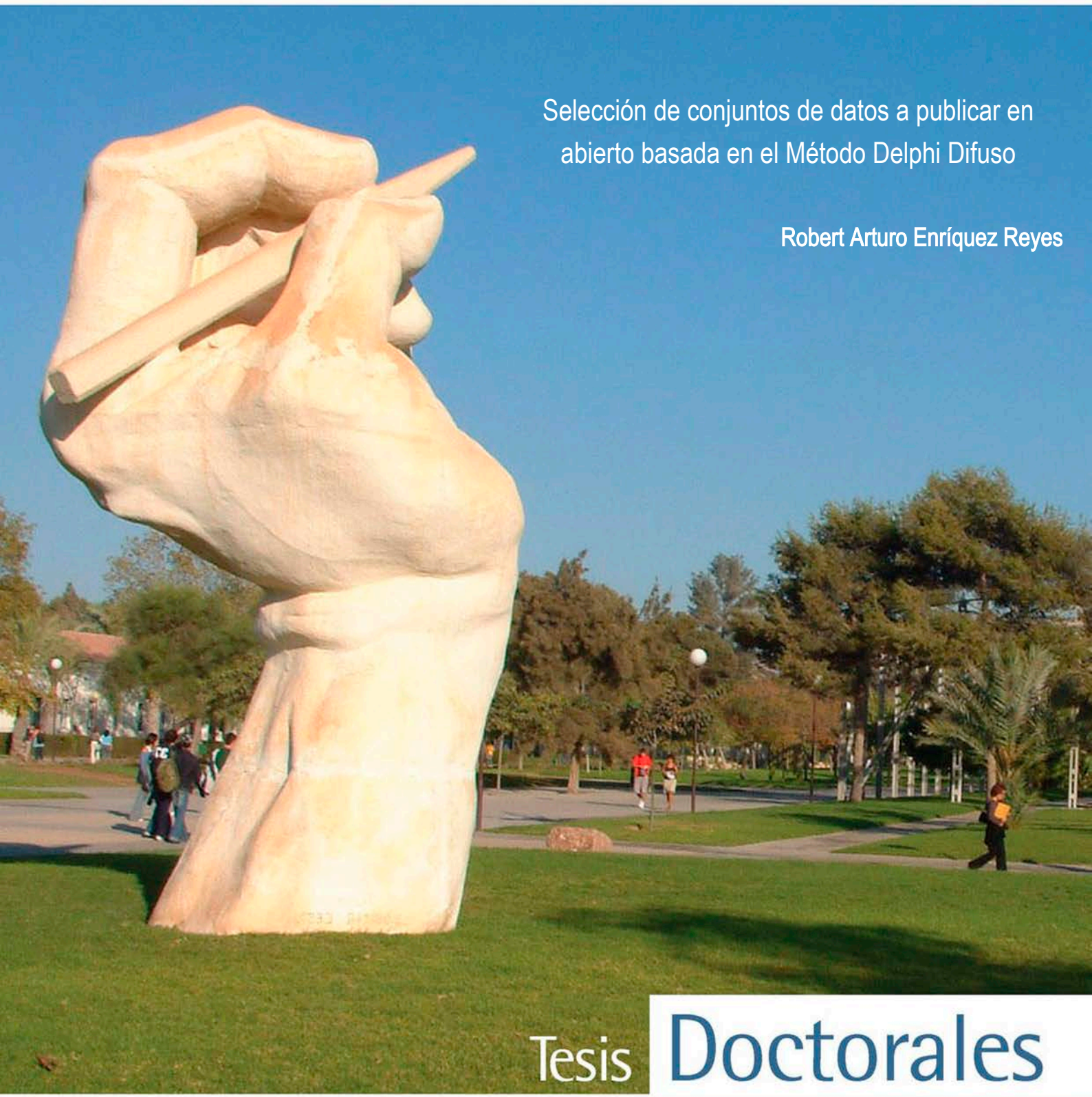
Universitat d'Alacant
Universidad de Alicante

Selección de conjuntos de datos a publicar en abierto basada en el Método Delphi Difuso

Robert Arturo Enríquez Reyes



Tesis Doctorales

UNIVERSIDAD de ALICANTE

Unitat de Digitalització UA
Unidad de Digitalización UA

Universitat d'Alacant
Universidad de Alicante

Instituto Universitario de Investigación en Informática
Escuela Politécnica Superior

Selección de conjuntos de datos a publicar en abierto basada en el Método Delphi Difuso

ROBERT ARTURO ENRIQUEZ REYES

Tesis presentada para aspirar al grado

DOCTOR POR LA UNIVERSIDAD DE ALICANTE

DOCTORADO EN INFORMÁTICA

Dirigida por:
Dr. Jose Norberto Mazón López
Dr. Andrés Fuster Guilló

Alicante, noviembre 2019

Tesis doctoral realizada en el seno del proyecto de investigación "Publi@City: plataforma para la publicación y consumo de datos abiertos para una ciudad inteligente" (TIN2016-78103-C2-2-R) financiado por el Ministerio de Economía, Industria y Competitividad de España

"Carpe Diem quam minimum credula postero."

- Horacio (Odas, I, 11)

Agradecimiento

A Dios,
por la vida.

A mi Familia

por ser mi luz y la fuerza que me impulsa a mejorar siempre.

A mis Amigos
por compartir mis momentos más importantes.

Al pueblo ecuatoriano
por creer en los ecuatorianos y a través de la Universidad Central permitirnos prepararnos

A mis directores, José Norberto Mazón y Andrés Fuster Guilló,
por su apoyo incondicional en el desarrollo de esta investigación.

Resumen

Este trabajo de investigación se enfoca inicialmente en la realización de un mapeo sistemático para clasificar y analizar la investigación de datos abiertos realizada en la comunidad científica desde un punto de vista tecnológico, proporcionando una categorización de los trabajos de investigación basada en cinco facetas clave: impacto, tema, dominio, fases y tipo de investigación. Por lo tanto, este documento proporciona una visión general del área de datos abiertos que permite a los lectores identificar temas bien establecidos, tendencias y líneas de investigación abiertas. Además, se ofrece una extensa discusión cuantitativa y cualitativa que puede ser de utilidad para futuras investigaciones. La primera fase de identificación resultó en 671 artículos relevantes revisados por pares, publicados entre 2006 y 2017 en una amplia variedad de lugares. Se observa que el debate actual de la apertura de datos se centra en abrir los datos, especialmente públicos, teniendo en cuenta la ley de transparencia y la de protección de datos de cada país. Sin embargo, el mero cumpliendo de estas leyes no asegura la generación de valor a partir de los datos abiertos, ya que uno de sus beneficios más importantes se consigue cuando se reutilizan para crear productos y servicios TI de valor agregado. Uno de los problemas que necesariamente deben abordar las organizaciones que emprenden procesos de apertura es la selección de los conjuntos de datos a abrir. Se debe conocer qué datos de origen de la institución serían los más usados para generar valor a la sociedad con el fin de seleccionarlos para su apertura. Sin embargo, conocer los conjuntos de datos solicitados por la comunidad reutilizadora no es suficiente, pues publicar los datos tiene un costo en cuanto a hardware, software y recursos humanos que es necesario valorar. Por tanto, es necesario encontrar un equilibrio entre el interés reutilizador de los conjuntos de datos y su coste de publicación. Las propuestas estudiadas no combinan el criterio de los reutilizadores y el de los publicadores, ni utilizan un método formal conduciendo a resultados poco objetivos. En este trabajo de tesis doctoral se propone paliar esta problemática mediante la aplicación del Método Delphi Difuso con el fin de determinar qué conjuntos de datos son más susceptibles de ser reutilizados y qué conjuntos de datos tendrán un costo asumible para su publicación para proceder a su apertura. Se establece, además, a lo largo del trabajo de investigación un caso concreto de aplicación en el ámbito de las universidades ecuatorianas. Estas se encuentran en un proceso constante de innovación y buscan la participación y colaboración de los estudiantes y la comunidad universitaria en general, para generar productos y servicios TI de valor agregado a través de la apertura de sus conjuntos de datos.

Resum

Este treball d'investigació s'enfoca inicialment en realisar un mapage sistemàtic per a classificar i analisar l'investigació de senyes obertes realisada en la comunitat científica des d'un punt de vista tecnològic, proporcionant una categorisació dels treballs d'investigació basada en cinc facetes clau: impacte, tema, domini, fases i tipo d'investigació. Per lo tant, este document proporciona una visió general de l'àrea de senyes obertes que permet als llectors identificar temes ben establits, tendències i llínees d'investigació obertes. Ademés, s'oferix una extensa discussió quantitativa i qualitativa que pot ser d'utilitat per a futures investigacions. La primera fase d'identificació va resultar en 671 artículs rellevants revisats per parells, publicats entre 2006 i 2017 en una àmplia varietat de llocs. S'observa que el debat actual de l'obertura de senyes se centra en obrir les senyes, especialment públics, tenint en conte la llei de transparència i la de protecció de senyes de cada país. No obstant, el mer complint d'estes lleis no assegura la generació de valor de l'obertura de senyes, ya que un dels seus beneficis més importants es conseguix quan es reutilisen per a crear productes i servicis TI de valor agregat. Un dels problemes que necessàriament deuen abordar les organisacions que mamprenen processos d'obertura és la selecció dels conjunts de senyes a obrir. Per un costat, és necessari conéixer qué senyes d'orige de l'institució serien els més usats per a generar valor a la societat en la finalitat de seleccionar-los per a la seua obertura. No obstant, conéixer els conjunts de senyes solicitades per la comunitat que reutilitza no és suficient, puix publicar les senyes té un cost sobre hardware, software i recursos humans que és necessari valorar. És necessari trobar un equilibri entre l'interés reutilizador dels conjunts de senyes i el seu cost de publicació. Les propostes estudiades no combinen el criteri dels reutilizadores i el dels publicadores, ni utilisen un mètode formal conduint a resultats poc objectius. En este treball es propon paliar esta problemàtica per mig de l'aplicació de la Mètode Delphi Difús en la finalitat de determinar qué conjunts de senyes són més susceptibles de ser reutilisats i quins conjunts de senyes tindran un cost assumible per a la seua publicació en la finalitat de procedir a la seua obertura. S'establix, ademés, a lo llarc del treball d'investigació un cas concret d'aplicació en l'àmbit de les universitats equatorianes. Estes es troben en un procés constant d'innovació i busquen la participació i colaboració dels estudiants i la comunitat universitària en general, per a generar productes i servicis TI de valor agregat a través de l'obertura dels seus conjunts de senyes.

Abstract

This research work initially focuses on a systematic mapping to classify and analyze open data research performed in the scientific community from a technological viewpoint, providing a categorization of research work based on five key facets: impact, topic, domain, phases and type of research. This research work therefore provides a consolidated overview that allows readers to identify well-established topics, trends, and open research issues. Additionally, an extensive quantitative and qualitative discussion is offered that may be useful for future research. The first identification phase resulted in 671 relevant peer-reviewed articles published between 2006 and 2017 in a wide variety of venues. In addition, it identifies the current data open-up debate that focuses on opening the data, especially public data, taking into account the transparency law and the data protection law of each country. However, only complying with these laws does not ensure that value is generated from open data, as one of the most important benefits of open data is achieved when it is reused to create value-added IT products and services. One of the problems that must necessarily be addressed by organizations that undertake opening processes is the selection of the data sets to be opened. On the one hand, it is necessary to know which data from the origin of the institution would be the most used to generate value for society in order to select them for opening. However, knowing what data sets requested by the reusing community is not enough, since publishing the data has a cost in terms of hardware, software and human resources that needs to be valued. It is necessary to find a balance between the reusing interest of the data sets and their publication cost. The proposals studied do not combine the criteria of reusers and publishers, nor do they use a formal method leading to unobjective results. In this work it is proposed to alleviate this problem by applying the Fuzzy Delphi Method in order to determine which data sets are most likely to be reused and which data sets will have an assumable cost for publication in order to proceed, therefore, to their opening. In addition, a specific case of application in the field of Ecuadorian universities is established throughout the research work. These are in a constant process of innovation and seek the participation and collaboration of students and the university community in general, to generate value-added IT products and services through the opening of their data sets.

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1: Introducción

Este capítulo introductorio se ha organizado en varias secciones. La primera presenta la motivación y el contexto del trabajo de investigación realizado en esta tesis doctoral. La segunda sección describe los conceptos relacionados con el término Gobierno Abierto. La tercera sección introduce el concepto de Datos Abiertos. La cuarta sección presenta conceptos sobre la reutilización de datos. Finalmente, mientras en la quinta sección se describen el objetivo principal y los específicos, la sexta sección contiene la estructura del documento.

1.1. Motivación y contexto

El uso intensivo de las tecnologías de la información y comunicaciones, Internet, los dispositivos móviles al alcance de todos, los sensores, el internet de las cosas (IoT, Internet of the Things), el creciente impacto de las redes sociales y la Web 2.0, han generado un crecimiento exponencial en la producción de datos tanto en el sector público como en el privado [1]. Con el impulso del concepto de "Gobierno Abierto", las instituciones públicas están publicando sus datos en abierto para que sean reutilizados con el fin de estimular el crecimiento económico y social. La comunidad reutilizadora (ciudadanos, universidades, estudiantes, empresas, periodistas, desarrolladores de software, investigadores, ONGs, etc…) reconoce la importancia de que dichos datos sean publicados en abierto [2], [3]. Para que estos conjuntos de datos tengan la máxima efectividad en la reutilización, debe de asegurarse que estén disponibles de manera estable en el tiempo, así como un mantenimiento adecuado [4]. Esto implica costos de publicación a nivel económico, tanto en tiempo como a nivel presupuestario. Por tanto, las entidades públicas deben implementar estrategias que permitan publicar sus datos en abierto teniendo en cuenta un coste adecuado, evitando publicar datos que no se reutilicen. En la investigación desarrollada en esta tesis doctoral se plantea una propuesta formal que permita dotar a las organizaciones que quieran publicar sus datos en abierto del conocimiento adecuado para trazar este tipo de estrategias.

Son diversos los motivos que han incentivado mi curiosidad por los datos abiertos y que me han llevado finalmente al desarrollo de esta tesis doctoral. Pensando retrospectivamente, vienen a mi mente las primeras experiencias de datos abiertos que conocía. Luego del terremoto de Haití en el 2010, en su capital Puerto Príncipe, se logró levantar datos de geoposicionamiento de la ciudad, ubicando lugares de acopio de alimentos, agua y medicinas que permitieron salvar muchas vidas. Estos datos se levantaron en una plataforma abierta con la colaboración de voluntarios, una aplicación celular y una plataforma abierta de datos geoposicionados. En mi ejercicio profesional tuve la oportunidad de participar en proyectos de recolección de datos en la industria petrolera, a través de sistemas de sensores y programas de recopilación. Los datos se almacenaban en grandes repositorios, pero no se lograba realizar el análisis de estos, debido a varios factores como la identificación de las necesidades de los equipos de personas de análisis y reutilización de datos. Como profesor de la Universidad Central del Ecuador organicé con mis alumnos un proyecto para

apoyar en la recopilación de datos para mapear las áreas de afectación del terremoto de Pedernales en la costa ecuatoriana en el 2016. Paralelamente empecé el trabajo de investigación para el doctorado de informática en la Universidad de Alicante y revisé los trabajos de investigación que en materia de datos abiertos se estaban realizando. Estos trabajos me impulsaron a concursar por fondos para la creación de un proyecto semilla en la Universidad Central del Ecuador. El proyecto consistía en la creación de un portal de datos abiertos para la universidad. Como director del proyecto se gestionó y se publicaron dos artículos resultados del proyecto en una revista interna indexada en Latindex[1]. El trabajo realizado con la colaboración de estudiantes de grado permitió crear el portal de datos abiertos, donde se publicaron varios conjuntos de datos disponibles en ese momento, tales como los estudiantes por periodo académico, género, facultad[2], esta información está disponible en formato pdf, csv y excel.

Dentro de la sociedad civil he tenido la oportunidad de crear y participar en un colectivo de personas dedicadas a la difusión y reutilización de datos abiertos REDAM (red de datos abiertos y metadatos)[3], organización de la que soy su secretario y nos dedicamos a impulsar las iniciativas de datos abiertos en Ecuador. A través de la REDAM se realizó en marzo de 2018 el primer congreso de datos abiertos con la participación de 12 proyectos[4]. Como representante de la REDAM y la academia he participado en las mesas de discusión sobre la política de datos abiertos del Ecuador y la Ley de Protección de Datos Personales.

El Gobierno del Ecuador, en el año 2018 decidió entrar al Open Goverment Partnership (OGP)[5], lo que ha incentivado a que las entidades públicas estén preparando la agenda donde se contempla el apoyo a la publicación en abierto de sus conjuntos de datos. Las propuestas de ley y reglamentos los maneja la Presidencia de la República y el Ministerio de Telecomunicaciones y de la Sociedad de la Información. Las universidades como entes dinamizadores de proyectos tecnológicos y

[1] ***http://revistadigital.uce.edu.ec/index.php/RevFIG/article/view/865***
http://revistadigital.uce.edu.ec/index.php/RevFIG/article/view/73

[2] ***http://datosabiertos.uce.edu.ec/***

[3] ***https://datosabiertosecuador.org/***

[4] ***https://congreso.datosabiertosecuador.org/programa/***

[5] ***https://www.opengovpartnership.org/***

sociales se encuentran proponiendo la apertura de datos no solo por el cumplimiento de la ley de transparencia, sino también por la promesa de innovación y crecimiento productivo de esta acción.

1.2. Gobierno Abierto

El término Gobierno Abierto no es novedoso, sino que ya se tienen referencias del mismo desde los años 70 [5]. El gobierno abierto surge como un nuevo modelo de relación entre los gobernantes, las administraciones y la sociedad con el fin de garantizar que los servicios de las administraciones públicas puedan ser supervisados por la ciudadanía, de tal manera que (i) se incrementa la transparencia de la administración hacia la ciudadanía, (ii) se generan espacios de encuentro y participación entre administración y ciudadanía, y (iii) se canaliza el potencial innovador de la ciudadanía en colaboración con las organizaciones privadas y públicas para el beneficio de la sociedad a través de la reutilización de datos y la búsqueda de soluciones a los problemas públicos [5]. Recientemente, en el año 2009, el memorando Obama sobre Gobierno Abierto ha convertido este modelo de gobernanza en una corriente dominante [6] resaltando las tres componentes básicas del gobierno abierto: (i) transparencia (es decir, rendición de cuentas hacia la ciudadanía), (ii) participación (es decir, posibilidad de intervención de la ciudadanía en las actividades del gobierno), y (iii) colaboración (ciudadanía, sector público y privado ayudan a solventar los problemas públicos reutilizando los datos disponibles y su conocimiento). A continuación, se describen más en detalle cada uno de estos conceptos:

- **Transparencia:** un gobierno transparente pone a disposición de la ciudadanía todas sus fuentes de datos, con el fin de que pueda haber una verdadera rendición de cuentas y un permanente control social. Los gobiernos transparentes garantizan el derecho de acceso a la información, desarrollando acciones de publicidad activa [5].

- **Participación:** un gobierno participativo genera espacios de encuentro entre organismos públicos y ciudadanía que permiten la participación activa en la formulación de políticas públicas, compartiendo datos y beneficiándose del conocimiento y experiencia de la ciudadanía, que forma parte del proceso de toma de decisiones [5].

- **Colaboración:** un gobierno colaborativo compromete e implica a los ciudadanos y sectores sociales para trabajar conjuntamente para resolver los problemas nacionales [5] para lo cual es

necesaria la compartición de datos del sector público. Esta colaboración implica una gran oportunidad de emprendimiento, ya que aprovecha el potencial de la sociedad en la reutilización de datos de las administraciones públicas, contribuyendo a la mejora de los servicios públicos mediante la innovación basada en esos datos.

Según estas definiciones, se puede destacar, que el acceso a la información es fundamental para un gobierno abierto, es decir gobierno abierto implica disponibilidad de datos accesibles y reutilizables: los denominados datos abiertos [7],[8].

1.3. Datos abiertos

Según la definición dada por la OKFN (Open Knowledge Foundation), datos abiertos (open data) son aquellos que se pueden acceder, reutilizar y redistribuir libremente por cualquier persona, sujetos únicamente a la obligación de atribuir la procedencia de los mismos [9]. Los datos se publican en sitios de acceso público, conocidos como portales de datos abiertos, lo que permite su fácil organización y consulta.

Publicar datos abiertos es fundamental para que las organizaciones logren alcanzar altas cotas de transparencia, lo que implica que sean más colaborativas y participativas con su entorno [10]. Cabe destacar que la transparencia y el derecho de acceso a la información de las organizaciones públicas son actualmente funciones esenciales en la sociedad ya que permiten una adecuada rendición de cuentas [11]. Sin embargo, más allá de la importancia de la transparencia de los datos, con el único fin de la rendición de cuentas de dichas organizaciones, se encuentra la participación de los reutilizadores en el uso y la creación de información a través del trabajo de red colaborativo [12]. De hecho, la apertura de datos no es un producto sino un proceso que debe ser visto desde la perspectiva del usuario de estos datos, es decir, el colectivo reutilizador de datos abiertos [13].

Un ejemplo paradigmático de proyecto de apertura de datos de administraciones públicas es el proyecto APORTA [14], impulsado por el Ministerio de Industria, Energía y Turismo y el Ministerio de Hacienda y Gobernaciones Públicas, ambos del Gobierno Español. En este proyecto se llevó a cabo un trabajo de recopilación de conclusiones, síntesis y constitución de un decálogo, como elemento fundamental a tener en cuenta para potenciar la filosofía de datos abiertos en España. Este decálogo busca garantizar que los datos publicados en abierto tengan un nivel de calidad que satisfaga las expectativas de los consumidores de datos (es decir de la comunidad

reutilizadora). A pesar de que no es el único decálogo que existe, este está ampliamente difundido en España y varios ayuntamientos lo han incorporado en sus estrategias de apertura de datos, tal es el caso de Andalucía, el País Vasco entre otros. Un breve resumen de este decálogo extraído del proyecto Aporta se detalla a continuación:

0. Armonización entre administraciones. Todos los puntos del decálogo se basan en la premisa de que debe existir una armonización entre todas las administraciones públicas. Todas las iniciativas de datos abiertos deben compartir los mismos principios y definiciones que se listan en el decálogo.

1. Publicar datos en formatos abiertos y estándares. Cualquier iniciativa de datos abiertos debería publicar sus conjuntos de datos en formatos abiertos (no propietarios) y que sean adecuados para permitir la reutilización de los mismos por parte del colectivo destinatario.

2. Usar esquemas y vocabularios consensuados. Además de los formatos abiertos y estándares, la estructura de los datos debería seguir un convenio o unos esquemas definidos. Si se crean vocabularios o esquemas de representación de la información específicos, éstos se deberían exponer públicamente para que se pueda interpretar correctamente la información al reutilizarla.

3. Inventario en un catálogo de datos estructurado. Cualquier iniciativa de datos abiertos debe tener un punto de consulta donde se incluya un inventario con información descriptiva y técnica sobre los conjuntos de datos que se exponen. Los metadatos que informan sobre cada conjunto de datos deberían seguir una estructura común y estándar.

4. Datos accesibles desde direcciones web persistentes y amigables. Tanto las fichas de los conjuntos de datos, como la distribución de la propia información (volcado en un archivo, API de consulta, RSS, etc.) deberían de estar accesibles desde URLs que persistan en el tiempo y así evitar que se pierdan las referencias en el futuro. Además, deben seguir una estructura homogénea y bien definida, con información legible con el fin de facilitar su reutilización mediante el conocimiento del contenido referido por dichas URLs.

5. Exponer un mínimo conjunto de datos relativos al nivel de competencias del organismo y su estrategia de exposición de datos. Cada administración que impulse una iniciativa de datos abiertos debería crear una hoja de ruta donde especifique la estrategia de exposición de los conjuntos de datos y sus prioridades. Inicialmente, debería publicar los conjuntos de mayor interés según las competencias del propio organismo.

6. Compromiso de servicio, actualización y calidad del dato, manteniendo un canal eficiente de comunicación entre reutilizador y administraciones públicas. La administración debe mantener un mínimo de calidad y servicio en su iniciativa de datos abiertos, manteniendo lo expuesto en la estrategia de publicación y comprometiéndose con su colectivo reutilizador. Debe establecer un canal eficiente de comunicación que permita la interacción bidireccional entre organismo público y reutilizadores.

7. Monitorizar y evaluar el uso y servicio mediante métricas. La administración debe crear métricas y evaluar sus indicadores de uso y servicio de la iniciativa de datos abiertos. De esta forma puede monitorizar el funcionamiento y uso, y así analizar si se está cumpliendo el compromiso con la comunidad de reutilizadores y cuáles son las potenciales carencias del sistema o de la estrategia.

8. Datos bajo condiciones de uso no restrictivas y comunes. Las condiciones de uso deberían ser lo menos restrictivas posible y permitir la reutilización libre, incluso para fines comerciales. Se recomienda la creación y uso de licencias tipo autodocumentadas (por ejemplo ODbL[6]) y que sean comunes entre distintas administraciones.

9. Evangelizar y educar en el uso de datos. Es necesario educar en el uso de los datos, tanto a los colectivos de reutilización específicos (sector TIC, periodismo, investigación, etc.) como a la sociedad en general y así fomentar el conocimiento y la inquietud por procesar información de una forma autónoma.

10. Recopilar aplicaciones, herramientas y manuales para motivar y facilitar la reutilización. Cualquier iniciativa de datos abiertos debería recopilar ejemplos de uso y herramientas que faciliten y motiven la reutilización de los datos que se publican.

Cumplir con este decálogo no es una tarea sencilla para las administraciones públicas, por lo que desde diversas instituciones se trabaja para facilitar a las administraciones públicas la publicación de datos en abierto. Un ejemplo es la guía de la Federación Española de Municipios y Provincias (FEMP [7]) que ayuda a las administraciones locales y provinciales a publicar sus datos en abierto.

[6] ***https://opendatacommons.org/licenses/odbl/index.html***

[7] ***http://femp.femp.es/files/3580-1617-fichero/Gu%C3%ADa%20Datos%20Abiertos.pdf***

1.4. Reutilización de datos

Los gobiernos y el sector público en general poseen una amplia gama de información y contenidos, que es potencialmente reutilizable por los ciudadanos y por la industria como contenidos digitales de información social, económica, geográfica, estadística, turística, o meteorológica y sobre empresas y educación [14].

Cuando la información reutilizable procede del sector público surge el concepto RISP (Reutilización de Información del Sector Público). RISP incluye toda iniciativa en la cual personas físicas y jurídicas, con fines comerciales o no comerciales, usen documentos que obran en poder del sector público, siempre que dicho uso no constituya una actividad administrativa pública (Proyecto Aporta, 2012)[14]. RISP requiere de la participación de tres agentes fundamentales: Los infomediarios, los ciudadanos y la administración pública.

Infomediarios: se trata de emprendedores y empresas que participan en el modelo de colaboración mediante la combinación, enlace, procesamiento y elaboración de los datos reutilizables. Este sector está formado por empresas y organizaciones que aprovechan el potencial de la enorme cantidad de datos estructurados procedentes tanto del sector público como el privado, para crear nuevos servicios y productos de valor[14].

Ciudadanos: el principal beneficiario del modelo de colaboración RISP es el ciudadano que al usar los datos en forma de productos y servicios obtiene una mejora. Son numerosos los casos de éxito de estos nuevos servicios basados en la reutilización: aplicaciones y webs de servicios en varios sectores [14]. En el capítulo del estado del arte se amplía los temas donde se han trabajado con datos abiertos.

Administraciones Públicas: la reutilización por parte de las administraciones públicas de su propia información genera mayor efectividad en la toma de decisiones a la hora de aplicar políticas sociales o económicas. Por otra parte, existen beneficios al mejora la calidad del servicio y ahorros de costes desarrollando soluciones, necesarias para la sociedad [14].

La necesidad de estimular la reutilización es una tarea clave, pues permite impulsar la capacidad de innovación de desarrolladores e infomediarios. De acuerdo a la Fundación CTIC[8], el impulso de las políticas más la disponibilidad de estándares tecnológicos maduros, están posibilitando una gran revolución en la forma de distribuir y consumir información pública, donde no solo las webs estarán enlazadas entre sí, sino también los datos, reduciendo drásticamente el costo de la reutilización, haciendo muy sencilla su integración con futuras aplicaciones. Se debe conformar una suerte de ecosistema entre el gobierno, las empresas y la ciudadanía para fomentar el consumo de información pública y contribuir así al bienestar social [15].

Otro aspecto importante cuando se publican datos es conocer qué datos son más relevante para los ciudadanos, infomediarios y en general los reutilizadores de datos. Para lograr esto, una primera propuesta pudiera ser incorporar las redes sociales a los canales habituales de participación [16]. Hay que basar la oferta de datos en su demanda para evitar la publicación infinita e irrelevante y reducir el presupuesto de publicación. De hecho, "los proyectos de datos abiertos tienen éxito allí donde satisfacen la demanda existente de información y compromiso. Debe existir una asociación activa entre el gobierno y los actores privados" [15].

