content (ideas, publications, market data, ...) to make the appropriate decisions. They need internal data from the own enterprise but also external information. Interoperability among systems, repositories and process becomes essential to assure sufficient information and the quality of it. Without interoperability data is isolated and underused. Interoperability reduces the amount of time spend on swapping from a system to another providing additional relevant information to the user.

In the state of the art it has been proven the importance of using ICT in the Innovation and the Technology Watch processes. Innovation process stages (Idea generation, analysis, enrichment, selection, development and implementation) and Technology Watch process stages (planning, information gathering, data analysis, dissemination and feedback or evaluation process) have also been defined.

On the first stage of innovation process, *idea generation*, social collaboration platforms have been proven advantageous to achieve greater participation or collaboration. Sharing ideas among Social platforms guarantees reaching more collaborators. In some cases it opens the process to new participants which translates into greater amount of ideas. Interoperability can be achieved using Semantic Web and Linked Data by relating ideas with the data gathered in Technology Watch or any other information system or repository. This could support users work by adding automatically relevant data to ideas.

On the *Idea Analysis* and *Idea Selection* stages of the innovation process, Semantic Technologies could help experts in filtering ideas. NLP and classification tools can assist them finding similarities between ideas or reducing the amount of time spent analysing

them. Social tools, such as comments or ratings, can help giving clues to experts when choosing which of them will pass the filter.

On the *Idea Enrichment* stage, information from other systems can be used to automatically add relevant data to ideas. Therefore, Semantic Web technologies can be used for interoperability with other information systems.

During the last two Innovation process phases, *Idea Implementation* and *Idea Development*, the interoperability provided by Semantic Web could maintain ideas linked with data collected from the Technology Watch process or available in other repositories (internal or external).

For example it could be linked with patent data, in order to see if during the implementation a new patent related to the idea appears. It can also be related to other ideas in the early stages and see if additional content can be attached to the idea in the implementation stage.

The same technologies can be applied to the TW process. Semantic Web standards will assure interoperability by linking data with content from other systems, platforms or repositories, assisting TW process on its information gathering stage.

On the other hand, semantic technologies can be useful for the *data analysis* stage and the *dissemination* stage too, classifying data and helping experts to group information and give it to the correct people. Moreover, when companies have a great amount of data flow, there is a risk of info-intoxication. The information gathered through TW is very diverse and is extracted from multiple sources. Therefore, there is a need for filtering, classifying and distributing data without saturating the experts, optimizing productive time.

Having described the current state of the technology and its technical application in the Innovation and Technology Watch fields, now is time to describe which is the level of implementation or usage of these technologies in industry today. From the study we extract the following conclusions:

- There is a growing market for tools and platforms to support Innovation and TW.
- Most of the solutions consist of proprietary software, although there are open source platforms with similar characteristics.
- The level of implementation and use of these tools is uneven among companies and depends on management implication.
- Innovation and TW tools and platforms are mainly implemented in large companies. Implementation of these tools in medium and small enterprises is incipient.
- Most of the time those solutions are scattered and concentrate on specific issues of the processes they do not have a holistic vision of the problem.
- TW tools in particular concentrate on specific sources of information and provide solutions for only a specific aspect. Thus, the number of existent different tools is large and there is room for platforms that integrate those solutions.

- All analysed IMSs are idea centred gathering little context information about the process itself.
- Most IMSs understand the importance of collaboration among different agents in order to be successful. Nevertheless, not many of them share ideas with other platforms and in some cases they not even open the process to new participants. Any existent or new innovation support platform must consider the integration of its mechanism with other popular social web platforms.
- There is not a consensus among developers on providing standards that will support integration between platforms. This means that most systems are not interoperable with other systems or solutions.
- Although most of the references in the study understand the relation between the Innovation and the TW processes, there is not a semantic model, tool or platform that supports that integrated vision of both processes.
- In the last couple of years there is a trend to apply IMS solutions to specific domains. That is build solutions to gather ideas oriented to very particular areas (i.e. aeronautics or biotechnology).
- Neither innovation platforms or TW commercial tools employ Semantic Web annotation of content to encourage interoperability.

The state of the art on recent research and previous work from our research team provide answers to some of these issues although some of them remain unanswered and new questions arise. As a result of this research, issues such as participation and collaboration among agents have been addressed by means of constructing a platform based in Social Web technologies. Details of this platform can be found in section 3.3.4. The selection of open source software to build the testing platforms opened the room for the integration of the Semantic Web to boost interoperability. As has been shown throughout this chapter there are models to semantically represent the innovation process and provide interoperability. There is not a need for creating a new model. GI2MO ontology has been adopted in this work as the most representative model for the innovation process and consequently it will be the starting point for the research to be carried out. Along with the ontology Adam Westerski has contributed in his thesis work with additional semantic solutions and experiments in the innovation context but also has identified room for further experimentation. Among the lines opened for investigation it is worth mentioning the following:

- Usage of the newly discovered idea relationships for idea assessment. The integration with other systems or platforms and the relations obtained as a result could be an area of research.
- Idea annotation with domain ontologies. That is analyze the application of specific domain ontologies over ideas. Ideas will be linked to additional content and consequently richer ideas will be available.

- Further research on automatic idea annotation by means of applying semantic technologies (NLP or AI).
- Improved Enterprise Linked Data evaluation by testing in real industrial scenarios.

Taking into account previously mentioned drawback identified by Maynard et al., (1) *annotation complexity depending on the domain's complexity* and (2) *the need of experts for the construction of specific domain ontologies and taxonomies*, the first one needs to be addressed by means of semantic technologies such as NLP or AI. The second one must be addressed by enterprise experts in association with semantic web specialists. Both cases open the gate to further research.

As a summary it can be stated that there is not a study that analyses the application of the social and semantic technologies in the integration of the Innovation process and Technology Watch process with a holistic or systemic view. Although there are models that represent the innovation process and specific areas of the TW process, there is not a model that integrates both. The application of the Semantic Web in the Innovation field is incipient and there is still room for further research and improvement in the areas of idea enrichment, idea relations with other ideas and automatic annotation. Semantically linking specific content for a particular domain obtained during the TW process to ideas has not been research yet. Gather context information on the implementation of both processes is important to evaluate successful campaigns. The ontologies identified in both processes gather only little information about the context in which processes occur. Finally, it can be concluded that not only ICTs are needed in order to achieve successful innovation, management implication is much more important. The systematic application of an innovative culture in the company is one of the key factors for success. Proper application of the tools will only be valid if the company establishes the conditions (resources, methodologies...) necessary to encourage that innovation culture.

Chapter 3

Web Based Platform for Innovation processes

This chapter focuses on a the development of a Web Based Platform as a base application for testing the research proposals presented in this thesis. The name of the new developed platform is Innoweb. Innoweb is an IMS that enables interoperability by providing semantic technologies according to the architecture proposed in chapter 1. The platform addresses the 1st objective of the thesis defined in chapter 1: *Propose a conceptual model for the identification of successful idea contests and their replication*. Innoweb also provides the semantic tools and technologies necessary to evaluate the research presented in chapter 4 and chapter 5.

The first section makes a brief introduction to the research. Section 3.2 shows the research context introducing the innovation framework for the project and the state of the art on the technologies selected to create the platform. In section 3.3, the research approach followed by the team is shown. Next, section 3.4 exposes the extracted findings. Finally, section 3.5 presents the final conclusions obtained in this research and the future work withdrawn from the project..

Main contributions

- IMS platform to enable the identification and replication of successful idea contests.
- Gi2Mo Wave ontology to annotate semantically the background of idea contests.

3.1 Introduction

Innovation is extremely important for the growth strategy of most enterprises [24]. With the rise of emerging economies, business is entering a new era of extreme competition where the only way to survive is to innovate. Many companies and especially SMEs have problems applying innovation processes, due to the lack of resources, appropriated tools or innovation culture. Without innovation those enterprises are not able to grow

and their competitors take advantage of that weakness. Innovation allows enterprises to compete and evolve efficiently.

Many tools have been developed to support innovation processes. *Idea Management Systems* (IMSs) are employed in the early stages of the process, where ideas are generated. IMSs are idea centred. Thus they collect information about the ideas gathered in the platform. Existent IMSs hardly collect information about the context. That is, the conditions on which those ideas have been gathered are lost. Information such as the type of contest where the idea was conceived, the actions taken during the different stages, the idea contributors or the timing between stages are often forgotten since there is not a platform that gathers that information. Collecting that data is essential to reproduce the conditions and the context for successful ideas. Thus, in most IMS when an idea becomes successful there is no way to identify the context where that idea was conceived.

The work presented in this chapter describes the development of an open source community platform for the front end of the innovation process focused on the innovation context. The platform gathers and stores information about environment conditions for different types of innovation processes. The importance of collecting those conditions lies on the possibility of repeating those contexts where successful ideas have been created. Hypothetically, the re-creation of those conditions will turn into new ideas with higher probability of becoming successful.

In order to enable this context information and interoperability, a new ontology called Gi2Mo Wave has been also developed in this chapter, in collaboration with UPM. This ontology has been used inside the platform for idea context semantic representation.

The left column of the architecture proposed in chapter 1 (figure 1.1) is focused on the content related with ideas. Thus, all the developments of this chapter are concerned with the creation of tools that build the left column of that architecture.

3.2 Theory

In a thorough study on the innovation process [46], different innovation models, existent tools and technologies were applied to the different innovation process stages. The innovation process identified in that study is outlined next. According to the study, the requirements guide to the implementation of a collaborative platform to manage innovation. Therefore, section 3.2.2 presents a state of the art on the technologies selected for the development of that platform.

3.2.1 Innovation Process

Most innovation models show a similar baseline and the differences among them lay in the particularities incorporated to the model in each particular case [45]. This baseline is understood as a process with several stages. Four stages form the first part or front end of the innovation process (see figure 3.1).

1. **Idea Generation:** creation and collection of new ideas and comments.



Figure 3.1: Early stages of the innovation process

2. **Idea Analysis:** the study of created ideas and the search of relations among them, in order to merge, split or complement with other ideas.
3. **Idea Enrichment:** experts add more valuable information into chosen ideas.
4. **Idea Selection:** select the best ideas for their development into projects, using custom criteria and weights.

The front end of the innovation process where idea management is developed is one of the most critical stages. Thus, the main issue is to manage the innovation process and more specifically the front end of the Innovation process, providing an efficient platform for the Innovation process. The following requirements have been identified:

- A common space to represent and gather all the information related to innovation processes is needed. This common platform will be used to collect ideas, identify experts, introduce comments and follow idea progress or search for similar ideas.
- Idea context gathering is essential in order to reproduce successful idea contests.
- The advantage of using social networks in organizations is clear; employees could develop contacts, share knowledge, improve communication between experts, gather interest in new projects or ideas, enrich ideas using incremental collaborative contributions and identify professional opportunities.
- The platform has to be flexible enough to accommodate different types of innovation campaign or waves simultaneously; firm-centric innovation process or crowd sourcing contest with hundreds of participants.

3.2.2 Technology

This section presents a state of the art where technologies and tools used to enhance the innovation process are studied. A description of those technologies are presented, classified in three groups: *social software and innovation platforms*, *semantic web* and *real time web*.

Social software and innovation platforms

Social software is the term used to refer to the different applications and technologies associated to the Web 2.0 term, widely introduced and depicted in a seminal article

by Tim O'Reilly [96]. James Surowiecki demonstrated [121] that complex tasks can be solved more effectively by group collaboration than by any individual of the group.

On the other hand, the term enterprise social software (Enterprise 2.0) refers to the application of Social Web applications to enterprise environments. Most demanded functionalities for these applications are similar to those of Facebook or LinkedIn but with more control and governance. Besides, the increasing interest in IMS [31] has pushed community platform vendors to integrate idea management and community software into a single type of platform that includes idea management functionalities and social technologies.

As social networking started to grow in popularity a new breed of Web applications took on the market among enterprises; community platforms. Among the core features, community platforms offer all the functionalities inherited from social Web technologies like blogging, wikis or social networking [100].

An issue of the series of McKinsey reports on Web 2.0 adoption shows very positive results on the use of social technologies and a majority of respondents say their companies enjoy measurable business benefits from using Web 2.0 [20]. The use of social webs in the context of enterprise is very incipient. Many organizational barriers are detected, as no implication of managers, need of a cultural change or hierarchical structure. However, initial data show very quantifiable benefits and there is no doubt about the upward trend of adoption of these technologies. A high percentage of companies have planned to increase the investment in 2.0 technologies.

As seen in section 2.4.1, Errasti et al. [46] analyzed several architectures of participation. The main conclusions extracted from the analysis are:

- All analysed IMSs are idea centred, gathering little context information in the best cases.
- There are open source platforms with similar characteristics to proprietary software. The inclination to select open source is reinforced.
- Any existent or new innovation support platform must consider the integration of its mechanism with most popular social web platforms, such as Facebook or Twitter in order to be successful by means of participation. Share ideas among those platforms guarantees reaching collaborators and in some cases open the process to new participants.
- Drupal and Liferay obtained the highest score among studied platforms, but the dimension of community users and efforts to integrate semantic web technology currently favours Drupal.

Semantic Web

Semantic Web consists on transforming plain text found in the Internet's content into another one with sense and meaning. It is defined as the web of data that can be processed directly and indirectly by machines. The objective is to build a context around

information by adding categories, metadata and relations between things that add sense to that data. This way, the data is more understandable for the machines, enhancing the interoperability among different content.

In order to add semantic meaning to data, ontologies are used. Two are the relevant innovation ontologies encountered in the literature; the innovation management ontology presented by Christopher Riedl on 2009 [109] and the Gi2MO ontology presented by Adam Westerski [131]. The developers of Gi2MO ontology have also created a RDF metadata publishing module for Drupal called RDFme.

Other works have also involved semantic web at IMS context. Anadiotis et al. [4] (2012) analyzed how semantic web can be used to facilitate deliberation and collective decision making on IMS. Podeda, et al. [107] analyzed in 2012 how semantic web can be applied in order to enhance searches in IMS.

Real Time Web

Real-Time Web (RTW) consist on notifying events occurred in the web without the need of the active interaction of the users. Users should receive a notification when something happens in order to react as soon as possible. For example, the users can be notified when one of their ideas has been selected, using e-mail, twitter or a similar service. This makes the users aware of the events even if they are not checking the web actively.

One of the most important technologies related to the real time web is *Asynchronous JavaScript and XML* (AJAX) ¹. AJAX allows modifications to the content of a page or sending information without reloading a new version of that page. The most famous example of RTW is Twitter², that allows to follow users and see what they share in real time.

3.3 Material and Methods

The state of the art shows that although many idea management tools are available, community software platforms found are idea centred. They do not register the context where ideas are gathered. Thus, a platform that collects ideas and registers the context where they are created is needed. Additionally the platform must enhance integrability [109] using semantics.

In this section, firstly, the main objectives of the research can be read. Secondly, the metrics and data to be collected by the platform are described. Next, the methodology followed in the development of the platform is detailed. Finally, the platform itself is presented with all the functionalities.

3.3.1 Objectives

The main objective is to develop a baseline social networking platform to support the front end of the innovation process, gather context information and enhance integrability

¹<https://www.w3.org/standards/webdesign/script>

²<http://twitter.com>

of data.

The technological objectives of the research addressed in this chapter are the following:

- Deploy the baseline platform on Drupal representing each of the stages for the innovation process and gathering all the context related data (ideas, contests, participants and their skills, outcome, companies, events, etc.). The platform will be made flexible enough to accommodate and adapt to different scenarios.
- Provide a set of tools for each of the four stages identified.
- Be able to articulate a successful architecture of participation around the platform using the possibilities brought by social web and real time web technologies.
- Prepare information collected with semantic meaning enhancing integrability.
- Design a theoretical model for Idea contests and formalize it in an ontology.

The methodological objectives are the following:

- Apply the deployed platform launching idea generation campaigns within a set of companies.
- Gradually establish a cooperative culture in all aspects of innovation through the baseline platform.
- Measure those case studies with previously defined metrics. This will enable a better understanding of the issues related with the innovation process.

3.3.2 Metrics

In order to identify innovation success factors, metrics have to be defined. Environment conditions or context information is an important issue that needs to be addressed. Data and metrics about the innovation process, innovation campaigns, the activity and outcome has been identified and classified according to the following criteria. This way, metrics can be analyzed to see what factors have the greatest impact on the success of the contests.

General Wave Characteristics

Enterprises usually launch innovation contests (waves) in their innovation processes. Wave information shows the framework environment on which ideas are generated.

- Innovation type: stores the level of innovation of the wave (radical, incremental...).
- Stages: the stages the wave will follow on the Idea Management life cycle.
- Status: describes the current stage.

- Target: indicates the objective or target aimed by a wave.
- Topic: how the wave has been classified.
- Contest type: indicates the type of innovation searched by the wave.
- Fields of the idea: the fields will the user have to fill in order to submit an idea.
- Duration: the time the wave will last.
- Situation: environment in which the wave is created (relaxed, time or condition pressure...).
- Selection criteria: indicates the criteria used for idea ratings (set by experts).

Structural Metrics

These metrics measure the structural properties of the contest and their impact.

- People: groups, dedication, number of participants, active users, roles...
- Enterprise: time and resources assigned to R+D+I by companies.
- Resources: locations (e.g. meeting rooms), amount of resources spent on awards and prizes...

Activity Metrics

These metrics measure the activity in the platform.

- Traffic measures: number of views, unique visitants, average time spent, repeat visitors...

Stimulation Metrics

These metrics register actions or stimuli provided in the wave to boost participation and improve quality of ideas.

- Events: the number of events, participants, type, duration, location...
- Awards: the number of awards and the amount of money earned in each...

Outcome Metrics

The outcome metrics measure the result of the contests.

- Wave level: number of ideas that fit the target, become a product, create a spin off...
- Idea level: innovative ideas, innovation level, number of innovations introduced, generated sales...

3.3.3 Methodology

The agreed methodology approach followed for the creation of the web platform is based on an incremental development cycle, where requirements guide implementation (see figure 3.2). The experience collected in a cycle will help to improve the next one. The results and conclusions obtained with the developed prototypes can generate new requirements.

Next the phases that summarize each cycle in the development are presented.

1. Requirements: a field study is performed through a set of interviews to different representatives of the involved organizations in order to assess the use of Social Web technologies. This input, together with the state of the practice research, is used to depict the case studies.
2. Implementation: the necessary prototypes and the methodology are developed. The methodology will provide a stepwise approach for the adoption of the prototypes within an organization.
3. Validation: a set of piloting activities are carried out within real production scenarios and are based on the depicted case studies. A set of indicators are set up in order to evaluate the result and the success of the contest.

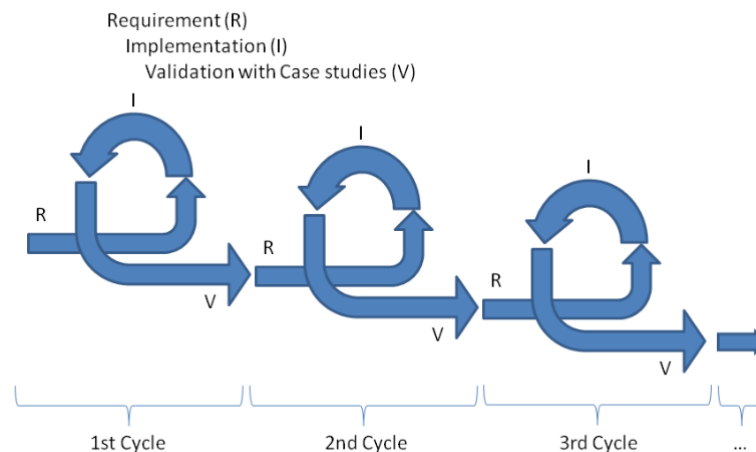


Figure 3.2: Incremental Development Cycle methodology

3.3.4 Platform

The innovation platform presented in this chapter has been developed in Drupal. Drupal is a powerful open source *Content Management System* (CMS). One of Drupal's main assets is its flexibility and modularity. Drupal is like a Lego kit. Skilled developers have already made the blocks or modules that Drupal users need to create their sites; news

site, an online store, a social network, blog, wiki, or something else (in our case, an innovation platform and it's early stages, see section 3.2.1).

Drupal's core includes basic community features like blogging, forums, and contact forms, and can be easily extended by downloading other contributed modules and themes. Drupal also provides a set of *Application Programming Interfaces* (APIs) that bring the possibility of creating new functionalities programmatically and has a very active community that develops and offers a wide variety of modules.

Innoweb platform not only adds some modules to Drupal in order to transform this CMS into a IMS but the addition of Wave module also allows administrators the personalization of idea campaigns and manage the workflow of the innovation process. Other modules with specific functionalities were also constructed. Next a description of the technological solutions, tools and modules employed in each of the stages of the innovation process is presented. Finally the ontology and tools used to store data in semantic format is outlined.

Idea Generation

For the first stage of the process a module that collects ideas has been developed. Ideas are collected in blog format (see figure 3.3) and their content stored in database (MySQL). The module also provides idea visualization in a list (see figure 3.4). That way, it enables the first stage of the innovation process, the *Idea Generation*.

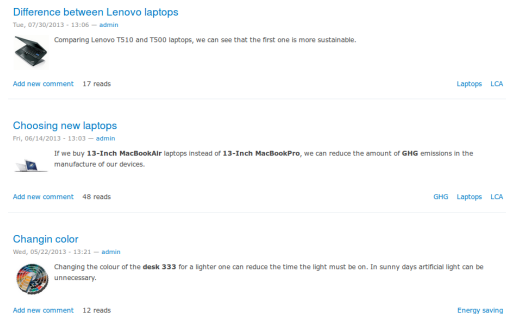
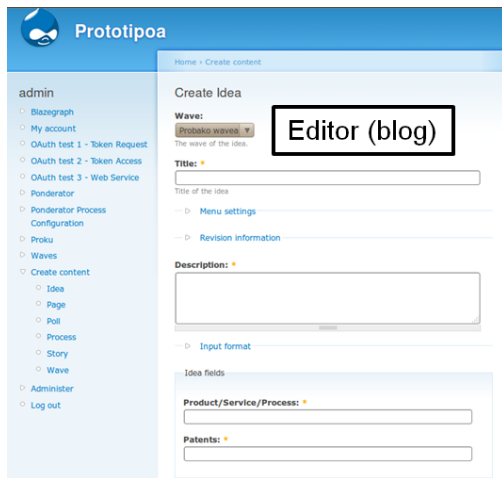


Figure 3.4: Idea list.

Figure 3.3: Idea generation in blog format.

Ideas can be commented by other users contributing or enhancing idea quality at this early stage. Users can also vote upon ideas and comments using votingapi and voteupdown modules provided by Drupal community. Comments are stored in databases along with related ideas. Innoweb also offers the option of voting upon ideas. Number of votes is saved providing administrators with valuable information for further stages

(analysis or selection). All ideas are linked to the wave they belong.

Idea Analysis

The second stage consists in analysing ideas and preparing them for the next stage. Here, the administrator converts ideas in blog format into wikis. For this stage Innoweb provides a set of graphical tools that help managers in the search and comparison of ideas. The following relations are explored; ideas using the same tags, ideas from the same user, most voted/readed/commented ideas? These tools are especially useful when administrators deal with a large amount of ideas and the relation among ideas is not clear. Finding which ideas are well considered in the community and the existent relations among ideas-users-companies make the filtering of ideas an easier task. This turns into an increase on productivity and a better coverage.

Finally another module has been developed at this stage to convert automatically ideas in blog format to ideas in wiki format. This way, the first filter is done. From all the ideas generated in the first stage, some experts select the ones that are going to be enriched (see figure 3.5) and the rest are discarded but stored, in case they can be exploited in the future.

To wiki	Idea	# of votes	# of comments
<input type="checkbox"/>	Ingeniería inteligente del residuo electrónico	8	9
<input type="checkbox"/>	City App	7	6
<input type="checkbox"/>	Espacios agro-urbanos sostenibles	3	1
<input checked="" type="checkbox"/>	GEO-LIFE	2	2
<input type="checkbox"/>	Sistema de compactación de residuos urbanos	1	1
<input type="checkbox"/>	OPERADOR DE MOVILIDAD (Smart Mobility) sistema inteligente de transporte	1	0
<input type="checkbox"/>	GARKITCH: garden in the kitchen	0	0

Figure 3.5: Idea analysis implementation with idea2wiki module.

Idea Enrichment

At this stage a wiki is provided for each filtered idea. Once ideas are in wiki format, experts add valuable information creating richer ideas. All contributions are saved in different revisions in order to identify contributors and idea evolution, and also give the possibility of restoring previous versions. A custom module has been developed to offer

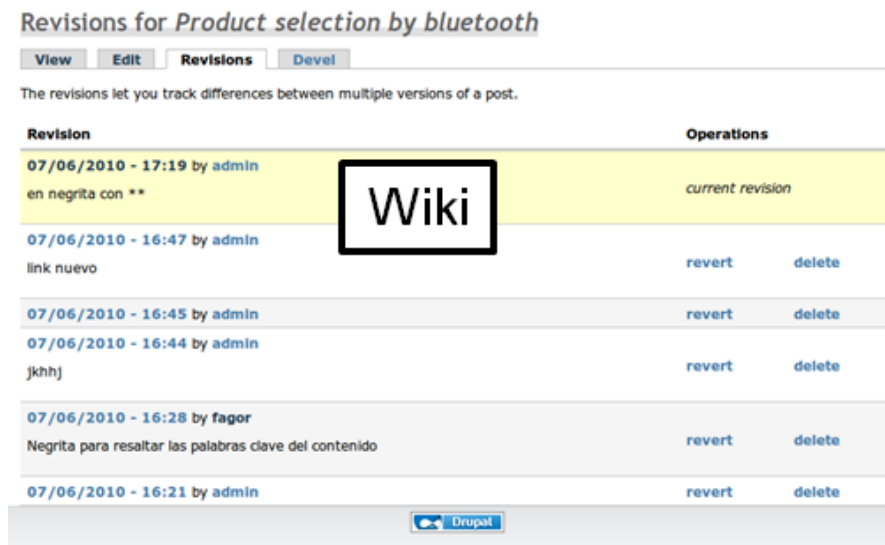


Figure 3.6: Wiki format for Idea enrichment.

wikis linked to the wave and the original idea (see figure 3.6). This module performs the third stage of the innovation process, the *Idea Enrichment* state.

Idea Selection

At this stage of the process Innweb provides tools to carry out the selection of ideas and to establish customize criteria that support the selection process. A new module called Innoselect has been created to help in this task (see figure 3.7).

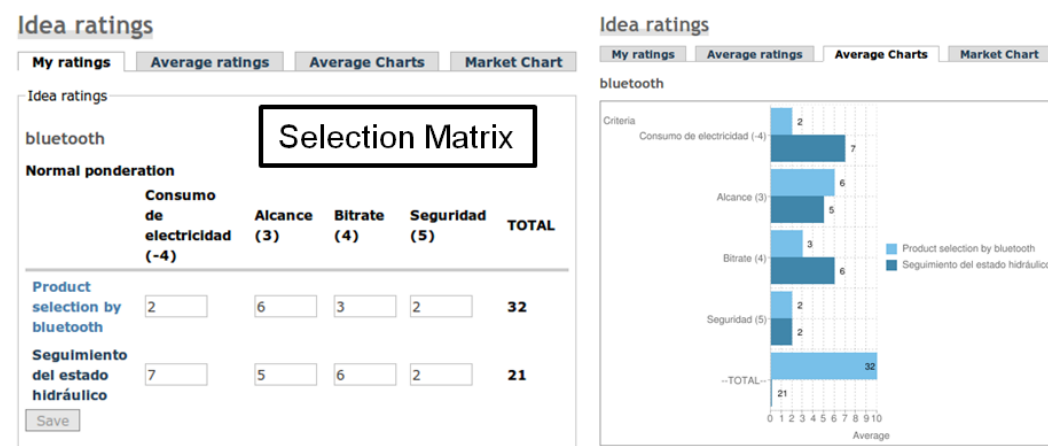


Figure 3.7: Innoselect module for Idea Selection.

Administrators configure the selection framework by adding conditions or questions

that will be considered when ideas are evaluated (selection criteria). Each of those conditions can be weighted. That is, different weights can be established depending on the relevance of the criterion. Once criteria are established, authorized users will have the possibility of rating ideas. There is a box for each criterion-idea relation where the user introduces his grade or rating. The module offers three rating possibilities:

1. Normal weighting: The user has to enter a value between 1 and 10.
2. Weighted selection: The user has to order ideas from the worse to the best.
3. Criteria matrix: The user can only enter predefined values in the ratings.