Los datos abiertos pueden encontrarse en cualquier formato, pero para su reutilización y automatización por los lenguajes de programación es ideal encontrarlos en formatos estructurados, no propietarios. Ejemplos de formatos reutilizables son CSV, JSON, XML, etc. Otros formatos comunes como el formato PDF proporcionan menores posibilidades de reutilización [14].

1.5. Objetivos

La apertura de datos puede dar beneficios significativos y concretos, tanto para el Gobierno como para los ciudadanos. Esto, solo si los datos a publicarse son cuidadosamente planeados y ejecutados [15]. Demasiado a menudo el Gobierno ha invertido recursos para publicar ciertos conjuntos de datos, pero ha encontrado a muy pocos creadores dispuestos a desarrollar nuevas herramientas y servicios para utilizarlos. De hecho, gran parte de los datos publicados aún no se han destinado a usos subsidiarios significativos. Data.gov, por ejemplo, dice tener más de 3,500

[8] ***https://www.fundacionctic.org/servicios/open-government***

conjuntos de datos de alto valor, pero afirma tener menos de 250 aplicaciones desarrolladas por ciudadanos sobre la base de sus datos [17].

Cabe destacar que en la mayoría de guías y trabajos sobre publicación de datos abiertos no se profundiza en desarrollar estrategias de selección de datos a publicar. Los datos abiertos que se publican son aquellos que marca la legislación vigente, normalmente siguiendo el principio de publicidad activa de la legislación sobre transparencia y que no presenten problemas de privacidad según la legislación de protección de datos.

Para asegurar que el retorno de la inversión pública sea óptimo, se debe disponer de una estrategia de selección de datos abiertos que priorice la publicación de aquellos datos que tengan más potencial para ser reutilizados (es decir, que los datos que se van a publicar sean relevantes para el colectivo reutilizador), a la vez que su coste de publicación sea asumible [18],[19].

Estudios recientes involucran al colectivo reutilizador en la selección de los conjuntos de datos a publicar en abierto. Por ejemplo en [6] se propone un proceso para desarrollar un portal de datos abiertos con varios conjuntos de datos seleccionados desde el punto de vista del reutilizador. Sin embargo, estos estudios actuales no consideran ningún método formal y, desafortunadamente, sólo se basan en la intuición de las personas responsables del portal de datos abiertos, por lo que las decisiones en cuanto a seleccionar datos a abrir pueden no ser acertadas. De igual manera, una vez definidos los conjuntos de datos desde el punto de vista del reutilizador es preciso determinar si van a tener un costo asumible para la publicación. Los costos de implementación incluyen hardware, software y recursos humanos, más allá del trabajo técnico para la implementación de un proceso [4].

El objetivo principal de este trabajo de tesis doctoral es proporcionar un método para la selección de los conjuntos de datos a abrir por parte de organizaciones que aborden procesos de apertura. El método se propone con el objetivo de ser general y aplicable en diferentes contextos que impliquen distintas organizaciones y colectivos. Además, se busca que el método tenga una base científica, debiendo estar dotada del formalismo necesario para garantizar una aplicabilidad y obtención de resultados sistemática. Finalmente, de acuerdo con las conclusiones del estado del arte, este

método deberá integrar el punto de vista del reutilizador de datos abiertos, así como el del publicador de los mismos.

Para conseguir este objetivo general se requiere abordar una serie de objetivos específicos que guiarán el desarrollo de esta tesis doctoral:

- Analizar y clasificar, de manera sistemática, la investigación de datos abiertos realizada por la comunidad científica desde un punto de vista tecnológico, identificando espacios de mejora en el proceso de apertura de datos.
- Estudiar métodos de selección de conjuntos de datos a publicar en abierto para identificar posibles mejoras, tanto en los métodos como en los participantes en la toma de decisiones.
- Analizar métodos científicos aplicables al problema de selección de conjuntos de datos a abrir con el carácter formal requerido. Investigar sobre el Método Delphi Difuso y su aplicabilidad en este problema.
- Proponer un método para la selección de conjuntos de datos a publicar en abierto basada en el Método Delphi Difuso, que integre el punto de vista del reutilizador de datos abiertos y el publicador de los mismos.
- Aplicar el método propuesto para la selección de conjuntos de datos a abrir en un caso de estudio como el de las Universidades Ecuatorianas, con dimensión y complejidad para la obtención de conclusiones.

1.6. Estructura

El presente trabajo se ha organizado en cinco capítulos. En el siguiente capítulo se presenta la metodología utilizada en las diferentes etapas de la investigación. En el capítulo 3 se expone el estado del arte en materia de datos abiertos, se clasifican trabajos de investigación atendiendo a diferentes criterios y se obtienen conclusiones sobre las contribuciones de la investigación realizada en relación con el estado del arte. En el capítulo 4 se describe el método propuesto para la selección de conjuntos de datos a abrir. También se describe la aplicación del método en la selección de conjuntos de datos abiertos en el ámbito universitario en Ecuador y la discusión sobre los resultados obtenidos. En el capítulo 5 se exponen las conclusiones, las principales contribuciones y líneas futuras de investigación. Los anexos incluyen materiales, herramientas y datos asociados al trabajo de investigación desarrollado.

Capítulo 2: Metodología de investigación

En este capítulo se describen las metodologías utilizadas en la investigación. En primer lugar, se describe el método utilizado para estudiar y clasificar la literatura relacionada con la investigación en el campo de los datos abiertos. En segundo lugar, se detalla el método formal utilizado como núcleo del método propuesto para la selección de conjuntos de datos. Se trata de una variante del método Delphi conocida como Delphi Difuso.

2.1 . Mapeo Sistemático

Un mapeo sistemático tiene como objetivo encontrar y clasificar las publicaciones primarias en un área temática específica siguiendo un método bien definido y repetible [20]. Consiste en varios pasos: (1) obtención de un esquema de clasificación, (2) recopilación de publicaciones pertinentes, (3) clasificación, y (4) análisis de los resultados [21]. El análisis se centra en responder a preguntas específicas de investigación, generalmente relacionadas con la identificación y cobertura del campo y sus subcampos, y en la evolución a lo largo del tiempo, así como en la discusión adicional sobre los desafíos y las tendencias en relación con el tema específico [22].

Para el presente trabajo, se aplicó el proceso introducido en [22] basado en la adaptación de los estudios de mapeo sistemático en el campo médico y su aplicación a la ingeniería del software [21] . El proceso consistió en definir: (1) alcance de la investigación, (2) proceso de búsqueda, (3) esquema de clasificación, (4) mapeo de las publicaciones según el esquema de clasificación y (5) análisis de datos. Una descripción detallada de este proceso se muestra en la Figura 1. Los pasos a seguir se detallan a continuación.

2.1.1. Alcance de la investigación

El objetivo general del mapeo sistemático es proporcionar una visión general consolidada de la investigación en el campo de estudio. Para ello se definen preguntas de investigación que guiarán el proceso. Las preguntas permitirán definir el alcance de la investigación, en forma clara y detallada. Ejemplos de preguntas son:

- ¿Cuáles son los lugares de publicación en los que se han publicado la investigación?
- ¿Qué impacto ha tenido la investigación?
- ¿Qué ámbitos recibieron mayor cobertura en la investigación, en qué medida y cómo está evolucionando la cobertura?
- ¿Cuáles son los temas que se abordan en la investigación, hasta qué punto se cubren y cómo evoluciona la cobertura?

2.1.2. Proceso de búsqueda

El objetivo de este paso es identificar las publicaciones científicas primarias que son relevantes dentro del ámbito de la investigación de una manera definida y repetible. Se realiza un proceso de búsqueda en tres etapas para identificar dichos estudios primarios:

ETAPA 1: Búsqueda inicial, para obtener el conjunto inicial de publicaciones científicas primarias potencialmente relevante, se realiza una búsqueda exhaustiva utilizando motores de búsqueda científica de los repositorios más importantes (por ejemplo, aquellos pertenecientes a editoriales dentro del campo de investigación). Las búsquedas se realizan en un periodo de tiempo definido de acuerdo a la cobertura temporal que se desea abarcar. Se utilizan cadenas de búsqueda con la (o las) palabra(s) clave(s) de la investigación. Se trabaja en el hallazgo de la cadena de búsqueda en el título y abstract o en el título, abstract y conclusiones, dependiendo del alcance del estudio.

ETAPA 2: Revisión de los resultados, al conjunto inicial de publicaciones se define y aplica criterios de inclusión y exclusión:

- ***Criterios de inclusión***. Los criterios pueden ser variados y dependerán de la investigación que se esté realizando, por ejemplo, se puede incluir entre otros: publicaciones revisadas por pares en inglés para las cuales el texto completo esté disponible, es decir, los trabajos de investigación completos publicados en revistas, conferencias, simposios o talleres revisados por pares. Se define un espacio de tiempo de publicación.

Criterios de exclusión. Igualmente los criterios de exclusión dependerán de la investigación, se puede incluir entre otros: fuentes que no pasaron por un proceso de revisión por pares o que no constituyeron contribuciones puras de investigación, por ejemplo, libros, tesis doctorales y de maestría, descripciones de patentes, normas y recomendaciones, resúmenes de libros o tesis, informes técnicos, libros blancos, charlas invitadas, documentos de demostración, documentos de tutoría, publicaciones de pósteres, pósteres, editoriales, prefacios, artículos o columnas en revistas que no han sido revisadas por pares, boletines informativos, entradas de enciclopedias, resúmenes o entradas de blog.

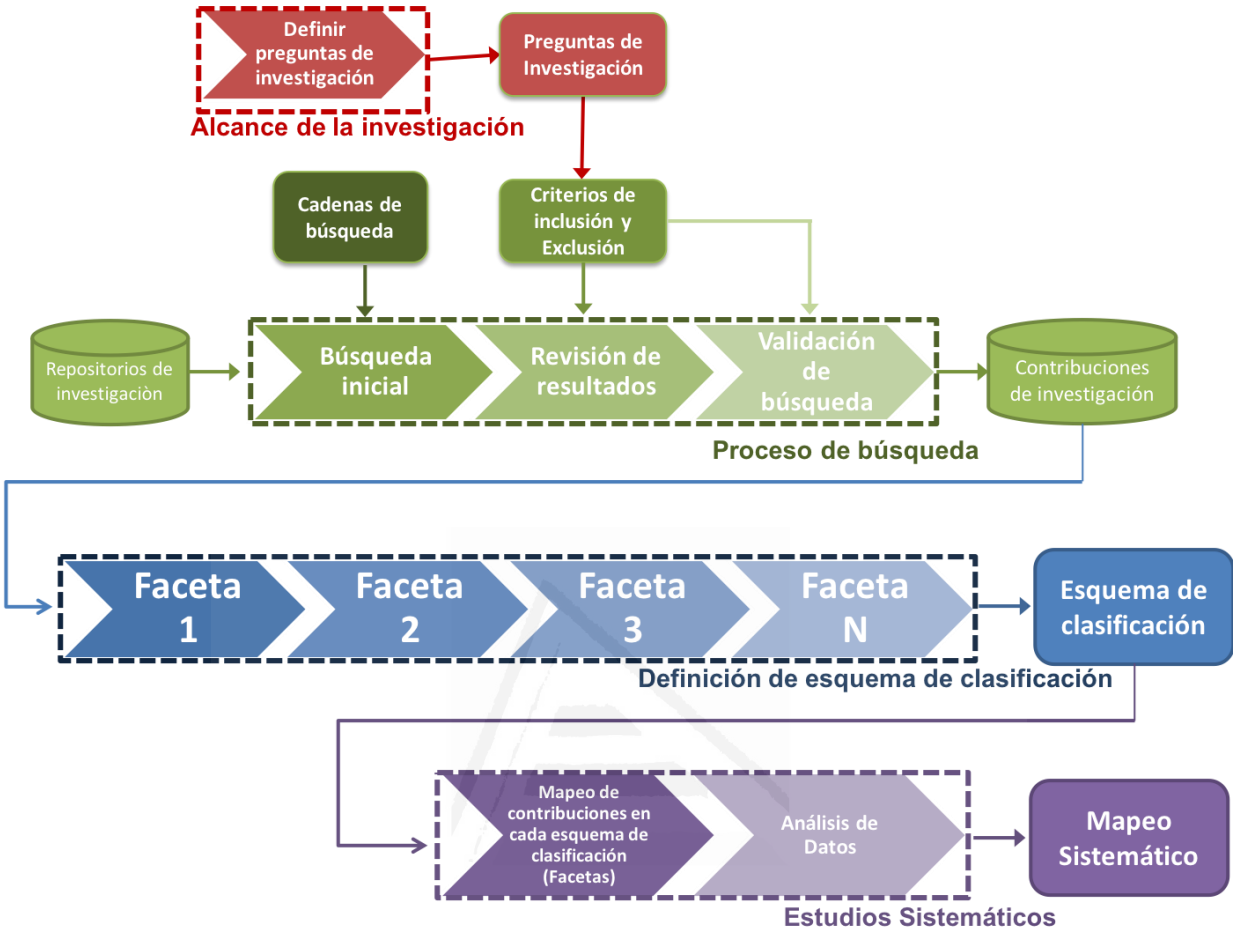


Figura 1. Proceso de investigación para el mapeo sistemático

ETAPA 3: Validación de la búsqueda, se debe validar el proceso de búsqueda con alguna metodología como revisión de pares, grupos focales, entre otras.

2.1.3. Definición del esquema de clasificación

Una vez obtenido el conjunto final de publicaciones científicas, se elabora un esquema de clasificación correspondiente al alcance de la investigación y a las preguntas planteadas en dicho alcance. Para esto se pueden considerar facetas para la clasificación de acuerdo a la investigación que se realiza. Las facetas están vinculadas a la temática de la investigación, pero se podrían utilizar entre otras: lugar de publicación, impacto, tipo de investigación, etc.

2.1.4. Esquema de clasificación

Basándose en el esquema de clasificación definido en el punto anterior se procede a aplicarlo a las publicaciones relacionadas con la investigación y se valida este proceso a través de la revisión por pares.

2.1.5. Amenazas a la validez del mapeo sistemático.

En la literatura, las principales dificultades de los mapeos sistemáticos se han identificado de la siguiente manera: sesgo en la selección de publicaciones [23], errores en la categorización de las publicaciones en categorías detalladas [24] y errores duplicados o contribución adicional débil de una publicación a otra (los llamados "delta papers", es decir, documentos que proporcionan sólo adiciones menores en comparación con trabajos publicados anteriormente por los mismos autores).

2.2 . Método Delphi Difuso

El Método Delphi tradicional se basa en un proceso iterativo para la construcción de consenso ante un panel de expertos que son anónimos o parcialmente anónimos entre sí [25], [26]. Con la variación realizada por Ishikawa en [27], donde usa la lógica difusa, el Método Delphi tradicional se convierte en el Método Delphi Difuso (MDD). Noorderhaben indicó que la aplicación del Método Delphi Difuso a la decisión del grupo puede resolver la imprecisión del entendimiento común de las opiniones de los expertos y de esta manera se utiliza en investigaciones cuantitativas [25],[28],[29].

Los pasos del Método Delphi Difuso son: (1) recopilar opiniones del grupo de expertos, (2) configurar números difusos triangulares y (3) desfuzzificación, convertir los números difusos en valores definidos. Estos pasos se muestran en la Figura 2.

2.2.1. Recopilar opiniones del grupo de expertos

Como primer paso, se debe identificar a los expertos. Según [25], de una revisión de 50 estudios que usaron el método Delphi, en el 95% de los casos los expertos fueron seleccionados por muestreo intencional. Para el muestreo intencional se buscan expertos en base a la revisión de la literatura, búsquedas en Internet, listados y referencias de organizaciones profesionales y simplemente expertos conocidos por el investigador. Siempre es recomendable verificar la experiencia a través del examen del currículum vitae de los expertos.

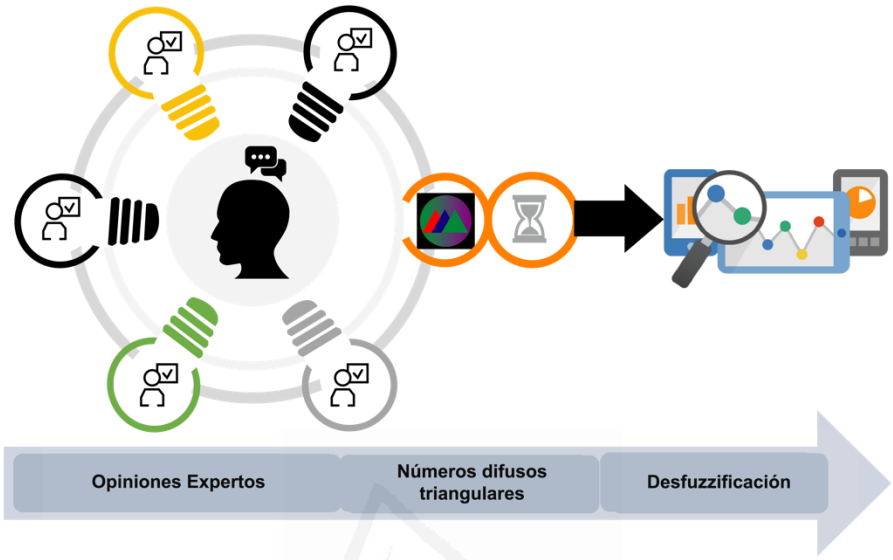


Figura 2. Método Delphi Difuso

El número de expertos dependerá de la naturaleza del estudio. De acuerdo con [25], [30],[31] se han realizado con la participación de 6 o más expertos. En cuanto a la selección de las funciones de membresía difusa, investigaciones anteriores [27], [32],[33] se basaron generalmente en el número difuso triangular, el número difuso trapezoidal y el número difuso gaussiano.

Para el caso de esta investigación se utiliza el número difuso triangular. En función de esta definición, se deben generar preguntas de investigación que generen un número triangular difuso. Es difuso porque se trata de un juicio subjetivo. Es triangular porque hay tres juicios. La misma matemática difusa puede extenderse a los casos en los que hay más de tres sentencias [25], [34].

Una vez que se han realizado las preguntas de investigación y se han definido los expertos, se recopila la puntuación de evaluación de la importancia de cada factor dada por cada experto mediante el uso de variables lingüísticas que aparecen en la construcción de las preguntas de investigación [32].

2.2.2. Configurar números difusos triangulares

Se calcula el valor de evaluación del número difuso triangular de cada factor alternativo dado por los expertos. Se averigua el número difuso triangular significativo del factor alternativo. Este estudio utilizó la media geométrica del modelo general medio propuesto por Klir y Yuan [35] para MDD, donde se averigua el entendimiento común de la decisión de grupo. La fórmula de cálculo se ilustra a continuación:

$\widetilde{W_{ij}}$ representa el peso difuso otorgado por el experto i al elemento evaluado j y que se representa como sigue:

$$\widetilde{W_{ij}} = ((l_{ij}, m_{ij}, u_{ij}) \quad (1)$$

Donde:

i representa un experto

j representa el elemento evaluado.

l_{ij} representa el valor mínimo otorgado por el experto i para el elemento evaluado j.

u_{ij} representa el valor máximo otorgado por el experto i para el elemento evaluado j.

m_{ij} representa el valor medio otorgado por el experto i para el elemento evaluado j.

En la investigación se utiliza la función de membresía triangular, esta nos indica el grado en que cada elemento de un universo dado pertenece a dicho conjunto.

De acuerdo a la función de membresía triangular, el grado de satisfacción de los extremos l_{ij} y u_{ij} se representa como 0, mientras que los segmentos entre estos valores tendrán un grado de satisfacción entre 0 y 1 [36],[28], como se muestra en la Figura 3.

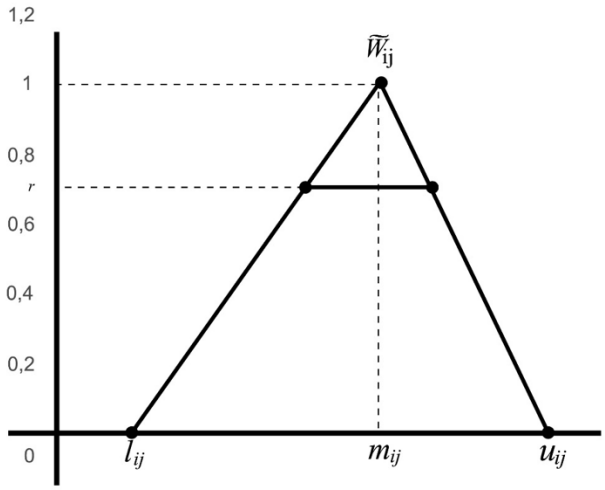


Figura 3. Función de Membresía

A continuación, se evalúan los pesos difusos de los expertos para obtener el peso promedio para el elemento evaluado j. El peso promedio se representa como sigue [37]:

$$\widetilde{W_j} = (l_j, m_j, u_j) \quad (2)$$

Donde,

$$l_j = \text{Min}(l_{ij}) \quad (3)$$

i = 1, 2, …, n; j = 1, 2, …, m.

$$m_j = \left(\prod_{i=1,j=1}^{n,m} m_{ij}\right)^{\frac{1}{n}} \quad (4)$$

i = 1, 2, …, n; j = 1, 2, …, m.

$$u_j = \text{Max}(u_{ij}) \quad (5)$$

i = 1, 2, …, n; j = 1, 2, …, m

n representa el número total de expertos.

m representa el número total de elementos evaluados.

La verificación de consensos puede realizarse, cuando se cumplen las siguientes inecuaciones:

$$u_{ij} > l_j \; \forall \; i,j \quad (6)$$

$$G\left(G=u_{ij}-l_j\right) > C\left(C=m_j-m_{ij}\right) \quad (7)$$

En la Figura 4 se representan estas condiciones. El cumplimiento de estas significa que los expertos han alcanzado un consenso en el elemento evaluado j [30], [38].

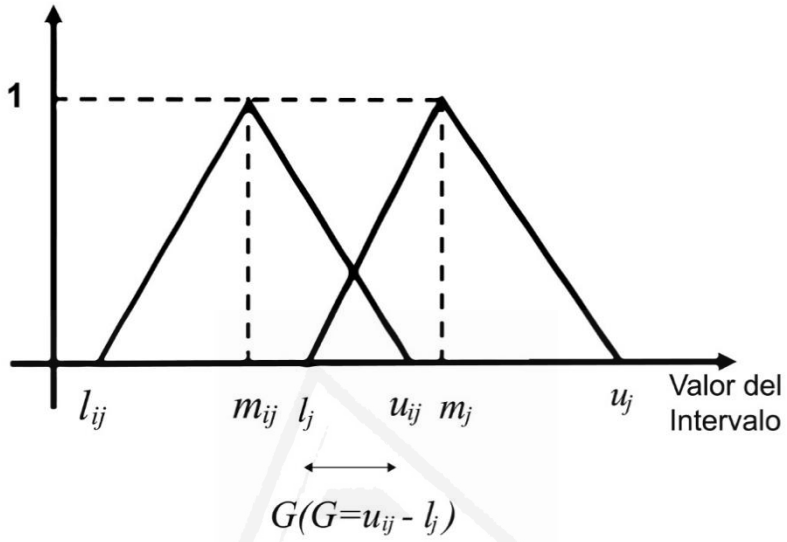


Figura 4. Verificación de consensos

2.2.3. Valores definidos (Desfuzzificación)

Dado que el peso difuso $\widetilde{W}_j$ no puede ser usado para comparación directa, es necesario utilizar el promedio difuso y el método disperso para transformar el resultado en un número definido o nítido S_j. El cálculo se realiza como sigue [33], [39]:

$$S_j = \frac{l_j+m_j+u_j}{3} \quad (8)$$

A continuación, se seleccionan los conjuntos de datos basados en las siguientes reglas:

Si $S_j \geq r$, el elemento evaluado deberá ser seleccionado.

Si $S_j < r$, el elemento evaluado deberá ser eliminado.

En estudios similares [27],[40], un valor de r de 0,7 ha sido seleccionado como referencia, pero esto dependerá del índice de confiabilidad que se necesite para cada estudio, pudiéndose tomar valores menores o mayores a 0,7.

2.2.4. Amenazas a la validez del Método Delphi Difuso

Son varios los factores que pueden amenazar la validez del Método Delphi Difuso, a saber:

- Se debe lograr el compromiso de los expertos acerca de permanecer en el estudio hasta lograr los resultados [25].
- Las formulaciones de las preguntas con las variables lingüísticas adecuadas constituyen un factor clave para que los expertos puedan desarrollar los cuestionarios de forma efectiva. La selección adecuada de preguntas del cuestionario sencillas y claras garantizan el éxito del estudio. Desafortunadamente, esta actividad puede llevar mucho tiempo.
- Finalmente, otro factor muy importante para el éxito es la selección de los expertos [32].

Capítulo 3: Estado del arte

En esta sección se presenta en detalle la aplicación de la metodología de mapeo sistemático en materia de datos abiertos. En concreto, se profundiza en los resultados de la investigación, relaciones entre las facetas de clasificación, discusión de los resultados y conclusiones que nos permiten identificar la necesidad de tener los métodos de selección de conjuntos de datos para su apertura de una manera formal, con base científica, y que permita relacionar a los reutilizadores y publicadores de datos abiertos.

3.1. Mapeo sistemático sobre investigación en datos abiertos

Para esbozar el estado del arte en lo referente a los datos abiertos, se utiliza la metodología de mapeo sistemático que ha permitido encontrar y clasificar las publicaciones científicas en esta área del conocimiento. El proceso consistió en: definir el alcance de la investigación, definir el proceso de búsqueda, el esquema de clasificación, el mapeo de las publicaciones según el esquema de clasificación y el análisis de datos. Una descripción detallada de este proceso aplicado a los datos abiertos se puede ver en la Figura 5. A continuación, en esta sección, se detalla el proceso llevado a cabo.

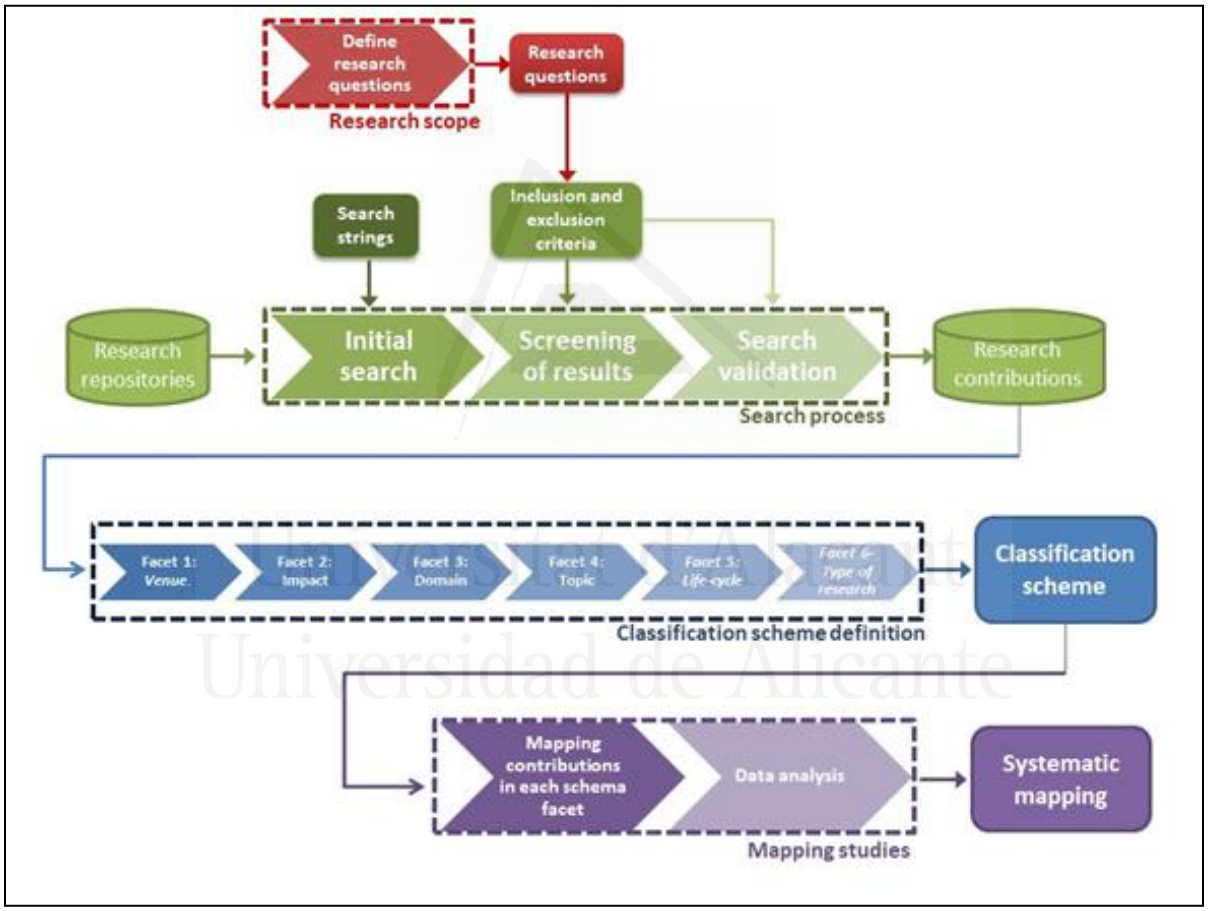


Figura 5. Proceso de investigación para el estudio sistemático de datos abiertos

3.1.1. Alcance de la investigación

El objetivo general del estudio sistemático fue proporcionar una visión general consolidada de la investigación en el campo de los datos abiertos desde una perspectiva tecnológica, a través de sus

lugares de publicación, dominios de impacto, fases de publicación de los datos, nivel de impacto y su evolución a lo largo del tiempo, de esta manera identificar las áreas con mayores oportunidades de investigación. El proceso de desarrollo respondió a las siguientes preguntas de investigación:

- ¿Cuáles son los lugares de publicación en los que se han publicado la investigación de datos abiertos?
- ¿Qué impacto ha tenido la investigación sobre los datos abiertos?
- ¿Qué ámbitos recibieron mayor cobertura en la investigación de datos abiertos, en qué medida y cómo está evolucionando la cobertura?
- ¿Cuáles son los temas que se abordan en la investigación de datos abiertos, hasta qué punto se cubren y cómo evoluciona la cobertura?
- ¿Qué fases del ciclo de vida de los datos se han considerado en la investigación de datos abiertos?
- ¿Qué tipos de investigación de datos abiertos se han reportado, hasta qué punto y cómo está progresando la evolución?