Wave

One of the objectives for this research work was to create a platform flexible enough to accommodate and adapt multiple different innovation campaigns or waves (see figure 3.8). Each wave has to be customized according to the requirements suitable for that campaign and yet the baseline platform has to be the same for different innovation experiments. Issues such as stages involved in the process, users allowed, topics, tag vocabularies or idea fields to be collected have to be fully configurable. Campaign owners must be able to configure that process before the campaign is launched and once it is activated they need tools to manage its progress. This module must gather most of the innovation process context information.

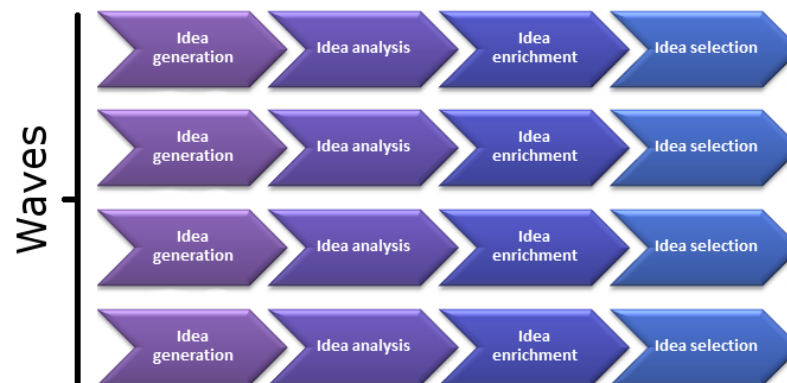


Figure 3.8: Multiple innovation processes (*Waves*) at once.

A new module that addresses these issues has been developed for the platform. The module, called Wave, allows the creation of different customized idea contests and enables their management simultaneously over the same baseline platform. The customized options included are the following:

- Stages of the innovation process. Depending on the wave the process can omit some stages.

- Dictionaries and tags. To customize vocabularies and tags to be used in a wave.
- Events. Specific organizational actions can be register in order to determine which management actions help in the innovation process.
- Permissions. Users are allowed the possibility of viewing/editing/creating content depending on their role.
- Idea fields. Administrators define the fields to be collected in each idea. Real time notifications. Represent the communication networks the platform will use to communicate with users when something relevant happens.

Another module has been developed, called WaveRT, to enable the possibility of real time notifications (see figure 3.9) of different actions. For example, it can notify to a user that one of his ideas has been promoted.

The image shows a user interface for selecting real-time notification options. At the top, there is a button labeled 'Real Time'. Below it, the text 'Real time notifications:' is displayed in bold. Underneath, there are two checkboxes: one for 'Twitter' and one for 'e-mail'. At the bottom of the section, there is a subtitle: 'What real time options to use in idea creation'.

Figure 3.9: Real time notification selection.

Finally a Wave Event module has been developed, that stores events happened within a Wave, so their impact can be analysed. A visual example can be seen on figure 3.10.

Semantic web

In order to represent the innovation process domain, Gi2MO ontology was selected for the platform. Additional classes and properties were added to fulfil the requirements innovation campaigns introduced into the domain. This way, in collaboration with UPM (see subsection 7.1.2), Gi2MO ontology was extended creating a new ontology called Gi2MO Wave (see figure 7.1).

Gi2MO Wave adds 8 new classes to the ontology, enabling the context data gathering:

- **Competitor:** An object of this class indicates the possible competitor affected or created within the idea or idea contest. This is, a new idea (or idea contest) can affect an existing competitor or bring new competitors to the company.
- **Customer:** This class indicates the targeted customer on the idea or idea contest.
- **IdeaContestType:** This class represents the objective type of innovation searched by the Idea Contest, such as: offer, product, value capture, supply chain...

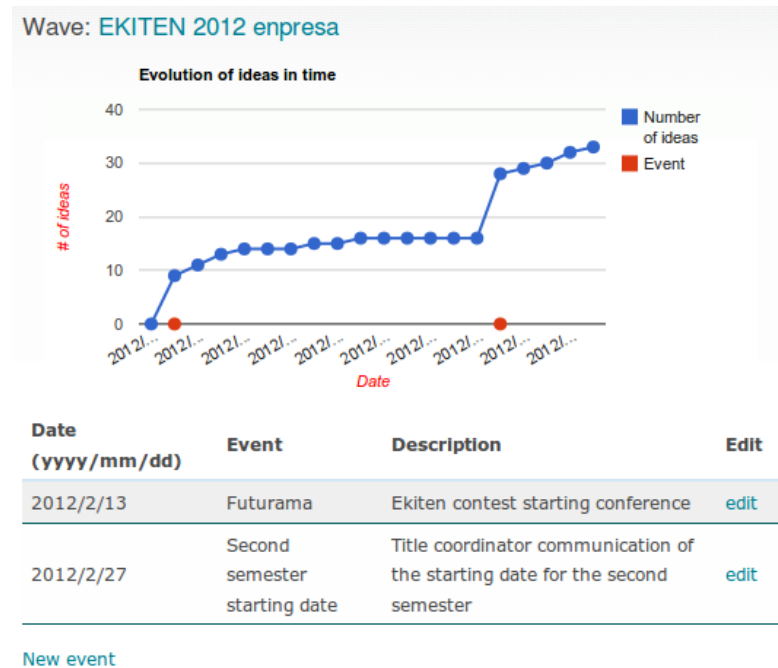


Figure 3.10: Wave Event visualization example.

- **Market:** An object of this class indicates the market aim by an idea or Idea Contest. It could be a place, a sector, a niche, etc.
- **Outcome:** This class indicates the expected final result or outcome for an idea or Idea Contest. This outcome or result can be a product, a service, a process, a strategy, an improvement, a company or spin-off or cooperation among companies, departments, etc.
- **Resource:** This class indicates the necessary resources for an idea or Idea contest. This resource can be of different types; general, knowledge, financial, income, etc.
- **Target:** An object of this type indicates the objective or target of an idea or idea contest.
- **Technology:** An object of this class indicates a technology involved with the idea or idea contest, whether it is a proposed technology or used one.

It also adds 19 properties: hasCompetitor, hasCustomer, hasMarket, hasOutcome, hasOwner, hasResource, hasSelectionCriteria, hasTarget, hasTechnology, hasType, isCompetitorOf, isCustomerOf, isMarketOf, isOutcomeOf, isOwnerOf, isResourceOf, isSelectionCriteriaOf, isTargetOf, isTypeOf. For more information read the ontology specification in the appendix (see appendix A) or in the web³.

³<http://purl.org/gi2mo/wave/ns>

Using Gi2MO Wave ontology, Innoweb brings the possibility of serving the ideas in RDF format, making semantically stored data interoperable. Not only ideas are stored in RDF, but also idea campaign metadata (or waves), ideas, events, tags... A basic example idea in RDF format can be seen on figure 3.11. In order to do this, a module called *IMS2RDF* was developed over Drupal.

```
<gi2mo:Idea rdf:about="http://gi2mo.org/idea/012345">
  <foaf:page rdf:resource="http://gi2mo.org/ideaView?id=012345"/>
  <gi2mo:hasCreator rdf:resource="http://gi2mo.org/user/pedro"/>
  <gi2mo:content>A new, nice and modern building for the department that would have a similar
  interior design as shopping malls.
  </gi2mo:content>
  <dcterms:title>Department building with shopping mall interior design</gi2mo:title>
  <dcterms:created>2012-04-23</gi2mo:created>
  <gi2mo:hasStatus rdf:resource="http://www.purl.org/gi2mo/ns#Implemented"/>
  <gi2mo:hasComment rdf:resource="http://gi2mo.org/comment/054321"/>
  <gi2mo:hasCategory rdf:resource="http://gi2mo.org/category/General"/>
  <gi2mowave:hasTarget rdf:resource="http://gi2mo.org/description/02345"/>
  <gi2mowave:hasCustomer rdf:resource="http://gi2mo.org/description/02346"/>
  <gi2mowave:hasOutcome rdf:resource="http://gi2mo.org/description/02347"/>
</gi2mo:Idea>

<gi2mowave:Outcome rdf:about="http://gi2mo.org/description/012345">
  <foaf:page rdf:resource="http://gi2mo.org/ideaView?id=012345"/>
  <gi2mo:hasCreator rdf:resource="http://gi2mo.org/user/pedro"/>
  <gi2mo:content>A new building would make people feel better in comparison to studding in the
  current one that has 40+ years without renovation.
  </gi2mo:content>
  <dcterms:title>Outcome: Department building with shopping mall interior design</gi2mo:title>
  <dcterms:created>2012-04-23</gi2mo:created>
  <gi2mowave:isOutcomeOf rdf:resource="http://gi2mo.org/idea/012345"/>
</gi2mowave:Outcome>
```

Figure 3.11: An example idea on RDF format.

Some community modules have also been used in order to give the main Semantic Web capabilities to the platform, offering an endpoint along with the previously mentioned *IMS2RDF* module (see section 3.3.6).

3.3.5 Visualization

To ease the understanding of the platform, some visualization modules have been developed. Two types of graphical APIs are provided in Innoweb developed by the team. This API enables using Google Chart Tools JavaScript API and Blazegraph Flash dynamic graph layout engine.

- *Blazegraph API*: Blazegraph⁴ is a dynamic graph layout engine implemented in Flash and ActionScript. Developed Blazegraph API module enables using this tool for data visualization (see figure 3.12).
- *JSchart API*: JSChart API module enables using dynamic charts from Google Charts⁵, such as Scatter Charts or Bar Charts.

⁴<http://blazegraph.sourceforge.net/>

⁵<https://developers.google.com/chart/>

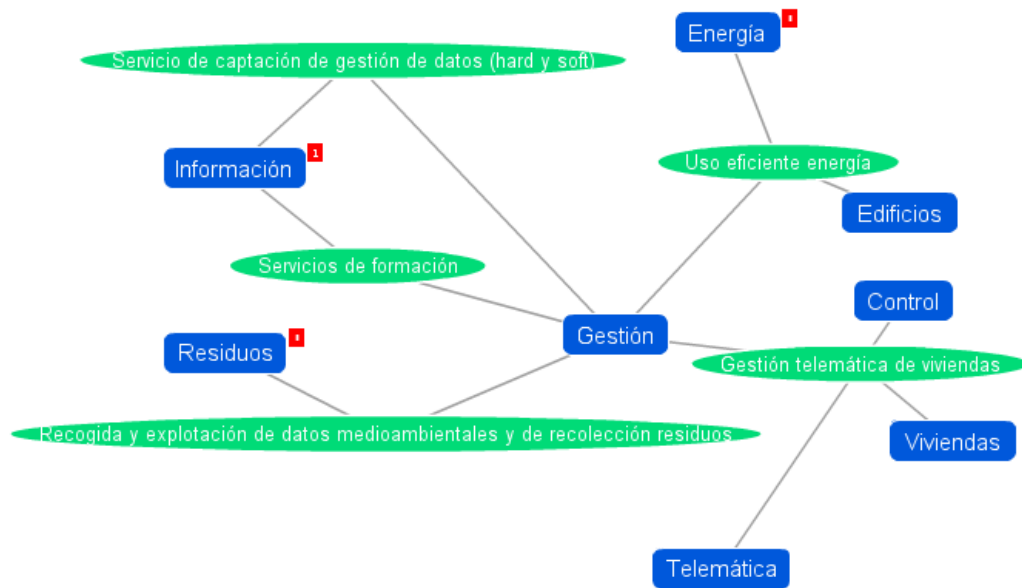


Figure 3.12: Blazegraph visualization tool.

Using these Drupal community modules and APIs, some visualization modules have been developed:

- Innoselect Chart module visualizes the expert ratings of the ideas from *Innoselect* module (see figure 3.13). The use of JSChart API enables interaction with the chart in order to visualize the data in a dynamic way.

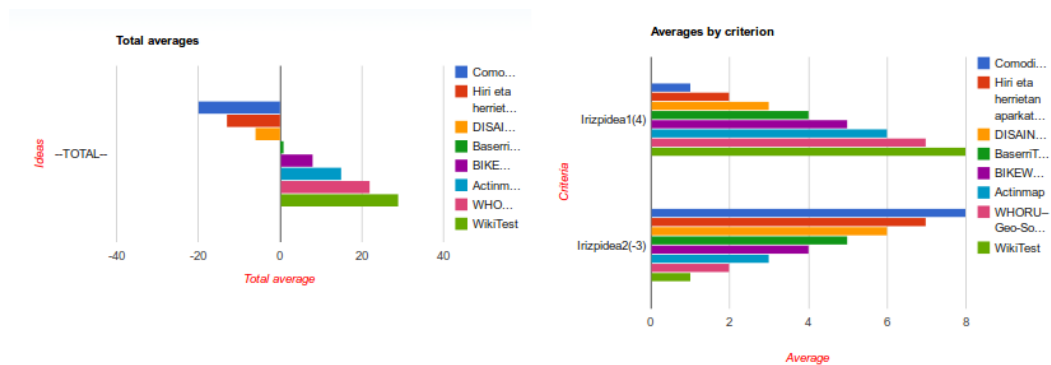


Figure 3.13: Idea Selection rating visualization example from *Innoselect Chart* module.

- *WaveChart*: This module uses JSChart API to show Wave data visualizations. A graphical example can be found on figure 3.14. This makes easier the wave administration tasks.

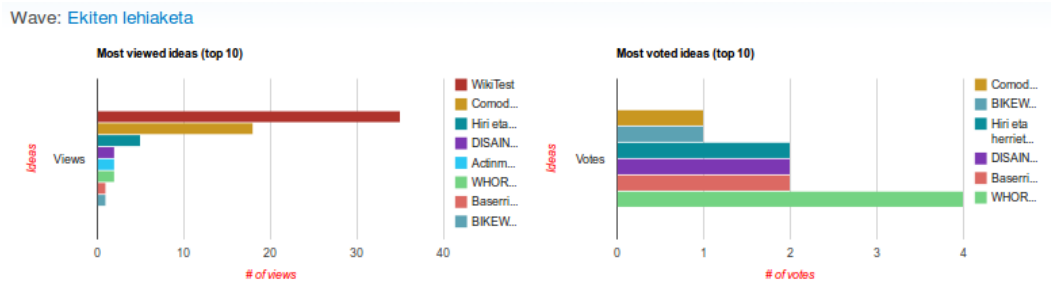


Figure 3.14: Wave Chart visualization example, showing idea related data.

- *WaveBG*: WaveBG uses Blazegraph API to show relations between ideas (see figure 3.12), making easier for the administrators the task of analysing ideas.

3.3.6 Community modules

There are some community modules that have been used in order to add functionalities to the platform. Those modules are configured in order to work along with developed modules. Below those modules are described:

- *VotingAPI*: This module enables functionalities to perform ratings in nodes and comments.
- *Vote Up Down*: VoteUpDown adds a rating interface in order to use VotingAPI. This way, the users will have some widgets to perform the ratings.
- *CCK*: This module adds different functionalities to add fields to any node inside Drupal.
- *Workflow*: Workflow module allows Drupal nodes to have different life-cycles. This way, nodes can have different states that change depending on the situation. Usually it is used to manage the content, if it is in a draft state, pending for publication or published. In our case it has been very useful in order to manage the ideas life-cycle through the innovation process along with the "Wave" module we built.
- *RDF*: this API makes use of the ARC2 library if available, and will integrate RDF capabilities with other modules.
- *SPARQL*: This is a module that enables the use of SPARQL queries with the RDF API.
- *RDFme*⁶: This is a Drupal extension that allows to publish RDF metadata attached to regular Drupal HTML pages. It also enables the possibility of publishing an SPARQL endpoint with any RDF data.

⁶<http://www.gi2mo.org/apps/drupal-rdfme-plugin/>

Cras tempor [View](#) [Voting details](#) [Edit](#) [Track](#)

Voting & Comments

Mon, 05/30/2011 - 16:02 — admin

1 [+](#) Cras tempor, elit eget hendrerit tincidunt, sapien quam rutrum purus, a fermentum nunc metus in tellus. Ut ultricies feugiat erat et dapibus. Cras blandit congue ornare. Vivamus velit purus, adipiscing sit amet mollis eu, aliquet ac eros. Aenean aliquet, metus vitae pretium accumsan, nibh neque luctus tortor, ut dictum orci lacus eu neque. Fusce vel odio ut nibh tincidunt vestibulum. Fusce eget diam eu arcu tristique molestie. Vestibulum tellus neque, facilisis ut elementum malesuada, ornare vitae nibh. Duis est nisi, cursus id eleifend vel, porttitor eget neque. Quisque aliquam lacus eu odio malesuada sodales. Duis neque quam, volutpat ut condimentum quis, viverra ut felis. Vestibulum congue nulla eu magna rutrum et blandit nisi pretium. Aenean et vulputate dui. Aliquam id consectetur turpis. Nam porta pellentesque fringilla. Donec hendrerit, enim egestas porta elementum, lorem lacus venenatis nisi, vehicula mattis odio nulla vitae leo. Aliquam vel posuere mi.

[More information](#)

[Add new comment](#) 8 reads [Reset your vote](#) Mobi

Comments

Donec non lobortis nisi Mon, 05/30/2011 - 16:13 — admin

1 [+](#) Donec non lobortis nisi. Vestibulum ultrices nunc id nunc interdum quis feugiat magna commodo. Proin massa felis, aliquam eu egestas vel, lobortis non arcu. Fusce vulputate adipiscing lacus, sed fermentum massa tempus quis. Sed dictum ultrices porttitor. Cras non lectus ac ante sodales eleifend. Donec lobortis, lacus a volutpat ornare, mi turpis adipiscing eros, vel vulputate urna justo quis lectus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Suspendisse sit amet mauris id elit tempus vestibulum. Phasellus a lacus lorem. Aenean fringilla, felis nec tempor consequat, ante elit ornare magna, luctus venenatis risus purus quis quam. Duis condimentum libero eu est semper non volutpat lorem sodales. Maecenas et lectus risus. Nullam at lacus id est fermentum tristique. Pellentesque quis turpis sit amet velit euismod sodales sit amet sed lacus. Aenean massa sem, sodales vel rutrum id, vestibulum vitae leo. Vestibulum nec eros eu leo venenatis eleifend. Maecenas lacinia dapibus nisi sit amet adipiscing. Suspendisse potenti.

[delete](#) [edit](#) [reply](#) [Reset your vote](#)

Figure 3.15: Voting widgets in the ideas.

3.4 Results

Innoweb platform has been used in 3 different real scenarios: (1) Elkarbide, conducted in ISEA; (2) Ideiak, conducted in Koniker and (3) Ekiten, conducted in Mondragon Corporation. Nevertheless, only the *Ekiten* case study had more than one iteration of the innovation process, therefore, the results of the case study conducted in Mondragon Corporation are going to be presented.

3.4.1 Ekiten (Mondragon Corporation)

This case study is carried out in Mondragon Corporation⁷, today the top Basque business group and the seventh biggest in Spain. Mondragon Corporation has a total of 256 companies and bodies, of which approximately half are co-operatives. The average number of employees at Mondragon Corporation is 83.859 and approximately 9000 students course their studies at *Mondragon Unibertsitatea* (MU).

The case study, named *Ekiten*, is an idea contest driven by the Engineering, Business and Humanities faculties of MU and with the sponsorship of Mondragon Corporation, SAIOLAN entrepreneurship development centre, Debagoiena commercial development centre, Gazteempresa foundation and Athlon enterprise. The objective of the contest is to promote entrepreneurship among students by collecting their ideas on the creation of new enterprises or business models. Information about Ekiten has been collected using

⁷<http://www.mcc.es>

Context parameter	2010-11	2011-12	2012-13	2013-14	2014-15
<i>Groups</i>	10	27	46	98	78
<i>Participants</i>	40	92	155	240	276
<i>Experts</i>	10	10	10	14	14
<i>Evaluators</i>	9	9	9	10	10
<i>Events</i>	11	16	24	24	15
<i>Activities integrated in regular courses</i>				5	13

Table 3.1: Inputs from Ekiten case study.

Outcome	2010-11	2011-12	2012-13	2013-14	2014-15
<i>Ideas</i>	10	27	49	105	84
<i>Promoted ideas</i>	3	2	3	6	8
<i>Spin-offs</i>	0	1	1	1	1
<i>Average rating of ideas</i>	-	3.76	4.77	5.9	5.76

Table 3.2: Outcomes from Ekiten case study.

Innoweb since 2010. Wave module for context data gathering has been used since 2011. Each year three main topics were withdrawn; rural development, youth-leisure-sports and innovation enterprise. A wave was launched for each main topic.

Every wave had a sponsor from outside the university as the owner or manager for that wave. An external selection committee was appointed by each sponsor company. Each committee was formed by five external experts that established the selection criteria for each wave and made the actual selection of ideas. The criteria considered in most of the waves was related to the level of innovation, definition and maturity of the idea, the technical and economic feasibility, the level of alignment of the idea with the strategy and the priorities set for the topics dealt with in each wave, and finally, the confluence and leverage of the proposal with the capacities and competences available in Mondragon Corporation's companies.

Wave administrators set the general parameters for the wave and configured participants, stages, permissions, vocabularies and time-lines prior to opening Innoweb idea management tools to users. Students were allowed to introduce their proposals including their description and the title of the idea, the outcome expected (become a new product, process, service or spin-off), the issues addressed with the proposal, the type of innovation and the objective market or customer. Experts used Innoselect to rate and select best ideas. Events were registered in the platform during the whole process. The events registered were of 4 types: success stories, workshops, information bulletins and coaching sessions. The amount of participants, events, experts, etc. can be found on table 3.1. The outcome values, such as amount of ideas, average or promoted ideas, can be found on table 3.2.

Context parameter	Rural Development	Innovation Enterprise	Youth-Leisure-Sports
<i>Ideas</i>	4	33	12
<i>Average grade of ideas</i>	3,17	2,78	6,68
<i>Ideas online contest</i>	2	15	12
<i>Promoted ideas</i>	0	1	2
<i>Spin-offs</i>	0	0	1
<i>Events (success stories)</i>	6	6	8
<i>Event (workshops)</i>	3	3	5

Table 3.3: 2012 wave comparison

As an example of the type of information that can be gathered with the platform, an extract for the three waves hold in 2012 is presented on table 3.3. The influence of events in the quality of ideas can be observed. Further tracking of those events can be done in the platform. Thus, while 9 events were common to all campaigns, 4 were specific to the Youth-Leisure-Sports wave that obtained better grades and where all ideas were aligned with the contest objectives.

3.5 Conclusions

Having presented the developed platform and the case study where it has been tested, some conclusions have been drawn:

- This chapter has addressed the 1st research question of this thesis: *Can a conceptual model help on replicate successful Idea Contests?*. According to the requirements a platform to support the innovation process has been built. The platform gathers context data that can be further analysed to determine the influence in the outcome and detect success factors. This way, and in relation with the 1st objective of the thesis (*Propose a conceptual model for the identification of successful idea contests and their replication*), a model has been defined and implemented with the developed platform making easier to identify and replicate successful campaigns.
- The platform offers data in semantic format. This enables interoperability, exploitation of semantic meaning and the possibility to incorporate ideas to the Linked Data. The platform and its semantic capabilities represent the left column of the architecture described in chapter 1 (the column related to the idea content).
- The platform presents not only activity or traffic metrics, but also quality measures such as idea grades.
- Visual tools developed in the platform ease the way to analyze data to the managers. Raw data is more complicated to manage, so visual tools enable data analyzation in faster views.

- The impact of management decisions can also be measured. That is, campaigns can determine if a workshop or brainstorming session translates into more (quantity) and better (quality) ideas.
- In the research it has been identified that the complexity in managing the innovation process is not trivial and managers involvement is imperative. The more involved the managers are in the process the better results are gathered in the idea contest.
- The platform provides a better campaign control. Campaigns can be easily stopped, paused, shorten, expanded or re-launched depending on activity or environment conditions.
- Previous examples and experience recorded in the platform enable a better design of new campaigns. For example, data from tables 3.1 & 3.2 show that in the 2014-15 wave there was a reduction on the amount of the ideas and their average ratings. This happens in a context where there were more participants with less amount of groups and less events. Thus, the managers identified that the events and groups were more *course oriented* in that year. This is, less people was adding particular ideas outside the defined groups. Therefore, it was concluded that more generic events with open groups had better results than *course oriented* groups. In conclusion, future idea contest should not align their events and groups to the courses of the university in order to collect more particular ideas again.
- The platform allows active participants identification and co-creation traceability. This is, if multiple users collaborate in the creation of an idea, their inputs can be traced in the platform.

The aim now is to enhance the platform with new functionality taking into consideration the incremental development cycle approach. Thus next cases studies will be conducted at Mondragon Corporation.

The next steps on the research are the following:

- Identify the ways to exploit the semantic possibilities already available, linking it with other repositories. This step is analyzed in chapter 4.
- Enhance the platform to contemplate other aspects of the innovation process; Technology Watch, decision making or outcome traceability. In chapter 5 the relation of the IMSs with TW platforms is analyzed.
- Assist companies' managers to automate tasks in order to reduce the non productive efforts. In chapter 6 non productive task automation is analyzed in the TW field.
- Keep on collecting ideas and measuring the performance of the platform.

Finally, with all the work described in this chapter, 2 papers were published (for more info see section 7.3): (1) *A case study on the use of community platforms for*

inter-enterprise innovation in the 17th Concurrent Enterprising International Conference (ICE2011) and (2) *INNOWEB: Gathering the context information of innovation processes with a collaborative social network platform* in the 19th Concurrent Enterprising International Conference (ICE2013).

Chapter 4

Semantic Web to Link the Innovation Process with Internal and External Repositories

This chapter focuses in a proof of concept on how to link content from IMS with internal and external repositories. A new system was developed over the IMS platform developed on chapter 3 to enable the interoperability with internal and external repositories. Tests were conducted in order to analyze the system, using Sustainability as the main topic for the research. This chapter addresses the 2nd objective of the thesis defined in chapter 1: *Identify semantic web methods that provide additional content to IMSs from repositories inside and outside the company.*

Firstly, a brief introduction is presented, on sustainability, Innovation processes and the research made for this chapter. Secondly, the potential use cases that can use the interoperability of sustainability repositories and IMSs are presented. Thirdly, the developed platform and it's functionalities is described. Next, the potential benefits of the platform are presented. Finally, the summary and future challenges of the research are shown.

Main contributions

- Proposal of a method to enable interoperability of IMS platforms with semantic repositories internal and external to the company.

4.1 Introduction

Sustainability is the responsible management of resources encompassing the triple bottom line of environmental, economic, and social dimensions. Many organisations are starting to make serious commitments towards incorporating sustainability into their own organizational logics [34] to maximise profits in an environmentally and socially responsible manner. Sustainability is not only about Corporate Social Responsibility, Sustainability

is an important business issue, affecting new products and services, compliance, cost reduction opportunities, the organization's reputation, and revenue generation often derived from technological innovation [129]. Porter recognises the role sustainability can play as part of an organization's Competitive Strategy with the concept of "innovation offsets" where companies can "not only lower the net costs of meeting environmental regulations, but can lead to absolute advantages" over competitors [106].

Sustainability requires information on the use, flows and destinies of energy, water, and materials including waste, along with monetary information on environment-related costs, earnings, and savings. This type of information is critical if we are to understand the causal relationships between the various actions that can be taken, and their impact on sustainable performance.

Innovation is key to articulate knowledge management by means of effective processes and methodologies. The phase of the innovation process where idea management is developed is one of the most critical stages [46]. IMSs support this stage providing the necessary tools to collect, enrich, store, present and select ideas. IMS manage ideas through their life-cycle from the time of creation until they are selected for implementation. During this life-cycle it is crucial to gather as much relevant information as possible in order to collect quality-relevant ideas. Users can enrich ideas with opinions, other ideas and additional content. This task can be cumbersome and its automation is fundamental.

This chapter aims to find technologies that enable the interoperability between IMS and sustainability repositories and to define how they can benefit from each other.

The hypothesis is that more precise relations among ideas and richer content can be automatically achieved if Semantic Web and Linked Data technologies are employed in IMS, linking sustainability ideas with data from different data sources. Managing innovation for sustainability needs to address some major challenges; the emergence of radical new technologies and markets, constant shift in the regulatory conditions, the involvement or participation of many agents, the large volume of ideas for screening and evaluation, and in particular the need to acquire, assimilate and exploit new knowledge [117].

An increasing number of organizations worldwide have adopted innovation contests not only for innovation purposes, but also for other reasons such as promoting sustainability [1]. A proof to this can be seen in the annual reports and sites of several energy providers and enterprises. In some of these experiences IMS have proven beneficial. IMS provide the workflow tools necessary to launch and manage innovation contest or waves, a common platform where different agents can collaborate, a repository where ideas are gathered and tools for editing, commenting or voting upon ideas. IMSs also encourage collaboration among people and enterprises.

One of the biggest problems in IMS is the difficulty of enriching these ideas. It is a manual and time consuming task that includes the searching and gathering of additional knowledge in different sources. Most of the time ideas are not linked with other data the enterprises may have in their systems or data available outside the company [36]. This causes disinformation and generates duplicates or poor quality ideas. A system that links



Figure 4.1: Energy Reduction and LCA ideas (automatically added data widgets)

generated ideas to stored data automatically may be an improvement. Semantic Web and Linked Data technologies propose a set of good practices to publish and link structured data in the Web. Many datasets and repositories are already available adopting this philosophy enabling machines in the understanding of the data they store.

4.2 Use Cases

This section describes 3 different sustainability use cases that can benefit from a system that links ideas with additional information. An internal sustainability repository from DERI has been used in order to test the use cases. The first use case aims to enrich ideas for energy reduction. The second one addresses products' life cycle and how data can be linked. The last use case shows how similar ideas can be identified helping administrators management tasks.

1. **Energy Reduction:** Imagine a user involved in an idea contest oriented to sustainability that proposes a new idea (graphically on Figure 4.1): *I would change the incandescent bulb in desk #333 for a LED bulb in order to save energy.*

The system could identify the concepts *incandescent bulb in desktop #333* and *LED bulb*, find information about both bulbs on the sustainability data space (external to the IMS) and show it in a widget or block next to the idea. If the system is able to identify the domain of the idea and annotate it semantically, the idea can be linked to data stored in a data space [35] (results on grey and blue widgets in figure 4.1) or searched on other data sources [37]. That way, the user would create an idea with automatically added information. If someone reads the idea, they will know if it is worth the effort of changing it or not. The reader could comment on the idea and discuss about it, and the decision makers will have richer information.

2. **Life-cycle Assessment (LCA):** This use case links IMSs with LCA data stored in a sustainability data space. Imagine a user concerned about the *Greenhouse Gas* (GHG) emissions discovers that new laptops are going to be bought. He could write the following idea (graphically in Figure 4.1):*If we buy 13-Inch MacBook*

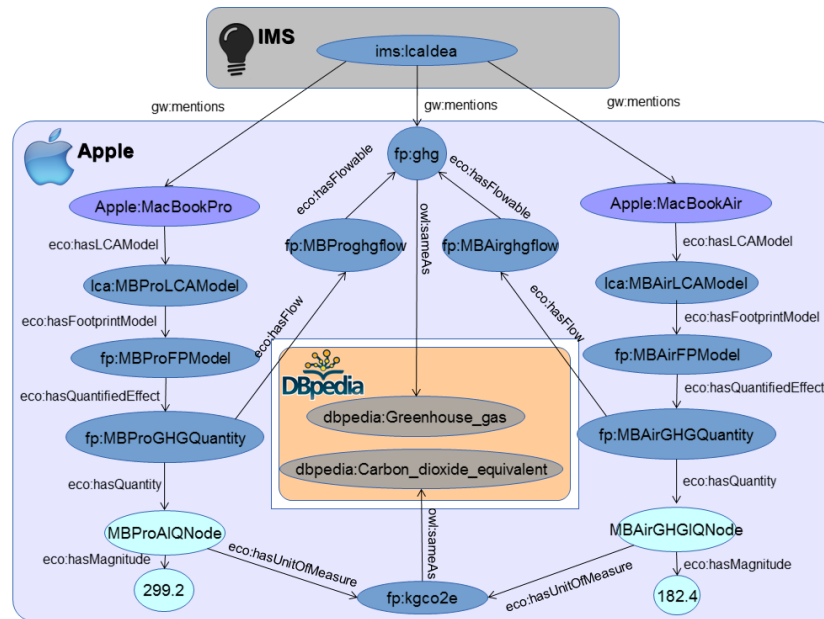


Figure 4.2: Data links example

Air laptops instead of 13-Inch MacBook Pro, we can reduce the amount of GHG emissions in the manufacture of our devices.

If we can identify that the idea talks about 2 different laptops and their GHG emissions, we could link the idea with that data and show the amount of GHG emissions each laptop has and the savings of the idea.

In order to link the data we have to annotate it semantically, for example using RDF. In RDF, the statement *LCA Idea mentions MacBook Air* is expressed in triple format as:

$$(Subject - \mathbf{LCA\ Idea}) \Rightarrow (Predicate - \mathbf{mentions}) \Rightarrow (Object - \mathbf{MacBook\ Air})$$

Using this semantic annotations some links can be found between different data in the system. That data can be found in the IMS or in some internal and external data spaces. On Figure 4.2, a graphical representation of those links can be seen. The figure shows an idea (*ims:lcaidea*) that mentions 2 different laptop models (*apple:MacBookPro* and *Apple:MackBookAir*) and the GHG (*fp:ghg*). If the system identify those mentioned elements, they can be searched inside the repository using a SPARQL query. Then, the information about the amount of GHG emitted to the atmosphere during the laptops' manufacturing process can be extracted from the repository and shown to the users of the system.

3. **Similar ideas recognition:** We can imagine an enterprise that has an innovation process for new idea gathering. Sometimes ideas can be repeated in the same or past idea contests. Having a system that identifies similarities can help innovation administrators in identifying relations or knowing the reason of rejection.

For instance in a previous idea contest, the idea of changing an incandescent bulb for a led one was rejected because the led bulb was too expensive. If someone generates a similar idea it can be linked to the previous one, and see the reasons for the rejection. If now the led bulbs are cheaper, that idea may be interesting.

In order to perform that task the identification of main concepts of the idea is needed. If the system can find similarities in those main concepts, the ideas should be linked. The idea with mentioned concepts can be compared with the other ones and similar ideas can be identified and presented to the users.

4.3 Platform

In order to enable this interoperability, extra tools were added to the platform developed in chapter 3, providing the possibility of linking IMS ideas with sustainability data. These tools were centred on enabling the interoperability of the Innovation process platform with repositories external and internal to the organization (see figure 4.3). This helps exploiting the data from other systems, such as sustainability data, adding it to the Innovation process.

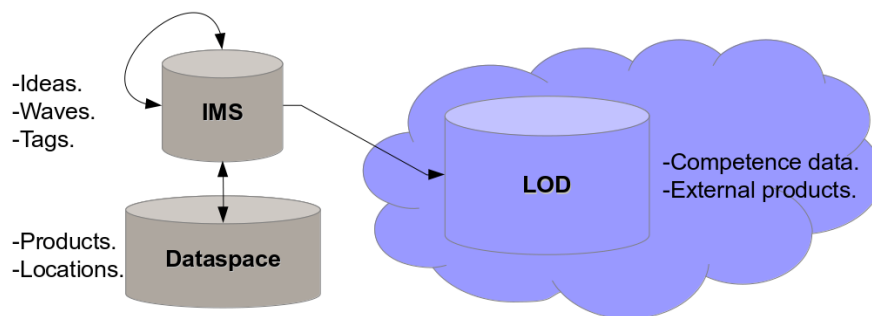


Figure 4.3: Linking IMS with external and internal repositories.

The process to find related elements in the repositories and exploit them is represented on figure 4.4. That process consists in 5 steps:

1. **Text extraction:** the plain text of the ideas from the IMS is extracted in this first step.
2. **NLP for noun extraction:** The plain text is used as an input for a NLP tool called FreeLing¹ in order to extract the elements mentioned in the ideas. Those

¹<http://nlp.lsi.upc.edu/freeling/>

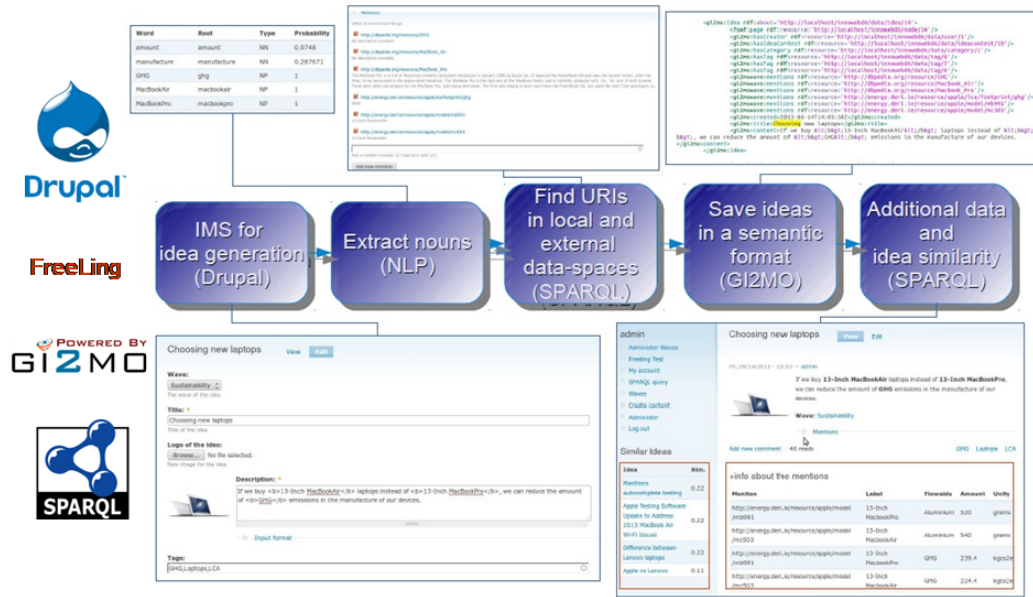
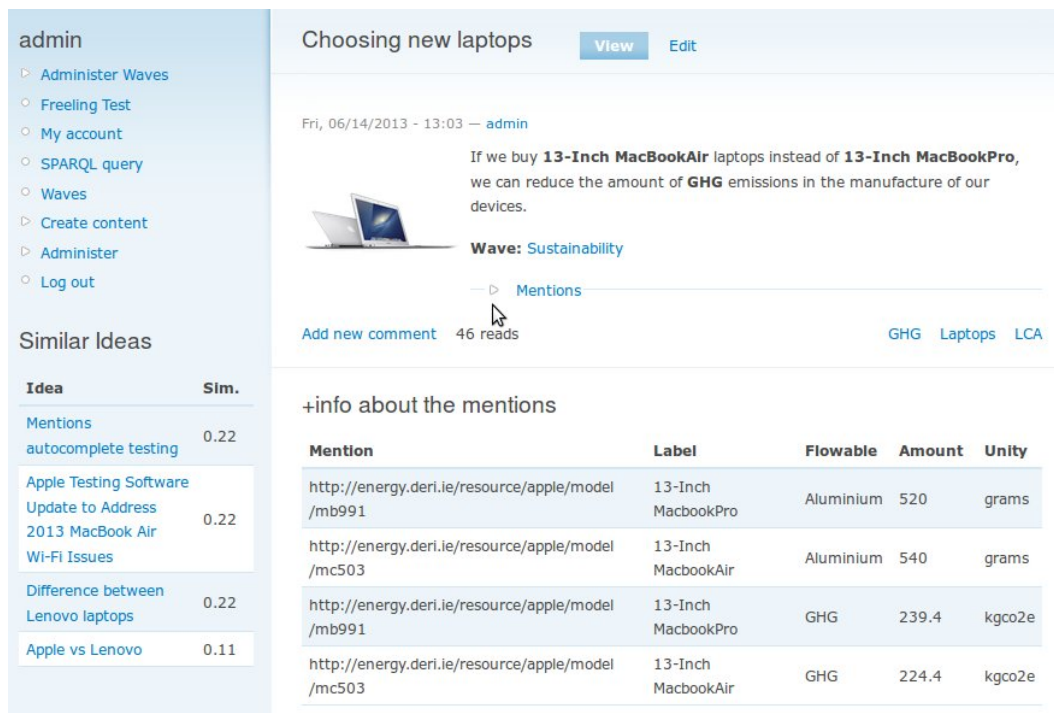


Figure 4.4: Process to find related elements in Internal and External repositories.

elements can be nouns, compound nouns, numbers (for reference numbers for example).

- URI identification**: those elements are used to identify if there is any item inside the repositories related to them, using SPARQL queries. This way, the items related to the ideas can be identified, and the URIs of those items are saved.
- Save ideas in a semantic format**: the ideas and their related item URIs are stored in a semantic format in order to enable the interoperability with other platforms, the ideas are stored using GI2MO ontology and also the URIs of the related items.
- Exploiting the interoperability**: with all the data in semantic format, the data of internal and external repositories can be shown to the user. Therefore, the last step of the process is showing that data to the users, enabling them to use the extracted knowledge. This data can be extracted using SPARQL queries and showed with different visual representations. An example of this representations can be found on figure 4.1 that shows the information about the items identified in the ideas and also on the right-bottom side of figure 4.5. The visualization of the most similar ideas can be found on figure 4.5 (left-bottom side), also showing their similarity score.

In order to perform all this steps, 2 modules were developed and added to Innoweb (the IMS platform developed in chapter 3). Below, those new modules are described:



The screenshot shows a Drupal interface for an idea titled "Choosing new laptops". The main content area displays the idea text: "If we buy **13-Inch MacBookAir** laptops instead of **13-Inch MacBookPro**, we can reduce the amount of **GHG** emissions in the manufacture of our devices." Below the text is a "Wave: Sustainability" section with a "Mentions" link. A table titled "+info about the mentions" lists the following data:

Mention	Label	Flowable	Amount	Unity
http://energy.deri.ie/resource/apple/model/mb991	13-Inch MacBookPro	Aluminium	520	grams
http://energy.deri.ie/resource/apple/model/mc503	13-Inch MacBookAir	Aluminium	540	grams
http://energy.deri.ie/resource/apple/model/mb991	13-Inch MacBookPro	GHG	239.4	kgco2e
http://energy.deri.ie/resource/apple/model/mc503	13-Inch MacBookAir	GHG	224.4	kgco2e

Figure 4.5: Example of automatically added information from IdeaMentions module.

- **FreeLing API:** FreeLing API module makes possible integrating FreeLing NLP libraries with Drupal. This allows other Drupal modules to use FreeLing's NLP services.
- **IdeaMentions:** This module performs the 5 steps of the previously described process. First, it extracts the text of the ideas from the IMS. Then, it interacts with *FreeLing API* to identify the mentioned elements. Using those elements, it performs some SPARQL queries to search for items in external (DBpedia² for example) or internal repositories (using dataspace such as Virtuoso³) that are related with the mentioned elements. If the module finds any items, their URIs are stored in the database relating the items with the ideas. Once the mentioned items are identified and stored, this module shows any found information to the users when they visualize the ideas.

4.4 Benefits

The proof of concept described in this chapter can have several potential benefits. Three of those benefits have been identified and are described below.

²<http://dbpedia.org/>

³<http://virtuoso.openlinksw.com/>

Firstly, IMS ideas would be enriched automatically with relevant data provided by LOD/SW repositories. In the use case presented in this chapter, sustainability data has been added to ideas when the concepts mentioned in the ideas are identified. This functionality helps users and decision makers to see how important an idea is. Although sustainability repositories have been considered in this case, other repositories can be also useful in different contexts using the same system.

Secondly, the developed tools should help users in order to understand the context of the ideas. If the additional information related to the mentioned items have a description, the problems of the idea could be understood more easily, even if the idea itself does not describe the information explicitly.

Finally, the tools are expected to help decision makers to perform their task faster and in an easier way. On one hand, showing them the relationships among ideas can help them identifying in what are the users concerned (many ideas about the same topics for example) and see if there are repeated ideas. On the other hand, measuring the possible impact of the ideas (cost, power consumption, GHG emissions...) could help them recognizing the most important ideas and select the best ones. The data could help decision makers identify if the idea is feasible or not, for example, showing them the price of a new product they can decide if it worthwhile.

4.5 Conclusions

Addressing the 2nd research question of this thesis (*Can semantic methods enable content linking among IMS platforms and semantic repositories internal and external to the companies?*), this chapter has tested different semantic technology methods in order to find the ones that would improve the Innovation and TW platforms (more specifically in the topic of sustainability), enabling the interoperability of the IMS with internal and external repositories.

The main innovation of this chapter strives in the application of semantic web and LOD technologies to interlink Sustainability repositories with IMS in such a way that ideas are enriched with relevant content.

As a first approach, the system developed on this chapter identifies all the possible elements mentioned inside an idea. Then, a user must select manually which of them is really mentioned in the idea. A future challenge can be selecting the correct element automatically in order to help users linking the idea. Identifying the domain of the idea can be helpful too if additional information is wanted to be added. Knowing the domain of the idea can help recognizing specific data sources where more data can be found and linked. Finally, some case studies should be implemented in order to obtain results that validate the functionalities proposed in this chapter.

This chapter has used real sustainability repositories in order to get data related with the ideas. Therefore, a future challenge could be using these tools in real scenarios or Innovation processes to see if they are found useful. Next chapter (Chapter 5), aims to use one of the defined potential use cases (the idea relations use case) in a real case scenario, enabling interoperability among platforms.

Therefore, we can conclude that this chapter has accomplished the 2nd objective of this thesis: *Identify semantic web methods that provide additional content to IMSs from repositories inside and outside the company.*

Finally, the work in this chapter has been done in collaboration with *Digital Enterprise Research Institute* (DERI) of the *National University of Ireland* (NUI). It was published in the *Modelling and Knowledge Management for Sustainable Development* (MoKMaSD) conference in 2013 (see section 7.1.1 and 7.3.3 for more information).

Chapter 5

Semantic Web to enhance Innovation and Technology Watch linking

This chapter focuses on a case study that enable the interoperability between Innovation and *Technology Watch* (TW) platforms discovering mentioned resources inside plain text of the content of the platforms. The work is based one of the use cases identified in chapter 4. The content of the platform are ideas and news items in the field of Ubuntu GNU/Linux distribution.

The chapter addresses the 3rd objective identified in chapter 1: *Propose a concept model to enable the interoperability among platforms linking content*. Moreover, it represents the left column (related to the idea domain) and the central column (related to the news domain) of the architecture. It also represents the 2 horizontal layers relating the content among different domains.

Firstly, in section 5.1 a brief introduction is presented about the problematic and motivation of the chapter. Section 5.2 describes the theory of the research along with the problematic found. Section 5.3 shows the material and methods used in the development of the research, describing the used tools, a developed ontology for mention annotation, the experiments performed for the research and the evaluation method for those experiments. Section 5.4 shows the results and discussions extracted within the experiments. Finally, section 5.5 enumerates the conclusions of the research.

Main contributions

- Architecture to enable interoperability among Innovation and TW platforms.
- *Mentions Ontology* (MO) to annotate semantically mentioned elements inside text based content.

5.1 Introduction

There is no doubt that in this new era of emerging economies *innovation* is the valid alternative for achieving competitiveness and the only chance to survive. This is precisely one of the bases of the Lisbon Strategy of the European Commission [30], innovation as a driving force for change in business.

Innovation and the *Technology Watch* (TW) processes are closely related. In any Innovation process, both *TW* and *Competitive Intelligence* are key instruments, and require the creation of a collaborative environment. There are several *Information and Communication Technology* (ICT) and platforms that support these processes, but each of them focuses on specific data or areas of the process. There is little *data interoperability* among the tools and platforms to support the flow of information during the processes and their stages.

Thus, there is an urgent need to design a model that takes into account *data interoperability* among tools and systems involved in Innovation and TW, linking content together and finding relationships among that content. It is also important to measure the impact data interoperability has on these processes.

The primary objective of the research is to deliver a solution that would aid IMS managers to reduce the amount of work on idea assessment and help them judge the ideas and select the most useful ones for their organization. The global objective of the research is divided into 3 specific goals:

- **(1) Propose a conceptual model for linking IMS with TW platforms:** Our objective is to identify a way to link information from TW platforms, such as news, patents, designs, etc. with ideas created in the IMS. The formalization of the model should improve data interoperability and portability.
- **(2) Deliver definitions from external repositories:** This objective focuses on the understandability of the ideas. Innovation platform users should receive information related to the ideas in order to understand them, such as definitions about mentioned concepts.
- **(3) Show content similarities:** This objective focuses on the discovery of relationships between content of Innovation and Technology Watch processes. It aims to automatically discover similarities in order to find related or duplicated data.

5.2 Theory

This section describes the base theory of the research. It defines Innovation and Technology Watch processes, and the problems they face in order to extract the largest amount of information possible between them.

5.2.1 Innovation

In 1934, Schumpeter [115] made one of the first definitions of innovation. From its traditional definition, innovation encompasses the following five cases:

1. Market introduction of a new good.
2. A new method of production.
3. The opening of a new market in a country.
4. The conquest of a new source of supply of semi-finished products or raw materials.
5. The implementation of a new structure in a market.

Half a century later, Padmore, Schuetze and Gibson [98] summarized Schumpeter's definition by saying that innovation is any change in inputs, methods, or outputs that manages to improve the trading position of a company and are new to the existing market.

Innovation is more effective in organizations that combine two features: (1) the control of direct initiatives at the different layers of the organization and (2) the strong commitment of the participants of the organization to the process [95].

Igartua confirmed the relationship between the use of Innovation Management Tools and innovation activity, showing the need of ICT tools in the process [68].

In reference to the operational part of the innovation process, innovation requires a flow of ideas, obtained through formal and informal processes [75]. IMSs are some ICT tools that manage that flow of ideas, and therefore, they are key tools for innovation process management.

Westerski [132], in relation to the flow of ideas, identified 3 data categories IMSs could be linked with:

1. Internal assets
2. Data across enterprise
3. Public data

But when it comes to implementing the innovation process, one of the biggest problems is *interoperability*. Many systems containing that data, are not interoperable. Users are frequently required to switch from one system to another and waste time getting the needed information.

ICT can be used to create links between the data and generating an interoperable data space where related information can be found and exploited. Igartua, J.I. [68] confirmed on 2010 the relationship between the use of Innovation Management Tools and innovation activity, showing the need of ICT tools in the process. Feigenbaum et al. [48] said that "The Semantic Web thus permits workers in different organizations to use their own data labels instead of trying to agree industry-wide on one rigid set", so it could be useful in order to achieve interoperability among data in different platforms.

Westerski et al. [133] wrote about innovation, interoperability and linking. They described how Semantic Web and ontologies can be used for fully describing the Idea Management domain in pair with other existing ontologies. They focus their efforts on assisting IMS managers, classifying idea relationships. They also identify the need to enable interoperability among different platforms.

The work on this chapter focuses on enabling interoperability and finding idea relationships not only to assist IMS managers, but also support other users in different platforms. Therefore, a content relationship architecture is proposed in order to enable interoperability among text based content, finding mentioned elements inside the plain text.

Another innovation process problem identified in previous studies is the large amount of data available in the platforms. A company with a large idea repository would require extensive resources to manage the gathered data. Geffen and Judd [56] said that “duplicates are a significant part of the gathered ideas and make idea assessment a time consuming, tedious process”. Ford and Mohapatra [52] added that “innovators do not find it worthwhile to review the huge database of existing ideas in search of duplicates before posting their own idea” and that “idea authors are not presented with proper tool support for duplicate search”. A key aim of these researches is to find similarities between ideas, not only to show duplicated ideas, but also to give the users as much related information as possible.

As R&D is closely associated with the Innovation process, TW and its implication within innovation was studied.

5.2.2 Technology Watch

As outlined in the standard UNE 166006:2011 (*R&D&i management: Technological watch and competitive intelligence system*), Technology Watch is an *organized, selective and permanent process*, to capture information from outside and inside the organization about science and technology. This information is then analyzed, selected and disseminated within the organization, enabling decision making with less risk and anticipating change. In this way, Technology Watch represents a key tool in the R&D&i process.

Within the TW process, Durán et al. [43] identified the importance of the classification of the information and proposed to use lists of controlled terms, classified and grouped according to different points of view. Fernandez Fuentes et al. (2009) indicated that TW is the starting point of the innovation process and identified the need of establishing key words in order to tag the information gathered.

Companies use TW platforms for collecting data relevant to their work. But usually there is a data-overflow that make complicated to transform that data into usable information. So enabling interoperability of TW platforms can be beneficial in order to exploit that data and make it usable in other systems.

As reflected in the previously mentioned standard (UNE 166006:2011), the information that must be watched in a TW process comes from different domains and formats:

- Patents, utility models, industrial designs (national, European or global). Often

the time of the presentation is important, other times, the expiring time.

- Legislation and Regulations that may affect the activity of the company, its customers or suppliers.
- Socio-economic situation in the home or target countries of the company.
- Scientific and technical news on specialist journals, symposia, conferences and other scientific events.
- Doctoral theses and scientific and technical publications from universities, research centres and agencies.
- Sector news (without neglecting other sectors that can have positive or negative interference with the business of the company).
- Information on grants and subsidies.
- Products, prices, quality and sale conditions of competitors.
- Trade Shows: emerging industries, new competitors, distribution strategies, new products, etc.
- Direct personal contacts with competitors, suppliers, research centres, universities, etc.

Maynard et al. [88] described the news domain as a “clear area where it is important for companies to keep a close eye on technological developments in their field”. Thus, the news domain of Technology Watch was used in the experiments of this research aiming to test the linking ability of the defined architecture. Nevertheless, any other Technology Watch domain content could have been used.

5.3 Material and methods

This section provides the details to reproduce this work. First, it describes the solution architecture and tools to tackle the problems described in the previous section. Finally, it shows the evaluation method used to test the architecture.

5.3.1 Solution architecture

The solution architecture employed in the research is shown on figure 5.1. With this architecture the research faces the 1st objective of the chapter “Propose a conceptual model for linking Idea Management Systems with Technology Watch platforms”.

At the lowest level of the architecture the content can be found. This content was composed by ideas from IMSs and news items gathered using Technology Watch platforms.

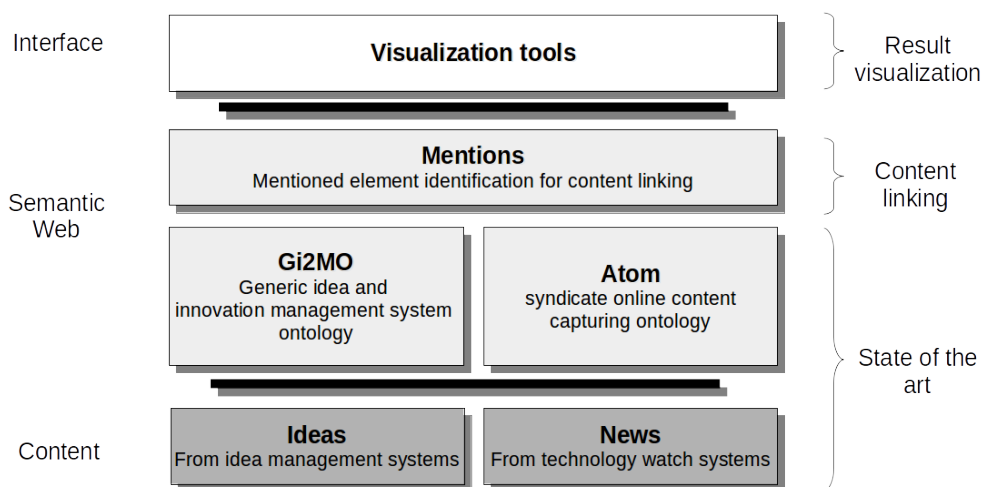


Figure 5.1: Solution architecture

At the second level, the content is described semantically. On one hand, a generic idea and innovation management system ontology called Gi2Mo¹ [131] was used for the semantic representation of ideas. On the other hand, a syndicate online content capturing ontology called AtomOwl² was employed to annotate the news items.

At the third level content from both systems is linked using an ontology. The ontology is fed with annotations from DBpedia [8] [79]. Those annotations link ideas from IMSs with other ideas and also with news items from TW platforms. In order to represent those annotations semantically, an ontology has been developed (more information can be found in section 5.3.3). A simplified version of the mentioning and how it can be used to find relationships between content can be seen in figure 5.2.

Finally, the visualization tools are located in the last level. In this research, those visualization tools have been developed over a Drupal platform. Those tools enable the visualization of most related content when a user reads an idea. They also provide the definitions of the identified mentioned concepts in both the title and the body of the content.

5.3.2 Tools

This subsection describes the different tools used in this research. Firstly, 2 different platforms are presented: (1) the IMS platform used to gather ideas and (2) the TW platform used to collect news items. Finally, the API used in order to identify mentioned elements from ideas and news items is described.