3.1.2. Proceso de búsqueda

El proceso de búsqueda tuvo como objetivo identificar las publicaciones científicas que son relevantes dentro del ámbito de la investigación de una manera definida y repetible. Se realizó un proceso de búsqueda en tres etapas para identificar dichas publicaciones científicas.

PASO 1: Búsqueda inicial, para obtener el conjunto inicial de las publicaciones científicas potencialmente relevantes, se realizó una búsqueda exhaustiva utilizando motores de búsqueda científica de los repositorios de editores dentro del campo de la tecnología de la información (de acuerdo con la perspectiva tecnológica de este estudio); por lo tanto, se seleccionaron los motores de búsqueda incluidos en la Tabla 1.

Dado que el objetivo era estudiar las publicaciones sobre datos abiertos, los títulos de las publicaciones tenían que incluir la cadena de búsqueda específica: "datos abiertos" (“open data” en inglés). Se incluyeron publicaciones entre 2006 y 2017. Se tuvieron en cuenta las publicaciones en inglés. Los motores de búsqueda general de publicaciones científicas como Google Scholar no fueron considerados por dos razones:

Tabla 1. Motores de búsqueda.

Repositorios científicos.	**Url**
Springer RD para investigación y desarrollo	http://rd.springer.com/
IEEE Xplore	http://ieeexplore.ieee.org/.
ACM Digital Library	http://dl.acm.org
IOS Press	http://content.iospress.com
Science Direct	http://www.sciencedirect.com/

1) Indexan las publicaciones de los repositorios mencionados anteriormente
2) Enumeran una gran cantidad de literatura gris (es decir, investigaciones producidas fuera de los canales tradicionales de publicación y distribución académica, tales como informes, documentos de trabajo, documentos gubernamentales, libros blancos, etc.).

Los motores de búsqueda produjeron 671 resultados: éstos fueron clasificados, cuantificados y presentados en el análisis de resultados, basados en las diferentes facetas derivadas de las preguntas de la investigación.

PASO 2: Revisión de los resultados, el conjunto inicial de publicaciones contenía algunos resultados irrelevantes que había que descartar. En el proceso de selección se comenzó por eliminar duplicados. A continuación, se examinó la relevancia de cada publicación con respecto a los objetivos de investigación. Para ello, se definió criterios de inclusión y exclusión, tanto en la forma como en el contenido que se describen a continuación:

- ***Criterios de inclusión***, todas las publicaciones revisadas por pares en inglés para las cuales el texto completo estaba disponible. Esto incluyó todos los trabajos de investigación publicados en revistas, conferencias, simposios o talleres revisados por pares. Se consideraron todos los documentos publicados entre 2006 y 2017.
- ***Criterios de exclusión***, fuentes que no pasaron por un proceso de revisión por pares o que no constituyeron contribuciones puras de investigación, por ejemplo: libros, tesis doctorales y de maestría, descripciones de patentes, normas y recomendaciones, resúmenes de libros o tesis,

informes técnicos, libros blancos, conferencias invitadas, documentos de demostración, documentos de tutoría, publicaciones de pósteres, pósteres, editoriales, prefacios, artículos o columnas en revistas que no han sido revisadas por pares, boletines informativos, entradas de enciclopedias, resúmenes o entradas de blog. Además, se excluyeron las fuentes para las que no se publicó el texto completo, como los resúmenes, los resúmenes ampliados y las presentaciones (presentaciones de diapositivas). Documentos que no se centraban en la publicación o reutilización de datos abiertos, por ejemplo, los relacionados con la publicación de datos institucionales dirigidos únicamente a usuarios internos o los que utilizaban datos abiertos en algunos experimentos, es decir, el contenido no identificaba un problema de investigación relacionado con la reutilización o publicación de datos abiertos, sino que sólo mencionaba de forma incidental los datos abiertos.

PASO 3: Validación de la búsqueda, para validar el proceso de selección, cada uno de los resultados de la búsqueda, los criterios de inclusión y exclusión se verificaron de la siguiente manera: el equipo de investigación que participó en la realización del mapeo sistemático se dividió en dos grupos para asegurar las decisiones correctas; cada publicación del conjunto inicial se asignó a un grupo mientras que el otro grupo revisó todo el proceso. En caso de desacuerdo, ambos grupos discutieron los resultados hasta que llegaron a un consenso. De esta manera, se redujo en 5% la cantidad de publicaciones relevantes.

3.1.3. Definición del esquema de clasificación

Una vez obtenido el conjunto final de publicaciones, se elaboró un esquema de clasificación correspondiente al alcance de la investigación y a las preguntas planteadas en el alcance. Se consideraron seis facetas para la clasificación: lugar, impacto, dominio, temas, fase del ciclo de vida de los datos y tipo de investigación.

Faceta 1: Lugar, identifica el lugar de publicación para lo cual, tras aplicar la cadena de búsqueda, cada motor de búsqueda encontró la fuente de la que procede la publicación, clasificándola por congreso o revista científica y cuantificando el número de publicaciones por congreso o revista científica.

Faceta2: Impacto, describe el campo de acción de la investigación, el esquema de clasificación se divide en:

- Impacto local, que describe un área pequeña y focalizada a la que se refieren los datos abiertos, por ejemplo: una ciudad, un museo, una universidad, un hospital.
- Impacto regional: describe un área mayor relacionada con los datos abiertos, como una provincia, un estado.
- Impacto nacional: describe un país, un estado o una nación.
- Impacto internacional: describe un área internacional.

Faceta 3: Dominio, relacionado con la cobertura en la investigación de datos abiertos. Los dominios se tomaron de las 14 categorías de datos sugeridas por la Carta de Datos Abiertos del G8 [41], y se describen a continuación:

- Agricultura: Seguridad alimentaria, agricultores y consumidores finales, desarrollo agrícola sostenible, nutrición.
- Biología: Artículos que tratan del estudio de la vida y de los organismos vivos.
- Química: Artículos que tratan sobre la materia, la composición material y la reacción.
- Cultura: Artículos que tratan sobre el comportamiento social y las normas que se encuentran en las sociedades humanas.
- Periodismo de datos: Enfoques para la narración y el periodismo, actividades periodísticas.
- Economía: Movimientos financieros, ingresos, gastos y presupuesto.
- Educación: Aplicaciones educativas, rendimiento escolar, habilidades digitales.
- Medio ambiente: Medio ambiente, cambio climático, recursos naturales, información medioambiental, paisajes forestales, meteorología/clima, agricultura, silvicultura, pesca y caza, niveles de contaminación, consumo de energía.
- Geoespacial: Topografía, códigos postales, mapas nacionales, mapas locales, seguridad alimentaria, lagos, zonas geográficas.
- Salud: Trabajos relacionados con el estudio del estado de completo bienestar físico, mental y social y no simplemente la ausencia de enfermedad.
- Humanitario: Asistencia humanitaria, gestión de desastres, actividades de socorro y reconstrucción, desastres, epidemias.
- Infomediarios: Empresarios e inversores que crean empresas para identificar y aprovechar el valor de mercado de la información al consumidor [42].

- Innovación: Conectar datos abiertos con impacto social o para impulsar las economías.
- Ciencia: Investigación y descubrimientos científicos, métodos innovadores para abrir datos científicos y crear nuevas herramientas para manipularlos, financiación de proyectos científicos, colaboración entre grupos de interesados.
- Transporte: Horarios de transporte público, puntos de acceso a la banda ancha.
- Energía: Aplicaciones orientadas al desarrollo energético.
- Turismo: Aplicaciones de datos abiertos aplicadas al desarrollo turístico.

Faceta 4: Tema, relacionado con los aspectos que se abordan en la investigación, los temas fueron tomados de la "Call for Papers" de dos de las conferencias científicas más importantes sobre datos y web semántica. La International Conference on Very Large Data Bases (VLDB) y la International Semantic Web Conference (ISWC). En relación con estas dos conferencias, se realizó una cuantificación y se establecieron los temas más recurrentes en las convocatorias de ambas conferencias en los últimos tres años (2015, 2016 y 2017), de acuerdo a lo siguiente:

- Emprendedor: Uso empresarial de datos abiertos. Uso e impacto de los datos abiertos en países o sectores específicos con el fin de potenciar el negocio y la economía.
- Gobierno: Elaboración, aplicación e institucionalización de políticas de datos abiertos; creación de capacidad para una mayor disponibilidad y utilización de datos abiertos; conceptualización de los ecosistemas e intermediarios de datos abiertos; vínculos entre la transparencia, la libertad de información y las comunidades de datos abiertos; medición de las políticas y prácticas de datos abiertos, incluidos los métodos para evaluar el impacto de los datos abiertos; y ubicación de los datos abiertos en el contexto de la gobernanza mundial y el desarrollo.
- Recuperación de información: Base de datos, recuperación de información, extracción de información, procesamiento del lenguaje natural y técnicas de inteligencia artificial para la web semántica, búsqueda y consulta de la web semántica, bases de datos difusas, probabilísticas y aproximadas, recuperación de información, texto en bases de datos.
- Infraestructura: Web semántica y datos enlazados para entornos de nube, métodos de acceso, control de concurrencia, recuperación, transacciones, indexación y búsqueda, gestión de datos en memoria, aceleradores de hardware, procesamiento y optimización de

consultas, gestión de almacenamiento, ajuste, análisis comparativo, medición del rendimiento, administración y gestión de bases de datos, base de datos como servicio.

- Sistemas inteligentes: Gestión de datos gráficos, redes sociales, sistemas de recomendación, inteligencia de negocio.
- Internet de las cosas (IoT): Flujos de datos e Internet de las cosas, crowdsourcing, bases de datos integradas y móviles, bases de datos en tiempo real, sensores e IoT, bases de datos de flujos.
- Calidad: Limpieza, aseguramiento de la calidad y procedencia de los datos, servicios y procesos de la web semántica, limpieza de datos, filtrado y difusión de información, integración de información, gestión de metadatos, descubrimiento de datos, gestión de datos web, web semántica, sistemas de bases de datos heterogéneos y federados, minería y análisis de datos, almacenamiento.
- Seguridad: Confianza, privacidad y seguridad de la web semántica, privacidad y seguridad en la gestión de datos, retos críticos para los datos abiertos: privacidad, exclusión y abuso.
- Web semántica: Creación de gráficos de conocimiento, razonamiento, uso, representación del conocimiento y razonamiento en la web, gestión escalable de semántica y datos en la web, incluyendo datos enlazados, análisis de datos de web semántica, lenguajes, herramientas y metodologías para representar y gestionar la semántica y datos en la web, arquitecturas y algoritmos para volúmenes grandes de datos, heterogeneidad, dinamicidad y descentralización de datos de web semántica, acceso a datos basados en ontologías e integración/intercambio en la web, ingeniería ontológica y patrones ontológicos para la web, modularidad ontológica, mapeo, fusión y alineación para la web, apoyo al multilingüismo en la web semántica, interfaces de usuario e interacción con la semántica y los datos en la web, visualización de información y métodos de análisis exploratorio para los datos de la web semántica, acceso personalizado a los datos y aplicaciones de la web semántica, métodos y aplicaciones de la semántica social.
- Ingeniería de software: Tecnologías semánticas para plataformas móviles, sistemas de bases de datos distribuidos, gestión de datos en nube, NoSQL, análisis escalables, transacciones distribuidas, consistencia, p2p y gestión de datos en red, desarrollo de software y redes de entrega de contenidos.

- Visualización: Modelos de datos y lenguajes de consulta, gestión y diseño de esquemas, usabilidad de bases de datos, interfaces de usuario y visualización.

Faceta 5: Ciclo de vida de los datos abiertos, describe el proceso de convertir datos en datos abiertos, es decir, preparar los datos que se van a publicar, utilizar los datos publicados y conservar los datos publicados. Por lo tanto, se ocupa principalmente de tres temas: el pretratamiento, la explotación y el mantenimiento.

Para este propósito se tomó en cuenta varias publicaciones de investigación [9], [43], [41] que definen el ciclo de vida de los datos del gobierno abierto, y se describen a continuación:

- Creación de datos: Se refiere a la generación de datos, así como a la recopilación de datos con el propósito específico de publicarlos.
- Selección de datos: Es el proceso de selección de los datos a publicar. Esto requiere la eliminación de cualquier dato privado o personal, así como la identificación de las condiciones bajo las cuales estos datos serán publicados, potencialmente a través de la especificación de políticas de datos abiertos (gubernamentales).
- Armonización de datos: Este paso implica la preparación de los datos que se publicarán para que se ajusten a los estándares de publicación, como los Ocho Principios de Datos del Gobierno Abierto (Eight Open Government Data Principles).
- Publicación de datos: Es el acto concreto de abrir los datos mediante su publicación en portales de datos abiertos.
- Interconexión de datos: Este es el paso final en el esquema de cinco estrellas propuesto por Tim Berners-Lee, para datos abiertos, es decir, la obtención de la interconexión de datos. Esto permite que los datos publicados tengan un valor adicional, ya que proporciona el contexto de la vinculación de los datos.
- Descubrimiento de datos: La publicación de los datos no es suficiente para permitir su reutilización. Los consumidores de datos deben descubrir la existencia de datos abiertos para poder consumirlos.
- Exploración de datos: Este paso es la forma más normal de consumir datos, un usuario examina pasivamente los datos abiertos visualizándolos o investigándolos.
- Explotación de datos: Este paso es una forma más avanzada de consumir datos. La explotación de datos permite a un usuario utilizar, reutilizar o distribuir los datos abiertos

de forma proactiva mediante la realización de análisis, la creación de emprendimientos o la innovación basada en los datos abiertos.

- Curado de datos: Aunque no necesariamente ocurre en una etapa fija, la curaduría de datos es vital para asegurar que los datos publicados sean reutilizables. Esto implica una serie de procesos, incluida la actualización de datos obsoletos, el enriquecimiento de datos y metadatos, la limpieza de datos, etc.

Faceta 6: Tipo de investigación, no es específico para el tema particular de los datos abiertos, sino que es aplicable en forma general. Como en el caso de otros mapeos sistemáticos realizados en ingeniería de software, utilizamos el esquema de clasificación propuesto en [44] y en [22]. El esquema de clasificación incluye:

- Propuesta de solución: Describe una solución normalmente ilustrada con un ejemplo, un estudio de caso, un ejemplo de ejecución, etc. El trabajo está poco o nada validado (véase el siguiente punto); la propuesta sólo explica y describe su aplicación.
- Validación de la investigación: Se lleva a cabo en la práctica, por ejemplo, mediante un experimento, la realización de tipos de pruebas, estudios de laboratorio, etc. Normalmente sigue a una propuesta de solución. Responde a la pregunta: ¿la solución propuesta es "buena"?
- Investigación de evaluación: Una evaluación de la investigación, generalmente observando cómo funciona la solución en la práctica o comparándola con otras soluciones, señalando puntos positivos y negativos. Es más extensa que la validación y a menudo se lleva a cabo en un entorno industrial. Responde a la pregunta: ¿es la solución propuesta la solución "correcta"?
- Propuestas filosóficas o conceptuales: Esbozan una nueva forma de ver las cosas existentes, proporcionando una visión o visión filosófica sobre un tema.
- Documento de opinión: Describe la opinión de los autores, que suelen adoptar una postura positiva o negativa. También puede presentar una visión general de un campo o una comparación de técnicas desde el punto de vista del autor. Generalmente no se basa en un trabajo relacionado o en una metodología de investigación.

- Documento de experiencia: Describe la experiencia de los autores, generalmente en la práctica, utilizando un determinado método, tecnología, etc. Los autores suelen ser personas que trabajan en la industria [22].

3.1.4. Esquema de clasificación

Basándonos en el esquema de clasificación mencionado anteriormente, se clasificó cada una de las 671 publicaciones relevantes de acuerdo con cada faceta. Para asegurar la corrección y consistencia del proceso de clasificación, los investigadores implicados en el mapeo sistemático fueron divididos en dos grupos y cada una de las 671 publicaciones fue asignada a un grupo para ser revisada y clasificada dentro de cada faceta. En el caso de resultados conflictivos, se pidió al otro grupo que realizara una clasificación y los resultados se discutieron hasta que se llegó a un consenso final. Un total del 10% de las 671 publicaciones requirieron discusión. Cada artículo está clasificado en una sola categoría dentro de una faceta. Cuando los integrantes del grupo no se pusieron de acuerdo en la categoría elegida intervino el otro grupo y en base a consensos se definió la categoría.

3.2. Amenazas a la validez

De acuerdo a la metodología propuesta en el capítulo 2, las principales deficiencias de los mapeos sistemáticos se han identificado de la siguiente manera: sesgo en la selección de publicaciones [23], errores en la categorización de las publicaciones en categorías detalladas [24] y errores duplicados o contribución adicional débil de una publicación a otra (los llamados "delta papers", es decir, documentos que proporcionan sólo adiciones menores en comparación con trabajos publicados anteriormente por los autores).

Para mitigar el riesgo de la primera amenaza, se trabajó con una selección de motores de búsqueda para cubrir el área de investigación específica en profundidad. Se seleccionó motores de búsqueda específicos para cubrir los mayores editores en el campo de la investigación (ACM, IEEE, Springer, IOS PRESS, SCIENCE DIRECT). En segundo lugar, se tuvo que asegurar que se encontraran todas las publicaciones sobre el tema seleccionado. Para ello, se realizó un paso de validación de la búsqueda diseñado para identificar las publicaciones que faltan después de revisarlas. En cuanto a la segunda amenaza identificada, se siguió un proceso de revisión formal como se indica en esta sección.

Finalmente, para mitigar el riesgo de error duplicado o de poca contribución, se trabajó en los resúmenes, discutiendo las contribuciones y si se debía incluir la publicación en el estudio. Se encontraron 6 artículos delta de un total de 671 publicaciones (= 1%), sobre todas las categorías en cada faceta. Se considera, por tanto, que esta amenaza es insignificante en el mapeo sistemático. De hecho, de estas 6 publicaciones, sólo una fue encontrada como duplicada; por lo tanto, esto no afectó la investigación.

3.3. Resultados del mapeo sistemático sobre datos abiertos

La metodología de mapeo sistemático utilizada permitió identificar todas las publicaciones relacionadas con datos abiertos, definir un esquema de clasificación, categorización, análisis y descubrimiento de la investigación desde una perspectiva tecnológica. Se realizó un análisis de datos de 671 publicaciones, se utilizaron diferentes tipos de gráficos para diferentes propósitos, respondiendo así a diferentes preguntas de investigación.

- Se utilizaron gráficos de barras apiladas para representar los resultados por año. Estos gráficos permiten: visualizar las publicaciones más frecuentes y la periodicidad de las diferentes categorías dentro de una misma faceta, visualizar el peso relativo de cada faceta y la evolución de cada categoría en el tiempo.
- Los gráficos de barras tienen por objeto visualizar la distribución de las publicaciones por lugar de publicación y por año.
- Los gráficos de burbujas se utilizan para mostrar las relaciones entre las diferentes facetas y, como tales, representan los mapeos sistemáticos de la investigación de datos abiertos. Estos gráficos se utilizan tradicionalmente para este propósito porque representan cuatro dimensiones de datos donde cada burbuja representa la frecuencia de publicación con respecto a dos facetas.

3.3.1. Faceta 1: Lugares de publicación

En la Figura 6 se muestran los lugares donde se publicaron la mayor parte de las investigaciones relacionadas con los datos abiertos, incluyendo revistas y conferencias. Se pueden identificar 11 revistas con 3 o más publicaciones sobre datos abiertos, de los 671 artículos. Las revistas Semantic

Web y la Government Information Quarterly fueron las más importantes para la difusión de la investigación de datos abiertos desde una perspectiva tecnológica.

En la Figura 7 se puede visualizar 13 conferencias donde se publica el mayor número de los 671 artículos revisados. Se destacan las dos primeras conferencias: DG. O (Conferencia Internacional de Investigación sobre Gobierno Digital) y WOD (Taller Internacional sobre Datos Abiertos)

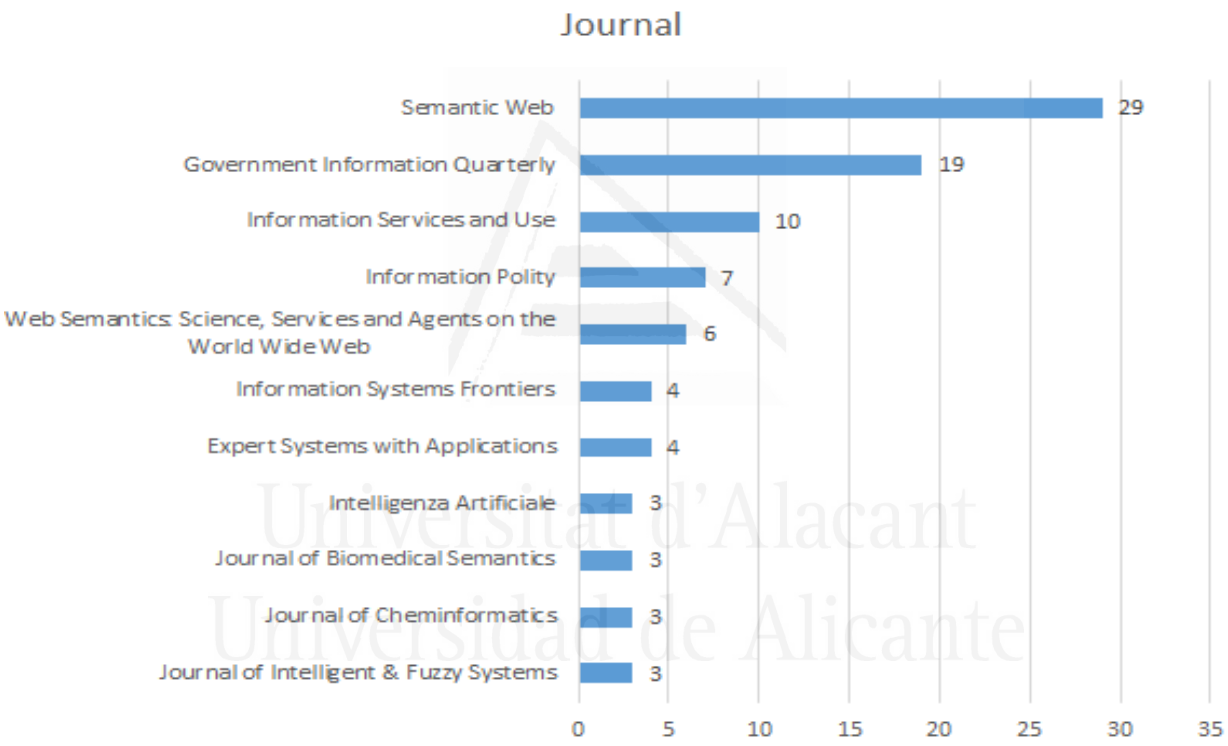


Figura 6. Revistas con mayor número de publicaciones

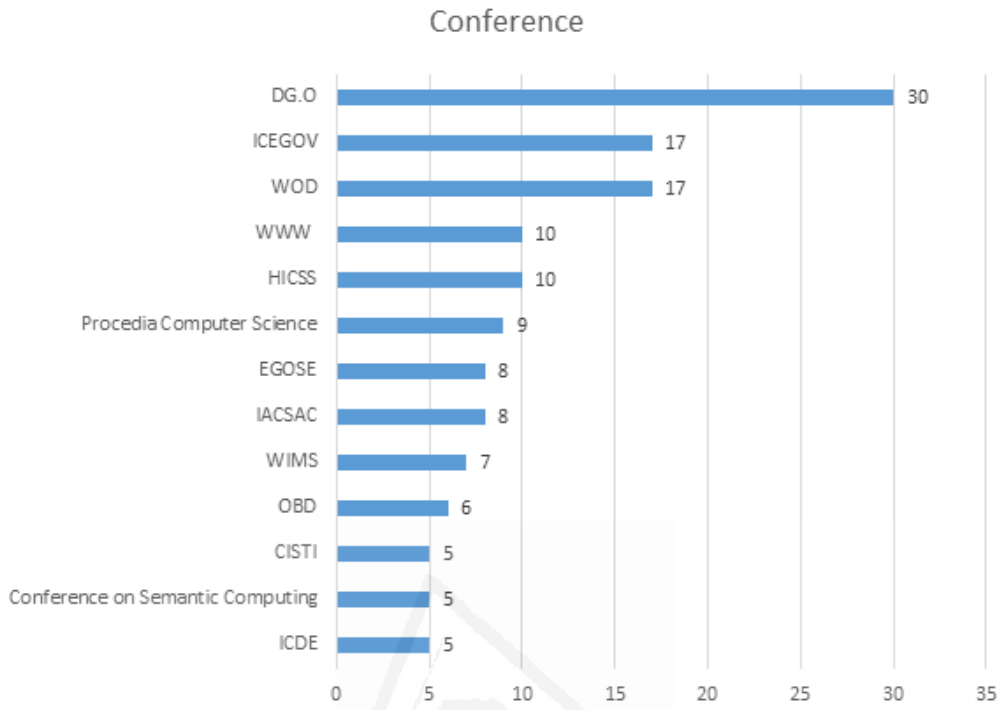


Figura 7. Conferencias donde se publican las investigaciones

En la Tabla 2 se presentan los acrónimos utilizados en los lugares de publicación.

Tabla 2. Acrónimos de las Conferencias

Conferencia	**Acrónimo**
International Digital Government Research Conference	DG.O
Hawaii International Conference on System Sciences.	HICSS
International Conference on the World Wide Web Companion.	WWW
IEEE Annual Computer Software and Applications Conference.	IACSAC
International Conference on Web Intelligence, Mining and Semantics.	WIMS
International Conference on Open and Big Data.	OBD
IEEE International Conference on Data Engineering.	ICDE
Iberian Conference on Information Systems and Technologies.	CISTI

La Figura 8 muestra un gráfico circular que representa la distribución de las publicaciones que se encuentran en los repositorios. El repositorio científico que más contribuye es IEEE con un 37%,

seguido por ACM con un 33%. El resto representa en conjunto el 30% de las publicaciones. Las dos primeras bases de datos científicas fueron las más relacionadas con las tecnologías de información y comunicaciones (TIC): esto confirma la relevancia de una perspectiva tecnológica cuando se examinan los estudios de datos abiertos.

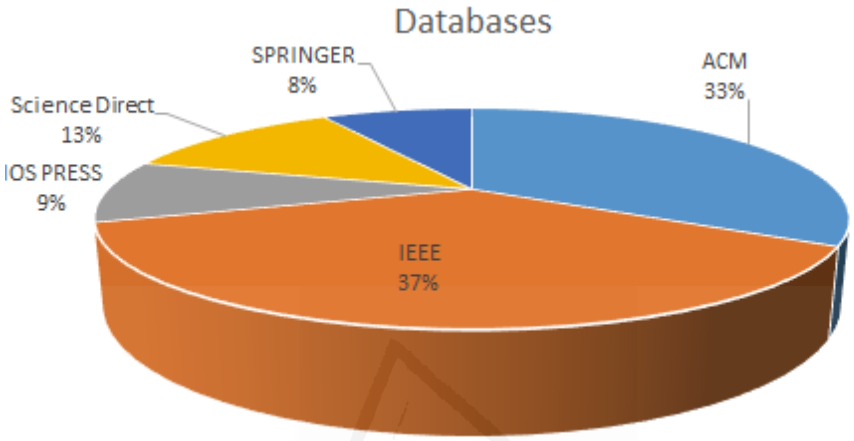


Figura 8. Porcentaje de publicaciones por base de datos científica

3.3.2. Faceta 2. Impacto

En la Figura 9 se muestran los porcentajes relativos al impacto de las publicaciones. Las publicaciones internacionales fueron las más frecuentes con un 69%.

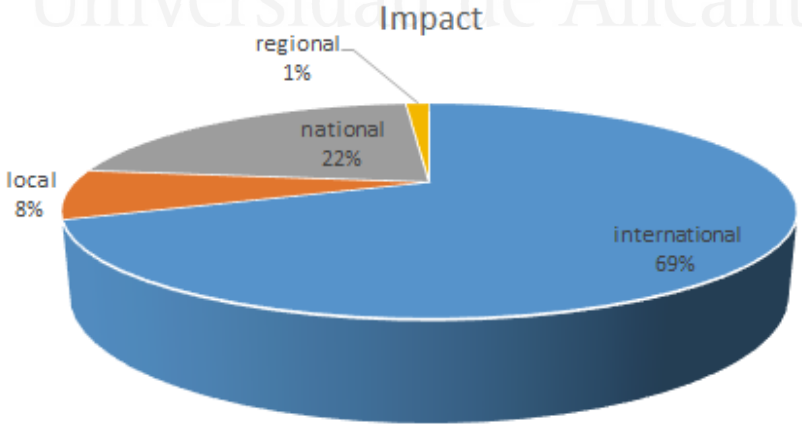


Figura 9.Publicaciones por impacto

3.3.3. Faceta 3. Dominio

En la Figura 10 y Figura 11 se presenta la distribución de las publicaciones según el dominio (sección 3.1.3, faceta 3), mostrando así el progreso de las publicaciones entre 2006 y 2017. Los infomediarios fueron el dominio más relevante con un 50%. Los otros dominios contribuyeron con menos del 8% cada uno, lo que sugiere que las publicaciones se dirigieron principalmente a los consumidores de datos abiertos.

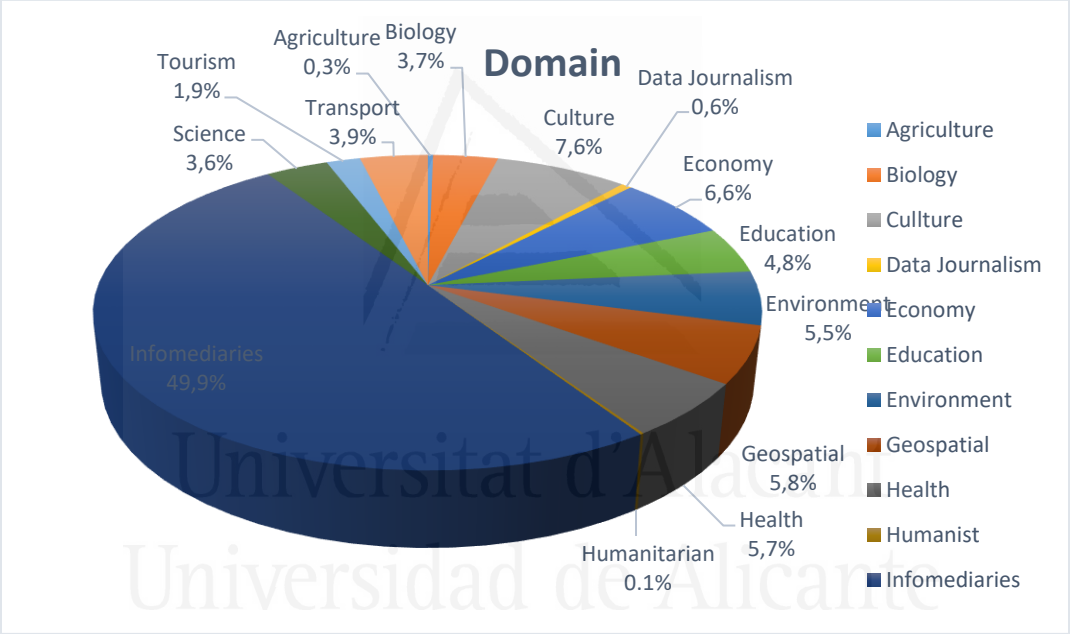


Figura 10. Porcentajes por dominio

Las cifras muestran que los campos del transporte, la economía, la educación, el medio ambiente, la cultura, la salud y el campo geoespacial han aumentado significativamente en la misma proporción en los últimos siete años. No hubo investigación en estos campos entre 2006 y 2009. Los datos del campo del periodismo y el humanitarismo no parecen haber generado interés en la comunidad investigadora.

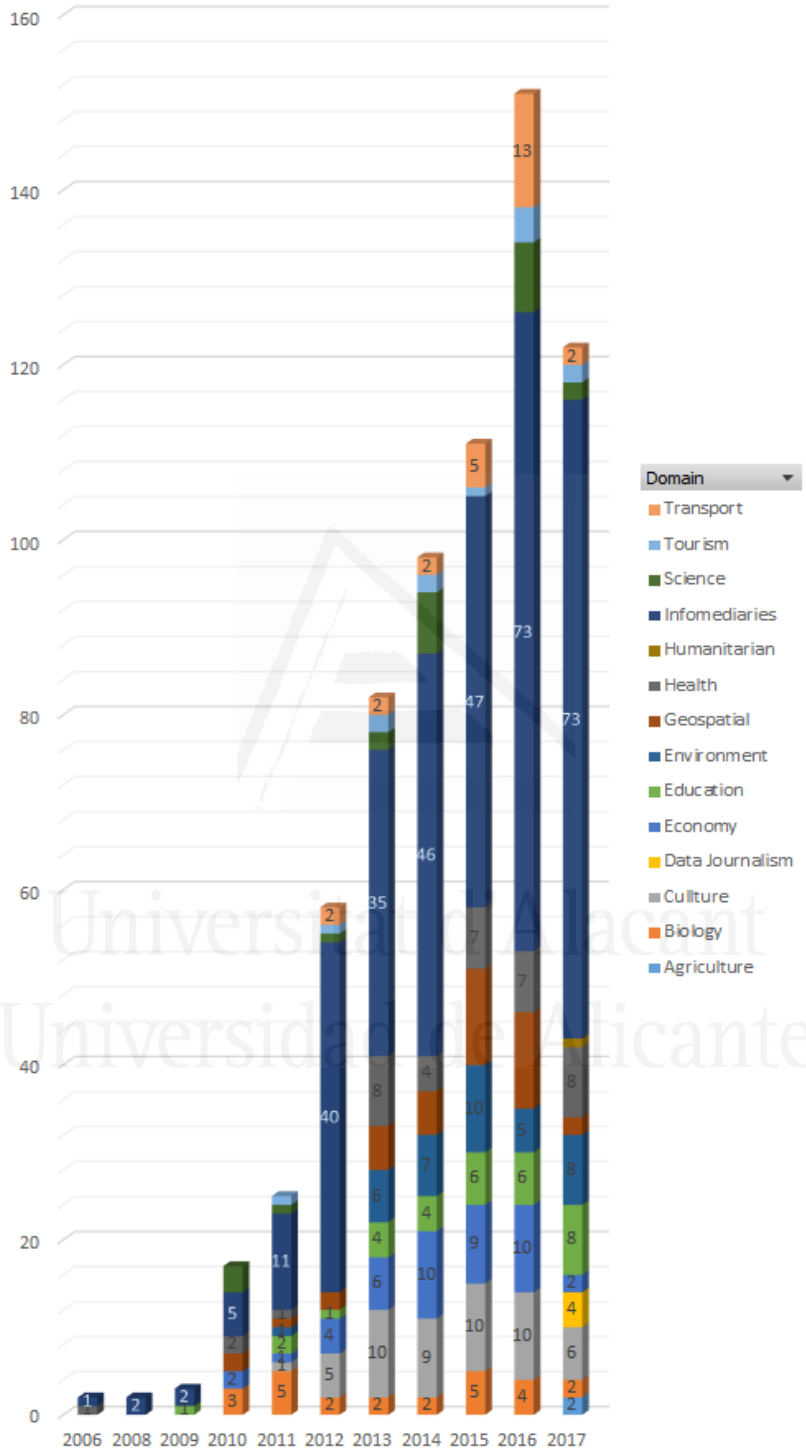


Figura 11. Distribución por año por dominio

3.3.4. Faceta 4. Tema

La Figura 12 y Figura 13 ilustran la distribución de las publicaciones según el tema y muestran cómo evolucionaron las publicaciones entre 2006 y 2017. El tema de la Web Semántica recibió mayor atención (24%), seguido por la Ingeniería de Software (22%) y el Gobierno (16%).

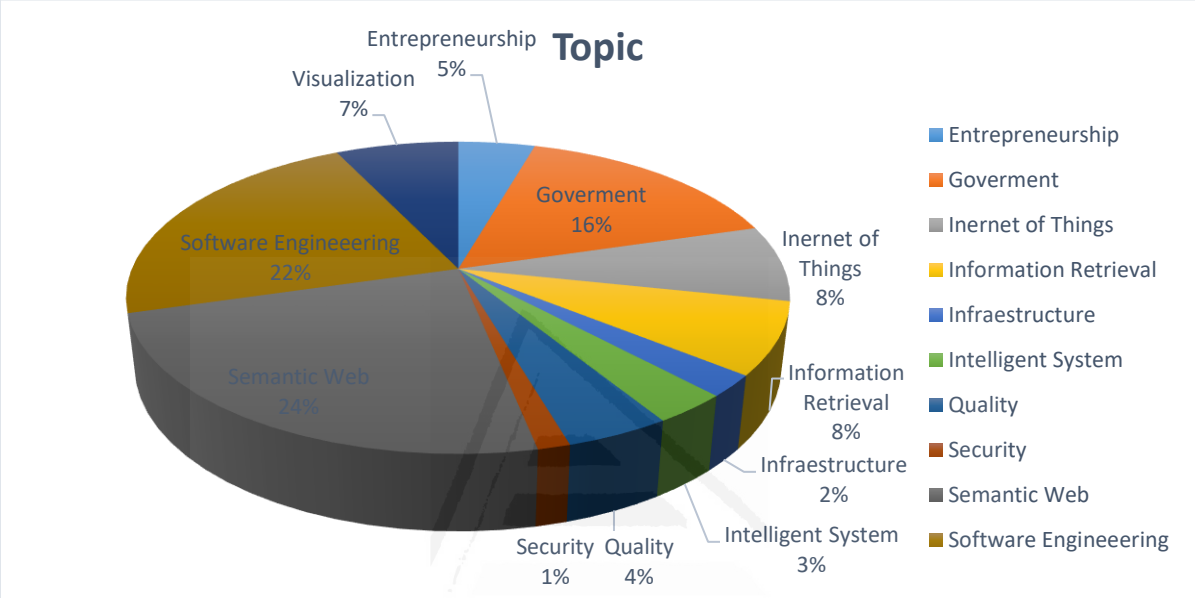


Figura 12. Porcentaje por tema

Sin embargo, los temas calidad y seguridad fueron los menos abordados con un 4% y 1% respectivamente.

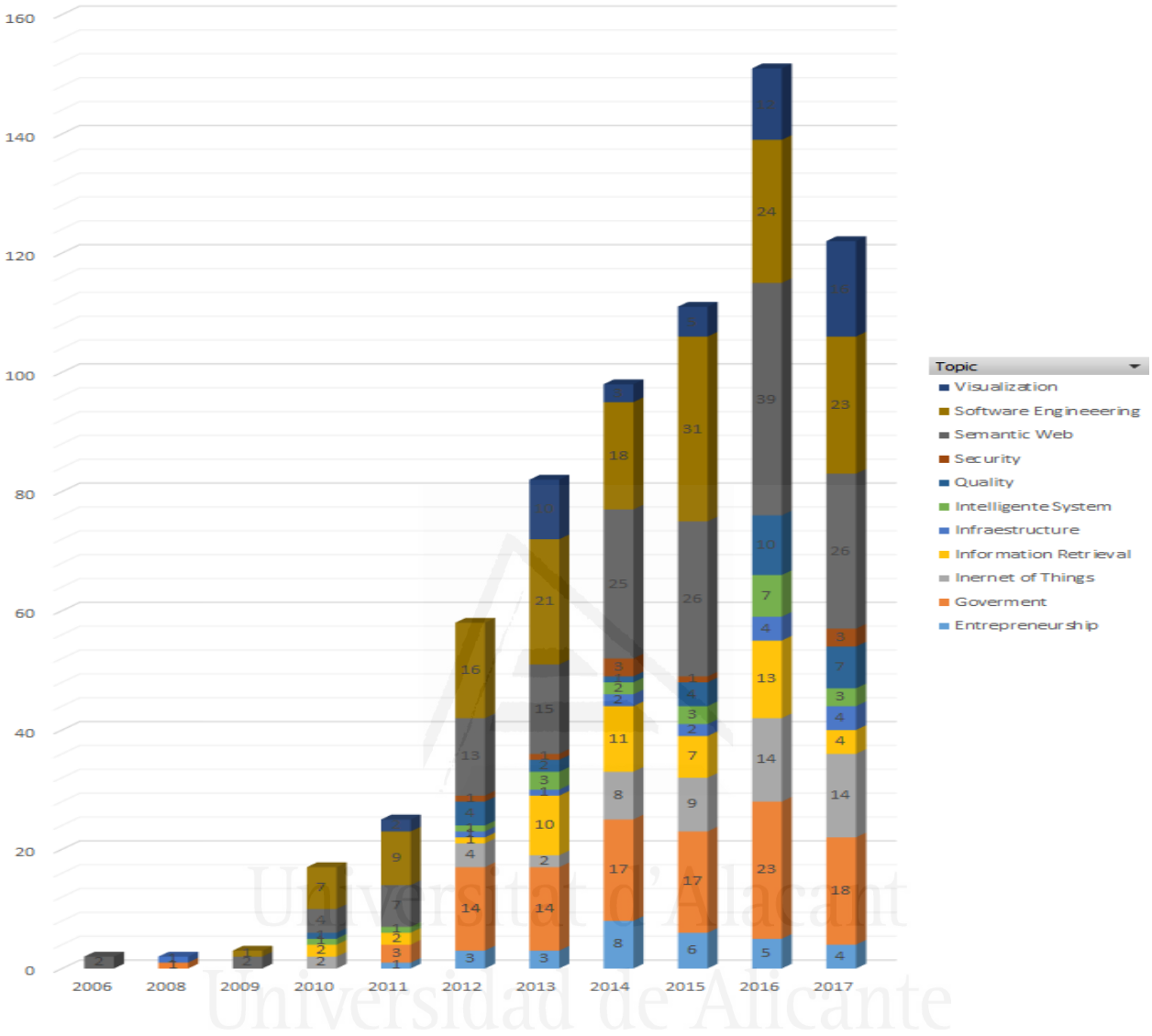


Figura 13. Distribución de publicaciones por tema entre el 2006 y 2017

3.3.5. Faceta 5. Clasificación del ciclo de vida de los datos

La Figura 14 ilustra el porcentaje de las publicaciones relacionadas con cada fase del ciclo de vida. La mayoría de los estudios se centraron en la fase de Explotación de datos (43% de las publicaciones). La siguiente fase fue la Exploración de datos (19% de las publicaciones). Esto sugiere que las publicaciones que buscan extraer valor de los datos y los resultados aplicables se multiplicaron.

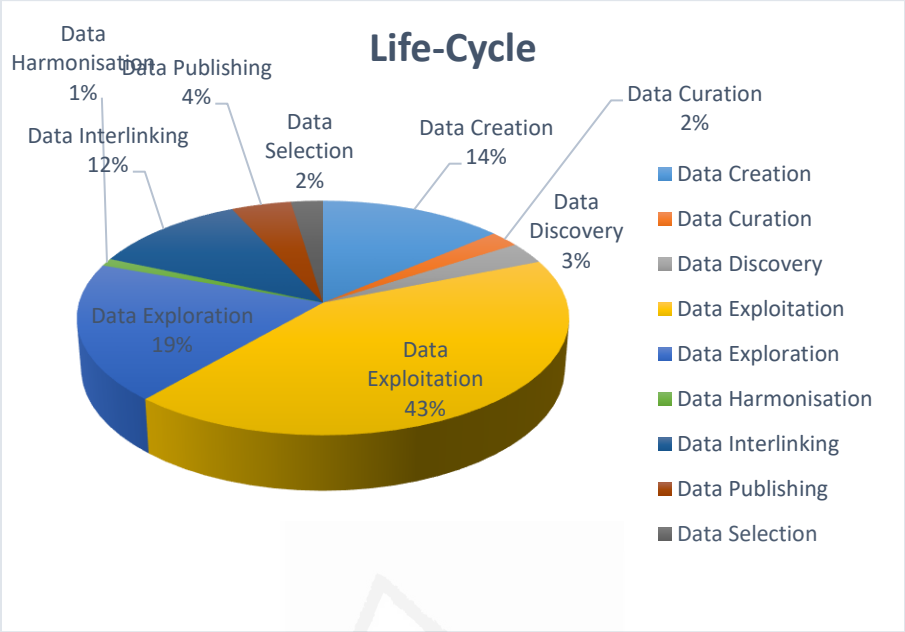


Figura 14. Porcentaje de las publicaciones según las fases del ciclo de vida de los datos de 2006 a 2017

La Figura 15 muestra la distribución de las publicaciones según las fases definidas en el ciclo de vida entre 2006 al 2017. La fase de Explotación creció notablemente. En 2009 comenzó con dos artículos científicos y llegó a un total de 62 en 2017. Las fases de Armonización de datos, Selección de datos y Curación de datos tuvieron una tasa de progresión de menos del 3%.

3.3.6. Faceta 6. Tipo de investigación

La Figura 16 y Figura 17 presentan la distribución de las publicaciones según el tipo de investigación, mostrando el progreso de las publicaciones desde 2006 hasta 2017. Se puede observar que el tipo más frecuente es el de Propuesta de Solución con un 53%, que representa la mayoría de las publicaciones, seguido por Validation Research con un 27%, un porcentaje muy inferior pero importante.

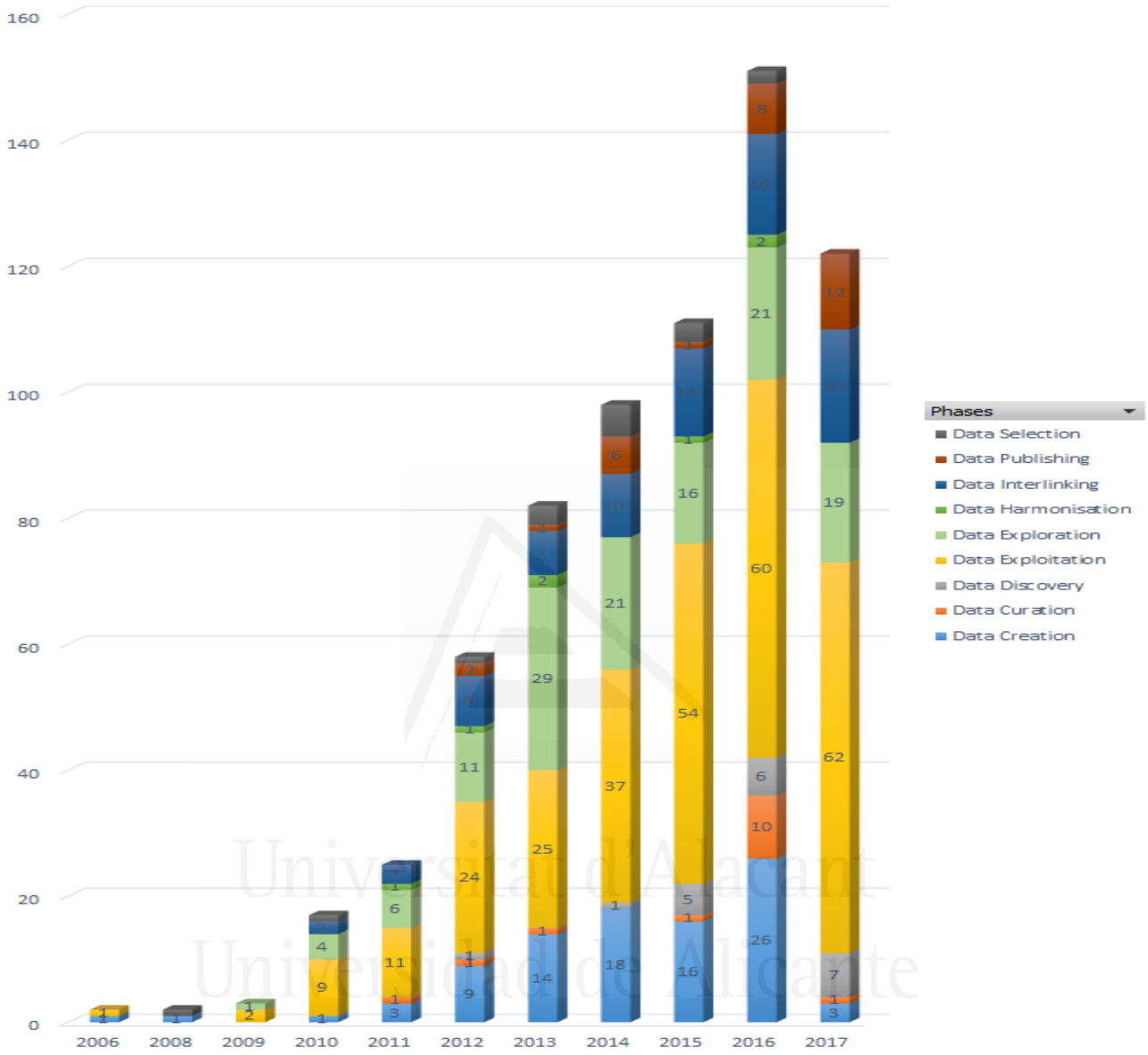


Figura 15. Distribución de las publicaciones según las fases del ciclo de vida de los datos de 2006 a 2017

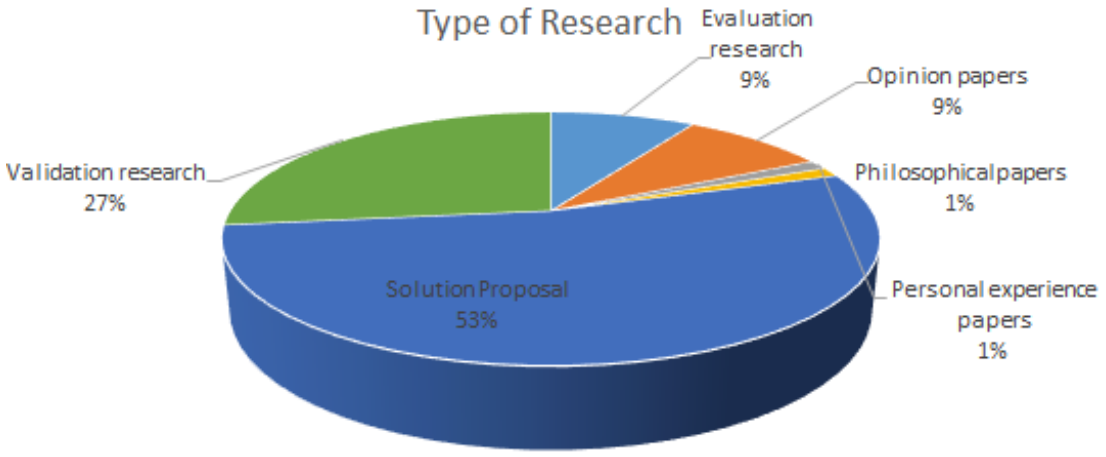


Figura 16. Porcentajes por tipo de investigación

Además, los documentos de evaluación han crecido en los últimos años, a medida que se ha establecido y desarrollado el campo de los datos abiertos.

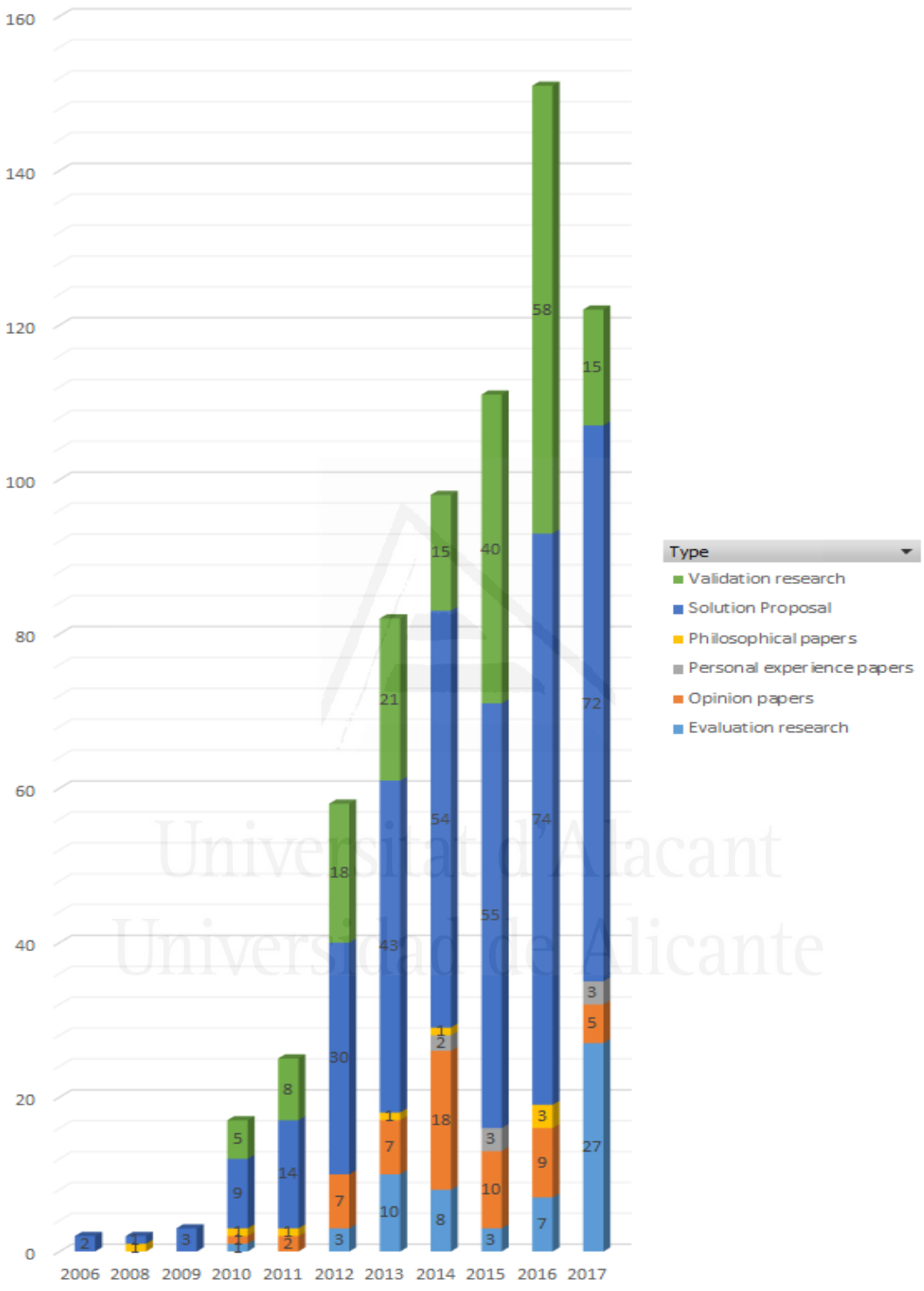


Figura 17. Distribución de las publicaciones por tipo

3.3.7. Combinación de facetas. Mapeo(s) sistemático(s)

Para un análisis completo de los resultados del mapeo sistemático, se combinaron facetas. Los resultados se presentan en 6 figuras, cada uno de los cuales representa diferentes combinaciones de facetas posibles.

La Figura 18 es el resultado de combinar la faceta Dominio con la faceta Fase del Ciclo de Vida de los Datos. La mayoría de las publicaciones de Explotación de datos (123) y de Exploración de datos (74) estaban relacionadas con el dominio de Infomediarios. También se observa que las publicaciones que incluyen una fase de Explotación de datos se distribuyen en casi todos los dominios.