The IMS used in this research was based on a system called *Innoweb*, built over *Drupal 6* developed in chapter 3. This platform covers the left side of the model, related

¹<http://purl.org/gi2mo/ns>

²<http://bblfish.net/work/atom-owl/2006-06-06/AtomOwl.html>

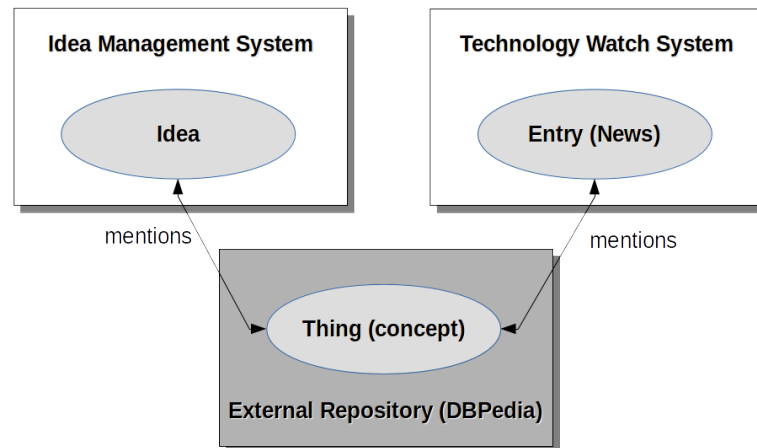


Figure 5.2: Simplified mentioned concept identification to find content relationships.

to the ideas. The platform allows adding different functionalities in a modular way. For this research, a module called IdeaMentions was added in order to take the text from the title and body of the ideas, send that text to the DBpediaSpotlight API [39] and save the annotations in the database. Data is also semantically structure so it can be consumed by third parties using SPARQL queries. That way, content similarity can be found using those annotations. This module provides the functionality to cover the third level of the model. Moreover, the definition of those elements can be gathered automatically from DBpedia articles.

On a similar way, a TW platform was build over Drupal 7, gathering news using RSS content from different sources. This covers the right side of the previously described model, focused on TW and the news domain. A module for getting annotations using DBpedia Spotlight API was also developed for this platform providing the same capabilities as those of the IMS.

This way, IMS and TW content can be linked using semantic annotations fullfilingthe functionality expected in the 3rd level of the model. The link among content through annotations is achieved making SPARQL queries between both platforms.

Furthermore, both platforms (IMS and TW) provide visualization tools that cover the upper level of the model presenting related content to an idea in the form of other related ideas or news and content from DBpedia.

In order to get previously mentioned annotations, *DBpedia Spotlight API* is used. This API allows us to configure the annotations through quality measures such as prominence, topical pertinence, contextual ambiguity and disambiguation confidence. We provide text fragments and DBpedia Spotlight returns URIs for found resources mentioned within the text. Those URIs are resources from **DBpedia**, which contains encyclopedic knowledge from Wikipedia for about 3.5 million resources, enabling access to many data sources in the Linked Data cloud.

The output of DBpedia Spotlight API provides this information for each annotation

found in the giving text:

- **URI**³: Unique resource identifier of the entity (mentioned element) from DBpedia.
- **Support**: Amount of in-links the resource has at DBpedia.
- **Surface form**⁴: Resource's String in the input text.
- **Offset**⁵: The position of the surface form in the input text.
- **Similarity Score**⁶: The topical relevance of the annotated resource for the given context.
- **Percentage of second rank**: The percentage of the similarity score of the next best entity compared to the current one.
- **Types**: Found categories of the entity.

In order to represent this information semantically, an ontology called *Mentions Ontology* (MO) has been developed. This ontology specification can be seen in section 5.3.3.

5.3.3 Mentions Ontology (MO)

Based on the simplified semantical represented shown in figure 5.2 and taking into account the information given by *DBpedia Spotlight API*, an ontology has been developed to represent the mentioned elements within a content.

Mentions Ontology (MO)

Gomez-Perez [58] defined some guidelines for ontological engineering and Uschold et al. [123] described a 5 step methodology for ontology modeling:

1. Identify the purpose and scope of the ontology.
2. Build the ontology by capturing knowledge, coding knowledge and integrating such knowledge with existing ontologies.
3. Evaluate the ontology.
4. Documentation.
5. Guidelines for each phase.

³These properties are includes in our semantic annotation representation seen section 5.3.3

⁴See footnote 3

⁵See footnote 3

⁶See footnote 3

Following those guidelines and steps, MO was built. For the first step of the methodology the objectives and scope described in the background of this paper were considered. The *scope* is then already defined: TW and Innovation *scopes*. Nevertheless, the ontology has been developed for a more generic scope, so it can be applied to any text based content. The *purpose* of the ontology is to enable the semantic annotation of mentioned resources in plain text, this is, to identify the elements mentioned in a giving text.

The 2nd step of the methodology is the technical building of the ontology. Due to Mentions similarity with Tagging, it was decided to base the ontology in an existing Tagging ontology. *Modular Unified Tagging Ontology* (MUTO) [81] ontology was selected due to its unification view, modularity and compatibility with other tagging ontologies. On figure 5.3, a class diagram of the *Mentions Ontology* (MO) can be found. 2 classes, 3 data properties and 1 object property have been added to represent Mention functionalities.

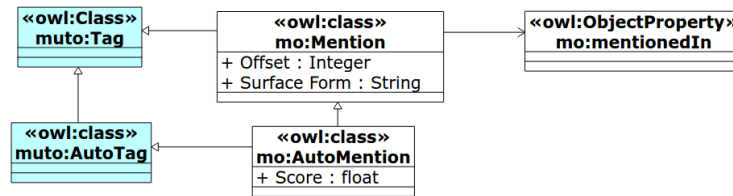


Figure 5.3: Mentions Ontology’s Class Diagram.

The 2 classes added in this Ontology are the ones that add Mention functionalities to the Tagging ontology. The first class (the *Mention* class) is the class that represents manual mentioning (manual annotation). This occurs when humans manually define mentioned resources in a text. Therefore, to add that mentioning functionality, 3 properties are needed: (1) Surface form, the base text that has been identified as the one mentioning the resource; (2) Offset, the position of the surface form in the given text; and (3) Mentioned In, the property where the text has been extracted from (e.g. the title or the body of an idea).

The second class (the *AutoMention* class) is the class that represents an automatic mentioning using a text annotation system, such as *DBPedia Spotlight API*. This is a subclass of the previously described Mention Class and the AutoTag class from MUTO. Apart from the 3 properties added by Mention class, a new property is added: *Score*. This property represents the similarity score given by the text annotation system to the resource, the bigger the number the more probable it is to be the correct resource. In the manual Mention class no human error is considered, so the *Score* property is not necessary.

The 3rd step of the methodology is the evaluation. The semantic annotation performed in this research has been done using the Mentions Ontology. This means that the work done in this research can be considered as the evaluation of the ontology.

Finally, 4th and 5th steps have been implemented publishing *Mentions Ontology* in

the web⁷, following the W3C best practices⁸, enabling ‘*text/html*’ access along with ‘*application/rdf+xml*’.

Basic Mentions Ontology example

The diagram in figure 4 illustrates an example of adding mentions to an idea from an IMS at *example.org*. The idea has a title (“*Card-lock for laboratory door*”) and a body (“*I often forget keys to the laboratory, so i think it would be great to install a card access lock with a reader for chips in spanish ID*”).

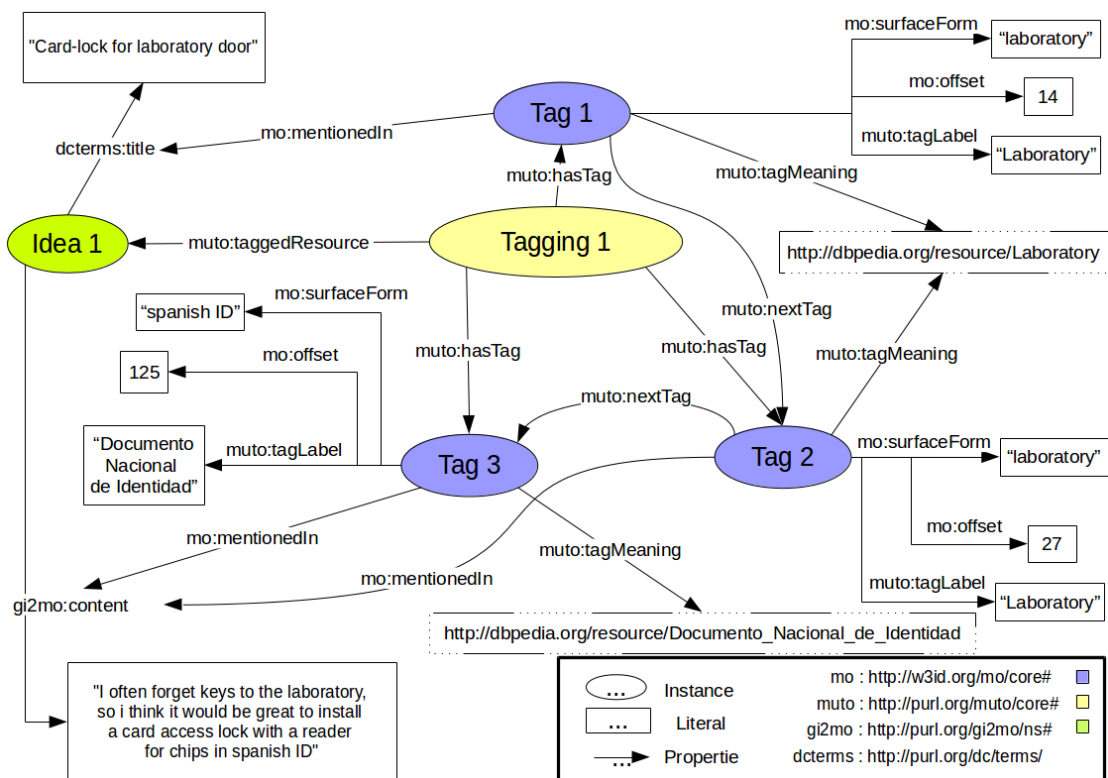


Figure 5.4: A graphic of a basic example of a Mentions Ontology (MO) use case.

According to figure 5.4, a user has identified that the idea has mentioned 2 times the same resource: “laboratory”. Therefore, the user has added 2 mentions of the same *resource* (“<http://dbpedia.org/resource/Laboratory>”) with the same *surface form* (“*laboratory*”), one to the title of the idea (*dcterms:title*) with *offset 14* and another to the content of the idea (*gi2mo:content*) with *offset 27*. He has also labeled the mention as “*Laboratory*”.

⁷<http://w3id.org/mo>

⁸<https://www.w3.org/TR/2008/WD-swbp-vocab-pub-20080123/>

The automatic mentioning system has automatically recognized the *resource* “http://dbpedia.org/resource/Documento_Nacional_de_Identidad” from the surface form “*spanish ID*” within the content of the idea, at *offset 125*. Therefore, it has created a new AutoMention linking it to the resource. According to the system, there is a 95% of probability for the surface form to be the retrieved resource, so it has set the *Score* to *0.95*. He has also identified the label of the resource and has set it to “*Documento Nacional de Identidad*”.

Different resources mentioned in the idea have been identified, both manually and automatically. This can be done with all the ideas of the IMS.

5.3.4 Experiments and population

During the evaluation stage of the project three experiments were conducted. These experiments were focused on (1) defining concepts found in the idea corpus; (2) finding similarities between ideas, and (3) discovering relationships between ideas and news items. The 1st experiment is related with the 2nd objective outlined in the introduction (“Deliver definitions from external repositories“) and the last two are related with the 3rd objective of this chapter (“Show content similarities”).

The ideas used for the experiments were taken from the *Ubuntu Brainstorm Idea Management System*⁹. This dataset was also used by Westerski [132] in some of his experiments related to IMSs. The dataset has over 27000 ideas about Ubuntu GNU/Linux distribution.

The system used all 27000 ideas, identifying the mentioned elements of all ideas. But as evaluating the whole population would take too much time and effort, all experiment evaluations were performed using 200 ideas. This population was composed by ideas following these criteria:

- **10 most commented ideas:** the 10 ideas with the greatest number of comments were selected.
- **10 not commented ideas:** from all ideas with no comments, 10 were randomly selected.
- **10 highest rated ideas:** the 10 ideas with the highest ratings were selected.
- **10 not rated ideas:** from all ideas of the dataset with not ratings, 10 were randomly selected, not repeating any of the ideas previously selected.
- **40 implemented ideas:** from all implemented ideas of the dataset, 40 were randomly selected, not repeating any of the ideas previously selected.
- **120 random ideas:** from all the ideas of the dataset, 120 were randomly selected, not repeating any of the ideas previously selected.

⁹<http://brainstorm.ubuntu.com> (closed)

The news content for the news similarity experiment was sourced from *OMG! Ubuntu!*¹⁰, *Ubuntu Security Notices*¹¹, *I heart Ubuntu*¹², *Web Upd8*¹³ and *Ubuntu Manual*¹⁴. That repository has 550 news items. As in the previous population, a population of 200 news items were selected in order to perform the experiment evaluation. Those news items were selected randomly within the whole population, with no other specific criteria taken into account.

5.3.5 Evaluation method

An expert in the field of Ubuntu was employed to evaluate the effectiveness of the system. They were given the set of 200 ideas each of which included the following information:

- The title of the idea.
- The corpus of the idea.
- All the concepts identified by the system and their definitions.
- 10 most related ideas identified by the system.

They were also given the set of 200 news items each of which included the following information:

- The title of the news item.
- The corpus of the news item.
- All the concepts identified by the system and their definitions.
- 10 most related ideas identified by the system.

The content was then analysed as follows by the expert:

1. Concept definition experiment:

The objective with this experiment was to determine if the concept definitions automatically provided by the platforms were aligned with the content of the idea and if those definitions agreed with the list of critical concepts identified by the expert. Expert conducted to main tasks:

- (a) *Found concept definition accuracy*: the expert analysed the list of all concepts the system recognized automatically for each idea and indicated the accuracy of those definitions. Accuracy in this context means whether the definition agrees with the content of the idea (is relevant to an idea).

¹⁰<http://feeds.feedburner.com/d0od?format=xml>

¹¹<http://www.ubuntu.com/usn/rss.xml>

¹²<http://feeds.feedburner.com/IheartUbuntu?format=xml>

¹³<http://feeds.feedburner.com/webupd8?format=xml>

¹⁴<http://feeds.ubuntumanual.org/ubuntumanual?format=xml>

- (b) *Critical concept accuracy*: previous to automatic idea concept annotation, the expert annotated manually the most relevant concepts with the main problem each idea attempts to solve. These concepts consolidated the list of Critical Concepts. Later, they compared the manually annotated definitions (Critical concept list) with the concepts automatically identified by the system, registering the accurately and the non-accurately identified critical concepts.

2. Idea similarity experiment:

The objective with this experiment was to identify relationships between ideas (Idea relationship). The procedure followed to conduct this experiment considered that for each idea, the system identified a further 10 most related ideas. The expert classified those idea-idea relationships according the following quality criteria:

- (a) Same: the idea focuses on the same problem.
- (b) Closely related: the idea focuses on a similar problem.
- (c) Related: the idea focuses on a different problem, but on the same main topics.
- (d) Somehow related: the idea focuses on different problems and main topics, but the same minor topics.
- (e) Not related: the idea does not mention any related topic.

3. News similarity experiment:

The objective with this experiment was to identify relationships between ideas and news items (News relationship). The procedure followed to conduct this experiment considered that for each news item, the system identified a further 10 most related ideas. The expert classified those news item-idea relationships according the following quality criteria:

- (a) Same: they focus on the same main topics.
- (b) Closely related: they both focus on one of the main topics, but not on all of them.
- (c) Related: they are about similar secondary topics.
- (d) Somehow related: they mention similar topics but not the same.
- (e) Not related: they do not mention any related topic.

With the data of the 2nd and 3rd experiments, a statistical inference was performed to reject a null hypothesis, see table 5.1. This way, we proposed an alternative hypothesis to be accepted if the null hypothesis is rejected.

"*Same*" and "*Closely related*" relationships were considered as high quality; "*Related*" relationships were considered as medium quality; and "*Somehow related*" and "*Not related*" relationships were considered as low quality. A significance level of 0.05 (5%) was used in the statistical inference.

Having collected the discrete values for the quality of the relationships associated for each content and even being a discrete variable, due to the size of the sample (200 ideas

H0 (null hypothesis)	The system finds low quality relationships
H1 (alternative hypothesis)	The system does not find low quality relationships

Table 5.1: Null and alternative hypothesis for similarity experiments.

* 10 links/idea = 2000 links) such distribution is approximated by a normal variable of unknown parameters.

For this distribution, it is possible to perform a hypothesis test about whether the mean of this distribution is "low" (low=0).

5.4 Results

This section, describes the results and discussion of the experiments performed in this research. First, *Concept definition* experiment results will be shown, followed by *Idea similarity* and finally *News similarity*.

5.4.1 Concept definition

During the found concept definition accuracy task the expert read each idea and reviewed all the concept definitions the system found automatically for that idea registering whether they were accurately identified. That is, whether the definition was aligned with the content of the idea or not. Figure 5.5 and figure 5.6 present the accuracy classification for all the concept definitions identified in the body and in the title of all ideas respectively.

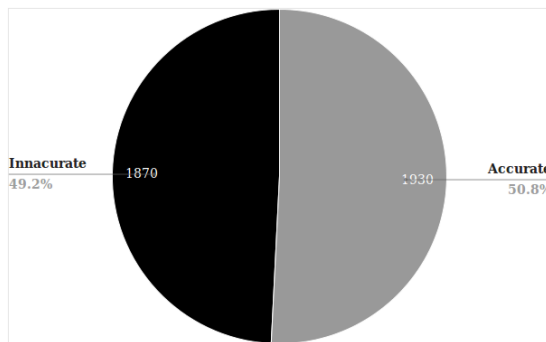


Figure 5.5: Accuracy of identified concepts from the body.

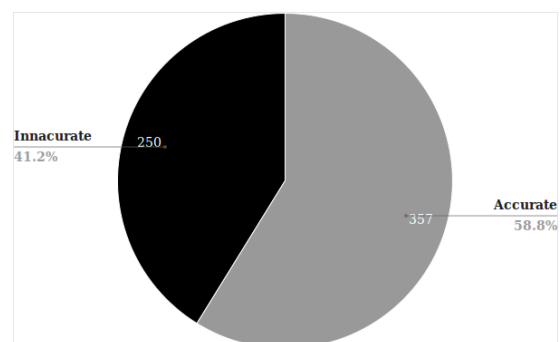


Figure 5.6: Accuracy of identified concepts from the title.

Results in figure 5.5 and 5.6 show that between 50 and 60 % of the concept definitions were aligned with the ideas they supported and that DBpedia Spotlight API identified better the concepts in the title than those in the body.

During the *critical concept accuracy* task, expert annotated manually every critical concept contained in each idea before reading ideas and then checked how many of those

critical concepts were identified by the system automatically analyzing all concept definitions. Figure 5.7 and figure 5.8 show how many concept definitions were in the list of critical concepts elaborated by the expert (black + dark grey), how many of those were accurate aligned with the idea (black), how many were not aligned with the idea (dark grey) and how many critical concepts identified by the expert were not identified (light grey). Results in figure 7 and figure 8 concerning critical concepts, show that DBpediaSpotlight API performs better with the text from the body than the text from the title.

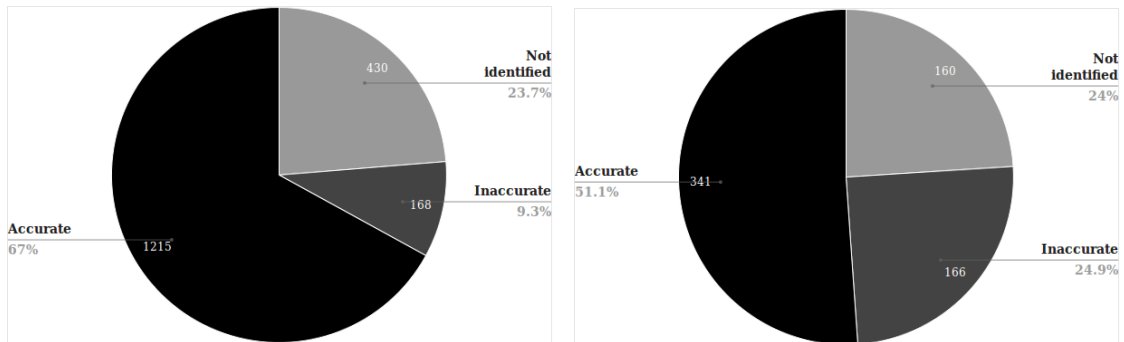


Figure 5.7: Accuracy of critical concepts from the body.

Figure 5.8: Accuracy of critical concepts from the title.

Further analysis comparing identified concepts (figures 5.5 and 5.6) and critical concepts (figures 5.7 and 5.8) show that the percentage of critical concepts identified in the title is bigger than the ones identified in the body. That percentage is calculated as follows:

- The system identified 607 concepts in the title text (sum of all values in figure 5.6), and 507 were critical concepts (sum of all black and dark grey concepts in figure 5.8). Thus, 83.52% of the concepts identified in the title were critical (out of those, 67.26% accurately aligned with the idea).
- The system identified 3800 concepts in the body text (sum of all values in figure 5.5), and 1383 were critical concepts (sum of all black and dark grey concepts in figure 5.7). Thus, 36.39% of the concepts identified in the body were critical (out of those, 87.85% accurately aligned with the idea).

Although the body concept identification has been more accurate, the percentage of critical concepts in the title is larger. Therefore, it is possible that using the title for content relationship identification could perform better using the title than using the body.

5.4.2 Idea similarity

Expert reviewed for each idea the 10 most related ideas automatically identified by the system and classified each relationship according to the quality criteria outlined in

section 5.3.5 (Idea similarity experiment). Considering that this experiment aims to identify relationships between ideas four iterations of the experiment were made, using different strategies. These strategies are outlined in the following subsections. The summarized results can be seen on figure 5.9. Each bar presents the expert evaluation on the relationships between ideas for each iteration. Relationships between two ideas considered the same are shown in black. Closely related idea relationships are colored in the darkest grey. The shade of grey decreases as the relationship between ideas weakens. Not related ideas are colored with the lightest grey. Next, each iteration will be described in detail and their results will be discussed.

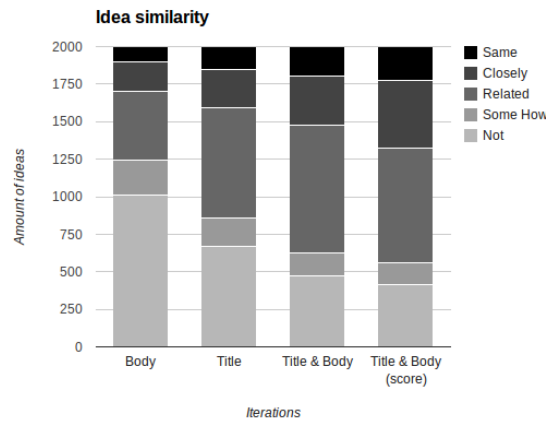


Figure 5.9: Idea relationships in each iteration.

Body iteration

For the 1st iteration idea relationships were calculated automatically considering only mentions extracted from the body section of the ideas. If two ideas had the same mentioned element in the body, a matching was found. The ideas with more matchings were higher in the related ideas list. Expert reviewed each of those matchings classifying them according the established quality criteria.

The results for this iteration are shown in figure 5.9 (*Body* bar) presenting the following numbers:

- 1008 *Not related* ideas.
- 233 *Some how related* ideas.
- 463 *Related* ideas.
- 194 *Closely related* ideas.
- 102 *Same* ideas.

Further we analysed the relevance of idea relationships by position. That is, the idea with the most matchings was placed in the first position of the list. The idea with the second most matchings was placed second and so on. Figure 5.10 shows the expert evaluation considering the position in detail. The idea presented in the first position obtained the best results. That is the idea considered by the system as the closest to the reference idea was also evaluated by the expert as being the idea with the highest relationship. According to the expert, ideas in higher positions (more matchings) with the reference idea present better results (stronger relationships) than those with lower positions in the relationship list.

As seen in the concept definition experiment, the system found many matchings of not critical concepts in the body. Too much noise was added in order to find high quality relationships. Therefore, a second iteration was planned using the title section of the idea, which usually contains a larger percentage of critical concepts.

Title iteration

The 2nd iteration of the experiment was made using only mentions extracted from the title section of the ideas. If two ideas had the same mentioned element in the title, a matching was found. As in the previous iteration the ideas with more matchings were higher in the related ideas list. The results for the 2nd iteration are shown in figure 5.9 (*Title bar*) presenting the following numbers and between brackets, the differences with the 1st iteration are outlined:

- 666 *Not related* ideas (342 less).
- 193 *Some how related* ideas (40 less).
- 733 *Related* ideas (270 more).
- 255 *Closely related* ideas (61 more).
- 153 *Same* ideas (51 more).

It can be noted that High and Medium quality relationships increase and Low quality relationships decrease when comparing this iteration with the 1st one. The numbers also show that there is a significant reduction on “Not related ideas” and an increase on “Related ideas”.

As in the previous iteration a second analysis focused on the results by position in the relationship list. Figure 5.11 shows the results by position. For this iteration the expert also determine that ideas in higher positions (more matchings) with the reference idea present better results (stronger relationships) than those with lower positions in the relationship list. Moreover it can be seen that the quality of related ideas has increased with respect to the first iteration specially in the first positions.

The iteration significantly improves the quality of the idea relationships, but many times ideas with the same number of matchings were found. They were ordered by idea identification number, which is the sequential number given by the system when it is

introduced. This identification number has not relation with the matchings. Therefore, a third iteration was planned, using the body matchings to order the ideas with the same number of matchings in the title.

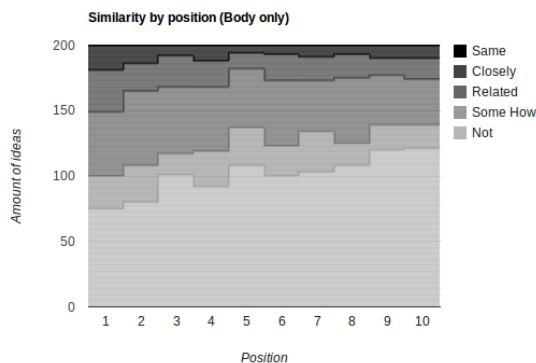


Figure 5.10: Idea relations by position on 1st iteration.

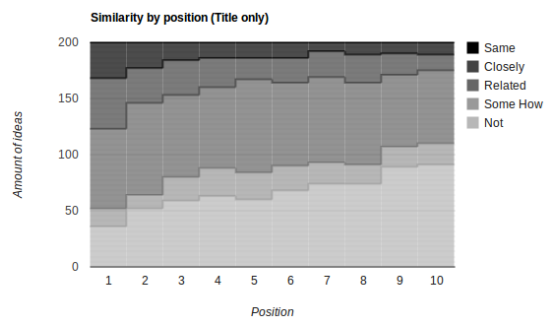


Figure 5.11: Idea relations by position on 2nd iteration.

Title & body iteration

The 3rd iteration of the experiment was made using mentions from the title and body sections of the ideas. The ideas with the largest amount of matchings in the title were higher in the related ideas list. If there was a tie on the amount of matchings in the title, the elements mentioned in the body were used to order those results.

The results for the 3rd iteration are shown in figure 5.9 (*Title& Body* bar) presenting the following numbers, and between brackets the differences with the 2nd iteration are presented:

- 475 *Not related* ideas (191 less).
- 147 *Some how related* ideas (46 less).
- 851 *Related* ideas (118 more).
- 330 *Closely related* ideas (75 more).
- 197 *Same* ideas (44 more).

When comparing this iteration with the previous it can be noted that again High and Medium quality relationships increase and Low quality relationships decrease. The numbers also show that there is a significant reduction on “Not related ideas” and an increase on “Related ideas”.

Figure 5.12 shows the results in more detail, showing the quality of the relationships by position. As for the previous case it can be seen that the quality of related ideas is higher for the first positions and that it has improve with respect to the previous iterations.

Title & body score iteration

In order to obtain a qualitative relationships between ideas, a fourth iteration was planned. The iteration used the similarity score given by *DBPedia Spotlight API*. The similarity score is a value of relevance of the annotated resource for the given context. In other words this value shows us how confident is the tool about the annotation it has provided. It is a confidence measure (numerical) that allows the comparison among annotations (mentions).

Thus, the 4th iteration of the experiment was made using mentions from the title and body sections of the ideas. If two ideas had the same mentioned element, the multiplication of the given similarity score was used as a scored matching. The ideas with the higher score in the title matchings were higher in the related ideas list. If there was a tie, the score from the body matchings were used to order them.

The results for the 4th iteration are shown in figure 5.9 (*Title& Body (score) bar*) presenting the following numbers, and between brackets the differences with the 3rd iteration are outlined:

- 414 *Not related* ideas (61 less).
- 143 *Some how related* ideas (4 less).
- 767 *Related* ideas (84 less).
- 448 *Closely related* ideas (118 more).
- 228 *Same* ideas (31 more).

This iteration presents the best results of the four iterations. When comparing this iteration with the previous one it can be noted that the improvement occurs mainly on High quality relationships while Medium and Low quality relationships decrease. The numbers also show that there is a significant increase on “Closely related ideas” while “Not related ideas” and “Related ideas” reduce their number in higher degree.

The analysis by position confirmed that the system provides better relationships when that relationship is higher in the position list. Figure 5.13 shows the results in more detail.