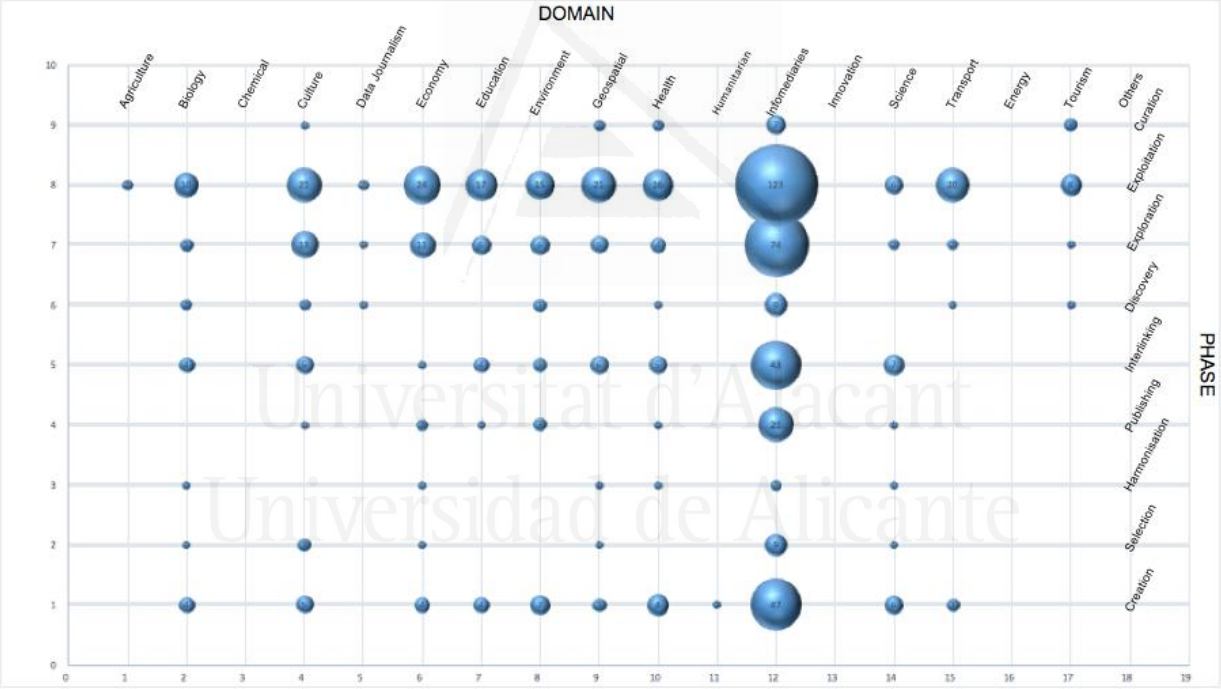


Figura 18. Dominio y ciclo de vida de los datos

La Figura 19 combina la faceta de Dominio y Tema. El mayor número de publicaciones de la web semántica (77) estaban relacionadas con el dominio de infomediarios. También encontramos que el principal dominio en las publicaciones gubernamentales era infomediarios (79). Clasificados en el dominio infomediarios existen (26) en calidad de datos y (8) en seguridad.

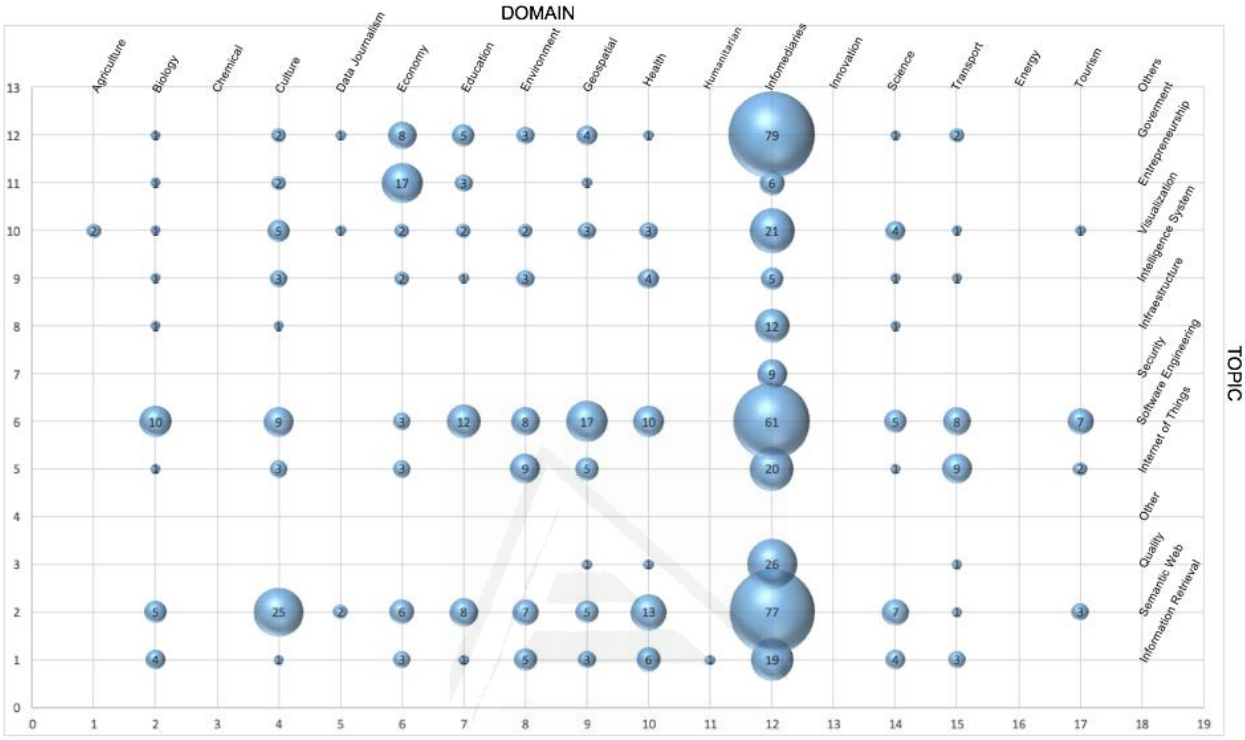


Figura 19. Dominio y Tema

En la Figura 20 la faceta Dominio se combinó con el Tipo de Investigación. La mayor cantidad de publicaciones fueron clasificadas como Propuestas de Solución (182), y estaban relacionadas con el dominio de los Infomediarios. Las publicaciones clasificadas como Validación de Investigación (71) estaban relacionadas con el ámbito de los Infomediarios. Las propuestas de soluciones y las publicaciones de investigación sobre Validación se distribuyeron en casi todos los ámbitos. Sin embargo, no había publicaciones sobre agricultura, química, periodismo de datos, humanitarismo, innovación y energía.

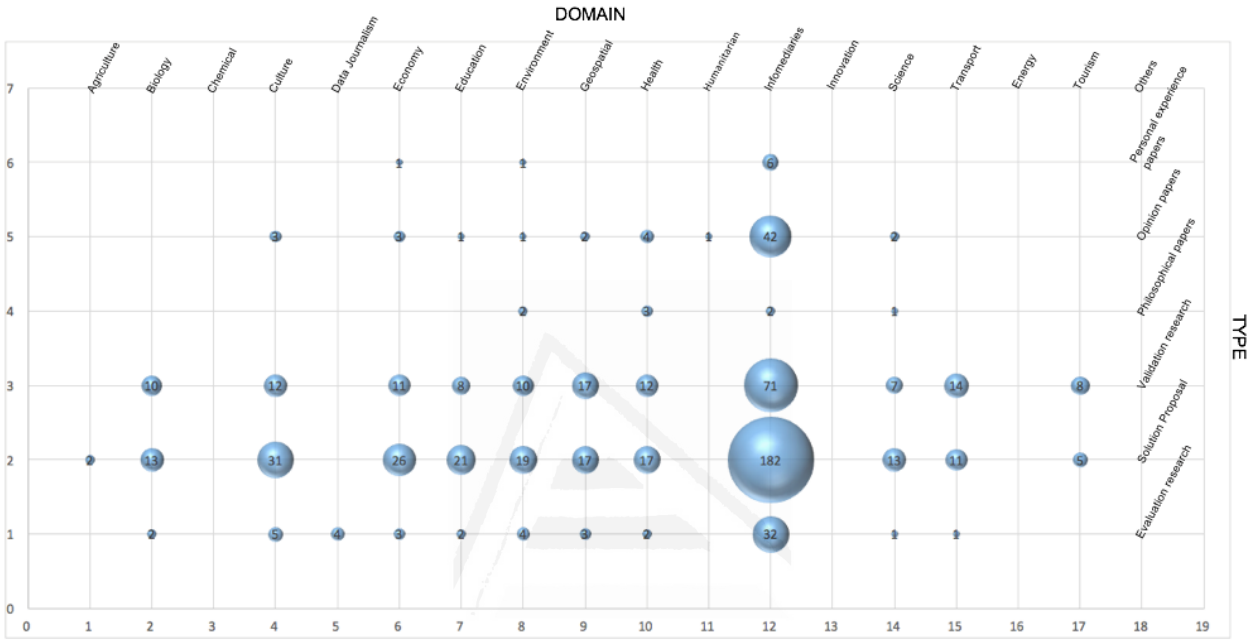


Figura 20. Dominio y tipo

La Figura 21 combina las facetas del ciclo de vida con tema. El mayor número de publicaciones clasificadas como Ingeniería de Software (88) estaban relacionadas con la fase de Explotación de datos, y la mayoría de las publicaciones clasificadas como Web Semántica (51) estaban relacionadas con la fase de Explotación de datos. El tema de la Web Semántica contó con un total de 159 publicaciones; Ingeniería de Software con 150, y Gobierno con 107, siendo estos últimos los temas más relevantes. En el ciclo de vida de los datos, los temas más relevantes fueron: Explotación de datos (285 publicaciones) y Exploración de datos (128).

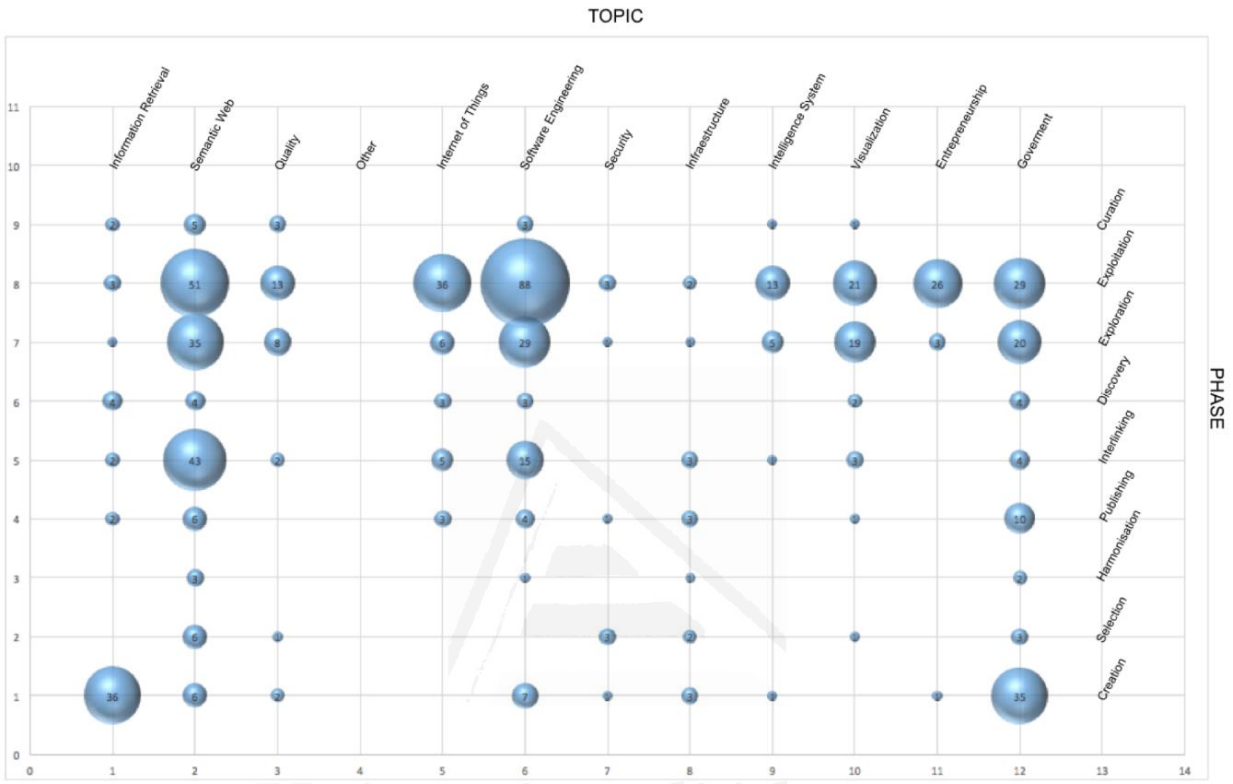


Figura 21. Fase del ciclo de vida del dato y tema

La Figura 22 combina la faceta del tema con el tipo de investigación. El mayor número de publicaciones relacionadas con el tema de la Web Semántica, con un total de 94 y sobre Ingeniería del Software (92), todas ellas fueron clasificadas como Propuestas de Solución. El tipo de investigación más frecuente fue Propuesta de Solución con 357.

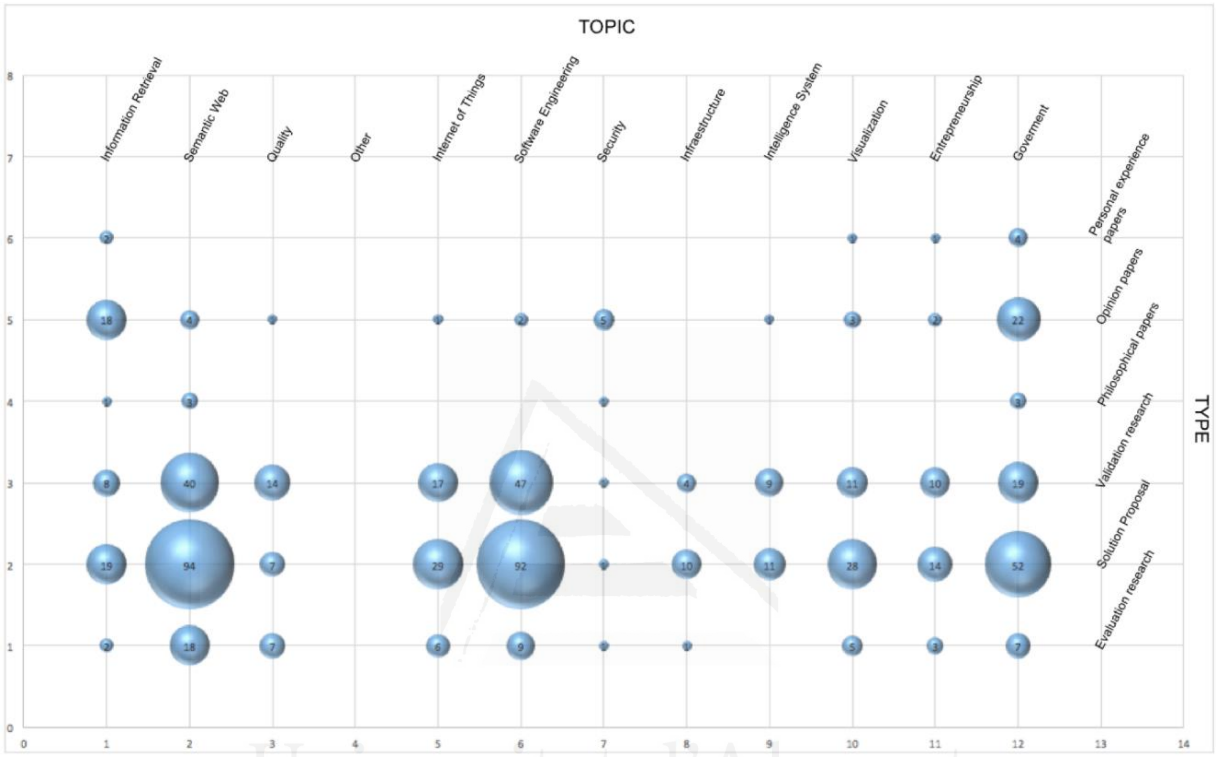


Figura 22. Tema y Tipo

La Figura 23 combina la faceta del Ciclo de Vida de los Datos con la del Tipo de Investigación. El mayor número de publicaciones pertenecía a la fase de Explotación (149) y se clasificaron como Propuesta de Solución. Además, las publicaciones en la fase de Explotación de datos (94) se clasificaron como Validación de la Investigación. Un total de 285 publicaciones estaban relacionadas con la Explotación, seguidas de 128 publicaciones sobre Exploración. Las Propuestas de Solución (357) fueron predominantes.

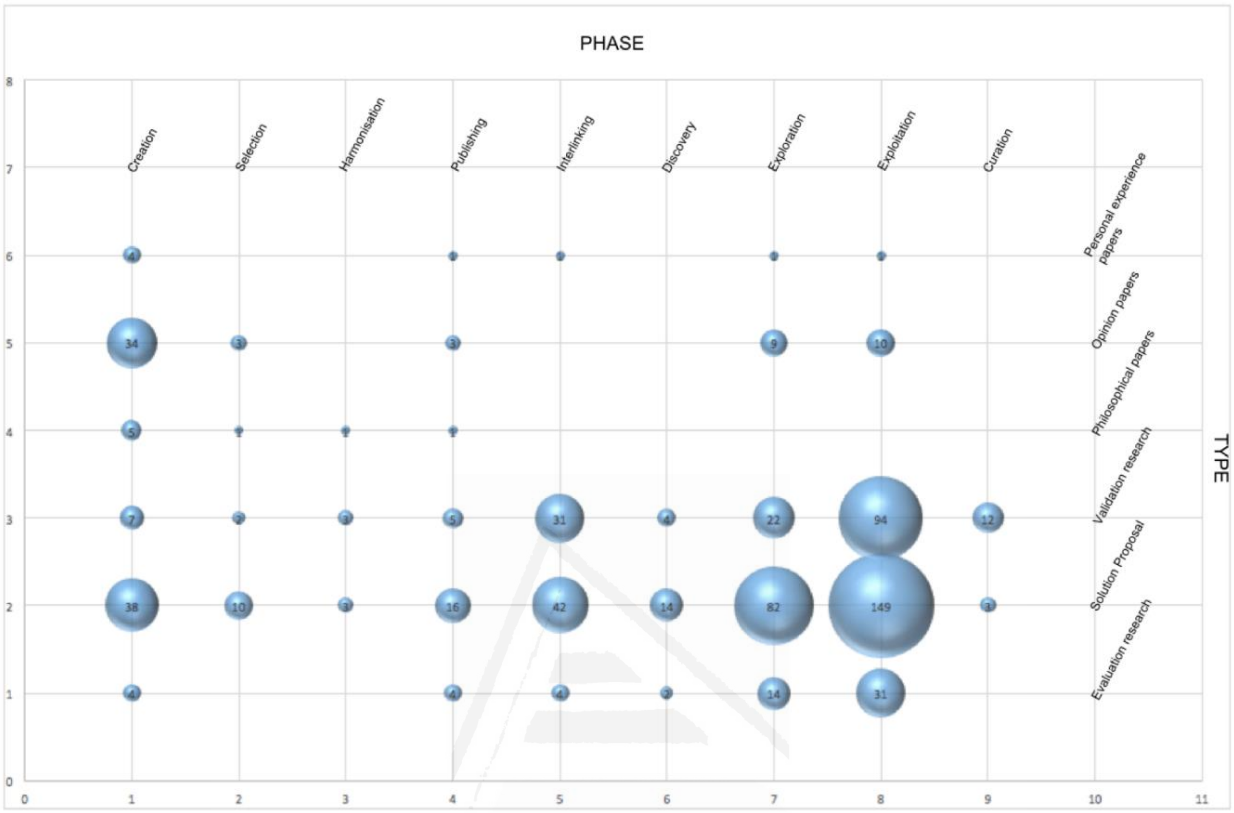


Figura 23. Ciclo de vida de los datos y tipo de investigación

3.4. Discusión de resultados del estudio sistemático.

En la sección anterior, el análisis de los datos describía objetivamente el número de publicaciones por faceta, basándose en la clasificación aplicada y teniendo en cuenta la evolución a lo largo del tiempo. En esta sección, se presenta una discusión de los resultados para responder a las preguntas de investigación e identificamos áreas de investigación, vacíos, tendencias y temas de investigación abiertos.

3.4.1. Lugares en los que se ha publicado investigaciones de datos abiertos.

Después de analizar los resultados, se puede observar que existen dos comunidades importantes en la investigación de datos abiertos:

- Una dedicada a temas relacionados con la web con especial énfasis en la Web Semántica (así como temas relacionados) como la revista Semantic Web o la conferencia Hawaii International Conference on System Sciences.
- Otra dedicada a la administración electrónica y su relación con los datos abiertos, es decir, la intersección entre la tecnología de la información y las publicaciones gubernamentales como Government Information Quarterly o conferencias como la DG.O.[16]. Por otra parte, hubo otros lugares interesantes con comunidades emergentes, como la ingeniería de datos, los sistemas de información o la inteligencia artificial. Estas novedosas comunidades enriquecerán el campo de los datos abiertos más allá de las tecnologías de la web semántica, contribuyendo a la integración de datos abiertos para el Business Intelligence [17].

3.4.2. Impacto que han tenido estos estudios

La Figura 9 muestra que el 69% de las publicaciones clasificadas tuvieron un impacto internacional, seguido por el 21% de las publicaciones con impacto local y el 9% con impacto nacional y un 1% de impacto regional. Por lo tanto, aunque los datos abiertos están estrechamente relacionados con las instituciones públicas que tienen regulaciones locales, regionales o nacionales, la mayoría de las publicaciones tienen una aplicabilidad internacional. Por ejemplo, los investigadores trabajaron con bibliotecarios del ZBW-Leibniz Information Centre for Economics y en reuniones internacionales de bibliotecas [45], en las que se trataron temas de investigación actuales de Linked Open Data (LOD) como la integración de datos y la integración de esquemas, la gestión de datos distribuidos, entre otros, que pueden utilizarse no sólo a nivel local sino también internacional. Según A. Ramo [46], los datos se tomaron de dos servicios relevantes proporcionados por MeteoGalicia -la generación de informes climáticos con descripciones mensuales del comportamiento climático y la generación de predicciones meteorológicas- que se convirtieron en el núcleo de un marco para generar descripciones lingüísticas del tiempo. El marco puede ser utilizado por cualquier servicio meteorológico del mundo. En un estudio de S. Chakraborty [47], las tecnologías de la web semántica se utilizaron para comprender la descripción de las entidades asociadas con la creación y el mantenimiento del patrimonio histórico en Bangladesh, pero los investigadores pueden utilizar estos datos en otras regiones del mundo.

Para concluir, en base a nuestro análisis, las publicaciones sobre datos abiertos a menudo se refieren a una ubicación geográfica específica promovida por instituciones públicas. Sin embargo, estas publicaciones tienen un alcance internacional, ya que se convierten en una referencia para otros contextos geográficos más amplios.

3.4.3. Dominios que han sido considerados por los investigadores

Tras un fuerte aumento en 2012, el número de documentos relacionados con los infomediarios se mantuvo constante hasta 2015, aumentando de nuevo en 2016 y estabilizándose en 2017. De hecho, la mayoría de las publicaciones se centraron en la búsqueda del valor de los datos abiertos, es decir, buscaron usos prácticos de los datos abiertos para beneficiar a la sociedad (por ejemplo, propuestas de investigación que resolvieran los problemas experimentados por los infomediarios, facilitando la reutilización de los datos y estudios sobre cómo realizar esta reutilización). Por ejemplo, el movimiento de los Recursos Educativos Abiertos (Open Educational Resources, OER) plantea retos inherentes al descubrimiento y reutilización de materiales educativos digitales procedentes de repositorios digitales altamente heterogéneos y distribuidos. En Towards a Learning Analytics Approach for Supporting discovery and reuse of OER [48], los autores presentaron las especificaciones de una plataforma de datos orientada al consumidor para datos abiertos, el Data-TAP. Data-TAP proporciona una interfaz fácil de usar y entender para hacer que los datos abiertos sean más amigables para los consumidores [49].

Sobre la base de los resultados de los datos clasificados, aproximadamente el 50% correspondió a infomediarios, el 8% a Cultura, el 7% a Economía, el 6% a Geoespacial y el 6% a Salud.

Los dominios del transporte, la economía, la educación, el medio ambiente, la cultura, la salud y el campo geoespacial han aumentado considerablemente en la misma proporción en los últimos siete años. Entre 2006 y 2009 no hubo investigación en estas áreas. Este crecimiento podría deberse al hecho de que estos ámbitos utilizan datos abiertos para proporcionar soluciones y beneficios a los ciudadanos, es decir, retos sociales que requieren investigación. Los datos de los campos del periodismo y el humanitarismo no generaron interés en la comunidad investigadora hasta 2017, aunque son campos relevantes:

1) Existen muchos datos heterogéneos (no sólo textuales sino también multimedia) en el ámbito humanitario (como los datos de gestión de desastres).

2) El periodismo de datos está relacionado con las tecnologías de visualización de datos y la gestión de datos fácil de usar, lo que merece ser abordado más a fondo por la comunidad investigadora.

La Figura 18 muestra la relación entre el dominio y la fase de datos del ciclo de vida, con 123 publicaciones que relacionan la explotación de datos abiertos con los infomediarios, revelando así la necesidad de resolver problemas reales de la sociedad.

En el dominio de la cultura, los avances se derivan de la política de apertura de los datos procedentes de fuentes bibliográficas que se reutilizan en las aplicaciones museísticas. El dominio de la economía está creciendo, y este aumento se atribuye a la utilización de datos por parte de las empresas. El dominio geoespacial se ha desarrollado debido al creciente uso en aplicaciones marítimas, de seguridad y de salud. Estos temas corresponden a mejoras en la calidad de vida, especialmente en el dominio de la salud. El dominio de la salud se ocupa de las necesidades básicas de la sociedad, por lo tanto, existe una investigación considerable basada en datos abiertos.

En la Figura 19 las Propuestas de Solución (357 de 671) predominan y la Validación de la Investigación (180 de 671) se distribuye en todos los dominios, pero claramente se aplica principalmente a los Infomediarios.

3.4.4. Fases del ciclo de vida de la publicación de los datos que han sido consideradas en las investigaciones

En la Figura 14 muestra las contribuciones de los trabajos clasificados por ciclo de vida de los datos: El 43% correspondió a Explotación de Datos, el 19% a Exploración de Datos, y en conjunto, estas dos fases del ciclo de vida representaron el 62% de los trabajos publicados. La Figura 15 muestra que desde 2010, ambas fases han crecido, respondiendo a la necesidad de obtener información a partir de datos publicados en aplicaciones de la vida real. Este proceso ha sido liderado por las necesidades de la comunidad de infomediarios con respecto a la reutilización de datos (lo que concuerda con la discusión anterior sobre el dominio de investigación más importante sobre datos abiertos, a saber, los infomediarios). Las primeras fases del ciclo de vida de los datos son la publicación, la creación, la interconexión y la armonización; al examinar los resultados, se

observa la necesidad de investigación adicional sobre estas primeras etapas del ciclo de vida de los datos.

La Figura 23 muestra las Fases del Ciclo de Vida de los Datos según el Tipo. La mayoría de los trabajos (149) se refieren a Propuestas de Solución dentro de la fase de Explotación de Datos, confirmando que la comunidad científica se está centrando en presentar soluciones para la reutilización de datos.

La Figura 14 muestra que el 14% de los artículos están dedicados a la Creación de Datos y el 4% a la Publicación de Datos. Esto contrasta con el 43% de los documentos relacionados con la explotación de datos, el 19% con la exploración de datos y el 12% con la interconexión de datos. Esto nos lleva a la conclusión de que se requiere un mayor esfuerzo de la comunidad investigadora en las fases del ciclo de vida de los datos relacionadas con su publicación. Hasta la fecha, los datos publicados no son de suficiente calidad, ya que se publican y explotan sin pasar por ciclos de vida que puedan añadir calidad, como la conservación de datos. Por lo tanto, se requiere investigación adicional sobre la calidad de los datos abiertos [9]. Por lo tanto, la fase de publicación de datos del ciclo de vida debe ser enfatizada en la investigación de datos abiertos. Cubrir las fases del ciclo de vida de los datos propuestas por [9] es hoy muy relevante, ya que un proceso de preparación previo a la publicación garantiza una mejor reutilización. Sin embargo, también cabe señalar que la investigación sobre datos abiertos se centró en la Explotación y la Exploración (es decir, en temas directamente relacionados con el consumo de datos).

Según varios autores, una de las razones por las que los investigadores no abren sus datos se debe a la compleja naturaleza técnica inherente a su publicación [50].

Las fases relacionadas con la publicación de los datos son cruciales debido al rápido proceso de digitalización de los datos, que plantea problemas de normalización de formatos y de conexión a fuentes de datos heterogéneas [51], [52],[53], [54]. Estas contribuciones son importantes para entender que es necesario trabajar todas las etapas propuestas en el ciclo de vida de los datos, con especial énfasis en las fases que preparan los datos para su publicación.

Sin embargo, el estudio muestra que Data Interlinking representó el 12% de todos los documentos científicos revisados. Esta fase de datos ha crecido constantemente con el paso del tiempo debido al auge de la web semántica. Estas aportaciones de la investigación sobre la interconexión de datos

podrían mejorar los procesos de interoperabilidad necesarios para la publicación de datos abiertos. Por último, las fases de armonización, selección y conservación de datos tuvieron una tasa de desarrollo inferior al 3%. La disminución del número total de trabajos en 2017 se debe básicamente a la disminución del número de trabajos relacionados con la conservación y creación de datos. Descubrimiento de Datos y Publicación de Datos crecieron a un ritmo más lento.

Por tanto, se establece la necesidad de mejorar los procesos de publicación considerando los diferentes orígenes de datos y la calidad de los datos para su publicación.

3.4.5. Tipo de investigaciones que se realizan

La Figura 16 muestra que el 53% de los trabajos de investigación consistían en Propuestas de Soluciones, seguido por Validación de la Investigación con un 27%, Artículos de Opinión con un 9% y Evaluación de la Investigación con un 9%.

Los tipos de investigación menos frecuentes fueron Experiencia personal y Artículos filosóficos, ambos con un 1%.

La Figura 22 relaciona el tema de la investigación con el tipo de investigación. Una vez analizados los datos, un gran número de publicaciones se clasificaron como Propuesta de Solución and Validación de la Investigación. Muchas propuestas de soluciones se orientaron hacia la web semántica (94), respondiendo a la necesidad de relacionar datos abiertos y crear aplicaciones más eficientes. Las soluciones propuestas tienen su origen en las necesidades de los infomediarios (182). Los infomediarios proporcionaron soluciones basadas en datos abiertos, pero, según la revisión, cada vez se necesitan más datos de calidad. Según [55], la calidad de los Datos Abiertos es una de las principales amenazas para alcanzar los objetivos del movimiento de Datos Abiertos.

La validación permite identificar científicamente la mejor solución a un problema. Se registraron muy pocos artículos como artículos de opinión o filosóficos, ya que los esfuerzos se dirigen hacia soluciones prácticas destinadas a ser validadas y evaluadas.

Un ejemplo de evaluación de datos abiertos es la propuesta de análisis de informes de ensayos clínicos, donde la evaluación de la investigación juega un papel importante antes de que la solución sea implementada en la sociedad [56]. Otro ejemplo de una propuesta de solución de web

semántica es el de un marco unificado para convertir datos heredados abiertos en imágenes gráficas y en un formato legible por máquinas y reutilizable mediante crowdsourcing [57]. En este documento, el beneficio del acceso abierto está directamente relacionado con el acceso de los usuarios a la información. El acceso se debe proporcionar de varias maneras, a través de aplicaciones de acuerdo a sus necesidades. Este autor propone una solución de ingeniería de software que responde a las necesidades de datos abiertos de las personas mayores aplicando tamaños de letra más grandes y comprensibles. Por otra parte, otros autores han abordado las necesidades de información de las personas con discapacidad motriz en relación con las distancias de marcha y la existencia de rampas, peldaños, asientos entre otros [58].

Estos ejemplos nos permiten entender que las soluciones propuestas son importantes y están estrechamente relacionadas con el desarrollo del concepto de datos abiertos. Sin embargo, el rápido aumento asociado de los volúmenes de datos digitales no se correlaciona automáticamente con los nuevos conocimientos y avances evaluados o validados [59].

Cuando se iniciaron los datos abiertos, se necesitaban propuestas sobre el uso de los datos abiertos. Estos requisitos se convirtieron entonces en la necesidad de evaluaciones y propuestas de validación para mejorar su aplicación. La cantidad de documentos de evaluación y validación aumentó a partir de 2010, lo que demuestra que la comunidad de investigación de datos abiertos ha alcanzado un cierto grado de madurez: la comunidad está cada vez más centrada no sólo en el desarrollo de soluciones, sino también en la evaluación y validación de las mismas.

3.4.6. Temas que se trataron en la investigación

La Figura 12 muestra que la Web Semántica representa el 24% de la investigación. El segundo tema más popular es la Ingeniería de Software con un 22%, seguido por el Gobierno con un 16%. El tema de la Recuperación de Información representó el 8% y el Internet de las Cosas el 8% de las publicaciones.

La Figura 19 relaciona el dominio y el tema. La Web Semántica (77), el Gobierno (79) y la Ingeniería de Software (61) representan conjuntamente el 32 % del total de publicaciones. El notable desarrollo de las publicaciones de Web Semántica y Desarrollo de Software se basa en la tendencia a Explotar y Explorar datos, en los diferentes dominios discutidos en la pregunta de investigación relacionada con el Dominio. El gran número de publicaciones relacionadas con el

gobierno revela la necesidad de un marco jurídico con políticas y reglamentos para la gestión de datos abiertos, pero desde una perspectiva tecnológica. Por lo tanto, la investigación multidisciplinaria sobre cuestiones jurídicas y tecnológicas de datos abiertos es muy necesaria para establecer políticas que faciliten el desarrollo de datos abiertos y generen servicios tecnológicos que utilicen datos abiertos.

La razón principal por la que el tema de la Web Semántica en el gobierno puede haber atraído la mayor parte de la investigación es la necesidad de esquemas para intercambiar datos. Este proceso permite generar automáticamente información para desarrollar aplicaciones y facilitar plataformas, por ejemplo, en las áreas de salud [60], o la visualización de temas ambientales, mapeo de gestión de servicios públicos, evaluación de cabildeo político, beneficios sociales, cierre de la brecha digital, biología y otros [61].

El desarrollo de la investigación en Ingeniería del Software aplicada a diversas áreas es notable en aplicaciones de captura de datos para la enseñanza o API para aplicaciones turísticas [34], análisis de datos generados por smartphones, captura abierta de datos para presentar información útil a los ciudadanos para eventos públicos [62], [63], [59], [64], [65], [66].

La Ingeniería de Software representó 148 artículos y fue el segundo tema más frecuente. Esto puede explicarse por la necesidad de los usuarios de reutilizar los datos. El mayor número de contribuciones se produjo en 2015, con un 21% del total. Sin embargo, este resultado contrasta con el desarrollo de proyectos empresariales. Las aplicaciones informáticas se orientaron hacia los servicios de reutilización, no hacia el espíritu empresarial. Debemos recordar que el espíritu empresarial no depende necesariamente de la investigación. Esto podría explicar por qué había pocos estudios científicos de este tipo.

Es importante destacar que el desarrollo de software de código abierto permite generar aplicaciones de servicios para los ciudadanos en diferentes ámbitos como el transporte y el turismo, entre otros. Por último, la investigación y el desarrollo en materia de seguridad han sido limitados, a pesar de que el tema es relevante desde el punto de vista de los datos abiertos. El tema de la internet de las cosas (IoT) creció constantemente hasta 2017, en respuesta al desarrollo de la IoT en el mundo de las TI.

3.5. Más allá del mapeo sistemático: proyectos de innovación de datos abiertos

Además de un análisis cuantitativo de la investigación sobre datos abiertos de los últimos años, presentada en las secciones anteriores, ahora vamos más allá del estudio de mapeo sistemático y presentamos una revisión de los proyectos de innovación de datos abiertos. Nos centramos en dos fases del ciclo de vida de los datos: la publicación (por ejemplo, portales de datos abiertos) y el consumo (por ejemplo, datos abiertos como facilitador del negocio) de datos abiertos.

3.5.1. Abrir portales de datos y búsqueda de conjuntos de datos

El primer paso crítico del ciclo de vida de los datos cuando se consideran los agentes externos que pueden consumir los datos es su publicación. Normalmente, esto se hace mediante la creación de un portal de datos abierto (PDA), donde la división administrativa en particular, o la organización pública que produce o recopila los datos los hace accesibles al público [67], propone una clasificación de los PDAs de acuerdo con el número de funcionalidades que proporcionan, ordenadas por la cantidad de esfuerzo que se requiere para su creación.

- Un "registro de conjuntos de datos", es una simple lista de enlaces hacia conjuntos de datos, no necesariamente alojados en el portal. Un registro responde a la pregunta "¿Quién tiene qué conjunto de datos (abierto) y dónde puedo encontrarlo?"
- Un "proveedor de metadatos", es un registro de conjuntos de datos que también contiene metadatos sobre conjuntos de datos, por ejemplo, licencias, contexto espacio-temporal, frecuencia de actualización.
- Una "plataforma de co-creación", es un proveedor de metadatos que incluye herramientas de participación de los ciudadanos y los consumidores de datos para actuar sobre los conjuntos de datos: generar ideas, plantear cuestiones, contribuir con ejemplos de reutilización, y/o participar en debates.
- Una "plataforma de publicación de datos", es una plataforma de co-creación que permite a múltiples proveedores de datos publicar sus propios conjuntos de datos. También soporta la fase de interconexión entre los conjuntos de datos alojados.

- Un "centro de datos común", es una plataforma de publicación de datos que también soporta la implementación de otras fases del ciclo de vida de los datos, permitiendo a los editores de datos implementar sus propios ciclos de datos.

Los sistemas de software de código abierto y comerciales como CKAN, Socrata y OpenDataSoft permiten a los proveedores de datos de código abierto configurar fácilmente un portal entre los niveles 3 y 4. Proporcionar una visión unificada que facilite la búsqueda a nivel nacional y supranacional desde los portales regionales y locales, los meta portales rastrean los conjuntos de datos y los indexan hasta una ubicación central en la que también se pueden realizar enlaces adicionales. Un ejemplo de formalización de tal modelo es el marco MODA (Middleware for Open Data), descrito en [68]. Muchos portales nacionales proceden de esta manera, agregando datos de portales regionales y de agencias gubernamentales. El proyecto Europeana [69], proporciona un único punto de acceso a millones de libros, pinturas, películas, objetos de museo y archivos que han sido digitalizados en toda Europa. El Portal Europeo de Datos recoge los metadatos de la información del sector público disponibles en los portales de datos públicos de todos los países europeos. También se incluye información sobre el suministro de datos y los beneficios de su reutilización.

Una vez publicados los conjuntos de datos, el siguiente paso que debe dar el usuario es proporcionar una funcionalidad de búsqueda adecuada. La búsqueda de conjuntos de datos es un campo de investigación relativamente nuevo que se encuentra en la intersección de la recuperación de información, las bases de datos, la web semántica y la gestión de datos empresariales. El primer esfuerzo fue la Búsqueda Internacional de Conjuntos de Datos de Gobierno Abierto [70], una interfaz de navegación para buscar en más de un millón de conjuntos de datos de gobierno abierto de todo el mundo. Los desafíos para la búsqueda de conjuntos de datos se esbozaron por primera vez en [71] en el contexto de los datos científicos, proponiendo un proceso Crawl-Read-Extract similar a los motores de búsqueda de documentos. Más recientemente, se han realizado esfuerzos para comprender las sutilezas de la búsqueda de conjuntos de datos desde la perspectiva del usuario, ya sea entrevistando a profesionales [72] o analizando el análisis de los registros de consultas y las solicitudes de datos [73]. La investigación actual se centra en el uso de estos conocimientos para desarrollar modelos de aprendizaje automático para clasificar conjuntos de datos en un portal según una consulta [74],[75].

3.5.2. La innovación de datos abiertos como facilitador del negocio

Varias obras han identificado el potencial de los datos abiertos como catalizadores de la innovación, permitiendo así la creación de valor y de servicios que, en última instancia, benefician a los ciudadanos. En el contexto de Open Government Data (OGD), el trabajo de [76] identifica 4 mecanismos para la generación de valor socioeconómico mixto: Transparencia (que mejora la visibilidad), participación (compromiso con todos los grupos de interés), eficiencia (reducción de costes y tiempo) e innovación (generación de nuevas ideas), destacando la apertura de datos como facilitador tanto de la generación como de la apropiación de valor. De una investigación realizada se encuestó a 138 empresarios suecos de TI y encontró que el acceso a datos públicos abiertos se considera muy importante para muchos de ellos; el 43% encuentra que los datos abiertos son esenciales para la realización de su plan de negocios y el 82% afirma que el acceso apoyaría y fortalecería el plan de negocios [75].

Los empresarios también mostraron interés y voluntad de pagar por los datos de información del sector público para apoyar o probar otros modelos de negocio. En otra investigación se analizó datos analizados de 500 empresas con sede en EE.UU. que utilizan datos de gobierno abierto en su modelo de negocio y propuso una taxonomía de arquetipos de modelos de negocio: Facilitadores de la recopilación, gestión y divulgación de datos públicos; Facilitadores que apoyan o aceleran el acceso y el intercambio de datos entre el lado de la oferta y el de la demanda; e Integradores que hacen uso de los datos de gobierno abierto combinándolos con datos internos u otros tipos de datos de propiedad exclusiva a fin de aumentar sus capacidades empresariales [77]. Por otro lado, [78] identifica ventajas competitivas, ya que los datos abiertos son accesibles a todos, su estudio sugiere que la generación de una ventaja competitiva con datos abiertos requiere que una compañía tenga capacidades y recursos internos para el uso de datos abiertos.

Por lo que se refiere a las actividades de innovación, dos recientes acciones de innovación financiadas por la UE se centraron en cómo liberar el potencial de los Datos Abiertos como herramienta de apoyo a las empresas, proporcionando a las PYME las capacidades necesarias para procesar datos abiertos, además del apoyo general a las empresas, de acuerdo con las recomendaciones de [79]. A continuación, se describe brevemente su funcionamiento y sus resultados:

1) El proyecto denominado Internet del Futuro y la Expansión de los Datos Abiertos (FINODEX) estuvo en marcha desde mediados de 2014 hasta 2016. Su objetivo era la promoción y el apoyo de servicios innovadores de TIC que reutilizan datos abiertos, utilizando como anclaje técnico la plataforma FI-WARE. Se organizaron dos convocatorias abiertas para que las PYME presentaran sus ideas sobre productos y servicios. Se proporcionó financiación, apoyo, formación a medida, oportunidades de creación de redes y conexión con los inversores a determinadas PYME. El tiempo y la financiación de las PYMEs dentro de la incubadora fueron dictados utilizando un enfoque de "embudo". Inicialmente, todas las PYMES recibieron una cierta cantidad de financiación para realizar un hito inicial, que se evalúa para decidir qué PYMES pasan a la siguiente fase, en la que reciben una cantidad adicional de financiación en función de un nuevo hito, y así sucesivamente, se han alcanzado hasta 4 hitos. Los importes más elevados de financiación directa: 170.000, 135.000 y 115.000 euros, se asignaron al primer, segundo y tercer proyecto en ambas rondas de aceleración.
2) La Incubadora de Datos Abiertos para Europa (ODINE) funcionó desde mediados de 2015 hasta mediados de 2017, con el objetivo de incubar ideas empresariales centradas en datos abiertos. En comparación con FINODEX, no utilizó el método de incubación en embudo, sino un período de incubación de 6 meses para todas las empresas seleccionadas, que fue seleccionado en una convocatoria abierta de 8 iteraciones y no obligó a las PYME a utilizar la plataforma FI-WARE, y con una financiación máxima de 100.000 euros por PYME. ODINE incubó 57 empresas que crearon 278 nuevos puestos de trabajo y 23,7 millones de euros en ventas e inversiones. El impacto de ODINE en las perspectivas de crecimiento de las empresas financiadas fue relevante, resultando en unos 110 millones de euros de ingresos acumulados en el periodo 2016-2020, más 784 puestos de trabajo creados. Un estudio independiente de evaluación de impacto sobre ODINE preparado por IDC [80] descubrió que la mayoría de las empresas financiadas eran jóvenes y desempeñaban el papel de "experimentadores", es decir, combinaban varias fuentes de datos abiertas para mejorar sus productos y servicios. En la evaluación también se observó una correlación positiva entre el nivel o la madurez a nivel de país del mercado de datos abiertos y el número de solicitantes exitosos de ODINE por país, lo que sugiere que un rico entorno de datos abiertos ofrece condiciones favorables para los innovadores en este campo.

3.6. Aspectos relevantes del estudio sistemático

En este trabajo se ha realizado un estudio sistemático para reunir, clasificar y analizar toda la investigación sobre datos abiertos desde una perspectiva tecnológica realizada entre 2006 y 2017 por la comunidad científica con el objetivo de: (i) proporcionar una visión general consolidada del campo de investigación, e (ii) identificar, entre otros, temas bien establecidos, tendencias y cuestiones de investigación abiertas. Este estudio reveló varios hechos interesantes:

- La mayor parte de la investigación sobre datos abiertos desde una perspectiva técnica provino de los repositorios científicos del IEEE y la ACM.
- Los trabajos sobre datos abiertos publicados antes de 2009 no fueron significativos. Hay un impulso a partir de que fue en 2009 cuando Estados Unidos estableció una estrategia de Gobierno Abierto bajo la administración de Obama.
- Las publicaciones de 2006 a 2009 fueron incipientes (de hecho, no hubo publicaciones sobre datos abiertos en 2007).
- La Web Semántica, la Ingeniería de Software y el Gobierno fueron algunos de los temas más importantes abordados en la investigación. Esto es lógico porque (i) las tecnologías de la Web Semántica están intrínsecamente relacionadas con la reutilización de datos abiertos, (ii) el desarrollo de software requiere nuevos enfoques para resolver problemas técnicos relacionados con la apertura de datos, y (iii) se necesita legislación y estandarización para introducir datos abiertos en la sociedad.
- Los temas de Internet de las Cosas aún no se habían desarrollado, pero dadas las evoluciones tecnológicas actuales, estos temas podrían adquirir una gran importancia en un futuro próximo con la implantación de ciudades inteligentes y la importancia de una calidad de datos suficiente para apoyar la toma de decisiones óptima.
- Infomediarios fue el dominio más desarrollado en las publicaciones. Los otros dominios representaban menos del 8% cada uno; esto sugiere que las publicaciones estaban dirigidas especialmente a los canales de información; sin embargo, los dominios geoespaciales, de salud y cultural resultaron ser un objeto de investigación en curso y estos estudios continúan creciendo.
- Al analizar las fases, la Explotación y la Exploración fueron las más frecuentes, lo que justifica la necesidad de la comunidad de aplicaciones prácticas de los datos abiertos.

Sorprendentemente, las fases relacionadas con el consumo de datos no tienen mucho interés en las publicaciones. Estas fases no han recibido suficiente atención y deberían investigarse más en el futuro.

- De la misma manera, los tipos de Investigación de Propuesta de Solución y Validación fueron los más frecuentes, mostrando que el campo está alcanzando la madurez.
- En cuanto al impacto, las publicaciones internacionales fueron las más frecuentes porque los desarrollos de software y las publicaciones de la web semántica son aplicables y aceptadas internacionalmente.
- En 2017, el interés de la investigación por los datos abiertos desde una perspectiva tecnológica en general disminuyó. Este hecho puede indicar que la investigación se está estabilizando, que el pico de la investigación de datos abiertos ha llegado a su fin y que este campo de investigación está alcanzando la madurez. Aunque también deben tenerse en cuenta otros factores para evaluar el nivel de madurez, por ejemplo, el tipo de publicación puede considerarse que la cantidad de investigación realizada se está consolidando, ya que el número de soluciones propuestas disminuyó en 2017, mientras que al mismo tiempo se están llevando a cabo más estudios de validación y evaluación.
- De acuerdo a los resultados de las fases en el ciclo de vida de la publicación de datos, el 14% de los artículos están dedicados a la Creación de Datos y el 4% a la Publicación de Datos. Se requiere un mayor esfuerzo de la comunidad investigadora en las fases del ciclo de vida de los datos relacionadas con su publicación para garantizar una mejor reutilización.
- El estudio revela que el 62% de las publicaciones corresponden a la explotación y la exploración de datos respondiendo a la necesidad de obtener información de los datos publicados. La tarea de reutilización ha sido liderada por los informes diarios donde cerca del 50% de las publicaciones corresponde a ellos. Por otro lado, existen pocos trabajos de investigación que propongan una metodología formal de selección de conjuntos de datos para abrir, tomando en cuenta tanto las necesidades y prioridades de los reutilizadores, como el criterio de costo de publicación. Por tanto, se identifica la necesidad de presentar métodos formales y sistemáticos que satisfagan las necesidades de los reutilizadores y el criterio de los publicadores para apertura conjuntos de datos, que tengan un alto potencial de reutilización con un coste adecuado de publicación.

Capítulo 4: Método para la Selección de Conjuntos de Datos a Publicar en Abierto

En este capítulo se presenta un método general para la selección de conjuntos de datos a publicar en abierto basado en el método Delphi Difuso e incorporando el punto de vista del reutilizador y el publicador. También se plantea un caso de estudio del método propuesto aplicado a la selección y publicación de conjuntos de datos abiertos en el ámbito universitario, sus resultados, discusión de los mismos y las conclusiones de la aplicación.

En este apartado 4.1 se describe el método de selección de conjuntos de datos a abrir aplicable a cualquier conjunto de datos del sector público como ayuntamientos, ministerios, universidades públicas, empresas públicas o en el sector privado en general. En el apartado 4.2 se aplica el método para un caso de estudio analizado al detalle.

4.1. Método de selección de conjuntos de datos a abrir

Como se ha analizado en el estado del arte, una de las cuestiones clave para las instituciones que abordan procesos de apertura de datos, es la selección de los conjuntos de datos a abrir. Se observa en este estudio previo que esta selección debe incorporar criterios para maximizar el potencial reutilizador de los conjuntos, minimizando los costes de apertura. El método propuesto toma en cuenta lo analizado en el estado del arte, donde para lograr una alta efectividad, tanto en uso como en costo de publicación, se debe tomar en cuenta a los reutilizadores y a los publicadores. Una vez con los conjuntos de datos seleccionados por los reutilizadores se debe optimizar la publicación de acuerdo al coste por los publicadores que son los que colocarán los conjuntos de datos en los portales de datos abiertos. Se realiza la consulta en primer lugar a los reutilizadores debido a que, de los conjuntos de datos que pueden estar en abierto solo se reutilizan aquellos en los que más interés tengan los reutilizadores.

Para conseguir esta selección efectiva de conjuntos de datos para abrir se propone un método que consiste en cuatro pasos, ver Figura 24. Este método utiliza como base de selección a los expertos, tanto reutilizadores como publicadores, quienes con el Método Delphi Difuso logran consensos para seleccionar los conjuntos de datos a abrir, debido a que serán los que más potencial tengan para crear productos y servicios TI de valor agregado, además de un menor costo de apertura.

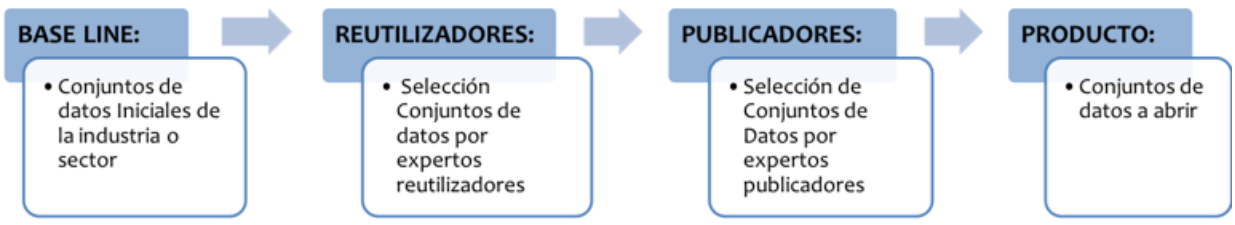


Figura 24. Método de Selección de conjuntos de datos a abrir

El método de selección de conjuntos de datos a abrir (ODSM del inglés Open Data Selecting Method) empieza con: (i) Definición de conjuntos de datos iniciales, dependiendo de la industria o sector donde se necesite abrir conjuntos de datos. Se puede tomar como referencia instituciones similares de otros países o experiencias de expertos. (ii) Aplicación a estos conjuntos de un método de selección basada en la consulta a expertos reutilizadores y utilizando el Método Delphi Difuso (MDD). (iii) Aplicación de un método de selección basados en la consulta a expertos publicadores que seleccionarán, entre los conjuntos seleccionados por los reutilizadores, los que sean factibles desde el punto de vista del costo asumible para su publicación, utilizando el mismo MDD para su definición. (iv) Se obtiene como producto los conjuntos de datos a abrir tomando en cuenta el punto de vista del reutilizador y el publicador haciendo efectiva su selección y apertura.

4.1.1. Método de selección de conjuntos de datos según el reutilizador

Para su aplicación se propone un proceso dividido en fases, cuyo objetivo es lograr identificar los conjuntos de datos a abrir según el criterio de los reutilizadores. A continuación, se describen estas fases (ver también la Figura 25).

De acuerdo a lo descrito previamente en el método, se selecciona los conjuntos de datos a abrir desde el punto de vista del reutilizador, en el cual se aplica la teoría difusa al Método Delphi y las funciones de membresía triangular, explicadas a detalle en el capítulo de metodología. La imprecisión del consenso de expertos se resuelve utilizando la teoría difusa, evaluándose en una escala más flexible. Con este método se consigue una manera formal y sistemática de definir los conjuntos de datos a abrir desde el punto de vista del reutilizador.

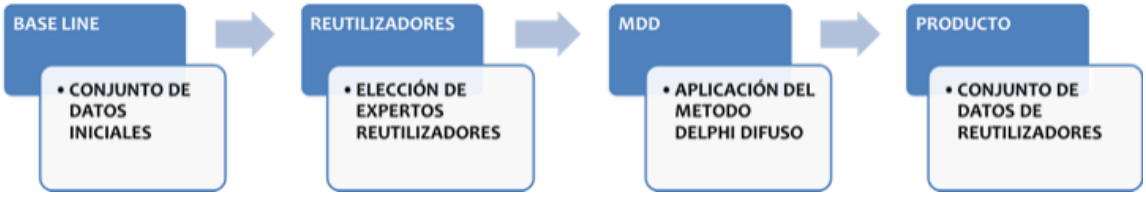


Figura 25. Método de Selección de conjuntos de datos por los reutilizadores

4.1.1.1. Fase 1: Conjunto de datos iniciales.

Se realiza una búsqueda de un mínimo conjunto de datos representativos publicados en el sector público o privado a analizar. Los pasos que se ejecutan son:

1) Definición: Se determinan conjuntos de datos abiertos publicados en el sector. Para esta definición se puede buscar por sector de industria en los portales abiertos publicados o consultar a expertos del área, también es importante la experiencia del investigador que va a aplicar el método.

2) Eliminación: Se eliminan los conjuntos de datos duplicados.

3) Consolidación: De los conjuntos de datos sin duplicación se realiza una depuración consolidando en grupos de conjuntos de datos bajo algún criterio, por ejemplo, de información contenida

4.1.1.2. Fase 2: Selección de los expertos.

Los expertos son personas del entorno del sector de la industria analizada que deben estar relacionadas con el ámbito de actuación y que tengan conocimiento de datos abiertos, preferiblemente que hayan participado como infomediarios de los mismos.

4.1.1.3. Fase 3: Aplicación del Método Delphi Difuso

Una vez definida la línea base y seleccionado el grupo de expertos, se procede a aplicar el Método Delphi Difuso. Se elabora un cuestionario que será respondido por los expertos reutilizadores y servirá para la primera iteración. Se realizan iteraciones con las que se debe conseguir los consensos necesarios para definir los conjuntos de datos a abrir desde el punto de vista de los expertos reutilizadores.

Se debe estructurar la pregunta de tal manera que los expertos puedan dar su criterio de acuerdo a la probabilidad de uso de los conjuntos de datos de entrada.

Como ejemplo de preguntas tipo se propone las siguientes:

Primera iteración:

Pregunta 1: *¿Cuál es la probabilidad que los siguientes conjuntos de datos se reutilicen para generar aplicaciones o servicios de valor agregado para el sector? Indique el porcentaje más pesimista, moderado y optimista de acuerdo a su criterio de probabilidad de reutilización.*

Pregunta 2: *¿Piensa que existen conjuntos de datos abiertos importantes que faltan?, Si es así, ¿podría sugerir nuevos conjuntos de datos que puedan ser abiertos para ser implementados en aplicaciones del sector?*

Segunda iteración:

Para la segunda iteración se presentan los resultados obtenidos a cada experto, añadiendo la media geométrica de cada valor dado y los comentarios de los conjuntos de datos analizados. Con esta información se solicita que ajusten sus valores iniciales o se mantengan en los mismos. Se debe incluir los nuevos conjuntos de datos que ellos propongan.

4.1.1.4. Fase 4: Conjuntos de datos seleccionados por los reutilizadores.

Con los resultados obtenidos de la aplicación del MDD, se elabora una tabla resumen donde se ubican los conjuntos de datos con sus valores de $S_j > r$, queda a criterio del investigador tomar el valor de referencia de acuerdo al índice de confiabilidad que necesite, de acuerdo con [40]

4.1.1.5. Resultados del proceso de selección de conjuntos de datos según el reutilizador.

Al aplicar la metodología definida se deben logran seleccionar conjuntos de datos, con su número nítido $S_j > r$ para su análisis.

Se debe mantener un número mínimo de participación de expertos, según lo presentado en la metodología.

4.1.2. Método de Selección de Conjuntos de Datos según el publicador

Para realizar la selección teniendo en cuenta el punto de vista del publicador, se aplican las fases del método de Selección de conjuntos de datos mostradas en la Figura 26.