Performed the statistical inference described in section 3.5, the confidence intervals for mean distribution parameter were obtained. The region limits found (0.529 and 0.530) and not including this interval the value of 0.5, it can be inferred with a significance level of 0.05 (95% of probability) that the data rejects the null hypothesis. Thus, it is statistically demonstrated that the quality of the quality of the relationships found by the system is superior to “low”.

Table 5.2 shows the statistical inference results by the amount of relationships. All inferences were performed in the last iteration, using a significance level of 0.05 (5%). The 1st column shows the region limits using the 10 most related ideas, the 2nd column using the 5 most related ideas and the 3rd column using the 3 most related ideas. The results show that the higher ranked the ideas are the better results the statistical inference show.

	10 relationships	5 relationships	3 relationships
Iu1	0.529	0.582	0.608
Iu2	0.530	0.584	0.612

Table 5.2: Statistical inference results by amount of idea-idea relationships.

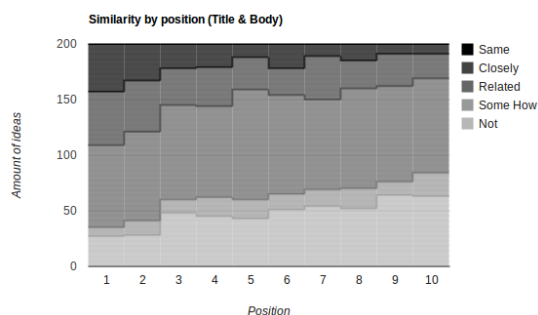


Figure 5.12: Idea relations by position on 3rd iteration.

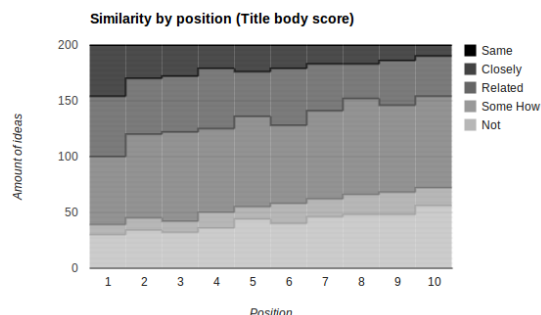


Figure 5.13: Idea relations by position on 4th iteration.

5.4.3 News similarity

For this experiment a single iteration was performed, using the one that presented the best results in the idea similarity experiment. Thus, the experiment was made using mentions from the title and body sections of the news items and ideas and the similarity score given by the annotation system. If a news item and an idea had the same mentioned element, the multiplication of the given similarity score was used as a scored matching. The ideas with the higher matching score in the title were higher in the related ideas list. If there was any tie, the matching scores from the body were used to order them. Figure 5.14 shows the results of the experiment and Figure 5.15 shows those results in more detail, showing the quality of the relationships by position.

This is the distribution of the quality of the relationships between ideas and news items:

- 518 *Not related* ideas.
- 86 *Some how related* ideas.
- 768 *Related* ideas.
- 550 *Closely related* ideas.
- 81 *Same* ideas.

It can be seen that these results are slightly worse than the idea similarity experiment. Nevertheless, taking into account that the found links are among different types of content, it can be seen that positive results have been gathered.

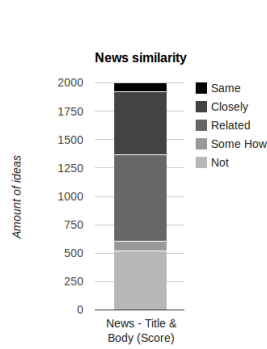


Figure 5.14: News items and ideas relationships.

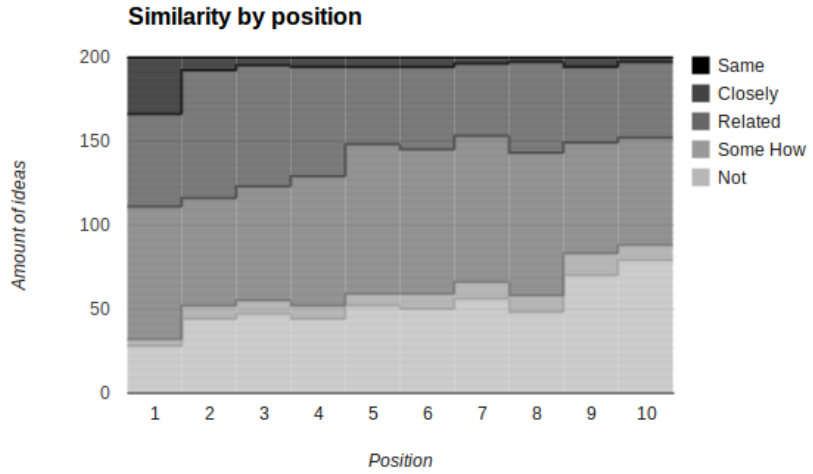


Figure 5.15: News items and ideas relationships by position.

	10 relationships	5 relationships	3 relationships
Iu1	0.506	0.560	0.591
Iu2	0.507	0.563	0.594

Table 5.3: Statistical inference results by amount of idea-idea relationships.

Similar to the previous experiment, a statistical inference was performed and the confidence intervals for mean distribution parameter were obtained. The region limits found (0.506 and 0.507) and not including this interval the value of 0.5, it can be inferred with a significance level of 0.05 (95% of probability) that the data rejects the null hypothesis. Thus, it is statistically demonstrated that the quality of the relationships found by the system is superior to “low”.

Table 5.3 shows the statistical inference results by the amount of idea-news item relationships. All inferences were performed using a significance level of 0.05 (5%). The 1st column shows the region limits using the 10 most related ideas, the 2nd column using the 5 most related ideas and the 3rd column using the 3 most related ideas. The results show that the higher ranked the ideas are the better results the statistical inference show.

5.5 Conclusions

In this chapter, the 3rd research question of the thesis has been answered (*Can interoperability among content from Innovation and TW platforms be generalized in a single model?*), defining an architecture to automatically link ideas between themselves and ideas with news items. This architecture also assists in identifying concepts mentioned in both ideas and news items and showing their description automatically. In order to achieve this, an ontology has been developed, representing semantically those mentioned

concepts. The results show that most critical concepts were correctly identified, and a definition was given to improve the understandability of those concepts and the ideas themselves. Moreover, this architecture has a more generalistic view, and any content based on plain text could be used for linking purposes.

According to the statistical inference performed, this research suggests that this architecture is feasible in order to link ideas among IMSs. It also suggests that this architecture is feasible in order to link ideas and news items from different platforms.

Taking into account the results of the different iterations performed in the research, the best way to relate ideas with new items or other ideas is using the mentioned elements from both, the title and the body, along with their similarity score.

The results of relationships by position show that the content higher in the related content list had better quality relationships than the ones lower in the list. This proves that the score given by the system follows a logical criteria and that in every iteration has achieved better results.

In this chapter the developed ontology has been used in specific domains (ideas and news items domains) and using specific semantic annotation tools (DBpedia Spotlight API and DBpedia content). Nevertheless the ontology has the capability to adapt to other domains and tools. Future experiments could be focused on testing the ontology with other domains and tools and test its behaviour.

Thus, we can conclude that the 3rd objective of the thesis (*Propose a concept model to enable the interoperability among platforms linking content.*) has been achieved in this chapter.

In future works, it would be interesting to test if adding automatic critical concept identification would help achieving a better performance on content linking quality. A way of achieving this could be to identify the ones that appear more frequently and those with the same categories. Thus, the idea would not only be related by the amount of mentioned elements and similarity scores, but also by frequency and category.

It could also be interesting to test the system in other use cases. Moreover, the system could be compared with other ways to find similar content. It could also be possible to combine several ways to test if it achieves better results. There are also many other areas that could benefit from this content relationships feature. Among others, patents or designs could be tested to see if they get similar results.

Chapter 6

Semantic Technologies to enhance Technology Watch productivity

This chapter focuses on a case study performed in a real TW scenario. Too much time was wasted in the classification of documents inside a company. Therefore, some task automation were performed in order to reduce the human workload, reducing the amount of readings of the documents in order to classify them.

This chapter addresses the 4th objective identified in chapter 1: *Reduce the workload of the experts in the TW process*. Moreover, it represents the upper layer of the central column (related to the news domain) of the architecture (*Task automation*).

Firstly, in section 6.1 a brief introduction is presented about the problematic and motivation of the chapter. Secondly, the background where the chapter has been developed is outlined in section 6.2. Thirdly, section 6.3 shows the theory analyzed in the chapter about different semantic technologies to assist automatic document classification. The used datasets and tools along with the experiments to test them are described in 6.4. Next, the results gathered from the experiments are outlined in section 6.5. Finally, the conclusions gathered in this chapter are outlined in section 6.6.

Main contributions

- Reduction of human workload on document classification tasks in the TW process.
- Identification of best semantic technologies for automatic document classification.

6.1 Introduction

Technology Watch (TW) is an organized, selective and permanent process, to capture information from outside and inside the organization about science and technology. The process consists on finding, extracting, selecting, analyzing, adding value and disseminating information. Up to 65% of the time expend in these tasks is considered repetitive and unproductive. *Information and Communication Technology* (ICT) enable the automation of parts of this process reducing human workload.

Emerging semantic technologies provide tools to classify, filter, discover or associate information. *Natural Language Processing* (NLP) and *Artificial Intelligence* (AI) technologies have attracted much of the scientific interests in turning plain text into valuable data for analysis. The process of deriving high quality information from plain text is called *Text Mining* (TM). Among other features TM enables document classification and domain identification. This article presents a case study where TM is applied to the TW process as proposed by Jacquenet and LARGERON [72].

The article leverages NLP and AI to automatically classify documents according to the criteria established by a group of TW experts in the domain of forming, manufacturing and assembly processes. A catalogue of previously categorized documents is used to generate a multi-class model that classifies documents automatically. The resultant model enables a reduction on the amount of readings necessary for correct classification.

The research aims to identify and analyze technological alternatives provided by artificial intelligence and natural language technologies for automatic classification of content in plain text format in the field of technology watch.

The main objectives are:

1. Identify the technologies and tools that enable automatic categorization of documents.
2. Analyze various alternative algorithms in a context of multi-class classification.
3. Reduce the amount of document readings TW human agents have to perform creating an automatic classification system.

Further, the article researches on improving automatic classification by including semantic annotation into the text available in the datasets.

6.2 Background

As reflected in the norm UNE 166006:2011 *R&D&i management: Technological watch and competitive intelligence system*, Technology Watch is an organized, selective and permanent process, to capture information from outside and inside the organization about science and technology. All of this in order to select, analyse, disseminate and communicate the information to turn it into knowledge, making decisions with less risk and anticipating change. This way, technology watch represents a key tool in the R+D+i process.

The process of transforming captured information into knowledge for the organization is known as *Competitive Intelligence* (CI). In its simplest form, it is a process of adding value to information, analysing and producing knowledge in an intelligent way [74]. The *Society for Competitive Intelligence Professionals* (SCIP)¹, identifies five steps in the process². Those stages can be seen on Figure 6.1 and are described below:

¹<https://www.scip.org/>

²http://www.dialog.com.tw/download/docs/63/CI_Handbook.pdf

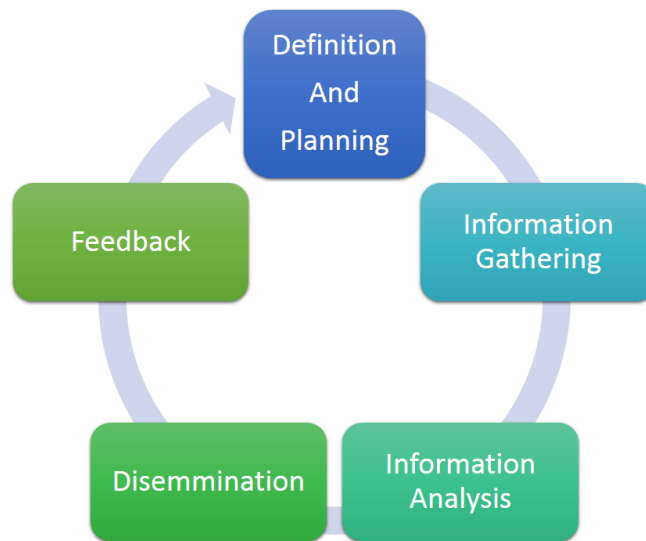


Figure 6.1: Five stages of the Technology Watch process according to SCIP.

- **Definition and Planning:** defining the intelligence question based on the intelligence customer's needs and planning your activities to meet those requirements.
- **Information Gathering:** collecting information from primary (people) and secondary (print) information sources, both internal and external to your organization.
- **Information Analysis:** analyzing the information and creating findings or scenarios.
- **Dissemination:** disseminating the intelligence (deliver the findings to decision makers).
- **Feedback:** receiving feedback on how the delivered product met the intelligence needs.

One of the most important issues TW processes face is the time spent on non productive stages or tasks of the process. This happens usually because of the great amount of data collected in the process and the burden of analysing, categorizing and filtering those data. Some of those tasks can be automatized.

The work presented in this chapter is part of a project which objective is to build a system that automatically classifies new content. The content is classified according to the categorization criteria used by a group of TW experts, specifically in the domain of forming, manufacturing and assembly processes. The work has been developed at Koniker S.Coop³ using a catalogue of documents previously categorized by a group of TW experts in those domains. Koniker provides TW services for several companies belonging

³<http://www.koniker.coop>

to Mondragon Corporation, one of the leading Spanish business groups, integrated by autonomous and independent cooperatives (about 250 companies and 74000 employees).

The technology watch process followed in this company is very similar to the one in figure 6.1. *Definition and Planning* and *Information Gathering* stages involve identifying data sources and collecting the information they provide, implicating the reading of a large amount of electronic texts in different formats. This information must be filtered, analyzed and categorized (*Information Analysis*), based on previously established criteria. TW experts complete the process by adding value, documenting and sending to the clients of the TW process (*Dissemination*).

Estimations made by Koniker indicate that only 35% of the experts working time can be considered as added value contribution, and therefore a 65% the dedication time is susceptible for automation.

Our goal is to save time by building an engine that allows filtering and classifying information from a technology watch system automatically, reducing the non productive time.

6.3 Theory

Emerging ICT have revolutionized the TW field by offering new options when seeking, treating and gathering information. One of the techniques that has attracted much of the scientific interest is *Text Mining* (TM) or *Text Data Mining* [84] [135] [18]. It is based on NLP and AI and it refers to the process of deriving high-quality information from text. TM is now a wide area of research that provides useful techniques that can be used in the context of technology watch [70] [41] [7]. According to Jacquenet and Largeron[72], the term appears for the first time in 1995 (Feldman [49]) and was defined by Sebastiani [116] as the set of tasks designed to extract potentially useful information, by analysis of large quantities of texts and detection of frequent patterns.

The NLP and AI technologies used for this research for TM are explained next.

6.3.1 Natural Language Processing (NLP)

Natural Language Processing (NLP)'s objective is to enable computers to make sense of human language.

In 2003, Chowdhury [28] described NLP as “*an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things*”. The human language has a structure, called grammar, and understanding that structure is one of the biggest efforts in NLP.

Electronic text is essentially a sequence of characters, some of which are content characters and other are control and formatting characters. Mitkov [90] proposes to perform several NLP tasks over electronic text: tokenization, elimination of function words and stemming.

First of all, text (sequence of characters), needs to be segmented into linguistic units, such as words, numbers, etc. This process is called *Tokenization* and segmented units

are called word tokens.

Among the main part of the speech, content words, such as nouns, verbs and adjectives, are the ones carrying most of the semantics, where as function words such as preposition pronouns and determiners have less impact on determining what a text is about. Elimination of those function words is another task commonly applied by the research community.

Stemming conflates morphologically related words to the same root. For some languages consist of stripping the end of words relating them with their stem or root.

Another of the uses of NLP is *Entity extraction* (or *Semantic annotation*). It is used to identify proper nouns and other specific information from plain text, mapping terms to concepts. For example, the text "Resource Description Framework" should map to the same concept as the text "RDF". On the contrary, the term "Apple" can be used to refer to a fruit or a company, so entity extraction tools should return a different *Uniform Resource Identifier* (URI) for each term. At DBpedia, the fruit's URI is <http://dbpedia.org/page/Apple> and the company's http://eu.dbpedia.org/page/Apple_Inc..

In conclusion, NLP can be used to identify concepts from text, enabling the identification of the elements appearing in that text. It can also be used to reduce the amount of text to be fed to learning algorithms, identifying non relevant words, articles, words with very few amount of occurrences...

6.3.2 Artificial Intelligence (AI)

Artificial Intelligence (AI) or *Computational Intelligence* "is the study of the design of intelligent agents" [104]. An agent is something that acts in an environment and an intelligent agent is an agent that acts doing something appropriate for its circumstances and its goals.

Baharudin et. al. [9] propose several AI learning algorithms for text mining. Among them, we selected three belonging to different approaches in order to compare their performance in our datasets:

1. **Support Vector Machine (SVM):** based on kernel equations that separates instances using a hyperplane on the multi-dimensional space. SVM classifier has been recognized as one of the most effective text classification method in the comparisons of supervised machine learning algorithms.
2. **Decision tree (J48):** a machine learning model that generates a decision tree where its branches preserve the possible values that the attributes can have in the observed samples. The main advantage of decision tree is its simplicity in understanding and interpreting, even for non-expert users. Besides, the explanation of a given result can be easily replicated by using simple mathematics algorithms, and provide a consolidated view of the classification logic, which is a useful information of classification.
3. **Naive Bayes:** a simple algorithm that observes attributes individually, independent from each other based on the rule of conditional probability. Naive Bayes

works well on textual data, easy to implement comparing with other algorithms.

In order to assist TW processes, automatic document classification can be achieved using AI and NLP technologies. Systems can be trained giving them some examples as training set. They can learn how to classify new documents based on some of the features of previous documents and their categorization, reducing the amount of non productive time in the TW process. One of the main objectives of the research is to test how well can an AI system replace that non productive work of an expert.

6.4 Material and methods

In this paper TM is employed for document classification using a previously categorized catalogue of documents available in Koniker. Subsection 6.4.1 explains the datasets used in this research based on that catalogue. Section 6.4.2 will describe the tools used in the research. Finally, the experiments performed are presented on section 6.4.3.

6.4.1 Datasets

Different nature datasets are used in this research. Three types of data are considered for the experiments:

1. A catalogue of previously categorized documents. That catalogue contains 7379 instances (or documents), categorized according to 14 classes. The text of those documents is referred as “Raw Text” in this paper.
2. Most of the documents of the catalogue have additional content added by experts on each class. That content is a title and a summary for each document with value-added information. The additional content is our second dataset, referred as “Experts’ information” in this paper. 6968 instances have this additional content stored in the database.
3. Finally, Semantic text annotations are extracted from both “Raw Text” and “Experts’ information”. Annotations are gathered using DBpedia Spotlight API (see subsection 6.4.2) for both “Raw text” and “Experts’ information”. Those annotations are URIs of elements mentioned in those texts. They are used as additional input attributes for the algorithms. These annotations are our third dataset, referred as “Semantic annotations” in this paper.

It is important to notice that the documents are written in three different languages (English, Spanish and Basque), making the classification problem more difficult.

6.4.2 Tools

To apply text mining algorithms to our datasets, we used the WEKA project. The WEKA project [59] *“aims to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers and practitioners alike. It allows*

users to quickly try out and compare different machine learning methods on new data sets". Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka is open source software issued under the *GNU General Public License*.

Different Weka filters were applied. A Weka *StringTowordVector* filter was applied to the input string with *IteratedLovinsStemmer*, *AlphabeticTokenizer* and *Spanish, English and Basque* stop-words included in one file in order to process the input. Default values for Weka were used for the rest of the parameters. In order to create the models, the 10 fold cross validation technique was used.

In order to get text annotations, DBpedia Spotlight API [40] was used. Providing plain text fragments to the service, it returns URIs of found resources mentioned within the text. Those URIs are resources from DBpedia [79] repository, that contains encyclopedic knowledge from Wikipedia for about 3.5 million resources, enabling access to many data sources in the Linked Data cloud.

DBpedia Spotlight returns the most feasible approach URIs of the elements that are mentioned in the giving text. Those URIs were given to our system as additional input data, adding more features to the algorithms. The objective was to test if adding semantic information to the plain text could improve the results of the created classification models.

6.4.3 Experiments

Using previously described datasets and modifying the factors outlined below, different experiments were performed.

The first modified factor was the learning algorithm, previously explained on section 6.4.2, namely "SVM", "J48" and "Naive Bayes".

The output given by the classification algorithms is not always a unique class, they usually give a probability distribution for the different classes. Therefore, the second factor modified was the amount of classes taken into account. As "1st hit" is considered the class with the highest probability, as "2nd hit" is the class with the 2nd highest probability, and the third class with the highest probability as "3rd hit" (see figure 6.2).

The third modified factor was the attributes given as an input to the algorithms. Those attributes were:

1. Text: these attributes are taken from the text of the instances, "Raw text" or "Experts' information".
2. Text + URI: same attributes as above plus the "Semantic Annotations" gathered from that text.

The attribute selection was made applying the following filtering steps: (1) Tokenizing, (2) Stemming, (3) Applying stop-words and (4) Word frequency (>3).

We identified two types of agents in the system: (1) *Experts*, agents that add value information to documents of an specific class; (2) *Deliverer*, agent that categorizes documents and sends them to the right experts (no human error was considered).

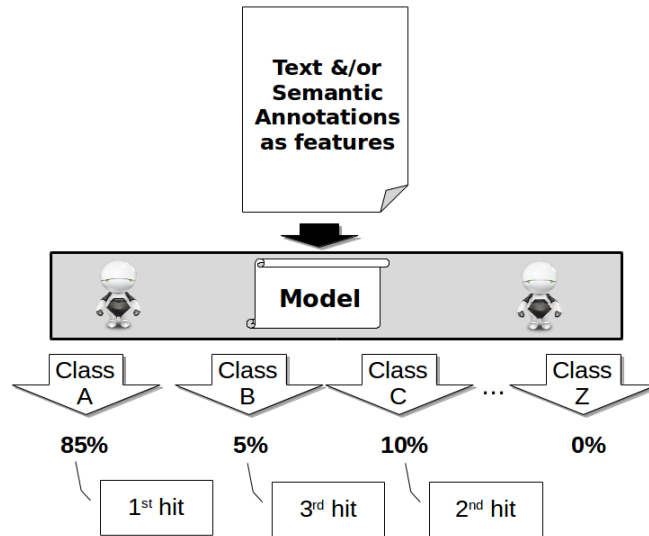


Figure 6.2: Single document's hits.

We quantify performance of different approaches in terms of precision and recall defined for multi-classification tasks [118] but also the human workload, measured as the amount of documents read by human agents. Each experiment considers 4 possible scenarios, which formulas to calculate the amount of readings are explained next:

- **No interactions:**

This is the base scenario. A human deliverer reads, categorizes and sends each document to the proper class expert. Then, the expert adds valuable information to that document (see figure 6.3).

$$\text{Amount of readings} = \text{'Amount of documents'} * 2$$

- **1st hit:**

In this scenario, a computer automatically classifies each document according to the NLP and AI models generated, replacing the work done by the deliverer in the “no interactions” scenario. If the document is correctly classified, the expert adds value information. Otherwise, it is sent to a human deliverer for correct re-classification. Finally, the correct expert adds value information to the document (see figure 6.4, steps 1-6-7).

$$\text{Amount of readings} = \text{'Amount of documents'} + \text{'Amount of misclassified on 1st hit'} * 2$$

- **2nd hit:**

In this scenario same process as in “1st hit” scenario is followed. However, in this case, wrongly classified documents are sent back to the computer for a second time (“2nd hit” classification). Wrongly classified documents for a second time are sent

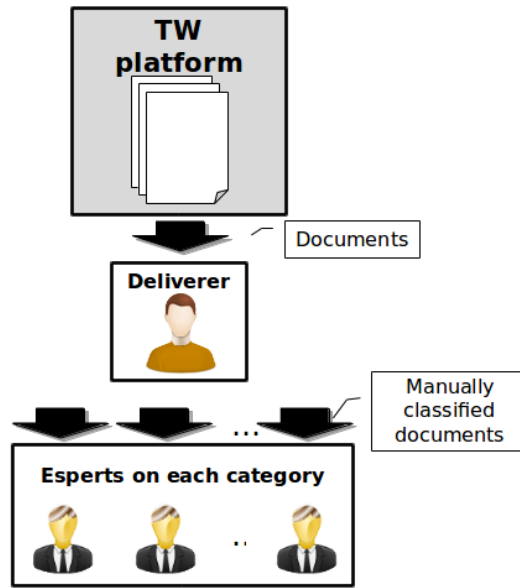


Figure 6.3: Information flow with no interactions.

to a human deliverer for correct re-classification. Finally, previously misclassified documents are read by the correct experts (see figure 6.4, steps 1-2-3-6-7).

*Amount of readings = 'Amount of documents' + 'Amount of misclassified on 1st hit' + 'Amount of misclassified on 2nd hit' * 2*

- **3rd hit:**

In this scenario, the same approach is followed for the third time (“3rd hit”). As in the previous scenarios, at the end, a human deliverer reads wrongly classified documents and sends them to the correct class experts (see figure 6.4, all steps).

*Amount of readings = 'Amount of documents' + 'Amount of misclassified on 1st hit' + 'Amount of misclassified on 2nd hit' + 'Amount of misclassified on 3rd hit' * 2*

Two experiments were performed using this approach. Each of them used a different input text for creating a different model and for testing (see figure 6.5). The models created in the first experiment are trained using the “Raw text” dataset. The models for the second experiment are trained using the “Experts’ information” dataset.

The objective of the second experiment is to compare its behaviour with that of the first experiment. Both use different datasets but a common categorization tree structure (see figure 6.6). We want to check if non previously treated information (“Raw text” dataset) is as useful as a previously treated information (“Experts’ information” dataset) for automatic classification.

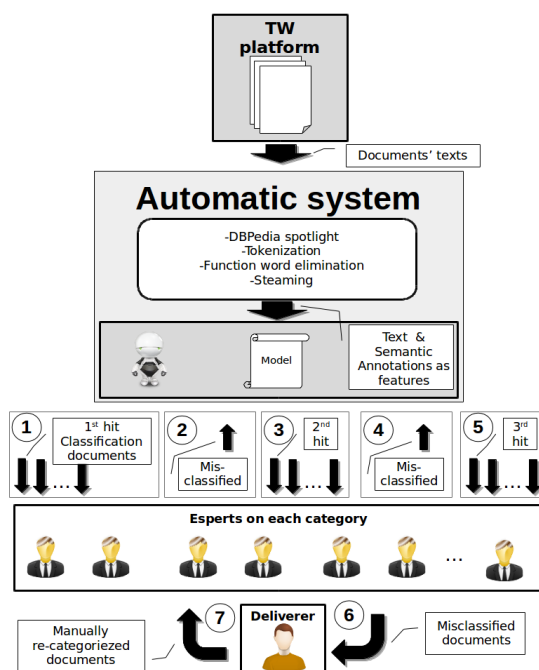


Figure 6.4: Information flow using automatic system.

6.5 Results

Table 6.1 shows the results of the tests for the first experiment, using the selected algorithms. The 3 columns under “Raw text” present the results for the three algorithms used in the research with the first dataset (7378 documents). In a similar way, the 3 columns under “Raw text + URI” use the same algorithms but adding the “Semantic annotations” of the “Raw text” to the previous dataset. Results for each performance indicator in each scenario are shown in each row.

Table 6.2 shows the results for the second experiment, that uses the second dataset (“Experts’ information”) as the input text for training the models. The results are shown in the same way as in the previous table.

Further data analysis revealed that most of the misplaced elements came from their own superclass (see figure 6.6 to see the class hierarchy). Table 6.3 shows the confusion matrix of the model with the best results, this is, the model generated using the J48 algorithm (1st hit and Raw text). Each column of the matrix represents the instances in the predicted class while each row represents the instances in the actual class. Diagonal values represent correctly predicted classes. The background color of each cell of the confusion matrix depends on the amount of elements in the cell, in order to see in an easier way the amount of migrated elements.

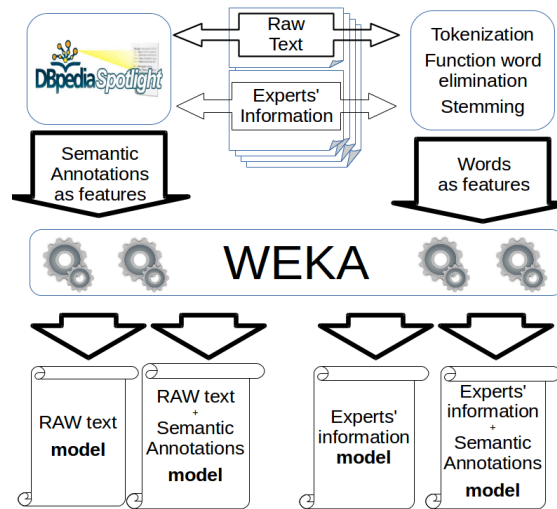


Figure 6.5: Model generation.

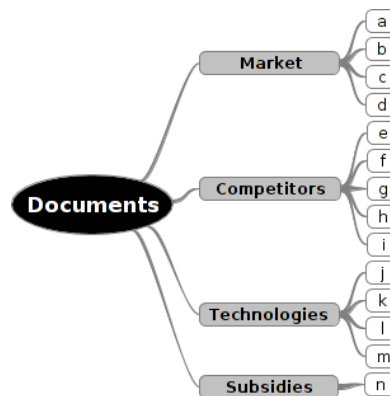


Figure 6.6: Class hierarchy.

Below the relation of the background color and the amount of element is defined for Table 6.3:

- 0 → White.
- 1-10 → Light grey.
- 11-25 → Grey.
- 26-50 → Dark grey.
- >50 → Black.

7379 Instances		Raw text (36635 attributes)			Raw text + Semantic annotations (46336 attributes)			No intervention
		J48	SVM	Naive Bayes	J48	SVM	Naive Bayes	
Accuracy	1st Hit	0,978	0,947	0,908	0,977	0,946	0,908	1
	2nd Hit	0,969	0,900	0,908	0,968	0,898	0,908	
	3rd Hit	0,963	0,837	0,908	0,964	0,836	0,908	
Precision	1st Hit	0,846	0,632	0,359	0,840	0,622	0,359	1
	2nd Hit	0,731	0,401	0,359	0,732	0,392	0,359	
	3rd Hit	0,687	0,287	0,359	0,691	0,283	0,359	
Recall	1st Hit	0,846	0,632	0,359	0,840	0,622	0,359	1
	2nd Hit	0,886	0,801	0,359	0,882	0,784	0,359	
	3rd Hit	0,893	0,860	0,359	0,888	0,849	0,359	
FScore	1st Hit	0,846	0,632	0,359	0,840	0,622	0,359	1
	2nd Hit	0,801	0,534	0,359	0,800	0,523	0,359	
	3rd Hit	0,776	0,430	0,359	0,777	0,425	0,359	
Amount of readings	1st Hit	9659	12807	16843	9736	12956	16838	14758
	2nd Hit	10201	13027	21565	10305	13347	21568	
	3rd Hit	10938	13626	26292	11085	13981	26298	

Table 6.1: Results using “Raw text”, without and with “Semantic Annotations” (1st experiment).

classified as ->	a	b	c	d	e	f	g	h	i	j	k	l	m	n
a	352	6	7	3	8	13	3	2	8	1	3	1	3	10
b	5	412	26	3	5	5	1	2	4	1	8	2	17	1
c	12	18	766	2	4	6	5	1	5	1	3	0	7	6
d	2	0	2	313	2	1	0	19	0	5	2	1	1	1
e	2	7	4	2	562	16	7	4	126	1	9	8	7	2
f	20	10	3	0	23	737	21	2	7	1	2	3	3	3
g	5	3	3	0	5	21	388	3	11	1	7	7	4	2
h	1	0	2	16	4	0	4	391	12	5	2	1	1	2
i	4	8	4	2	93	5	3	5	377	3	4	2	7	2
j	0	2	0	9	2	4	1	11	0	273	3	0	1	0
k	4	4	7	1	6	5	2	3	15	0	633	3	42	0
l	3	3	2	0	4	3	10	1	5	0	7	449	13	2
m	3	22	11	0	11	4	6	3	5	3	41	5	284	1
n	10	2	7	0	4	6	2	2	0	1	0	1	1	302

Table 6.3: Confusion matrix of the model with the best results (J48 algorithm - 1st hit - RAW text).

In a similar way, Table 6.4 shows the confusion matrix for the classes group accordingly to the superclasses they belong to (market, competitors, technologies and subsidies). Table 6.5 shows the classification process results in terms of accuracy, precision, recall and FScore. Those results use the same metrics as in table 6.1 and 6.2, but considering data grouped by superclasses.

6968 Instances		Experts' information (6804 attributes)			Experts' information + Semantic annotations (8483 attributes)			No intervention
		J48	SVM	Naive Bayes	J48	SVM	Naive Bayes	
Accuracy	1st Hit	0,975	0,958	0,935	0,974	0,962	0,931	1
	2nd Hit	0,958	0,912	0,934	0,960	0,914	0,931	
	3rd Hit	0,948	0,849	0,934	0,950	0,850	0,931	
Precision	1st Hit	0,824	0,709	0,542	0,819	0,732	0,519	1
	2nd Hit	0,655	0,442	0,536	0,667	0,450	0,515	
	3rd Hit	0,590	0,313	0,536	0,604	0,316	0,515	
Recall	1st Hit	0,824	0,709	0,542	0,819	0,732	0,519	1
	2nd Hit	0,873	0,884	0,548	0,868	0,900	0,524	
	3rd Hit	0,887	0,940	0,548	0,879	0,949	0,524	
FScore	1st Hit	0,824	0,709	0,542	0,819	0,732	0,519	1
	2nd Hit	0,748	0,589	0,542	0,754	0,600	0,519	
	3rd Hit	0,709	0,470	0,542	0,716	0,475	0,519	
Amount of readings	1st Hit	9421	11019	13355	9492	10708	13676	13934
	2nd Hit	9962	10615	16459	10072	10232	16962	
	3rd Hit	10646	10636	19606	10841	10245	20282	

Table 6.2: Results using “Experts’ information” , without and with “Semantic Annotations” (2nd experiment).

superclassified as ->	Market	Competitors	Technologies	Subsidies
Market	1929	94	56	18
Competitors	96	2827	78	11
Technologies	71	101	1757	3
Subsidies	19	14	3	302

Table 6.4: Superclass Confusion Matrix

Below the relation of the background color and the amount of element is defined for Table 6.4:

- 0-49 → White.
- 50-99 → Light grey.
- 100-250 → Dark grey.
- >250 → Black.

6.6 Conclusions

The 4th objective of the thesis was to identify ICTs that can reduce the workload of the experts in the TW process. More specifically, this chapter focused on automatically

	Market	Competitors	Technologies	Subsidies
Accuracy	0,952	0,947	0,958	0,991
Precision	0,912	0,931	0,928	0,904
Recall	0,920	0,939	0,909	0,893
FScore	0,916	0,935	0,918	0,899

Table 6.5: Superclass results.

classifying documents. In the state of the art, we have identified NLP and AI techniques that could help us classifying those documents. We also identified some tools (DBpedia spotlight and Weka) that enable us to develop a solution. The experiments carried out confirm that the solution automatically classifies the documents, giving satisfactory results.

Another objective of this chapter was to analyze various alternatives in order to test the best solution for the classification problem. The results show that J48 and SVM algorithms give positive outcomes in all cases, *J48* algorithm using *RAW text* is clearly the best solution. This way, we answer the 4th research question of the thesis defined in chapter 1: *Which are the best semantic technologies that help reducing the time spent on non-productive tasks inside the TW process?*

Taken “Semantic annotations” into account for multi-classification, the results do not seem to improve. The algorithms with just the text (“Raw text” or “Experts’ information”) seem to perform nearly identically to the ones with annotations. The classes may be too similar for text annotation to be relevant in this case. They may be useful for other cases where the classes have clearly different topics. Testing “Semantic annotations” with more heterogeneous data is proposed as future work.

The second experiment based on “Experts’ information” presents similar results to those of “Raw text”. This shows that using the “Raw text” of documents does not have to be previously treated by humans for a feasible automatic classification. Both models behave similarly.

If we consider second and third classes (or “hits”), there is only one case where the second hit gives better results than with the first, reducing the amount of readings: the second experiment using SVM. This proves that the first “hit” is the scenario with the largest reduction of human workload. Additional hits increase the recall, but they also augment the false positive classifications, generating a larger amount of readings.

The third objective of this chapter was to reduce the amount of readings performed by human agents. J48 and SVM reduce the amount of reading in all cases (a reduction of readings of 34.55-24.89% with J48 and 13.22-5.26% with SVM). Only Naive Bayes algorithm gives negative results, increasing the amount of readings. We conclude that the best results were achieved using J48 algorithm and “Raw text”, taken into account the first “hit” scenario, reducing the amount of readings a 34.55% (from 14758 to 9659).

Therefore, it can be concluded that the 4th objective of the thesis (*Reduce the workload of the experts in the TW process*) was achieved during the development of this chapter.

Chapter 7

Contribution

This chapter collects the work done during this research. It will describe the collaboration with universities and companies, scientific publications, projects involved during the research for founding and finally, the summer schools attended during the research period.

7.1 Cross university and company collaboration

During the research some collaborations with universities and companies have been made. In the following subsections those collaborations will be described and the works done in each of them. First, collaboration with some universities will be presented: DERI, from NUI; and UPM. Next, collaboration with some enterprises will be presented: ISEA and Koniker. Finally, the collaboration within the different parts of MU itself will be presented.

7.1.1 Collaboration with *Digital Enterprise Research Institute (DERI) of the National University of Ireland (NUI)*

DERI¹ is a Centre for Science, Engineering and Technology (CSET) established in 2003 with funding from the Science Foundation Ireland. It has become an internationally recognised institute in semantic web research, education and technology transfer. Therefore, and as DERI's researches are focused on the SW, it was choose as a good option in order to make a collaboration.

Tree months were spend in a collaborating project that linked some structured data spaces about sustainability with the innovation process. Thus, *IdeaMentions* module for *Innoweb* and some use cases were developed. As a result, a paper was published with the work done in DERI (see publication in subsection 7.3.3).

¹<http://www.deri.ie/>

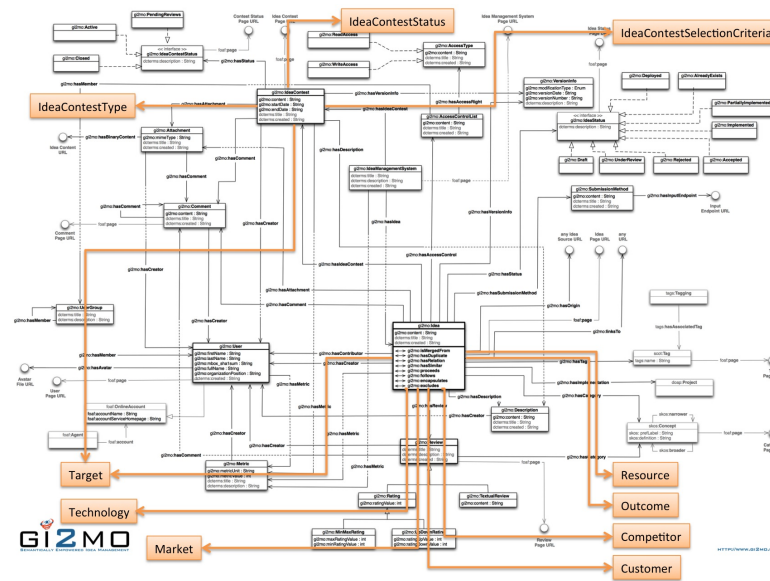


Figure 7.1: UML Class diagram for GI2MO Wave Ontology (based on Gi2MO Ontology diagram)

7.1.2 Collaboration with *Universidad Politécnica de Madrid (UPM)*

*Universidad Politécnica de Madrid (UPM)*² is a University founded in 1971 as the result of merging different Technical Schools of Engineering and Architecture, originated mainly in the 18th century. In the state of the art of this project, some works were found on Innovation process Ontologies involving UPM researching teams. Therefore, some contacts were done to collaborate in this researching areas.

As a result a new version of Gi2MO ontology was proposed, Gi2MO Wave (see Figure 7.1), more focused on the context of the ideas. The ontology has been and will be tested in future campaigns to collect context information about innovation processes. The specification of Gi2MO Wave ontology can be found at appendix A and it's latest version online³.

7.1.3 ISEA

Innovation in Advanced Business Services - ISEA S. COOP.⁴ is a private and non-profit innovation and entrepreneurship Centre, specialized in Business Services Sector, promoted by the Division of Engineering, and Business Services of MONDRAGON Corporation.

ISEA S.COOP. is a Science and Technology Agent, integrated into the Basque Sci-

²<http://www.upm.es/internacional>

³<http://purl.org/gi2mo/wave/ns>

⁴<http://www.iseamcc.net/isea>

ence, Technology and Innovation Network. Today is part of the Basque Innovation Agency - INNOBASQUE. Additionally, ISEA is a Certified Agent of the Basque Entrepreneurial Service of the Society for Competitive Transformation - SPRI.

In line with its corporate purpose, ISEA S. COOP. has promoted a Business Acceleration Center (BAC), a specialized structure designed to boost the process of launching new business initiatives in the Advanced Services Sector. In this context, ISEA used the developed platform to launch campaigns and collect ideas with an open innovation philosophy. The project is called *Elkarbide*⁵.

7.1.4 Koniker

Koniker S.COOP.⁶ is a non-profit Technology Centre of public utility, specialising in the research and development of new technologies related to forming and assembly processes.

Koniker provides services for the following companies: Fagor Arrasate, Batz, Mondragon Assembly, Onapres, Aurrenak, Matrici, Loramendi and Mondragon Corporation.

It was established in 2002 with the aim of giving a coordinated response to the R&D projects of all of these companies and of making the most of synergies present in the forming, assembly, machinery and tool areas. They have adopted Innoweb in a platform called *Ideak*, in order to collect ideas in the companies they work for. Using the findings of this case study two publications were made (see publications on section 7.3.1 and 7.3.2).

One of the main services provided by Koniker is consultancy on Technology Watch issues. Koniker possesses an important catalog or repository with data gather by means of TW actions. They have identified their TW workflow and the amount of time spent on each task (see figure 7.2). They consider that many of the tasks are unproductive and could be automatize by using ICTs. They consider that up to 55% of that time could be reduced.

7.1.5 Mondragon Unibertsitatea (MU)

Innoweb has also been used inside of the University, creating an idea gathering platform called *Ekiten*. The main objectives of EKITEN are two:

1. The development of entrepreneurial spirit and ability of students.
2. The launch of new business initiatives.

In Ekiten, multidisciplinary teams have been made. Those teams are made of students from different faculties of the University with different profiles working together and contributing with their own specific job profile. Also, all the equipment and therefore all projects will have a tutor who will guide the students and, moreover, will also feature the expertise of University faculty in the areas involved in the project .

⁵<http://www.elkarbide.com/>

⁶<http://www.koniker.coop/en>

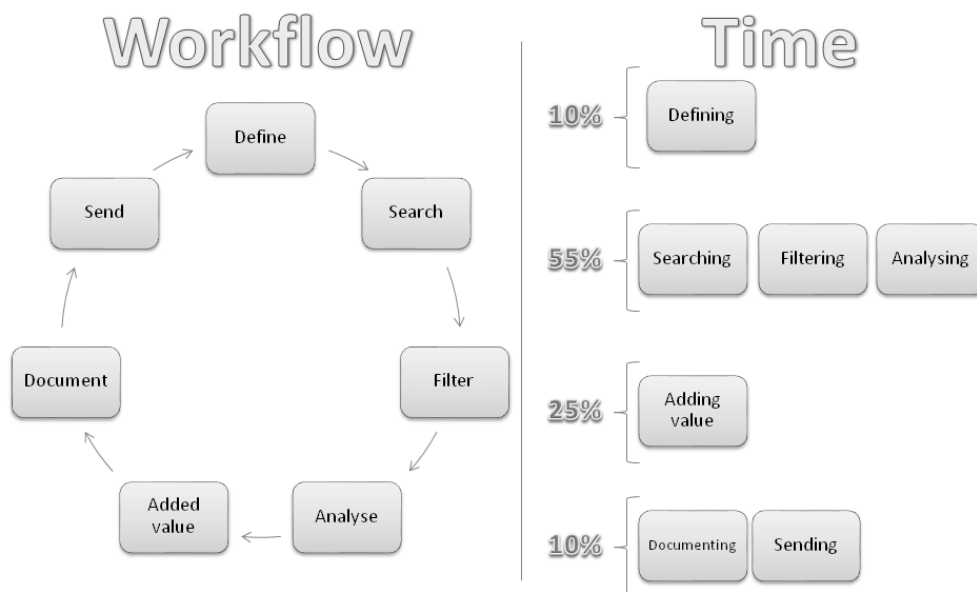


Figure 7.2: Current Technology Watch workflow and time spent in each step according to Koniker.

This project is a collaboration between MU and the Center for Business Innovation Saiolan, and has the support of the Department of Innovation and Information Society of the Provincial Council of Gipuzkoa and the Department of Industry of the Basque Government.

7.2 Projects

The team has been involved in many projects during the research for founding. Those project were and are being developed collaborating with some universities and enterprises described in section 7.1. Some collaboration have not been in direct contact, but with some of previously mentioned collaborators as intermediary. Below, those project are listed.

- **SELENE**: *Social and Semantic Web Environment for Network Innovation*, National Applied Research Program (Madrid) along with LKS, ISEA and Deusto Unibersity (2009-2011).
- **COLABORANOVA**: *Collaborative Innovation System Based on Participation*, Provincial Council of Gipuzkoa (2010).
- **ELKARWEB**: *Social and Semantic Web Platform to Support Collaborative Innovation*, GAITEK program from Industry Department of the Basque Government with KONIKER and DANOBAT (2011).

- **PATENT-AWARE:** *Support System for Strategic Decision and Patentability in the Field of Home Energy*, GAITEK program from Industry Department of the Basque Government with LKS INGENIERÍA (2012-2014).
- **INNES (Strategic Innovation):** *Smart Platform for Strategic Innovation in the Field of Health*, GAITEK program from Industry Department of the Basque Government with AURRENAK, LORAMENDI and PROSPEKTIKER (2013-2015).
- **K-INNES:** *Smart Platform for Strategic Innovation in the Field of Health*, Servicing for KONIKER.
- **ACCELERATE:** *A Platform for the Acceleration of GO-TO Market in the ICT Industry*. AEESD (Acción Estratégica Economía y Sociedad Digital) program with PLANET MEDIA STUDIO, SIVSA SOLUCIONES INFORMATICAS, S.A. and FAGOR ARRASATE (2013-2015).

7.3 Scientific Publications

Previously mentioned collaborations have been performed creating some tools and achieving some results for the research. Those tools and results have been gathered and 5 papers have been written so far. This section will describe those scientific publications reviewed by the community. Firstly, a publication of the base platform to support an innovation project is presented [78]. Next, a publication of the prototype adding semantic capabilities and context information gathering is described [103]. Thirdly, the publication where the work done in the abroad university internship is presented [102]. After that, a paper with the work done for Innovation and Technology Watch interoperability sent to *International Journal on Semantic Web and Information Systems* is shown. Finally, the paper sent to *Data and Knowledge Engineering* about Machine Learning techniques for supporting Technology Watch process is described.

7.3.1 *A case study on the use of community platforms for inter-enterprise innovation, 2011, 17th International Conference on Concurrent Enterprising (ICE 2011)*

This paper is enshrined in a larger research project which main goal is to determine key factors in the design, implementation and use of collaborative environments for the management of inter-enterprise innovation processes based on practical experiences. The paper presents an innovative approach to address the challenge of collaboration and participation in the submission of new ideas in the front end of the innovation process. The paper describes a case study on the adoption of an open source community platform based on Drupal in the context of a set of cooperative companies within Mondragon Corporation. The platform leverages social computing, real-time Web and semantic technologies to support collaboration, basic technology watch and idea management in the early stages of the innovation process. Preliminary field results show that the platform provides a

powerful collaborative tool that eases administration work and enhances collaboration among participants.

7.3.2 *INNOWEB: Gathering the context information of innovation processes with a collaborative social network platform, 2013, 19th International Conference on Concurrent Enterprising (ICE 2013)*

This paper describes the development of a collaborative social platform to support innovation process management. The Drupal based platform accommodates different types of innovation processes (also called waves or idea contests), enhances collaboration and eases management. The main contribution lies on the gathering of context parameters which helps enterprises on the detection of critical success factors, enabling later reproductions. The development leverages open source social computing, real-time web and semantic web technologies adding new functionality in a modular way. Blogs, wikis, graphical tools and voting systems support collaboration and idea management in the early stages of the innovation process. A workflow module launches customized innovation campaigns where topics, participants, stages, selection criteria and communication methods are optimized to enterprise needs.

7.3.3 *The Role of Linked Data and Semantic-Technologies for Sustainability Idea Management, 2013, 2nd International Symposium on Modelling and Knowledge Management for Sustainable Development (MoKMaSD 2013)*

Idea Management Systems (IMSs) manage the innovation lifecycle from the moment of invention until ideas are implemented in the market. During the lifecycle the IMS supports collaboration, allows idea enrichment with comments, contextual data, or connected to other relevant ideas. Semantic technologies can improve the knowledge management capabilities of IMSs allowing relevant information to be easily linked to ideas. Many Enterprises concerned with sustainability encourage employee's participation as a means to boost creative innovation within their Sustainability Initiatives. However little work has examined the role of an IMS within Sustainability. In this paper we analyse the impact of a semantic-enabled IMS within a sustainability innovation process. In particular, how ideas can be enriched with contextual *Linked Open Data* (LOD), especially *Life-cycle Assessment* (LCA) data, to improve the understanding, implication and value of the idea from the sustainability perspective.

7.3.4 *Semantic Annotations to enhance Innovation and Technology Watch Interoperability, pending at International Journal on Semantic Web and Information Systems (IJSWIS)*

Innovation and Technology Watch processes are very important matters for companies to keep up to date with competition. Having the data from these processes interoperable, understandable and linked to similar content is very important. This enables extracting

the largest amount of information in the shortest time. Therefore, this research focuses on automatically identifying concepts mentioned in ideas from Innovation platforms and news from Technology Watch platforms. Thus, three goals can be achieved: (1) to gather the definition of these concepts so the content can be understood, (2) to link similar ideas checking which of them do mention the same concepts and (3) to link ideas and news with similar topics. The experiment conducted in this research demonstrate that most of the mentioned critical concepts can be identified and that they can help to link related content.

7.3.5 *A case study on the use of Machine Learning techniques for supporting Technology Watch, pending at Data and Knowledge Engineering (DKE) journal*

Technology Watch human agents have to read many documents in order to manually categorize and dispatch them to the correct expert, that will later add valued information to each document. In this two step process, the first one, the categorization of documents, is time consuming and relies on the knowledge of a human categorizer agent. It does not add direct valued information to the process that will be provided in the second step, when the document is revised by the correct expert.

Machine learning tools and techniques can be used to learn from the manually pre-categorized data to automatically classify new content. For this work a real industrial context was considered. Text from original documents, text from added value information and Semantic Annotations of those texts were used to generate different models, considering manually pre-established categories. Finally, the results obtained were compared to select the best model in terms of accuracy and also on the reduction of the amount of document readings (human workload).

7.4 Summer Schools

Finally, as part of the research, two participations in different Summer Schools were performed. The first one aimed to increase the team's knowledge on Semantic Web Technologies and their application. The second one was more oriented to the doctoral research and information management. Below both summer schools are listed and their brief description:

- **The 9th Summer School on Ontology Engineering and the Semantic Web (SSSW 2012):** Presented by leading researchers in the field, it represented an opportunity for both students and practitioners to equip themselves with the range of theoretical, practical, and collaboration skills necessary for full engagement with the challenges involved in developing Ontologies and Semantic Web applications. To ensure a high ratio between tutors and students the school was limited to 50 participants.
- **Networks, Information, Technology and Innovation Management P.h.D.**

Summer School (NiTiM 2014): The objective of the NITIM School was to bring together highly talented and motivated young scholars in the field of Networks, Information, Technology and Innovation Management, to work collectively towards furthering their doctoral research. The School aimed to establish a forum for the development of the Ph.D. candidates' research, across various stages of the dissertation, through feedback from both peers and faculty members.

Chapter 8

Conclusions

This thesis has presented a number of solutions that contribute on the use of ICTs, more specifically Semantic Web and Semantic Technologies, to enhance SMEs' Innovation and TW processes. The main conclusions of the research performed in this thesis are gathered in this chapter. Moreover, possible future works are also outlined, taking into account the achievements of the thesis.

8.1 Conclusions

This thesis begun in late 2011 with the definition of the problematic and the goals of the research. After approximately four and a half years in development, this thesis has build some contributions to the state of the art of Innovation and Technology Watch processes. It should be noticed that the contribution area of this thesis has also changed in many ways during that period of time. Thus, concluding this thesis, a summary of the contribution and the significance of each one are outlined in this section.

Our motivation of the thesis was to improve the current situation of innovation and TW software solutions. In particular, we focus on the *data overflow* problem many systems have, the issues to *reproduce successful idea contests*, the lack of *interoperability among different platforms* and finally the *waste of time in non productive tasks* in TW processes. This way, a number of contributions have been delivered that can be gathered under four main contribution areas:

- **Model and platform for IMS context gathering.** Chapter 3 has focused on the development of a web platform to assist on the Innovation of companies, more specifically on IMSs. The platform, called Innoweb, has been built to gather context data that can be further analyzed to determine the influence of background information in the success of an idea contest. Moreover, the data of the platform has been also provided in semantic format, enabling better data interoperability. This platform has been tested in different real case studies, showing its ability to identify the elements that makes the idea contest successful and enabling their replication. Based on the research, the final contributions in the problem area of de-

scribing the contemporary state of Idea Management Systems are: (1) *the creation of a platform for multiple idea contests that eases the gathering of the background data for successful idea contest replication (Innoweb)*; and (2) *the formalization of the background data of idea contest in an ontology (Gi2Mo Wave)*. Due to the evaluation of those contributions, the main conclusions obtained are: (1) gathering background information of idea contests is key for successful idea contest identification and replication; and (2) formalization of this data using the ontology benefit in the area of system integration and eases future data usability and interoperability. The 2 contributions allow to gather more meta-data and show it to the Innovation process managers. This way, they have been provided with the tools to exploit that data and they have more information to make their work more accurately and wasting less time looking for meaningful context information. The contributions of this chapter represent the left column of the architecture defined in chapter 1, the column that represents the content related to the ideas.

- **Linking and reusing content from repositories inside and outside the company.** Chapter 4 has focused on analyzing the technologies that could interlink companies' data from internal and external repositories with IMSs, providing interoperability among their data and automate information extraction. In this research, real data from DERI's sustainability repositories has been used in order to test the identified methods in a real use case. The main contribution of this chapter has been identifying *how NLP, LOD and semantic web technologies can be used for the identification of mentioned elements inside ideas*. It has been also shown that the identification of those mentioned elements can be useful for cases such as: (1) energy reduction on sustainability IMSs, (2) Life-cycle Assessment repository data linking and (3) similar idea identification. This way can be concluded that using those technologies information from IMSs can be linked with companies' semantic repositories and exploited in different use cases. The contributions of this chapter represent the integration between the left and right columns of the architecture defined in chapter 1 through the 2 horizontal boxes that enable interoperability among different platforms (*Mentions implementation modules* and *Mentions Ontology*).
- **Linking text based content using mentioned element identification.** Based on the work done in the previous chapter, chapter 5 has focused on enabling interoperability among different platforms. More specifically, Innoweb and a TW platform have been selected to perform different tests, identifying the elements mentioned in the content of both platforms. Later on, those mentioned elements are used in order to find relationships between the content. Finally, an expert on the field of the content has rated those relationships to test the developed system. Based on the research, the main contributions of this chapter are: (1) *a method for linking text based content using mentioned element identification*; and (2) *the formalization of the mentioned elements in text based contents in an ontology (Mentions Ontology)*. Due to the evaluation of those contributions, the main conclusions obtained are: (1) the proposed method is useful for similar content identification; and (2) the

ontology represents the mentioned elements in a semantic way so they can be used for enabling interoperability among platforms. The contributions of this chapter represent the 2 horizontal boxes that enable interoperability among different platforms (*Mentions implementation modules* and *Mentions Ontology*).

- **Document auto-classification for reducing human workload on non-productive tasks.** Chapter 6 has focused on the data-overflow problem identified in this thesis. More specifically, it has been focused on the TW process and the classification of documents gathered in that process, considered as a *directly non-productive task*. Several AI and NLP technologies have been analyzed in order to automate the classification of documents. Those technologies have been fed by different inputs, testing and identifying the most useful ones in a real case study. According to the performed test the main contribution of this chapter is: *the identification of the best AI and NLP technologies on document classification tasks in a real case study for reducing the human workload*. Mentioning (or semantic annotation) function has also been tested in the case study, adding the identified elements as input features to the algorithms. Nevertheless, there has not been found any improvements in the results for this specific case study using the mentioning functionalities. In conclusion, the best results have been achieved using the J48 algorithm in an AI system, fed only by the RAW texts of the documents. This way, a reduction of the 34.55% has been achieved in the amount of readings made by TW human agents, representing a considerable reduction on the human workload in this case study.