Los conjuntos de datos obtenidos desde el punto de vista del reutilizador representan el conjunto de información de entrada para el proceso de selección desde el punto de vista del publicador. Los expertos publicadores seleccionan los conjuntos a abrir según el costo, en términos de hardware,

software y recursos humanos para preparar los datos para publicarlos en una plataforma de datos abiertos

Figura 26. Método de Selección de conjuntos de datos por los publicadores

4.1.2.1. Fase 1: Conjuntos de datos de entrada para los publicadores.

Los conjuntos seleccionados por los reutilizadores son los conjuntos de entrada para el método de selección de los expertos publicadores.

4.1.2.2. Fase 2: Selección de los expertos.

Los expertos serán escogidos por su participación en la implementación de portales de datos abiertos y que conocen los datos que se generan en el sector analizado y la estimación de su costo de publicación. Se debe cuidar un número mínimo de participantes según se propone en el capítulo sobre la metodología de este documento.

4.1.2.3. Fase 3: Aplicación del Método Delphi Difuso

Una vez definida la línea base se procede a aplicar el Método Delphi Difuso para lo cual se elabora el cuestionario contestado por los expertos publicadores. Para cada sector se podría escoger el cuestionario adecuado, como ejemplo se propone:

Primera iteración:

Pregunta: *La comunidad reutilizadora ha definido los siguientes conjuntos de datos para que sean publicados. Como usted conoce existe un costo involucrado, en términos de personas, infraestructura y aplicaciones, para preparar los datos para publicarlos en una plataforma de datos abiertos. Tomando en cuenta que estos datos no estuvieran digitalizados ¿Cuál es el costo más bajo, más probable y más alto que se tendría para publicar cada uno de los siguientes conjuntos de datos? Tomar una escala del 1 al 10 donde 10 es el costo más alto posible.*

Segunda iteración:

Para la segunda iteración se presentan los resultados obtenidos. Se entrega a cada experto la media geométrica de cada valor dado y los comentarios de los conjuntos de datos analizados. Con esta información se solicita ajusten sus valores o se mantengan en los mismos.

4.1.2.4. Fase 4: Presentación de resultados

Con los resultados obtenidos de la aplicación del MDD, se elabora una tabla resumen donde se ubican los conjuntos de datos que deberían publicarse.

4.1.2.5. Resultados Método de Selección de Conjuntos de Datos según el Publicador

Se debe realizar un análisis de los resultados obtenidos. Es conveniente construir tablas con la información necesaria para este análisis. Se deben presentar que conjuntos de datos alcanzan los consensos en la primera iteración y en las siguientes iteraciones.

4.1.3. Resultados del Método de Selección de Conjuntos de Datos

Considerando un costo razonable de publicación se debe construir una tabla con los resultados finales de los conjuntos de datos a publicar.

4.2. Aplicación del Método de selección de conjuntos de datos a abrir en el ámbito universitario

Actualmente, el concepto de datos abiertos es cada vez más popular en Ecuador, debido a la reciente adhesión al Open Goverment Partnership (OGP) que ha incentivado que las entidades públicas estén preparando la agenda donde se contempla el apoyo a la publicación en abierto de sus conjuntos de datos. Las universidades como entes dinamizadores de proyectos tecnológicos y sociales se encuentran proponiendo la apertura de datos, no solo por el cumplimiento de la ley de transparencia, sino también por la promesa de innovación y crecimiento productivo de esta acción. Por otro lado, se han creado colectivos civiles para apoyar a estas iniciativas y se tienen expertos tanto del lado de la reutilización como de la publicación.

Los siguientes aparatos detallan la aplicación del Método general de selección de conjuntos a abrir, tomando el ámbito universitario.

4.2.1. Método de selección de conjuntos de datos según el reutilizador.

Para realizar la selección teniendo en cuenta el punto de vista del reutilizador, se aplican las fases del proceso de definición del conjunto de datos mostradas en la Figura 25.

4.2.1.1. Fase 1: Conjunto de datos iniciales.

Se realiza una búsqueda de un mínimo conjunto de datos representativos publicados en universidades. Los pasos que se ejecutan son:

1) Definición: Se determinan universidades que han publicado conjuntos de datos abiertos de acuerdo a la organización Linked Universities (http://linkeduniversities.org/) y el proyecto de datos abiertos de la Universidad de Alicante (https://datos.ua.es/), incluyendo actividades como el concurso de ideas y uso de conjuntos de datos universitarios realizado en 2015.

2) Eliminación: Se eliminan los conjuntos de datos duplicados.

3) Consolidación: De los conjuntos de datos sin duplicación se realiza una depuración consolidando en grupos de conjuntos de datos bajo el criterio de información contenida como sigue:

- Académico
- Investigación
- Bienestar Estudiantil: Becas y Ayudas
- Organización e Infraestructura
- Datos en tiempo real, sensores

Como resultado, se han identificado 28 conjuntos de datos según esta clasificación y que se describen en la Tabla 3.

4.2.1.2. Fase 2: Selección de los expertos

Los expertos son personas del entorno universitario del Ecuador, relacionadas con la universidad por sus estudios académicos, que tienen conocimiento de datos abiertos y han participado en

colectivos ciudadanos de reutilización de datos abiertos. En este caso los participantes pertenecen a la REDAM (Red de Datos Abiertos y Metadatos del Ecuador) organismo que está acreditado en la SENECYT (Secretaría de Educación Superior Ciencia y Tecnología del Ecuador). Se consigue la participación de 10 expertos para la primera iteración. Un experto no participa en la segunda iteración, quedando 9 expertos, que es un número válido, como se describe en la metodología.

4.2.1.3. Fase 3: Aplicación del Método Delphi Difuso

Una vez definida la línea base y seleccionado el grupo de expertos, se procede a aplicar el Método Delphi Difuso. Se elabora un cuestionario que será respondido por los expertos reutilizadores y servirá para la primera iteración. Se realizan dos iteraciones con las que se consiguió los consensos necesarios para definir los conjuntos de datos a abrir desde el punto de vista de los expertos reutilizadores.

Para el caso de las universidades se tiene el siguiente cuestionario:

Primera iteración:

Pregunta 1: *¿Cuál es la probabilidad que los siguientes conjuntos de datos se reutilicen para generar aplicaciones o servicios de valor agregado para la comunidad universitaria? Indique el porcentaje más pesimista, moderado y optimista de acuerdo a su criterio de probabilidad de reutilización.*

Pregunta 2: *¿Piensa que existen conjuntos de datos abiertos importantes que faltan?, Si es así, ¿podría sugerir nuevos conjuntos de datos que puedan ser abiertos para ser implementados en aplicaciones en la universidad ecuatoriana?*

Segunda iteración:

Para la segunda iteración se presentan los resultados obtenidos a cada experto, añadiendo la media geométrica de cada valor dado y los comentarios de los conjuntos de datos analizados. Con esta información se solicita que ajusten sus valores iniciales o se mantengan en los mismos. Se incluye los nuevos conjuntos de datos que ellos propongan.

4.2.1.4. Fase 4: Conjuntos de datos seleccionados por los reutilizadores.

Con los resultados obtenidos de la aplicación del MDD, se elabora una tabla resumen donde se ubican los conjuntos de datos con sus valores de $S_j > 0,6$. Escogemos el 0,6 como referencia de acuerdo a [40]

Tabla 3. Conjunto de datos identificados en las universidades

Clasificación	Descripción
Académico:	
Unidades Organizativas.	Datos sobre Facultades, Institutos, Centros y Departamentos Académicos, Administrativos y de Investigación.
Titulaciones.	Carreras, asignaturas, plan de estudios.
Estudiantes.	Alumnos matriculados.
Seguimiento a Egresados.	Alumnos egresados.
Docentes.	Docentes que se encuentran con contrato.
Administrativos.	Trabajadores administrativos.
Investigación:	
Convenios.	Convenios suscritos con entidades universitarias, empresas, gobiernos.
Proyectos de Investigación.	Proyectos que se están ejecutando, investigadores.
Publicaciones Científicas.	Publicaciones científicas del personal docente investigador publicados e indexados.
Bienestar Estudiantil:	
Becas y Ayudas.	Becas y ayudas existentes.
Servicios para el Estudiante.	Servicios médicos, legales, administrativos.
Transporte.	Sistemas de transporte desde y hacia la Universidad.
Bibliotecas.	Sistema de bibliotecas dentro de la Universidad.
Oportunidades de trabajo.	Oportunidades de trabajo, pasantías y prácticas.
Eventos Culturales y Deportivos.	Descripción de Eventos culturales y deportivos.

Organización e Infraestructura:	
Datos Geo posicionados.	Datos geo posicionados de los edificios, dependencias, equipos de comunicaciones, máquinas expendedoras, zonas de jardinería y otros.
Organigrama de Dirección.	Datos de personal que dirige las unidades administrativas, académicas e investigación.
Restaurantes y Puestos de Comida.	Descripción de restaurantes, puestos de comida y bares.
Energía.	Consumo de energía tomado en forma física y digital.
Agua.	Consumo de agua tomado en forma física y digital.
Telefonía.	Consumo telefónico tomado en forma física y digital.
Colectores de Basura.	Sistema de recolección interna de basura, tachos de basura, horarios de recolección.
Jardinería.	Sistema de arreglo de jardinería, por zona, personal, herramientas, horarios de arreglo.
Guardianía.	Sistema de guardianía, por zona, personal, herramientas, rutas, cámaras.
Proyectos de Infraestructura.	Detalle de proyectos de infraestructura en mantenimiento a implementarse.
Presupuesto.	Presupuestos de la Universidad.
Datos Tiempo Real Sensores:	
Datos Meteorológicos.	Temperatura, Humedad, Rayos Ultravioleta, Lluvia, Presión atmosférica, Polución.
Datos de Sensores.	Datos de sensores ubicados en las dependencias, controles, equipos, registros, Cámaras, Wifi, accesos.

4.2.1.5. Resultados del proceso de selección de conjuntos de datos según el reutilizador

Al aplicar la metodología definida se logran seleccionar los conjuntos de datos presentados en la Tabla 4 con su número nítido $S_j > 0,6$ para su análisis.

En la segunda iteración se logró la participación de 9 expertos, donde 3 conjuntos de datos tienen un número nítido $S_j > 0,7$ y 9 conjuntos de datos un número nítido S_j entre 0,6 y 0,7.

4.2.2. Método de Selección de Conjuntos de Datos según el publicador

Para realizar la selección teniendo en cuenta el punto de vista del publicador, se aplican las fases del proceso de definición del conjunto de datos mostradas en la Figura 26.

4.2.2.1. Fase 1: Conjuntos de datos de entrada para los publicadores

Los conjuntos seleccionados por los reutilizadores son los conjuntos de entrada para el método de selección de los expertos publicadores.

4.2.2.2. Fase 2: Selección de los expertos

Los expertos son los directores y exdirectores de tecnología de las universidades que han participado en la implementación de portales de datos abiertos y que conocen los datos que se generan en la universidad y la estimación de su costo de publicación. Se consigue la participación de 11 expertos para la primera iteración, para la segunda iteración no participan 2 expertos, quedando 9, que es un número válido, como se describe en la metodología.

4.2.2.3. Fase 3: Aplicación del Método Delphi Difuso

Una vez definida la línea base se procede a aplicar el Método Delphi Difuso para lo cual se elabora el cuestionario contestado por los expertos publicadores. Para el caso de las universidades se tiene el siguiente cuestionario:

Primera iteración:

El cuestionario se compone de una pregunta, que definirá la prioridad de los conjuntos de datos de la línea base de acuerdo al costo de publicación en función de la experiencia de los expertos. Con dos iteraciones se consigue los consensos necesarios para definir los conjuntos de datos a abrir desde el punto de vista de los expertos publicadores.

Pregunta: *La comunidad reutilizadora ha definido los siguientes conjuntos de datos para que sean publicados. Como usted conoce existe un costo involucrado, en términos de personas, infraestructura y aplicaciones, para preparar los datos para publicarlos en una plataforma de datos abiertos. Tomando en cuenta que estos datos no estuvieran digitalizados ¿Cuál es el costo*

más bajo, más probable y más alto que se tendría para publicar cada uno de los siguientes conjuntos de datos? Tomar una escala del 1 al 10 donde 10 es el costo más alto posible.

Tabla 4. Conjunto de Datos Seleccionados por los Reutilizadores

No.	Conjunto de Datos	Si
1	Publicaciones Científicas.	0,78
2	Titulaciones.	0,76
3	Estudiantes.	0,75
4	Presupuesto.	0,69
5	Consumo de Agua.	0,67
6	Proyectos de Investigación.	0,67
7	Datos de Guardianía.	0,65
8	Colectores de basura.	0,64
9	Becas y ayudas.	0,64
10	Unidades Organizativas.	0,64
11	Oportunidades de trabajo.	0,63
12	Convenios.	0,61

Segunda iteración:

Para la segunda iteración se presentan los resultados obtenidos. Se entrega a cada experto la media geométrica de cada valor dado y los comentarios de los conjuntos de datos analizados. Con esta información se solicita ajusten sus valores o se mantengan en los mismos.

4.2.2.4. Fase 4: Presentación de resultados

Con los resultados obtenidos de la aplicación del MDD, se elabora una tabla resumen donde se ubican los conjuntos de datos que deberían publicarse.

4.2.2.5. Resultados del Método de Selección de Conjuntos de Datos según el Publicador

En la primera iteración se logra tener consenso en 2 de los 12 conjuntos de datos: Consumo de Agua y Oportunidades de trabajo que cumplen las ecuaciones de verificación de consenso (6) y (7).

En la Tabla 5 se presentan los resultados de la selección de los conjuntos de datos por los expertos en la primera iteración de la metodología. Se agrega una columna de la media geométrica del costo más probable, con el fin de notar la variación de costo dentro de los conjuntos de datos seleccionados en esta primera iteración. Se puede ver que los dos primeros conjuntos de datos lograron consenso tanto en la ecuación de verificación (6) como en la ecuación de verificación (7). Los 10 conjuntos restantes obtuvieron consenso en la ecuación de verificación (6) pero no en la ecuación de verificación (7), estos se ordenan de acuerdo al valor más probable en cuanto a costo más probable de forma descendiente, tanto para los conjuntos de datos donde se consiguió el consenso como en los que no se consiguió.

En la segunda iteración se logra el consenso de 9 de los 12 conjuntos de datos con los valores de las ecuaciones de verificación (6) y (7). En la Tabla 6 se ordena los 9 conjuntos de datos por el valor más probable en forma ascendente. También se añaden los 3 conjuntos de datos que no lograron un consenso de los expertos.

4.2.2.6. Resultados Proceso de Selección de Conjuntos de Datos

Considerando un costo razonable de publicación menor a 5 que es la media geométrica del costo más probable se construye la Tabla 7 con los conjuntos de datos mínimos a publicar en el ámbito universitario.

4.2.2.7. Discusión de los resultados de la aplicación del método de selección

Según los resultados obtenidos, la comunidad reutilizadora está poniendo mayor interés en los datos procedentes de las áreas Académicas y de Investigación (ya que los conjuntos de datos de estas áreas tienen $S_j > 0,7$). Es por ello, que las universidades deben poner más atención a la hora de la publicación de datos relativos a estas áreas, que, por otra parte, representan los dos objetivos más importantes de la universidad de cara a la sociedad: la investigación y la academia. No

obstante, de los resultados obtenidos se desprende que tan solo el 10% del total de los conjuntos de datos analizados tiene $S_j > 0,7$, esto es un número reducido de conjunto de datos y pone en evidencia que es de gran utilidad realizar un estudio como el presente para definir los conjuntos de datos a publicar. El 60% de los conjuntos de datos tiene un $S_j < 0,5$ que son los que van a tener menor posibilidad de reutilización según el presente estudio.

Tabla 5. Conjunto de Datos Seleccionados en la primera iteración por los publicadores ordenados por valor más probable

No.	**Conjunto de Datos**	**$u_{ij} > l_j \forall i,j$**	**G > C**	**Más Probable**
1	Consumo de Agua.	SI	SI	5,9
2	Oportunidades de trabajo.	SI	SI	5,8
3	Datos de Guardianía.	SI	NO	6,3
4	Colectores de basura.	SI	NO	5,7
5	Titulaciones.	SI	NO	4,5
6	Presupuesto.	SI	NO	4,4
7	Proyectos de Investigación.	SI	NO	4,3
8	Unidades Organizativas.	SI	NO	4,2
9	Publicaciones Científicas	SI	NO	4,1
10	Convenios.	SI	NO	4,0
11	Becas y Ayudas	SI	NO	3,9
12	Estudiantes.	SI	NO	3,5

La Tabla 4, con un orden descendente de números nítidos S_j de los conjuntos de datos evaluados, en combinación con la Tabla 7, con un orden ascendente en función del costo de publicación, se convierte en una herramienta para la estrategia de priorización en la evaluación desde el punto del reutilizador y del publicador. Usando esta idea se podría decir que esta técnica podría ser usada de manera periódica por las instituciones para trazar estrategias de priorización de publicación de datos abiertos, que complementen a otras como la descrita en [81].

Los resultados obtenidos en la estrategia de selección del conjunto de datos desde el punto de vista del publicador muestran que los expertos llegan en la primera iteración a un consenso de costo de 2 conjuntos de datos: Consumo de Agua y Oportunidades de trabajo con un valor más probable de costo de publicación de 5,9 y 5,8 respectivamente, lo que los hace muy costoso de publicar. En la segunda iteración se consigue consenso en 9 de los 12 conjuntos de datos, sin lograr consenso a los conjuntos de datos de Presupuesto, Becas y Ayudas y Colectores de Basura.

Tabla 6. Conjunto de Datos Seleccionados por los Publicadores

No.	Conjunto de Datos	$u_{ij} > l_j \; \forall \; i,j$	G > C	Más Probable
1	Estudiantes.	SI	SI	3,9
2	Proyectos de Investigación.	SI	SI	4,2
3	Unidades Organizativas.	SI	SI	4,4
4	Publicaciones Científicas	SI	SI	4,5
5	Convenios.	SI	SI	4,6
6	Titulaciones.	SI	SI	5,0
7	Consumo de Agua.	SI	SI	5,8
8	Oportunidades de trabajo.	SI	SI	5,9
9	Datos de Guardianía.	SI	SI	6,7
10	Presupuesto.	SI	NO	4,2
11	Becas y Ayudas	SI	NO	4,0
12	Colectores de basura.	SI	NO	5,9

El costo adecuado para la publicación depende de varios factores tales como presupuesto de cada institución, normativa externa (como la legislación en materia de transparencia) y normativa interna (como directivas de apoyo a la innovación). En este estudio se toma la media geométrica como un valor aceptable para la publicación, teniendo cada institución la capacidad de evaluar sus factores y publicar los conjuntos de datos que se ajusten a su realidad. El valor de la media

geométrica es 5 y se deberían publicar los siguientes 6 conjuntos de datos ordenados por costo: Estudiantes (3,9), Proyectos de Investigación (4,2), Unidades Organizativas (4,4), Publicaciones Científicas (4,5), Convenios (4,6), Titulaciones (5,0). Esto corresponde al 100% de los conjuntos de datos de Investigación, el 50% de los conjuntos de datos Académicos.

Tabla 7. Conjunto de Datos Seleccionados por los Reutilizadores y Publicadores

No.	Conjunto de Datos	Más Probable
1	Estudiantes.	3,9
2	Proyectos de Investigación.	4,2
3	Unidades Organizativas.	4,4
4	Publicaciones Científicas	4,5
5	Convenios.	4,6
6	Titulaciones.	5,0

Los conjuntos de datos iniciales que se usaron como línea base fueron 28. Los expertos reutilizadores propusieron 2 nuevos conjuntos, quedando 30 conjuntos de datos que tras completar el proceso de selección de los expertos publicadores quedaron en 6 conjuntos de datos, es decir el 20% de los conjuntos de datos propuestos. Por tanto, se comprueba que existe un número reducido de conjuntos de datos que tienen probabilidad de reutilización y que tiene costos adecuados de publicación.

Los reutilizares seleccionaron 3 conjuntos de datos con un $S_j > 0,7$: Publicaciones Científicas, Titulaciones y Estudiantes y tras la selección de los publicadores se determina que estos tres conjuntos de datos son los más factibles de ser publicados en función del costo, con lo que se consigue satisfacer en alto grado la necesidad del reutilizador y la necesidad de ser eficiente al publicar.

Los reutilizadores y los publicadores determinan que los conjuntos de datos relacionados con la investigación son los más factibles desde el punto de vista de reutilización y costo. La investigación es un objetivo estratégico de las universidades y con las cuales se efectiviza la vinculación con la

sociedad, otro objetivo estratégico. Por tanto, es fundamental que las universidades tengan implementados correctamente los procesos involucrados en la investigación, garantizando la calidad de los datos para que se puedan reutilizar a través de aplicaciones generadas por la misma comunidad y socializado a la ciudadanía a través de los infomediarios.

Otra estrategia fundamental en la universidad es la Academia, el estudio revela que se requiere la publicación del 50% de los conjuntos de datos clasificados en académicos, con la característica de que son los menos costosos, pues las universidades ya los tienen en sus sistemas académicos y son fruto de procesos y aplicación de sistemas informáticos.

Los expertos publicadores han llegado a un consenso respecto de los costos de los conjuntos de datos: Consumo de Agua, Oportunidades de trabajo y datos de Guardianía, pero con costos altos de publicación debido a que conlleva la implementación de infraestructura tecnológica. Es interesante ver que a juicio de los expertos es más costoso implementar los procesos adecuados para adquirir y publicar los conjuntos de datos de Oportunidades de trabajo que de Consumo de Agua, seguramente porque mantener con calidad el conjunto de datos de Oportunidades de trabajo depende de personas o empresas externas, en cambio el consumo de agua depende de la calidad de la tecnología implementada para la adquisición de los datos.

La metodología implementada para la selección de datos ha confirmado que existen conjuntos de datos cuya publicación es más demandada y menos costosa. Es importante que una vez determinado estos conjuntos de datos se proceda a asegurar la calidad de los datos para poder publicarlos y que se mantengan en el tiempo para que los reutilizadores tengan confianza en los datos.

El MDD se ha utilizado en primera instancia para determinar un grupo de conjuntos de datos mínimo a reutilizar, para ello nos sirve el valor nítido S_j que nos permite validar los conjuntos que los expertos han seleccionado. Al aplicar el MDD con los expertos publicadores nos permite determinar de una manera eficiente los consensos para publicar de acuerdo al costo. Son dos formas de utilización que combinadas nos permite tener un grupo final de conjuntos de datos que se debe publicar en la universidad ecuatoriana.

4.2.3. Conclusiones de la aplicación del método de selección de conjuntos de datos en el ámbito universitario

Este estudio presenta una metodología para la selección de conjuntos de datos a ser abiertos desde el punto de vista de su potencial de reutilización y el costo de publicación, consultando la opinión de expertos, para lo que se utiliza el Método Delphi Difuso.

El método permite generar una escala de prioridad de conjunto de datos a abrirse basados en los consensos de los expertos reutilizadores y los expertos publicadores.

Los expertos reutilizadores ponen interés en los conjuntos de datos para: Investigación y Academia. Los tres conjuntos de datos con mayor aceptación por los expertos están en estas categorías. Publicaciones Científicas, Titulaciones y Estudiantes.

El 40% de los conjuntos de datos superan el índice de confiabilidad del estudio con referencia mayor a 0,6 y solo el 10% del total de los conjuntos de datos analizados superan el índice de confiabilidad del estudio con referencia mayor a 0,7. Además, parece que existe un interés claro de los reutilizadores en la clasificación de conjuntos de datos de Organización e Infraestructura que corresponde al 33% de los conjuntos de datos seleccionados. Por otra parte, existen 18 conjuntos de datos que no han superado el índice de confiabilidad del estudio, correspondiendo al 60% de los conjuntos de datos analizados por los reutilizadores. Considerando, a los expertos publicadores, estos han definido que 6 conjuntos de datos tienen un costo adecuado para ser publicado, cumpliendo tanto la necesidad de los reutilizadores como el costo de publicación. Esto es el 20 % de los conjuntos de datos que evaluaron los expertos reutilizadores.

Cabe destacar que, los conjuntos de datos de la clasificación de Investigación se han seleccionado en su 100%, es decir, existe un interés de los reutilizadores por estos conjuntos de datos y el costo para publicarlos es adecuado. De los 6 conjuntos de datos de la clasificación de Académicos, 3 han sido seleccionado por los reutilizadores y los publicadores. Por ello, existe interés de los reutilizadores y la factibilidad de publicar se centra en los conjuntos de datos de la clasificación de Investigación y Académico que son los ejes estratégicos de las universidades.

El proceso implementado para determinar los conjuntos de datos a publicar en el ámbito universitarios ha permitido seleccionar una lista de 6 conjuntos de datos a publicar con costos adecuados, garantizando de esta manera que los esfuerzos de publicar tengan impacto en los reutilizadores, siendo eficientes en la utilización de recursos de las empresas públicas y privadas.

Con el método propuesto se ha logrado el consenso y decisión en dos iteraciones, esto ahorra mucho tiempo en comparación con el Método Delphi tradicional.

Capítulo 5: Conclusiones y trabajo futuro

En este capítulo se exponen las conclusiones de las diferentes etapas del trabajo de investigación en relación a los objetivos propuestos. También se desatacan las principales contribuciones en forma de publicaciones científicas y transferencia de investigación. Finalmente, se identifican las principales líneas abiertas de investigación que marcaran el trabajo futuro.

5.1. Conclusiones

Para lograr identificar espacios de mejora en el proceso de apertura de datos, un objetivo de esta tesis era analizar y clasificar la investigación de datos abiertos realizada en la comunidad desde el punto de vista tecnológico. Esto se logra implementando un estudio de mapeo sistemático a 671 publicaciones sobre datos abiertos con las siguientes aportaciones:

- Proporciona una visión general y consolidada del campo de los datos abiertos y sus tendencias.
- Determina las publicaciones y foros científicos más utilizadas sobre datos abiertos desde el punto de vista técnico.
- Define los espacios de tiempo donde han sido identificados los mayores aportes de publicaciones en el tema de datos abiertos.
- Encuentra que la Web Semántica, la Ingeniería de Software y el Gobierno son los temas más importantes abordados en la investigación. El gran reto de los datos abiertos es lograr una legislación adecuada que balancee la apertura de datos con la privacidad de datos personales, sin ser contrapuestos, sino más bien se complementan.
- Determina que los infomediarios es el dominio más desarrollado en las publicaciones, así como las fases de Explotación y Exploración de datos. Lo que confirma la necesidad de tomar en cuenta en la publicación de datos a los reutilizadores para aperturar los conjuntos de datos de las instituciones.
- Encuentra que las fases relacionadas con el consumo de datos están muy poco representadas y que se debe realizar un esfuerzo desde el lado de la investigación para publicar conjuntos de datos que finalmente se utilicen.
- Identifica que existen proyectos con financiamiento para impulsar iniciativas innovadoras en el campo de los datos abiertos y su impacto en la generación de empleo para la comunidad y beneficio social para la ciudadanía.
- Permite concluir la necesidad de proponer métodos formales de selección de conjuntos de datos a abrir por parte de las instituciones, teniendo en cuenta el punto de vista del reutilizador y el publicador.

Apoyado en las conclusiones del estado del arte, este trabajo de investigación también tenía como objetivo proporcionar un método para la selección de los conjuntos de datos a abrir por parte de organizaciones que aborden procesos de apertura. El método se propone con el objetivo de ser general y aplicable en diferentes contextos que impliquen distintas organizaciones y colectivos. Además, se busca que sea un método científico, debiendo estar dotado del formalismo necesario para garantizar una aplicabilidad y obtención de resultados sistemática. Finalmente, de acuerdo con las conclusiones del estado del arte, esta metodología deberá integrar el punto de vista del reutilizador, así como el del publicador. Dicho método ha sido implementado con la utilización de un proceso de selección de conjuntos de datos a ser abiertos desde el punto de vista de su potencial de reutilización y el costo de publicación, consultando la opinión de expertos, para lo que se utiliza el Método Delphi Difuso, presentando las siguientes aportaciones:

- Creación de un método general de selección de conjuntos de datos donde se consulta expertos reutilizadores que generan una escala de prioridad de conjuntos a abrirse. Estos conjuntos son evaluados por expertos publicadores tomando en cuenta el coste asumible para su publicación, dando como resultado un grupo de conjuntos de datos con mayor probabilidad de reutilización y coste adecuado para su publicación.
- Introducción del Método Delphi Difuso como mecanismo de generación de consensos en cada etapa. Este ha permitido la selección tanto desde el punto de vista del reutilizador como del publicador, con la rigurosidad matemática necesaria, que permite en pocas iteraciones dar los resultados de la selección.
- Definición del proceso de selección de expertos que trabajarán con el Método Delphi Difuso, las preguntas de investigación que generen números difusos y los conjuntos de datos iniciales que permiten aplicar el método en diferentes contextos que impliquen distintas organizaciones y expertos.
- Aplicación del método propuesto para la selección de conjuntos de datos a abrir en un caso de estudio en las Universidades Ecuatorianas, con dimensión y complejidad para la obtención de conclusiones. Donde se logró definir, de un grupo de 30 conjuntos de datos, 6 con potencial de reutilización alto y un coste adecuado de publicación. Esto permite a las universidades consensuar de forma objetiva, internamente entre distintos colectivos

usuarios y externamente con otras universidades, las prioridades de publicación de datos, con criterios de optimización del potencial reutilizador y el coste.