The order in which the contributions of the thesis have been presented is due to the fact that each of the contributions lead to the next one. The platform and model generated in the first contribution built the basis for idea gathering and a first approach on the semantic representation of content. This led us to the identification of methods for semantically linking the content in the IMSs with other content that could be useful in the Innovation process. The methods identified in the second contribution pushed us to identify a more automatic method for linking text based content and enable interoperability with content from the TW process. In time, analysing the TW process took us to explore the tasks in this process. We identified some classification tasks not providing added value (non-productive). Using semantic technologies we built automatic tools that reduce the time consumed by humans in those non-productive tasks. Finally, this led us to the identification of the non-productive tasks that made the TW process so time consuming for human agents and building some automatic tools that could reduce that non-productive time. Finally, chapter 7 shows the interest on the research from other researchers and companies and how they collaborated on building the ontologies, tools and case studies to test the impact of the research.

8.2 Future Work

The work performed on this thesis has opened new opportunities related to the Innovation and TW processes and the interoperability among platforms. The experiments carried

out proved the usefulness of some proposed solutions and opened new possibilities for future development. Due to time constraints some functionalities and tests have not been addressed. Therefore, some of those identified tests and functionalities have been gathered in this section as possible future challenges:

- **Test linking functionalities in other scenarios.** Thesis area: interoperability. One of the future challenges could be to keep testing the presented different solutions in other real case studies. This way, we could test if they are valid also in other cases or if they improve in other scenarios. For example, we could test the methods identified in chapter 4 in domains different to sustainability and see how they behave. This would show in which domains the solution automatically extracts knowledge improving the Innovation process.
- **Case study that implement the whole architecture at once.** Thesis area: interoperability. Another future challenge could be testing the whole architecture and the created tools at once in a real case study. It could use Mentions Ontology to relate the content from several platforms, AI and NLP technologies to automate content classification, Innoweb to manage an IMS and use the content extracted from internal repositories as done with the sustainability scenario. This could measure the impact of the whole architecture in a unified use case.
- **Interoperability among other type of content.** Thesis area: interoperability. Interoperability has been tested between ideas from IMS and news from TW platforms. It would be interesting to test the proposed solutions between ideas and other types of content (patents, designs or any other text based content). Mention identification tools could be used to extract mentioned elements and then test if useful related content is identified.
- **Addressing the implementation phases of the innovation process.** Thesis area: interoperability. The study has been focused in the first steps of the innovation process (from their generation to their selection). Nevertheless, keeping track of the implementation of the ideas is also important in order to see why the ideas have been implemented or cancelled in their life cycle. If all the process is connected through all its life cycle, that information could be used in future idea contests and can help on the identification of ideas that will not be possible to implement.
- **Testing alternatives to DBpedia Spotlight API.** Thesis area: interoperability. The focus around entity annotators has increased in the last years, with several algorithmic approaches to solve the mention-entity match problem, using knowledge bases such as DBpedia, Freebase or Yago. It could be interesting to test other mention-entity matching tools comparing the results with the ones found during this thesis. In 2012 DBpedia Spotlight was found as one of the system giving the best statistics [111]. Another study with new datasets and systems achieve different results in 2013 [33]. Therefore, systems such as *TagMe 2*, *Wikipedia Miner* or *AIDA* can be tested with the content of this thesis. Then, gathered results can be compared and determine the best tool for this approach.

- **Mentioning element identification in more generic case studies.** Thesis area: data-overflow. Mentioning element identification did not add any improvements to the results in document auto-classification in chapter 6. This could be due to the fact that the documents used in this case study were from a very specific topic. Therefore, we could test if mentioning functionalities could achieve better results in a case study with more generic content.
- **Clustering for enhance document classification.** Thesis area: wasted time reduction. Additionally to classification algorithms, clustering methods could be applied to the documents of the experiments in chapter 6. The results of the clustering could provide experts additional information on how classes group enabling the merging of some categories and the validation of others. This could help them making a more useful document classification.
- **More visual tools development.** Thesis area: successful idea content identification. One of the most useful tools for idea contest managers are the visual tools. Managers identify success factors easier if they are presented visually than if they have to navigate and compare all the data manually. Therefore, more graphical tools can be developed to assist on success factor identification. For example, idea contest can be shown in graphs linked to the related events, most mentioned elements, ideas, participants... This will ease the evaluation of idea contests and identification of similarities and differences among each other, assisting on successful idea contest replication.

Appendix A

Gi2MO Wave Ontology Specification

This appendix contains the specification of Gi2MO Wave Ontology. This ontology is based on Gi2MO Ontology, but with additions in the idea management context. This context is added in order to enable the reproducing of successful idea contests (or Waves). The specification can be found in the next pages, and it's lattes version can also be fount online¹.

¹<http://purl.org/gi2mo/wave/ns>

Gi2MO Wave Ontology Specification

V0.1 - 04 June 2012

This version: <http://purl.org/gi2mo/wave/0.1/ns> (RDF/XML, HTML)

Latest version: <http://purl.org/gi2mo/wave/ns>

Editors: Felix Larrinaga, Osane Lizarralde, Alain Perez

Authors: Felix Larrinaga, Osane Lizarralde, Alain Perez

Contributors: See [acknowledgements](#)

This work is licensed under a [Creative Commons Attribution License](#). This copyright applies to the Gi2MO Wave Ontology Specification and accompanying documentation in RDF. This ontology uses W3C's [RDF](#) technology, an open Web standard that can be freely used by anyone.



Abstract

Gi2MO Wave is a standardized data schema (also referred as "ontology" or "vocabulary") designed to use for annotation and describing resources gathered inside Idea Management facilities. The following document contains the description of ontology and instructions how to connect it with descriptions of other resources.

Table of Contents

1. [Introduction](#)
 1. [Idea Management Systems and InnoWEB project context](#)
 2. [The Semantic Web](#)
 3. [What is Gi2MO Wave for?](#)
2. [Gi2MO Wave Ontology at a glance](#)
3. [Gi2MO Wave Ontology overview](#)
 1. [Example](#)
4. [Cross-reference for Gi2MO Wave Classes and Properties](#)

Appendixes

- A. [Changelog](#)
 - B. [Acknowledgements](#)
-

1 Introduction

The following specification is a formal description of metadata schema proposal that can be applied to data gathered in the so-called Idea Management Systems. The goal of the following section is to introduce both Semantic Web and Idea Management experts to the topic and goals of the ontology and provide the basic knowledge to comprehend the technical part of the specification.

An important note is that Gi2MO Wave Ontology is not a complete model of the Idea Management System. It extends the concepts of the main [Gi2MO Ontology](#) and defines concepts that were particular for Idea Management pipeline of InnoWEB project. See [goals of the Wave ontology](#) to see if this particular setting is fit for reuse in your case.

1.1 Idea Management Systems and InnoWEB project context

The Idea Management Systems are referred most often as an application used by organizations to collect input about various ideas regarding their products and services; and manage them afterwards providing certain assessment and screening facilities. Those kind of systems are the main research interest of the InnoWEB project that evaluates new techniques for collaborative innovation in corporate environment.

The InnoWeb is an project by the [Mondragon University](#), [ISEA institute](#) and companies of [Mondragon Corporation](#). One of the outcomes of the project is a collaborative innovation platform based on social software. The specific characteristic of the platform is the introduction of 'innovation waves' - an element similar to the notion of Idea Contest but imposing a number of limitations on the innovators and as well as idea reviewers (e.g. elements of idea descriptions that users have to provide).

The presented ontology is an extension to a regular Idea Management model as a consequence of introducing this notion of innovation waves.

For more details about the research and case studies conducted during InnoWEB please refer to a related publication from the 17th International Conference on Concurrent Enterprising (ICE 2011): "[A Case Study on the Use of Community Platforms for Inter-Enterprise Innovation](#)" by Felix Larrinaga, Igor Santos, Osane Lizarralde, Alain Perez.

1.2 The Semantic Web

The [Semantic Web](#) is a W3C initiative that aims to introduce rich metadata to the current Web and provide machine readable and processable data as a supplement to human-readable Web.

Semantic Web is a mature domain that has been in research phase for many years and with the increasing amount of commercial interest and emerging products is starting to gain appreciation and popularity as one of the rising trends for the future Internet.

One of the corner stores of the Semantic Web is research on inter-linkable and interoperable data schemas for information published online. Those schemas are often referred to as ontologies or vocabularies. In order to facilitate the concept of ontologies that lead to a truly interoperable Web of Data, W3C has proposed a series of technologies such as [RDF](#) and [OWL](#). **Gi2MO Wave** uses those technologies and the research that comes within to propose an ontology set in the domain of Idea Management.

1.4 What is Gi2MO Wave for?

The goals of the Gi2MO Wave Ontology to achieve as a data schema are:

- extend the notion of Gi2MO's Idea Contests to 'innovation waves' that can have a status, type, and different idea selection criteria
- force (limit to) a certain type of idea descriptions: Target, Market, Customer, Competitor, Outcome etc.

2. Gi2MO Wave Ontology at a glance

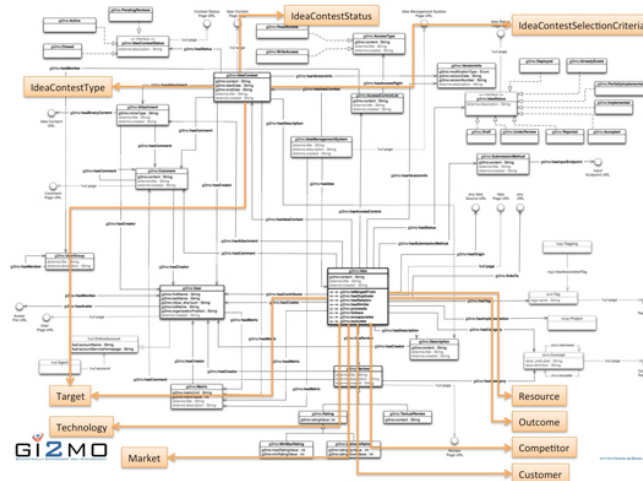
An alphabetical index of Gi2MO Wave terms, by class (concepts) and by property (relationships, attributes), are given below. All the terms are hyperlinked to their detailed description for quick reference.

Classes: | [Competitor](#) | [Customer](#) | [IdeaContestType](#) | [Market](#) | [Outcome](#) | [Resource](#) | [Target](#) | [Technology](#) |

Properties: | [hasCompetitor](#) | [hasCustomer](#) | [hasMarket](#) | [hasOutcome](#) | [hasOwner](#) | [hasResource](#) | [hasSelectionCriteria](#) | [hasTarget](#) | [hasTechnology](#) | [hasType](#) | [isCompetitorOf](#) | [isCustomerOf](#) | [isMarketOf](#) | [isOutcomeOf](#) | [isOwnerOf](#) | [isResourceOf](#) | [isSelectionCriteriaOf](#) | [isTargetOf](#) | [isTypeOf](#) |

3. GI2MO Wave Ontology overview

The GI2MO Wave UML diagram presented below shows connections between classes that implement the data model of Idea Management Systems.



UML Class Diagram for the GI2MO Wave Ontology (based on Gi2MO Ontology diagram), high resolution version: [PNG](#)

3.1. Example

A very basic example below shows a single idea annotated with Gi2MO and Gi2MO Wave metadata:

```
<gi2mo:Idea rdf:about="http://gi2mo.org/idea/012345">
  <foaf:page rdf:resource="http://gi2mo.org/ideaView?id=012345"/>
  <gi2mo:hasCreator rdf:resource="http://gi2mo.org/user/pedro"/>
  <gi2mo:content>A new, nice and modern building for the department that would have a similar interior design
  as shopping malls
  </gi2mo:content>
  <dc:terms:title>Department building with shopping mall interior design</gi2mo:title>
  <dc:terms:created>2012-04-23</gi2mo:created>
  <gi2mo:hasStatus rdf:resource="http://www.purl.org/gi2mo/ns#Implemented"/>
  <gi2mo:hasComment rdf:resource="http://gi2mo.org/comment/054321"/>
  <gi2mo:hasCategory rdf:resource="http://gi2mo.org/category/General"/>
  <gi2mowave:hasTarget rdf:resource="http://gi2mo.org/description/02345"/>
  <gi2mowave:hasCustomer rdf:resource="http://gi2mo.org/description/02346"/>
  <gi2mowave:hasOutcome rdf:resource="http://gi2mo.org/description/02347"/>
</gi2mo:Idea>

<gi2mowave:Outcome rdf:about="http://gi2mo.org/description/012345">
  <foaf:page rdf:resource="http://gi2mo.org/ideaView?id=012345"/>
  <gi2mo:hasCreator rdf:resource="http://gi2mo.org/user/pedro"/>
  <gi2mo:content>A new building would make people feel better in comparison to studding in the current one that has
  40+ years without renovation.
  </gi2mo:content>
  <dc:terms:title>Outcome: Department building with shopping mall interior design</gi2mo:title>
  <dc:terms:created>2012-04-23</gi2mo:created>
  <gi2mowave:isOutcomeOf rdf:resource="http://gi2mo.org/idea/012345"/>
</gi2mowave:Outcome>
```

4. Cross-reference for Gi2MO Wave classes and properties

Below see a comprehensive list of all Gi2MO Wave classes, properties and their descriptions.

Classes and Properties (full detail)

Classes

Class: gi2mowave:Competitor

Competitor - An object of this type indicates a competitor for an idea or IdeaContest

Status: unknown

Properties include: [isCompetitorOf](#)

Used with: [hasCompetitor](#)

OWL Class

[\[#\]](#) [\[back to top\]](#)

Class: gi2mowave:Customer

Customer - An object of this type indicates the customer targeted by an idea or IdeaContest

Status: unknown

Properties include: [isCustomerOf](#)

Used with: [hasCustomer](#)

OWL Class

[\[#\]](#) [\[back to top\]](#)

Class: gi2mowave:IdeaContestType

IdeaContestType - An object of this type indicates the type of innovation search by the IdeaContest. There are 12 dimensions or types identified. For a list of recommended instances of this class see the individuals list associated to this class in the ontology definition.

Status: unknown

Properties include: [isTypeOf](#)

Used with: [hasType](#)

OWL Class

[\[#\]](#) [\[back to top\]](#)

Class: gi2mowave:Market

Market - An object of this type indicates the market aim by an idea or IdeaContest. It could be a place, a sector, a niche, etc

Status: unknown

Properties include: [isMarketOf](#)

Used with: [hasMarket](#)

OWL Class

[\[#\]](#) [\[back to top\]](#)

Class: gi2mowave:Outcome

Outcome - An object of this type indicates final result or outcome for an idea or IdeaContest. This outcome or result can be a product, a service, a process, a strategy, an improvement, a company or spin-off or cooperation among companies, departments, etc. For a list of recommended instances of this class see the individuals list associated to this class in the ontology definition.

Status: unknown

Properties include: [isOutcomeOf](#)

Used with: [hasOutcome](#)

OWL Class

[\[#\]](#) [\[back to top\]](#)

Class: gi2mowave:Resource

Resource - An object of this type indicates a resource necessary for an idea or IdeaContest. This resource can be of different types; general, knowledge, financial, income, etc. For a list of recommended instances of this class see the individuals list associated to this class in the ontology definition.

Status: unknown

Properties include: [isResourceOf](#)

Used with: [hasResource](#)

OWL Class

[\[#\]](#) [\[back to top\]](#)

Class: gi2mowave:Target

Target - An object of this type indicates the objective or target aim by an idea or IdeaContest

Status: unknown

Properties include: [isTargetOf](#)

Used with: [hasTarget](#)

OWL Class

<#> [\[back to top\]](#)

Class: gi2mowave:Technology

Technology - An object of this type indicates the technology used or proposed by an idea or IdeaContest

Status: unknown

Used with: [hasTechnology](#)

OWL Class

<#> [\[back to top\]](#)

Properties

Property: gi2mowave:hasCompetitor

hasCompetitor - Property indicating an idea or other entity having a competitor in the market.

Status: unknown
Domain: <http://purl.org/gi2mo/ns#Idea>
Range: [Competitor](#)
Inverse property of [isCompetitorOf](#)
Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:hasCustomer

hasCustomer - Property indicating an idea or other entity having a customer or possible customer.

Status: unknown
Domain: <http://purl.org/gi2mo/ns#Idea>
Range: [Customer](#)
Inverse property of [isCustomerOf](#)
Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:hasMarket

hasMarket - Property indicating an idea or other entity aiming towards a market.

Status: unknown
Domain: <http://purl.org/gi2mo/ns#Idea>
Range: [Market](#)
Inverse property of [isMarketOf](#)
Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:hasOutcome

hasOutcome - Property indicating an idea or other entity having an outcome as an expectation.

Status: unknown
Domain: <http://purl.org/gi2mo/ns#Idea>
Range: [Outcome](#)
Inverse property of [isOutcomeOf](#)
Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:hasOwner

hasOwner - Property indicating an ideacontest or other entity having a user as an owner or responsible.

Status: unknown
Domain: <http://purl.org/gi2mo/ns#IdeaContest>
Range: <http://purl.org/gi2mo/ns#User>
Inverse property of [isOwnerOf](#)
Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:hasResource

hasResource - Property indicating an idea or other entity having (using, depending ...) a resource.

Status: unknown
Domain: <http://purl.org/gi2mo/ns#Idea>
Range: [Resource](#)
Inverse property of [isResourceOf](#)
Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:hasSelectionCriteria

hasSelectionCriteria - Property indicating an ideaContest uses one of the IdeaContestSelectionCriteria. For a list of the possible types of selection criteria available see the individuals for IdeaContestSelectionCriteria

Status: unknown
Domain: <http://purl.org/gi2mo/ns#IdeaContest>
Inverse property of [isSelectionCriteriaOf](#)
Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:hasTarget

hasTarget - Property indicating an idea or other entity having an objective or target.

Status: unknown

Domain: <http://purl.org/gi2mo/ns#:Idea> or <http://purl.org/gi2mo/ns#:IdeaContest>

Range: [Target](#)

Inverse property of [isTargetOf](#)

Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:hasTechnology

hasTechnology - Property indicating an idea or other entity using, employing or depending on a technology.

Status: unknown

Domain: <http://purl.org/gi2mo/ns#:Idea>

Range: [Technology](#)

Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:hasType

hasType - Property indicating an ideacontest has a type. For a list of the possible types available see the individuals for IdeaContestType

Status: unknown

Domain: <http://purl.org/gi2mo/ns#:IdeaContest>

Range: [IdeaContestType](#)

Inverse property of [isTypeOf](#)

Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:isCompetitorOf

isCompetitorOf - Property indicating an individual or entity being competitor of an idea or other entity.

Status: unknown

Domain: [Competitor](#)

Range: <http://purl.org/gi2mo/ns#:Idea>

Has inverse property [hasCompetitor](#)

Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:isCustomerOf

isCustomerOf - Property indicating a customer is the beneficiary of an idea or other entity.

Status: unknown

Domain: [Customer](#)

Range: <http://purl.org/gi2mo/ns#:Idea>

Has inverse property [hasCustomer](#)

Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:isMarketOf

isMarketOf - Property indicating a market is the objective of an idea or other entity.

Status: unknown

Domain: [Market](#)

Range: <http://purl.org/gi2mo/ns#:Idea>

Has inverse property [hasMarket](#)

Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:isOutcomeOf

isOutcomeOf - Property indicating an entity being the outcome or the expected outcome of an idea or other entity.

Status: unknown

Domain: [Outcome](#)

Range: <http://purl.org/gi2mo/ns#:Idea>

Has inverse property [hasOutcome](#)

Object Property

<#> [\[back to top\]](#)

Property: gi2mowave:isOwnerOf

isOwnerOf - Property indicating a user is the owner or responsible of an idea contest or other entity.

Status: unknown
Domain: <http://purl.org/gi2mo/ns#User>
Range: <http://purl.org/gi2mo/ns#IdeaContest>
Has inverse property [hasOwner](#)
Object Property

[\[#\] \[back to top\]](#)

Property: gi2mowave:isResourceOf

isResourceOf - Property indicating a resource is being used by an idea or other entity.

Status: unknown
Domain: [Resource](#)
Range: <http://purl.org/gi2mo/ns#Idea>
Has inverse property [hasResource](#)
Object Property

[\[#\] \[back to top\]](#)

Property: gi2mowave:isSelectionCriteriaOf

isSelectionCriteriaOf - Property indicating an ideaContestSelectionCriteria defines an IdeaContest selection method. For a list of the possible types of selection criteria available see the individuals for IdeaContestSelectionCriteria.

Status: unknown
Range: <http://purl.org/gi2mo/ns#IdeaContest>
Has inverse property [hasSelectionCriteria](#)
Object Property

[\[#\] \[back to top\]](#)

Property: gi2mowave:isTargetOf

isTargetOf - Property indicating an objective is being looked for by an idea or other entity.

Status: unknown
Domain: [Target](#)
Range: <http://purl.org/gi2mo/ns#Idea> or <http://purl.org/gi2mo/ns#IdeaContest>
Has inverse property [hasTarget](#)
Object Property

[\[#\] \[back to top\]](#)

Property: gi2mowave:isTypeOf

isTypeOf - Property indicating an ideaContestType defines the type of an IdeaContest.

Status: unknown
Domain: [IdeaContestType](#)
Range: <http://purl.org/gi2mo/ns#IdeaContest>
Has inverse property [hasType](#)
Object Property

[\[#\] \[back to top\]](#)

A Changelog

2012-06-04

- First version of the document

B Acknowledgements

This documentation has been generated automatically from the most recent ontology specification in OWL using a python script called [SpecGen](#). The style formatting has been inspired on [FOAF](#) specification.

Appendix B

Mentions Ontology Specification

This appendix contains the specification of Mentions Ontology. This ontology is based on MUTO Ontology, but adding Mention functionalities. This is added in order to enable interoperability among platforms with text based content. The specification can be found in the next pages, and it's lattes version can also be fount online¹.

¹<http://w3id.org/mo>



The Mentions Ontology (MO) 1.0

Specification of MO Core - 14 March 2016

Namespace URI:

<http://w3id.org/mo/core#>

Latest version:

Version 1.0 (OWL)

Authors:

Alain Perez (Mondragon Unibertsitatea),
Felix Larrinaga (Mondragon Unibertsitatea)

Abstract

Mentions Ontology (MO) is an ontology for annotating the elements mentioned in different texts. It is based in the *Modular Unified Tagging Ontology (MUTO)* due to its similarities with tagging and for reusability reasons.

Table of Contents

- Introduction
 - Background
 - What is Mentions Ontology (MO) for?
- Mentions Ontology (MO) at a glance
- Examples
- Mentions Ontology (MO) cross-reference
- External Vocabulary References
- Acknowledgments
- References
- Recent Changes

Introduction

The following specification is a formal description of metadata schema proposal that can be applied to any "text based" content. The aim of the following section is to introduce both Semantic Web and Content Management experts to the topic and goals of the ontology, and to provide the basic knowledge to comprehend the technical part of the specification.

An important note is that *Mentions Ontology* is not a complete model of *Mentioning*. It extends the concepts of the main *Modular Unified Tagging Ontology* (MUTO) and defines concepts for adding the mentioning capabilities to a Tagging ontology.

Background [\[back to top\]](#)

The work done with this ontology is based on a project called InnoWEB (at [Mondragon Unibertsitatea](#)). Part of this project aims to find *Information and Communication Technologies* (ICTs) that help to improve innovation processes. A first approach to this objective was made identifying the benefits that Semantic Web can bring to Innovation processes [2].

A case study of this ontology was to find resources mentioned in Ideas (from an Innovation Process platform, more specifically in an Idea Management System) and News Items (from a Technology Watch platform) about the same topics [1]. Using the methodology described in the research, it is possible to find 2 different type of relationships:

- **Idea-idea relationships:** we can identify the 10 most similar ideas for each idea.
- **News items-idea relationships:** we can identify the 10 most similar ideas for each news items.

An automatic way was proposed to find those relationships using the [DBPedia Spotlight API](#). The most successful approach was found taking *mentions* into account, both from the title and the body of the content, along with the similarity score provided by *DBPedia Spotlight API*. Therefore, and in order to represent this approach semantically, *Mentions Ontology* (MO) was built.

What is Mentions Ontology (MO) for? [\[back to top\]](#)

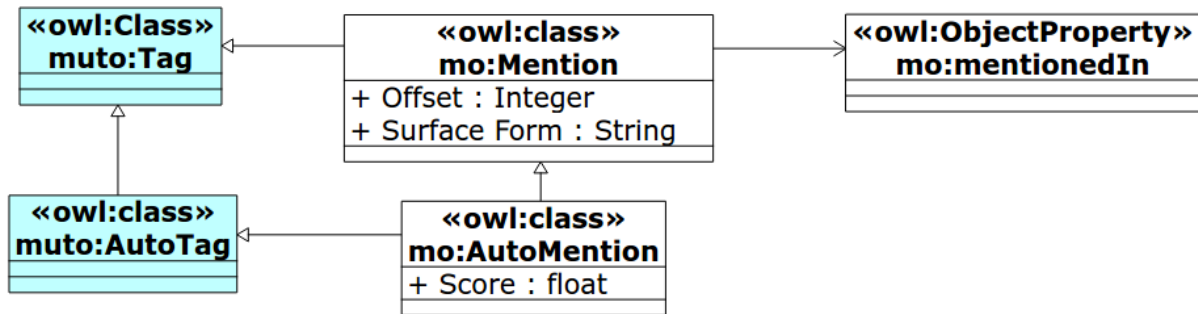
The goals of *Mentions Ontology* to achieve as a data schema are:

- Extend the notion of MUTO's Tagging to '*mentioning*', that can have a *Surface Form*, *Offset* and a *similarity Score* in order to identify where the resource has been mentioned in a text and how probable it is to be that resource.

- It is also possible to identify where the resource has been mentioned, for example, whether a resource has been mentioned in the title of an idea or its body.

Mentions Ontology (MO) at a glance [\[back to top\]](#)

An a-z index of *Mentions Ontology (MO)* terms, by class (categories or types) and by property.



The class diagram of *Mentions Ontology (MO)*.

Classes: | [Mention](#) | [AutoMention](#) |

Properties: | [mentionedIn](#) | [offset](#) | [score](#) | [surfaceForm](#) |

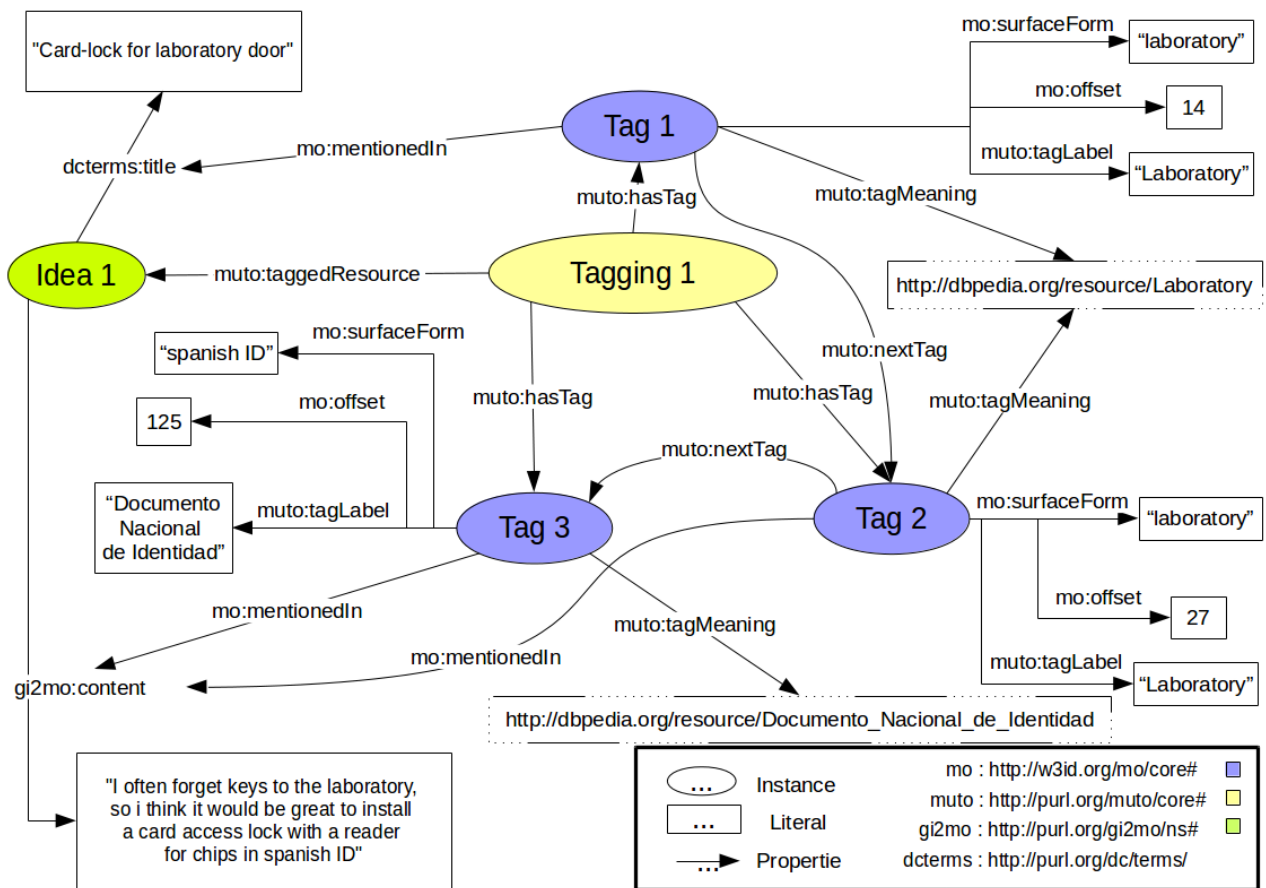
Example [\[back to top\]](#)

The following diagram illustrates an example of adding mentions to an idea from an imaginary *Idea Management System* (IMS) at *example.org*. The idea has a title (*"Card-lock for laboratory door."*) and a body (*"I often forget keys to the laboratory, so i think it would be great to install a card access lock with a reader for chips in spanish ID"*).

According to the diagram, a user has identified that the idea has mentioned 2 times the same resource: *"laboratory"*. Therefore, the user has added 2 mentions of the same resource (*"http://dbpedia.org/resource/Laboratory"*) with the same surface form (*"laboratory"*), one to the title of the idea (*dcterms:title*) with offset 14 and another to the content of the idea (*gi2mo:content*) with offset 27. He has also labeled the mention as *"Laboratory"*.

The automatic mentioning system has automatically recognized the resource *"http://dbpedia.org/resource/Documento_Nacional_de_Identidad"* from the surface form *"spanish ID"* within the content of the idea, at offset 125. Therefore, it has created a new *AutoMention* linking it to the resource. According to the system, there is a 95% of probability for the surface form to be the retrieved resource, so it has set the *Score* to 0.95. He has also identified the label of the resource and has set it to *"Documento Nacional de Identidad"*.

Different resources mentioned in the idea have been identified, both manually and automatically. This can be done with all the ideas of the IMS.



A graphic of a simple example of an use case of the Mentions Ontology (MO)

@prefix muto: <<http://purl.org/muto/core#>> .
@prefix mo: <<http://w3id.org/mo/core#>> .
@prefix gi2mo: <<http://purl.org/gi2mo/ns#>> .
@prefix dcterms: <<http://purl.org/dc/terms/>> .

<<http://example.com/ideas/idea1>> a <<http://purl.org/gi2mo/ns#Idea>>;
dcterms:title "Card-lock for laboratory door";
gi2mo:content "I often forget keys to the laboratory, so i think it would be great to install a card access lock with a reader for chips in spanish ID";
<<http://example.org/tagging/taggin1>> a <<http://purl.org/muto/core#Tagging>>;
muto:taggedResource <<http://example.com/ideas/idea1>>;
muto:hasTag <<http://example.org/tag/tag1>>,
<<http://example.org/tag/tag2>>,
<<http://example.org/tag/tag3>>;
<<http://example.org/tag/tag1>> a <<http://w3id.org/mo/core#Mention>>;
muto:tagLabel "Laboratory";
muto:tagMeaning <<http://dbpedia.org/resource/Laboratory>>;
mo:offset 14;
mo:surfaceForm "laboratory";
mo:mentionedInProperty <<http://purl.org/dc/terms/title>>;
muto:nextTag <<http://example.org/tag/tag2>>.
<<http://example.org/tag/tag2>> a <<http://w3id.org/mo/core#Mention>>;
muto:tagLabel "Laboratory";
muto:tagMeaning <<http://dbpedia.org/resource/Laboratory>>;
mo:offset 27;
mo:surfaceForm "laboratory";
mo:mentionedInProperty <<http://purl.org/gi2mo/ns#content>>;
muto:nextTag <<http://example.org/tag/tag3>>.
<<http://example.org/tag/tag3>> a <<http://w3id.org/mo/core#AutoMention>>;
muto:tagLabel "Documento Nacional de Identidad";
muto:autoMeaning <http://dbpedia.org/resource/Documento_Nacional_de_Identidad>;
mo:offset 125;
mo:surfaceForm "spanish ID";
mo:mentionedInProperty <<http://purl.org/gi2mo/ns#content>>;
mo:score 0.95 .

Mentions Ontology (MO) cross-reference: Listing the Mentions Ontology (MO) Classes and Properties [\[back to top\]](#)

The Mentions Ontology (MO) introduces the following classes and properties.

Classes: | [Mention](#) | [AutoMention](#) |

Properties: | [mentionedIn](#) | [offset](#) | [score](#) | [surfaceForm](#) |

Classes and Properties (full detail)

Classes [\[back to top\]](#)

Class: Mention

Mention - Mention is a tag associated to a text, an specific string (or surface form) in that text and to the offset of the surface form in the text.

URI: <http://w3id.org/mo/core#Mention>

Properties include: [Offset](#) | [Surface form](#)

Sub class of [muto:Tag](#)

OWL Class

[\[#\]](#) [\[back to top\]](#)

Class: AutoMention

Automatic Mention - An automatic mention is a mention that is automatically associated with a resource (e.g. by a text annotation system like DBPedia Spotlight), i.e. it is not entered by a human being.

URI: <http://w3id.org/mo/core#AutoMention>

Properties include: [score](#)

Sub class of [Mention](#) | [muto:AutoTag](#)

OWL Class

[\[#\]](#) [\[back to top\]](#)

Properties

Property: Mentioned In

Mentioned In - The property of the source text where the Mention has been mentioned.

URI: <http://w3id.org/mo/core#mentionedIn>

Domain: [Mention](#)

Object Property

[#] [\[back to top\]](#)

Property: Offset

Offset - The position (offset) of the surface form string in the source text.

URI: <http://w3id.org/mo/core#offset>

Domain: [Mention](#)

Range: [xsd:integer](#)

Datatype Property

[#] [\[back to top\]](#)

Property: Score

Score - The score the automatic system gave to the surface form to be the giving resource.

URI: <http://w3id.org/mo/core#score>

Domain: [Automatic Mention](#)

Range: [xsd:float](#)

Datatype Property

[#] [\[back to top\]](#)

Property: Surface Form

Surface form - The surface form is the string that has been identified to mention a resource in the text.

URI: <http://w3id.org/mo/core#surfaceForm>

Domain: [Mention](#)

Range: [xsd:string](#)

Datatype Property

[#] [\[back to top\]](#)

External Vocabulary References [\[back to top\]](#)

This ontology has been created to add "mentions" feature to the [Modular Unified Tagging Ontology \(MUTO\)](#). The example of this ontology also uses [Gi2MO Ontology](#) for idea annotation.

Acknowledgments [\[back to top\]](#)

This work has been conducted in the context of the Innoweb project in Mondragon Unibertsitatea.

References [\[back to top\]](#)

[1] Pending: *Semantic Annotations to enhance Innovation and Technology Watch Interoperability.*

[2] Perez, A., Larrinaga, F., Curry, E. (2013). *The Role of Linked Data and Semantic-Technologies for Sustainability Idea Management.* In *Software Engineering and Formal Methods* (pp. 306-312). Springer International Publishing. [\[BibTex\]](#)

[3] Lohmann, S., Díaz, P., Aedo, I.: *MUTO: The Modular Unified Tagging Ontology.* *Proceedings of the 7th International Conference on Semantic Systems (I-SEMANTICS 2011)*, pp. 95-104. ACM, New York, NY, USA, 2011. [\[BibTex\]](#)

[4] A. Westerski, *Gi2MO: Interoperability, Linking and Filtering in Idea Management Systems*, in *Extended Semantic Web Conference 2011. PhD Symposium Poster.*, Heraklion, Greece, 2011. [\[BibTex\]](#)

Recent Changes [\[back to top\]](#)

=====
Version 1.0 (Mar 2016)
=====

- *ADDED: mo:Mention - A class to represent the mentioned resources.*
- *ADDED: mo:AutoMention - A mention gathered using an automatic system.*

This work is licensed under a [Creative Commons Attribution License](#). This copyright applies to the Mentions Ontology (MO) Specification and accompanying documentation in RDF. Regarding underlying technology, the Mentions Ontology (MO) uses W3C's [RDF](#) technology, an open Web standard that can be freely used by anyone.



Bibliography

- [1] S. Adamczyk, A. C. Bullinger, and K. M. Möslein. Innovation contests: A review, classification and outlook. *Creativity and Innovation Management*, 21(4):335–360, 2012.
- [2] E. Almirall. El papel de la informática en la innovación. *Novática*, 1(195):12–18, 2008.
- [3] T. M. Amabile. A model of creativity and innovation in organizations. *Research in Organizational Behavior*, 10:126–167, 1988.
- [4] G. Anadiotis, K. Kafentzis, I. Pavlopoulos, and A. Westerski. Building consensus via a semantic web collaborative space. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 1097–1106. ACM, 2012.
- [5] N. Anicic, N. Ivezic, and A. Jones. An architecture for semantic enterprise application integration standards. In *Interoperability of Enterprise Software and Applications*, pages 25–34. Springer, 2006.
- [6] N. S. Argyres and B. S. Silverman. R&d, organization structure, and the development of corporate technological knowledge. *Strategic Management Journal*, 25(8/9):929–958, 2004.
- [7] M. G. Armentano, D. Godoy, M. Campo, and A. Amandi. Nlp-based faceted search: Experience in the development of a science and technology search engine. *Expert Systems with Applications*, 41(6):2886–2896, 2014.
- [8] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.
- [9] B. Baharudin, L. H. Lee, and K. Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20, 2010.
- [10] P. Beneito. The innovative performance of in-house and contracted r&d in terms of patents and utility models. *Research Policy*, 35(4):502–517, 2006.
- [11] T. Berners-Lee. Semantic web on xml. 2000.

- [12] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, volume 2006, 2006.
- [13] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):28–37, 2001.
- [14] D. Bibikas, D. Kourtesis, I. Paraskakis, A. Bernardi, L. Sauermann, D. Apostolou, G. Mentzas, and A. C. Vasconcelos. A sociotechnical approach to knowledge management in the era of enterprise 2.0: the case of organik. *Scalable Computing: Practice and Experience*, 9(4), 2008.
- [15] C. Bizer. Evolving the web into a global data space. In *28th British National Conference on Databases, BNCOD 28*, Advances in Databases, page 1, Berlin, Germany, 12-14 July 2011 2011. Freie Univ. Berlin, Berlin, Germany, Springer Verlag.
- [16] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.
- [17] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the web (ldow2008). In *17th International Conference on World Wide Web 2008, WWW'08*, pages 1265–1266. Affiliation: Freie Universität, Berlin, Germany; Affiliation: Talis, United Kingdom; Affiliation: OpenLink Software, United States; Affiliation: W3C, United States; Correspondence Address: Bizer, C.; Freie Universität, Berlin, Germany; email: chris@bizer.de, 21 April 2008 through 25 April 2008 2008.
- [18] S. Bolasco, F. Baiocchi, A. Canzonetti, F. Della Ratta, and A. Feldman. Applications, sectors and strategies of text mining, a first overall picture. In *Text Mining and Its applications*, pages 37–51. Springer, 2004.
- [19] J. G. Breslin, A. Passant, and S. Decker. *Social Semantic Web*. Springer, 2009.
- [20] J. Bughin, J. Manyika, and A. Miller. Building the web 2.0 enterprise. *McKinsey Quarterly*, 7:2008, 2008.
- [21] A. C. Bullinger. *Innovation and Ontologies: Structuring the Early Stages of Innovation Management*. Gabler, 1 edition, 2008.
- [22] V. Bush. *Science, the endless frontier*. AYER Co. Pub., 1945.
- [23] A. Cantisani. Technological innovation processes revisited. *Technovation*, 26(11):1294–1301, 2006.
- [24] M. M. Capozzi, B. Gregg, and A. Howe. Innovation and commercialization, 2010: Mckinsey global survey results, 2010.

- [25] B. Cassiman and R. Veugelers. In search of complementarity in innovation strategy: Internal r&d and external knowledge acquisition. *Management Science*, 52(1):68–82, 2006.
- [26] H. Chesbrough and A. K. Crowther. Beyond high tech: early adopters of open innovation in other industries. *R&D Management*, 36(3):229–236, 2006.
- [27] H. W. Chesbrough. *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business Press, 2003.
- [28] G. G. Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.
- [29] J. F. Christensen, M. H. Olesen, and J. S. Kjær. The industrial dynamics of open innovation—evidence from the transformation of consumer electronics. *Research Policy*, 34(10):1533–1549, 2005.
- [30] E. Commission. Improving knowledge transfer between research institutions and industry across europe: embracing open innovation - implementing the lisbon agenda. *European Commission*, 2007.
- [31] A. Conry-Murray. Can enterprise social networking pay off. *Informationweek. com*, 1224:23–29, 2009.
- [32] G. Convertino, Á. Sándor, and M. Baez. Idea spotter and comment interpreter: Sensemaking tools for idea management systems. In *ACM Communities and Technologies Workshop: Large-Scale Idea Management and Deliberation Systems Workshop*, 2013.
- [33] M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. 2013.
- [34] E. Curry, B. Guyon, C. Sheridan, and B. Donnellan. Developing a sustainable it capability: lessons from intel’s journey. *MIS Quarterly Executive*, 11(2):61–74, 2012.
- [35] E. Curry, S. Hasan, and S. O’Riain. Enterprise energy management using a linked dataspace for energy intelligence. In *Sustainable Internet and ICT for Sustainability (SustainIT), 2012*, pages 1–6. IEEE, 2012.
- [36] E. Curry, S. Hasan, U. ul Hassan, M. Herstand, and S. O’Riain. An entity-centric approach to green information systems. In *ECIS*, 2011.
- [37] E. Curry, J. O’Donnell, E. Corry, S. Hasan, M. Keane, and S. O’Riain. Linking building data in the cloud: Integrating cross-domain building data using linked data. *Advanced Engineering Informatics*, 27(2):206–219, 2013.

- [38] D. Czarnitzki and K. Kraft. An empirical test of the asymmetric models on innovative activity: who invests more into r&d, the incumbent or the challenger? *Journal of Economic Behavior & Organization*, 54(2):153–173, 2004.
- [39] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM, 2013.
- [40] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.
- [41] R. Dale. Industry watch. *Natural Language Engineering*, 11(1):113–117, 2005.
- [42] M. Dodgson, D. M. Gann, and A. Salter. *The management of technological innovation: strategy and practice*. OUP Oxford, 2008.
- [43] J. M. Durán, M. M. Martínez, and J. V. Triano. La vigilancia tecnológica en la gestión de proyectos de i+ d+ i: recursos y herramientas. *El profesional de la información*, 15(6):411–419, 2006.
- [44] E. Enkel and O. Gassmann. Driving open innovation in the front end. *The IBM Case.St.Gallen and Friedrichshafen, University of St.Gallen and Zeppelin University-technical report*, 2008.
- [45] N. Errasti. Impacto del programa agenda en el caso de las empresas industriales de las comarcas del deba y del urola. propuesta de un modelo estratégico para la innovación. *MGEP-MU*, 2009.
- [46] N. Errasti. Social software in support of collaborative innovation processes. *Projectics/Proyética/Projectique*, (2):81–104, 2010.
- [47] N. Errasti, N. Zabaleta, and A. Oyarbide. A review and conceptualisation of innovation models from the past three decades. *International Journal of Technology Management*, 55(3):190–200, 2011.
- [48] L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neumann, and S. Stephens. The semantic web in action. *Scientific American*, 297(6):90–97, 2007.
- [49] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In *KDD*, volume 95, pages 112–117, 1995.
- [50] B. Fernández Fuentes, S. Pérez Álvarez, and F. d. Valle Gastaminza. Metodología para la implantación de sistemas de vigilancia tecnológica y documental: El caso del proyecto inredis. *Investigación bibliotecológica*, 23(49):149–177, 2009.
- [51] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, and J. Prager. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

- [52] E. Ford and S. Mohapatra. Idea de-duplication in an innovation community. *Urbana*, 51:61801–2302, 2011.
- [53] T. Fredberg, M. Elmquist, and S. Ollila. Managing open innovation, present findings and future directions. 2008.
- [54] K. Galanakis. Innovation process. make sense using systems thinking. *Technovation*, 26(11):1222–1232, 2006.
- [55] S. Gee. *Technology transfer, innovation, and international competitiveness*. Wiley New York, 1981.
- [56] C. Geffen and K. Judd. Innovation through initiatives—a framework for building new capabilities in public sector research organizations. *Journal of Engineering and Technology Management*, 21(4):281–306, 2004.
- [57] B. Godin. The linear model of innovation. *Science, Technology & Human Values*, 31(6):639–667, 2006.
- [58] A. Gómez-Pérez. Ontological engineering: A state of the art. *Expert Update: Knowledge Based Systems and Applied Artificial Intelligence*, 2(3):33–43, 1999.
- [59] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.
- [60] B. A. Hamilton. No relationship between r&d spending and sales growth, earnings, or shareholder returns. 2005.
- [61] P. Heim, J. Ziegler, and S. Lohmann. gfacet: A browser for the web of data. In *Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08)*, volume 417, pages 49–58. Citeseer, 2008.
- [62] S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating nlp using linked data. In *12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*, 2013.
- [63] I. Herman. Introduction to the semantic web. In *Semantic Technology Conference (San Jose, CA, USA, 2009)*, page 140, 2009.
- [64] M. Hobday. Firm-level innovation models: perspectives on research in developed and developing countries. *Technology analysis & strategic management*, 17(2):121–146, 2005.
- [65] J. Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.
- [66] L. Huston and N. Sakkab. Implementing open innovation. *Research-Technology Management*, 50(2):21–25, 2007.

- [67] F. Ibekwe-Sanjuan and E. Sanjuan. Mining textual data through term variant clustering: the termwatch system. In *Recherche d'Information et ses Applications (RIA0 2004)*, pages 487–503. FID, 2004.
- [68] J. I. Igartua, J. A. Garrigós, and J. L. Hervas-Oliver. How innovation management techniques support an open innovation strategy. *Research-Technology Management*, 53(3):41–52, 2010.
- [69] B. Iyer and T. H. Davenport. Una ingeniería inversa a la máquina de innovación de google. *Harvard business review*, 86(4):63–73, 2008.
- [70] J. Izquierdo and S. Larreina. *Collective SME approach to technology watch and competitive intelligence: The role of intermediate centers*, volume 185 of *Studies in Fuzziness and Soft Computing*. 2005.
- [71] F. Jacquenet and C. Largeton. Discovering unexpected documents in corpora. *Knowledge-Based Systems*, 22(6):421–429, 2009.
- [72] F. Jacquenet and C. Largeton. Discovering unexpected documents in corpora. *Knowledge-Based Systems*, 22(6):421–429, 2009.
- [73] L. Kahaner. *Competitive intelligence: how to gather, analyze, and use information to move your business to the top*. Touchstone, 1997.
- [74] L. Kahaner. *Competitive intelligence: how to gather analyze and use information to move your business to the top*. Simon and Schuster, 1997.
- [75] T. Koc and C. Ceylan. Factors impacting the innovative capacity in large-scale companies. *Technovation*, 27(3):105–114, 2007.
- [76] B. I. Koerner. Geeks in toyland. *WIRED-SAN FRANCISCO-*, 14(2):104, 2006.
- [77] P. Kolari, T. Finin, K. Lyons, and Y. Yesha. Expert search using internal corporate blogs. In *Workshop on Future Challenges in Expertise Retrieval, SIGIR*, volume 8, 2008.
- [78] F. Larrinaga, I. Santos, O. Lizarralde, and A. Perez. A case study on the use of community platforms for inter-enterprise innovation. In *Concurrent Enterprising (ICE), 2011 17th International Conference on*, pages 1–8. IEEE, 2011.
- [79] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morse, P. van Kleef, S. Auer, et al. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.
- [80] B. Lent, R. Agrawal, and R. Srikant. Discovering trends in text databases. In D. Heckerman, H. Mannila, and D. Pregibon, editors, *KDD*, pages 227–230. AAAI Press, 1997.

- [81] S. Lohmann, P. Díaz, and I. Aedo. Muto: the modular unified tagging ontology. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 95–104. ACM, 2011.
- [82] L. Lorenzo, O. Lizarralde, I. Santos, and A. Passant. Structuring e-brainstorming to better support innovation processes. In *Social Innovation and Social Media*, 2011.
- [83] G. Lortal, N. Chaignaud, J.-P. Kotowicz, and J.-P. Pécuchet. Nlp contribution to the semantic web: Linking the term to the concept. In J. Velásquez, S. Ríos, R. Howlett, and L. Jain, editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 5711 of *Lecture Notes in Computer Science*, pages 309–317. Springer Berlin Heidelberg, 2009.
- [84] P. B. Losiewicz, D. W. Oard, and R. N. Kostoff. Textual data mining to support science and technology management. *J. Intell. Inf. Syst.*, 15(2):99–119, 2000.
- [85] J. H. Love and S. Roper. The determinants of innovation: R&d, technology transfer and networking effects. *Review of Industrial Organization*, 15(1):43–64, 1999.
- [86] A. Maedche, B. Motik, L. Stojanovic, R. Studer, and R. Volz. Ontologies for enterprise knowledge management. *Intelligent Systems, IEEE*, 18(2):26–33, 2003.
- [87] D. Maynard, M. Yankova, A. Kourakis, and A. Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *ESWC Workshop "End User Aspects of the Semantic Web"*, Heraklion, Crete, 2005.
- [88] D. Maynard, M. Yankova, A. Kourakis, and A. Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *ESWC Workshop "End User Aspects of the Semantic Web,"* Heraklion, Crete, 2005.
- [89] D. R. Millen, J. Feinberg, and B. Kerr. Dogear: Social bookmarking in the enterprise. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 111–120. ACM, 2006.
- [90] R. Mitkov. *The Oxford handbook of computational linguistics*. Oxford University Press, 2005.
- [91] P. Morcillo. La dirección estratégica de la tecnología e innovación. *Civitas*, 1997.
- [92] J. Morin. *L'excellence technologique*. Publi-Union: Éditions J. Picollec, Paris, 1985.
- [93] J. Morin, R. Seurat, and C. Marbach. *Le management des ressources technologiques*. Les Éditions d'Organisation, Paris, 1989.
- [94] A. Mukherjee and S. Marjit. R&d organization and technology transfer. *Group Decision and Negotiation*, 13(3):243–258, 2004.

- [95] E. J. Nijssen, B. Hillebrand, P. A. M. Vermeulen, and R. G. M. Kemp. Exploring product and service innovation similarities and differences. *International Journal of Research in Marketing*, 23(3):241–251, 2006.
- [96] T. O'Reilly. What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, (1):17, 2007.
- [97] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: a document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics and Ontologies*, 3(1):37–52, 2008.
- [98] T. Padmore, H. Schuetze, and H. Gibson. Modeling systems of innovation: An enterprise-centered view. *Research Policy*, 26(6):605–624, 1998.
- [99] F. Palop and J. M. Vicente. *Vigilancia tecnológica e inteligencia competitiva: su potencial para la empresa española*. Cotec Madrid, 1999.
- [100] M. Parrish. The forrester wave: Community platforms. Technical report, Q4 2010 2010.
- [101] J. Pavón and R. Goodman. El proceso de innovación. *La planificación del desarrollo tecnológico: el caso español, proyecto Modeltec, Centro para el desarrollo tecnológico industrial, Madrid*, pages 221–229, 1981.
- [102] A. Perez, F. Larrinaga, and E. Curry. The role of linked data and semantic-technologies for sustainability idea management. In *Software Engineering and Formal Methods*, pages 306–312. Springer, 2013.
- [103] A. Perez, F. Larrinaga, O. Lizarralde, and I. Santos. Innoweb: Gathering the context information of innovation processes with a collaborative social network platform. In *International Conference on Concurrent Enterprising (ICE)*, 2013.
- [104] D. Poole, A. Mackworth, and R. Goebel. *Computational Intelligence*. Oxford University Press Oxford, 1998.
- [105] M. E. Porter. *The competitive advantage of nations: with a new introduction*. Free Pr, 1990.
- [106] M. E. Porter and C. Van der Linde. Toward a new conception of the environment-competitiveness relationship. *The journal of economic perspectives*, 9(4):97–118, 1995.
- [107] G. Poveda, A. Westerski, and C. A. Iglesias. Application of semantic search in idea management systems. In *Internet Technology And Secured Transactions, 2012 International Conference for*, pages 230–236. IEEE, 2012.

- [108] K. Rajaraman and A.-H. Tan. Topic detection, tracking, and trend analysis using self-organizing neural networks. In D. W.-L. Cheung, G. J. Williams, and Q. Li, editors, *PAKDD*, volume 2035 of *Lecture Notes in Computer Science*, pages 102–107. Springer, 2001.
- [109] C. Riedl, N. May, J. Finzen, S. Stathel, V. Kaufman, and H. Krcmar. An idea ontology for innovation management. *International Journal on Semantic Web and Information Systems*, 5(4):1–18, 10 2009.
- [110] G. Rigau and E. Agirre. Meaning for the semantic web. 2004.
- [111] G. Rizzo, R. Troncy, S. Hellmann, and M. Bruemmer. Nerd meets nif: Lifting nlp extraction results to the linked data cloud. *LDOW*, 937, 2012.
- [112] R. Rothwell. Successful industrial innovation: critical factors for the 1990s. *R&D Management*, 22(3):221–240, 1992.
- [113] R. Rothwell. Towards the fifth-generation innovation process. *International marketing review*, 11(1):7–31, 1994.
- [114] D. Rouach. *La veille technologique et l’intelligence économique*. Puf, 1996.
- [115] J. Schumpeter. *The Theory of Economic Development: An Inquiry Into Profits, Capital, Credit, Interest, and the Business Cycle*. Economics Third World studies. Transaction Books, 1934.
- [116] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.
- [117] D. Seebode, S. Jeanrenaud, and J. Bessant. Managing innovation for sustainability. *R&D Management*, 42(3):195–206, 2012.
- [118] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.
- [119] M. Stankovic. Open innovation and semantic web:problem solver search on linked data. In *9th International Semantic Web Conference (ISWC2010)*, November 2010.
- [120] D. E. Stokes. *Pasteur’s quadrant: Basic science and technological innovation*. Brookings Institution Press, 1997.
- [121] J. Surowiecki. *The wisdom of crowds*. Random House Digital, Inc., 2005.
- [122] H. Tomita, Y. Ikeda, and H. Takeda. Correlation between r&d investment and sales growth of a company with 90 years in r&d operation. In *Management of Innovation and Technology, 2008. ICMIT 2008. 4th IEEE International Conference on*, pages 1021–1026. IEEE, 2008.

- [123] M. Uschold and M. King. *Towards a methodology for building ontologies*. Citeseer, 1995.
- [124] G. Vasco. Plan de ciencia y tecnología 2007. 2004.
- [125] G. Vasco. Plan de ciencia y tecnología 2010. 2007.
- [126] G. Vasco. Plan de ciencia y tecnología 2015. 2010.
- [127] E. von Hippel. *The sources of innovation*. Oxford University Press, New York, 1988.
- [128] E. Von Hippel. Democratizing innovation: The evolving phenomenon of user innovation. *Journal für Betriebswirtschaft*, 55(1):63–78, 2005.
- [129] R. T. Watson, M. Lind, and S. Haraldson. The emergence of sustainability as the new dominant logic: implications for information systems. 2012.
- [130] J. West and S. Gallagher. Challenges of open innovation: the paradox of firm investment in open-source software. *R&D Management*, 36(3):319–331, 2006.
- [131] A. Westerski. Gi2mo: Interoperability, linking and filtering in idea management systems. In *Extended Semantic Web Conference 2011. PhD Symposium Poster.*, Heraklion, Greece, May 2011.
- [132] A. Westerski. *Semantic technologies in idea management systems: a model for interoperability, linking and filtering*. Adam Westerski, 2013.
- [133] A. Westerski, C. A. Iglesias, and F. T. Rico. A model for integration and inter-linking of idea management systems. In *4th Metadata and Semantics Research Conference (MTSR 2010)*, volume 108 CCIS, pages 183–194, Alcalá de Henares, Spain, October 2010. Springer Verlag.
- [134] Y. Wilks and C. Brewster. Natural language processing as a foundation of the semantic web. *Foundations and Trends in Web Science*, 1(3 8211; 4):199–327, 2009.
- [135] D. Zhu and A. L. Porter. Automated extraction and visualization of information for technological intelligence and forecasting. *Technological forecasting and social change*, 69(5):495–506, 2002.