5.2. Contribuciones

En este apartado se presentan las principales contribuciones del trabajo de investigación en forma de publicaciones científicas y contribuciones de transferencia.

Publicaciones científicas:

Se relacionan tanto publicaciones científicas ya publicadas como pendientes de publicación:

- R. A. Enríquez-Reyes, A. Fuster-Guilló, J.N. Mazón, "Considering reusers when selecting datasets to open: a case of study from universities", IEEE LATINAMERICA TRANSACCIONS, JCR Journal Impact Factor: 0,804, 2019.
- Robert Enríquez-Reyes, Andrés Fuster-Guilló, José-Norberto Mazón, "Selección y Publicación de Datos Abiertos: Una Aplicación para Universidades", , Ponencia en 4th International Conference on Information Systems and Computer Science, INCISCOS, publicados por Conference Publishing Services y enviados a la biblioteca digital IEEE Xplore y CSDL (Computer Society Digital Library), noviembre 2019, indexada en SCOPUS.
- Robert Enríquez-Reyes, Susana Cadena-Vela, Andrés Fuster-Guilló, Luis Daniel Ibáñez, José-Norberto Mazón, Elena Simperl, "Systematic mapping of open data studies: classification and trends from a technological perspective", presentado para revisión en revista Information Systems Frontiers, JCR Factor de impacto: 2,539. Enviado en febrero de 2019.
- Robert Enríquez-Reyes, Susana Cadena-Vela, "Portal de Datos Abiertos de la Universidad Central del Ecuador", Revista FIGEMPA Investigación y desarrollo, Año V, Volumnen 2, Número 7, 2017, indexada en Latindex. publicado
- Robert Enríquez-Reyes, Susana Cadena-Vela, "Diseño e Implementación de una Universidad Abierta: Caso Universidad Central del Ecuador", Revista FIGEMPA Investigación y desarrollo, Año V, Volumnen 2, Número 7, 2017, indexada en Latindex. publicado

- Robert Enríquez-Reyes, Andrés Fuster-Guilló, Jose-Norberto Mazón**,** "Selecting datasets to open involving reusers and publishers with Delphi Fuzzy Method: a case of study from universities", Social Science Computer Review**,** JCR Journal Impact Factor: 2,922, enviado.
- Robert Enríquez-Reyes, Edison Loza-Aguirre, Gonzalo Proaño-Chicaiza, Jaime Guilcapi-Mosquera, Byron López Chávez, "Using a web system to facilitate Fuzzy Delphi studies", International Conference on Applied Technologies, Indexado en Scopus, SCImago (SJR: Q2), enviado.

Contribuciones de transferencia de investigación

- MINTEL. Ministerio de Telecomunicaciones y Sociedad de la Información (Ecuador). Gerencia de Gobierno Electrónico. Colaboración con el estado en los procesos de publicación de datos como política de gobierno. Propuesta de aplicación del método de selección de conjuntos de datos en diferentes ministerios.
- MINTEL. Ministerio de Telecomunicaciones y Sociedad de la Información (Ecuador). Gerencia de Gobierno Electrónico. Colaboración con el estado en los procesos de publicación de datos como política de gobierno. Grupo de trabajo científico entre MINTEL y la Universidad Central del Ecuador para el desarrollo de herramientas para dinamizar la política de apertura de datos abiertos del gobierno de Ecuador.

5.3. Trabajos futuros

Como todo proceso de investigación, el avance en la consecución de objetivos ha permitido identificar nuevas oportunidades de interés investigador para el futuro. A continuación, se relacionan algunas de las líneas futuras prioritarias. Algunas de ellas se plantean como continuación de la investigación básica y otras están vinculadas a la transferencia de conocimiento:

- Aplicación del método propuesto en casos de estudio diferentes con organismos de otros sectores que impliquen colectivos reutilizadores y expertos distintos. Los sectores de aplicación podrían ser en principio: Ayuntamientos, ministerios, organizaciones sanitarias, centros educativos, entre otros.
- Transferencia de conocimiento del método y su aplicación al Ministerio de Telecomunicaciones y de la Sociedad de la Información de Ecuador (MINTEL) a través de

la inclusión del método para instrumentar el proceso de datos abiertos en Ecuador. Este trabajo está ya definido y empezará desde finales del 2019.

- Creación de un grupo de trabajo científico entre el MINTEL y la Universidad Central del Ecuador para realizar pilotos para la aplicación en el ámbito del Gobierno Central y los Gobiernos Autónomos como Ayuntamientos y empresas públicas. Este trabajo está ya en marcha para empezar en 2020, como parte de la vinculación de la universidad con la comunidad.
- Aportación de los estudios de investigación científica como el mapeo sistemático en la definición de la co-creación de la política de datos abiertos y de la política de protección de datos personales, un trabajo de vinculación de la universidad con el Gobierno dentro de la agenda de Gobierno abierto iniciada en Ecuador.
- Creación de un grupo de trabajo multidisciplinario con varias universidades ecuatorianas para la creación de un software que automatice el Método Delphi Difuso. En este momento ya se tiene los primeros resultados y se está publicando el software con licencia Creative Commos. Actualmente está publicado la versión beta. Esta aplicación ha sido un esfuerzo conjunto de universidades públicas donde la investigación ha servido para realizar el estudio comparativo del manejo de datos en forma manual y automática propuesta.
- Aplicación del método en ámbitos universitarios latinoamericanos e internacionales para determinar un conjunto mínimo de datos a abrir y se incluya en los planes estratégicos de las universidades. Se ha realizado acercamientos con universidades mexicanas que están interesadas en aplicar el método. Para marzo del 2020 se concursará con fondos locales privados para realizar el proyecto.
- Aplicación del método para determinación de indicadores de acreditación en la universidad ecuatoriana. Se ha propuesto la colaboración dentro de un proyecto con financiamiento privado local de CEDIA (Corporación Ecuatoriana de Internet Avanzado). Esto se estará realizando dentro del último trimestre de 2019. Es un trabajo en conjunto con la academia, la investigación y el ente regulador de la acreditación en Ecuador.

Finalmente, cabe reseñar que se llevarán a cabo trabajos de investigación donde se comparen metodologías diferentes para determinar su aplicabilidad o complementariedad en la resolución de problemas donde se utilice consensos de expertos como es el caso del Método Delphi Difuso.

Capítulo 6: Anexos

En este capítulo se incorporan tres anexos:

A) Oficio Nro. MINTEL-SEGE-2019-0459-O Quito, D.M., 25 de septiembre de 2019, Agradecimiento por aportes en el proceso de construcción de la política de datos abiertos de Ecuador, donde la metodología de selección de conjuntos de datos abiertos se integrará a la guía que instrumentará todo el proceso de datos abiertos en Ecuador.
B) Reconocimiento por el aporte en las mesas de diálogo para la co-creación de la Política Nacional de Datos Abiertos.
C) Aplicación de la metodología, uso de herramientas de encuestas.

A) Oficio Nro. MINTEL-SEGE-2019-0459-O Quito, D.M., 25 de septiembre de 2019

Oficio del Ministerio de Telecomunicaciones y de la Sociedad del Conocimiento de Ecuador (MINTEL) sobre la transferencia de conocimiento de la metodología y su aplicación al a través de la inclusión de la metodología para instrumentar el proceso de datos abiertos en Ecuador.

MINISTERIO DE TELECOMUNICACIONES Y DE LA SOCIEDAD DE LA INFORMACIÓN

Oficio Nro. MINTEL-SEGE-2019-0459-O

Quito, D.M., 25 de septiembre de 2019

Asunto: Agradecimiento por aportes en el proceso de construcción del política de datos abiertos de Ecuador

Señor Ingeniero
Robert Arturo Enríquez Reyes
Docente
UNIVERSIDAD CENTRAL DEL ECUADOR
En su Despacho

De mi consideración:

Por medio de la presente, deseo agradecer su constante aporte en el proceso de construcción de la política de datos abiertos de Ecuador, en especial por compartir el modelo denominado " Selección y Públicación de Datos Abiertos: Una Aplicación para Universidades", instrumento referencial con base científica que permité de una manera técnica y eficiente priorizar el conjunto de datos a liberarse.

En función de lo anterior deseo indicar que el modelo compartido será evaluado previo a su integración a la guía que instrumentará todo el proceso de datos abiertos en Ecuador.

Con sentimientos de distinguida consideración.

Atentamente,

Documento firmado electrónicamente

Mgs. Juan Carlos Castillo Moreno
GERENTE INSTITUCIONAL - GOBIERNO ELECTRÓNICO

Copia:
Señora Magíster
Johanna Aleida Pazmiño Coba
Especialista en Fomento y Difusión del Plan Nacional de Gobierno Electrónico - PNGE - Gobernanza Electrónica

OC

Firmado electrónicamente por:
JUAN CARLOS CASTILLO MORENO

B) Reconocimiento por el aporte en las mesas de diálogo para la co-creación de la Política Nacional de Datos Abiertos

Transferencia de conocimiento con la aportación de los estudios de investigación científica como el mapeo sistemático en la definición de la co-creación de la política de datos abiertos, un trabajo de vinculación de la universidad con el Gobierno dentro de la agenda de Gobierno abierto iniciada en Ecuador.

PRESIDENCIA **DE LA REPÚBLICA** · MINISTERIO DE **TELECOMUNICACIONES Y DE LA SOCIEDAD DE LA INFORMACIÓN** · SECRETARÍA NACIONAL DE **PLANIFICACIÓN Y DESARROLLO**

OEA Más derechos para más gente

Confieren el presente reconocimiento a:

Robert Enríquez

Por su valioso aporte en las mesas de diálogo para la
"Co-creación de la Política Nacional de Datos Abiertos"
realizadas en la ciudad de Quito, 10 y 11 de diciembre de 2018.

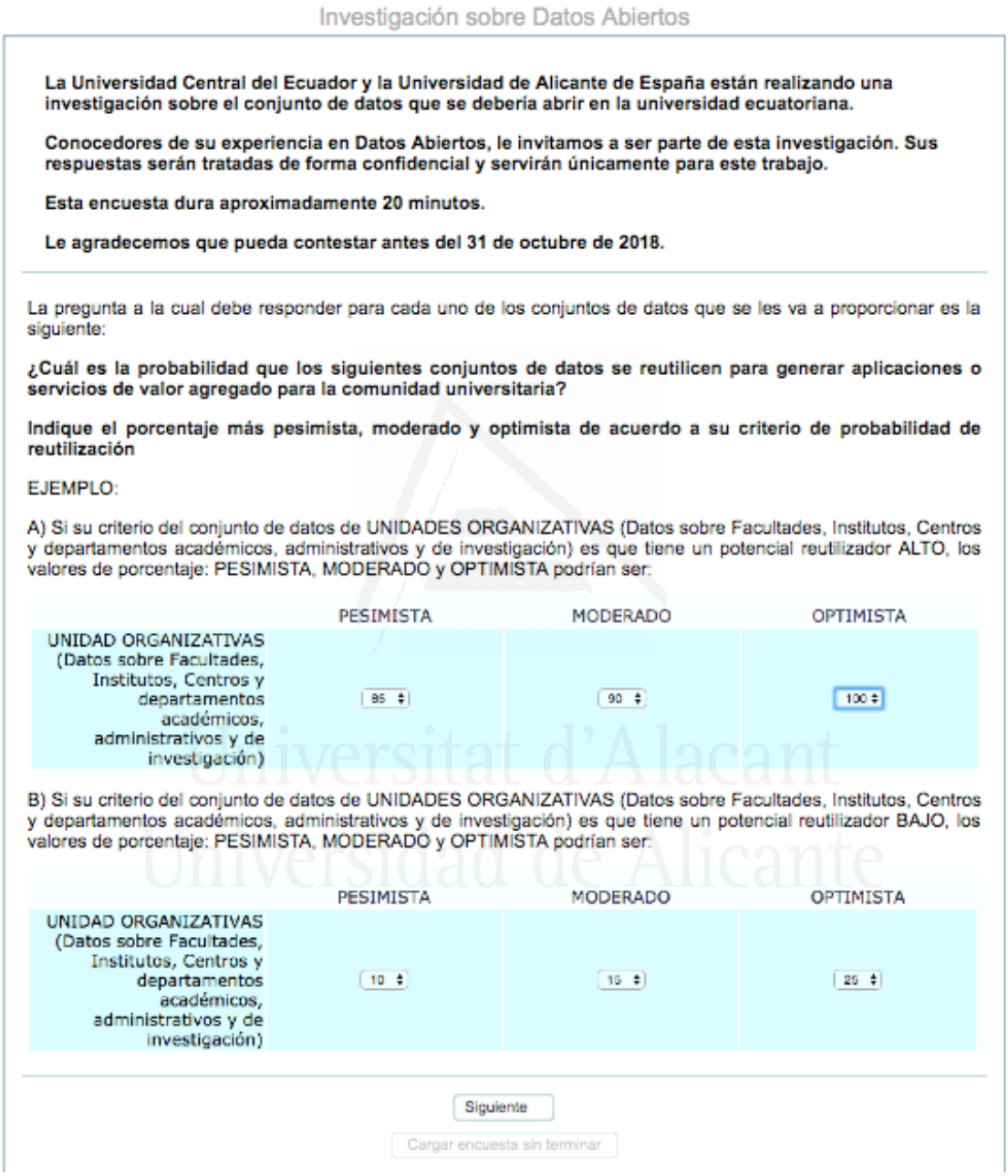
Quito, D.M., 27 de diciembre de 2018

CERTIFICACIÓN ELECTRÓNICA
Firmado Digitalmente por: RAUL JAVIER JARA NIQUEZ
Hora local: 09/01/2019 17:12

C) Aplicación de la metodología, uso de herramientas de encuestas

La ejecución de la metodología implicó el uso de herramientas de encuestas, las mismas que debía adaptarse a las características del Método Delphi Difuso. A falta de una herramienta especializada se utilizó una herramienta en software libre de encuestas, la misma que se adaptó para realizar la iteración del método. Se presentan a modo de ejemplo pantallas de la herramienta para la identificación gráfica del proceso.

Pantalla inicial que incluye la invitación a llenar la encuesta, la pregunta de investigación y un ejemplo para su correcto llenado.



Investigación sobre Datos Abiertos

La Universidad Central del Ecuador y la Universidad de Alicante de España están realizando una investigación sobre el conjunto de datos que se debería abrir en la universidad ecuatoriana.

Conocedores de su experiencia en Datos Abiertos, le invitamos a ser parte de esta investigación. Sus respuestas serán tratadas de forma confidencial y servirán únicamente para este trabajo.

Esta encuesta dura aproximadamente 20 minutos.

Le agradecemos que pueda contestar antes del 31 de octubre de 2018.

La pregunta a la cual debe responder para cada uno de los conjuntos de datos que se les va a proporcionar es la siguiente:

¿Cuál es la probabilidad que los siguientes conjuntos de datos se reutilicen para generar aplicaciones o servicios de valor agregado para la comunidad universitaria?

Indique el porcentaje más pesimista, moderado y optimista de acuerdo a su criterio de probabilidad de reutilización

EJEMPLO:

A) Si su criterio del conjunto de datos de UNIDADES ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y departamentos académicos, administrativos y de investigación) es que tiene un potencial reutilizador ALTO, los valores de porcentaje: PESIMISTA, MODERADO y OPTIMISTA podrían ser:

	PESIMISTA	MODERADO	OPTIMISTA
UNIDAD ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y departamentos académicos, administrativos y de investigación)	85	90	100

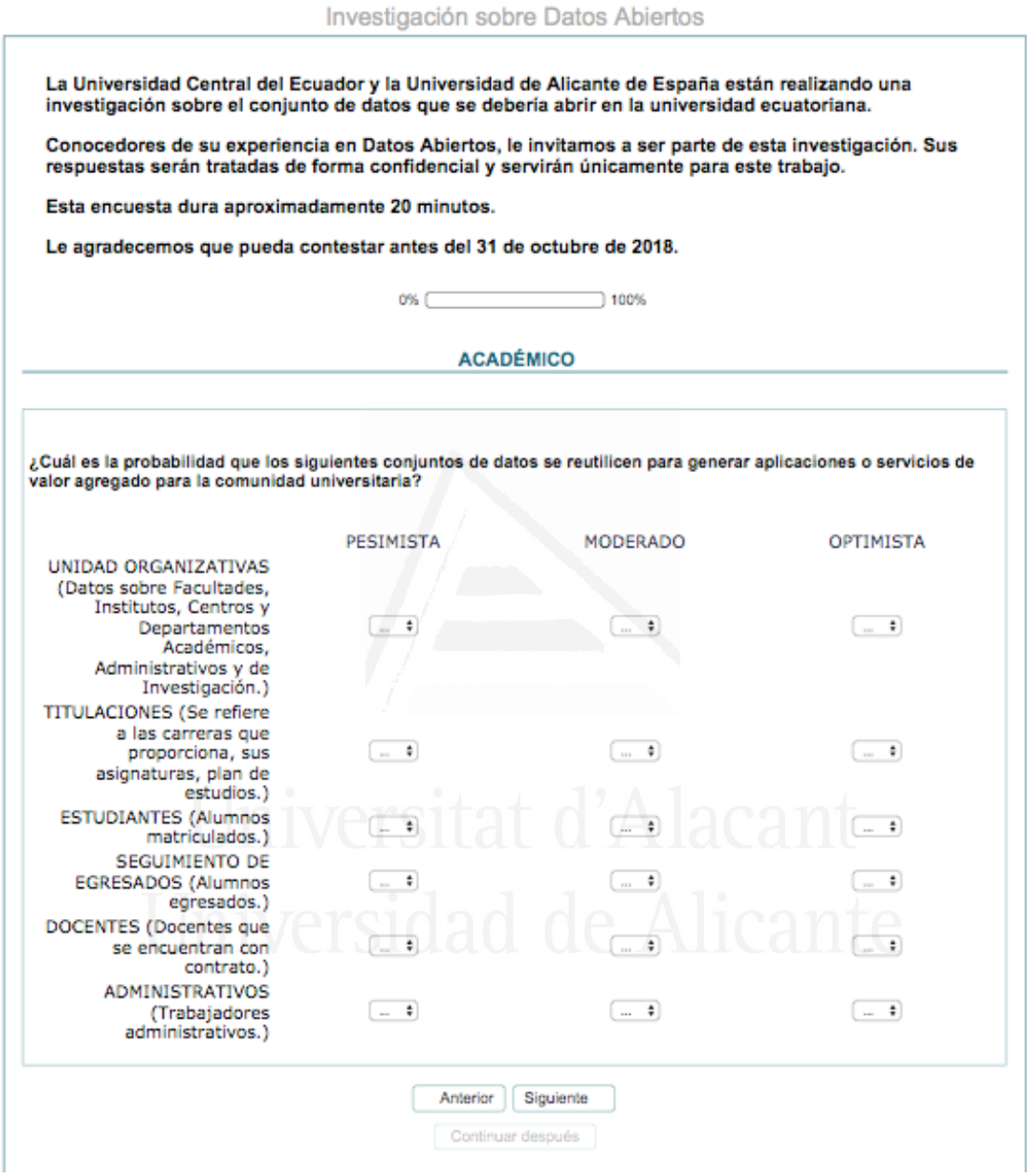
B) Si su criterio del conjunto de datos de UNIDADES ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y departamentos académicos, administrativos y de investigación) es que tiene un potencial reutilizador BAJO, los valores de porcentaje: PESIMISTA, MODERADO y OPTIMISTA podrían ser:

	PESIMISTA	MODERADO	OPTIMISTA
UNIDAD ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y departamentos académicos, administrativos y de investigación)	10	15	25

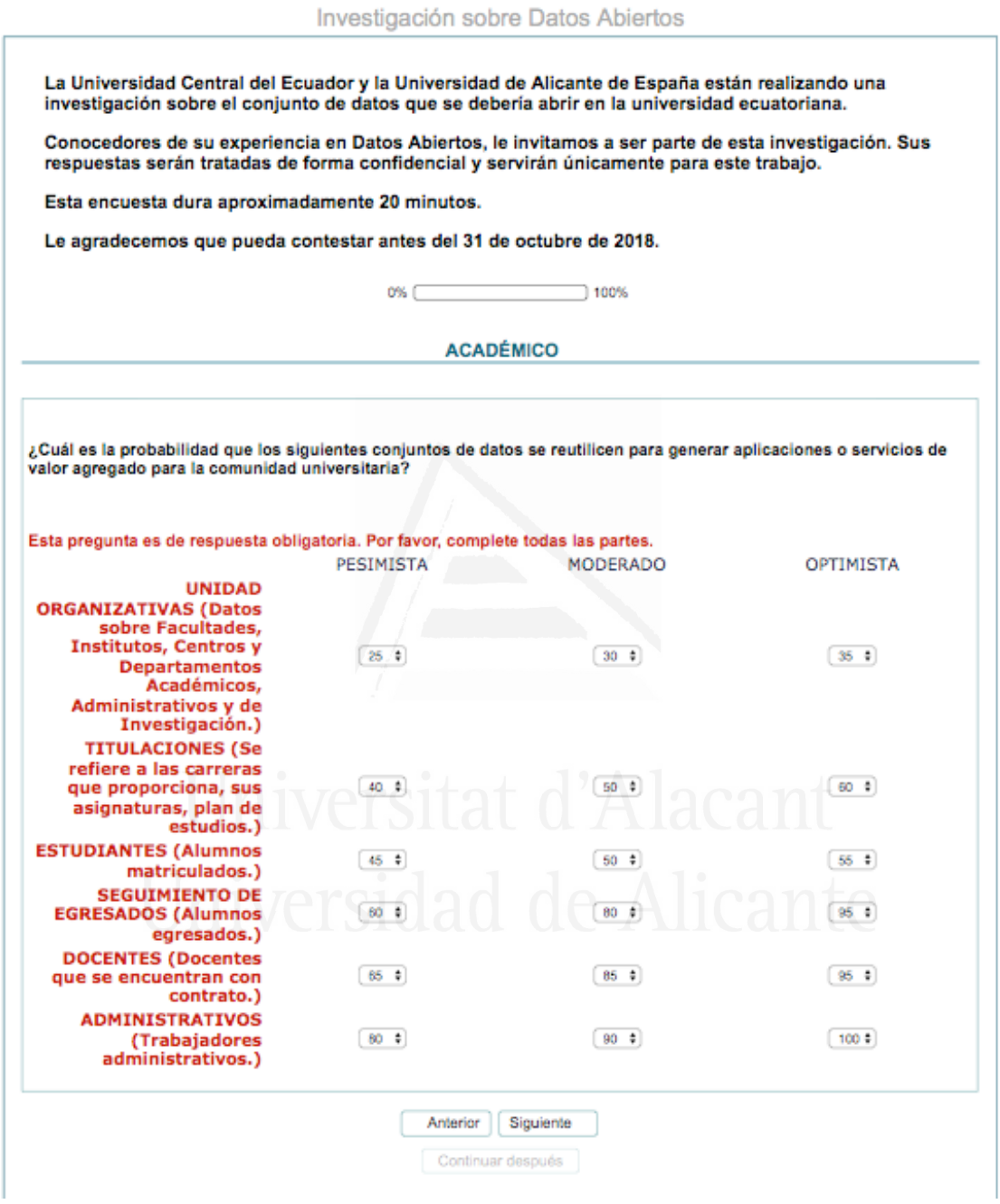
Siguiente

Cargar encuesta sin terminar

Pantalla con los primeros conjuntos de datos a seleccionar, casilleros para entrada de los números difusos de probabilidad: pesimista, moderada y optimista.



Pantalla con los primeros conjuntos de datos a seleccionar, con casilleros para entrada llenos



Investigación sobre Datos Abiertos

La Universidad Central del Ecuador y la Universidad de Alicante de España están realizando una investigación sobre el conjunto de datos que se debería abrir en la universidad ecuatoriana.

Conocedores de su experiencia en Datos Abiertos, le invitamos a ser parte de esta investigación. Sus respuestas serán tratadas de forma confidencial y servirán únicamente para este trabajo.

Esta encuesta dura aproximadamente 20 minutos.

Le agradecemos que pueda contestar antes del 31 de octubre de 2018.

0% 100%

ACADÉMICO

¿Cuál es la probabilidad que los siguientes conjuntos de datos se reutilicen para generar aplicaciones o servicios de valor agregado para la comunidad universitaria?

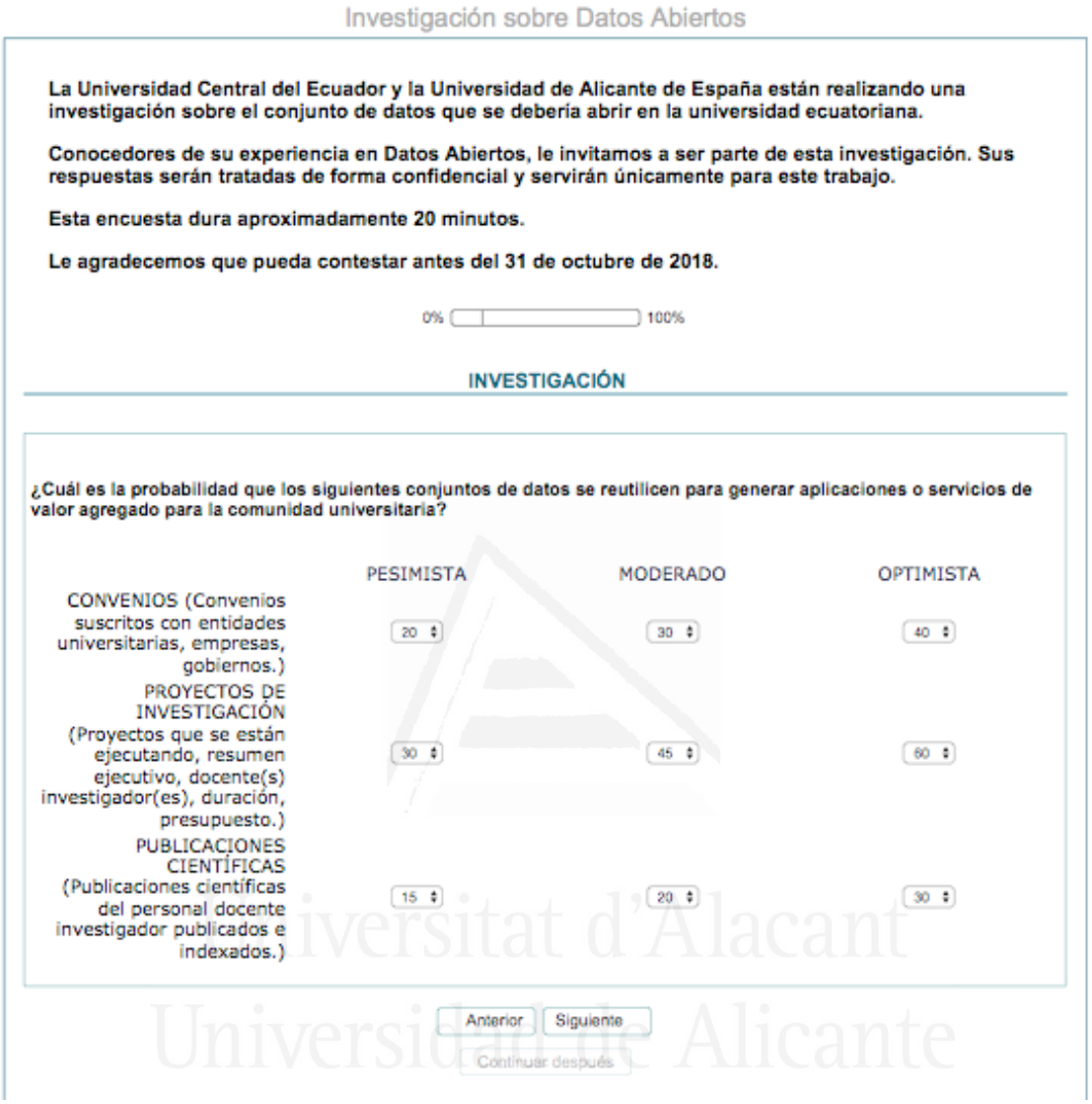
Esta pregunta es de respuesta obligatoria. Por favor, complete todas las partes.

	PESIMISTA	MODERADO	OPTIMISTA
UNIDAD ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y Departamentos Académicos, Administrativos y de Investigación.)	25	30	35
TITULACIONES (Se refiere a las carreras que proporciona, sus asignaturas, plan de estudios.)	40	50	60
ESTUDIANTES (Alumnos matriculados.)	45	50	55
SEGUIMIENTO DE EGRESADOS (Alumnos egresados.)	60	80	95
DOCENTES (Docentes que se encuentran con contrato.)	65	85	95
ADMINISTRATIVOS (Trabajadores administrativos.)	80	90	100

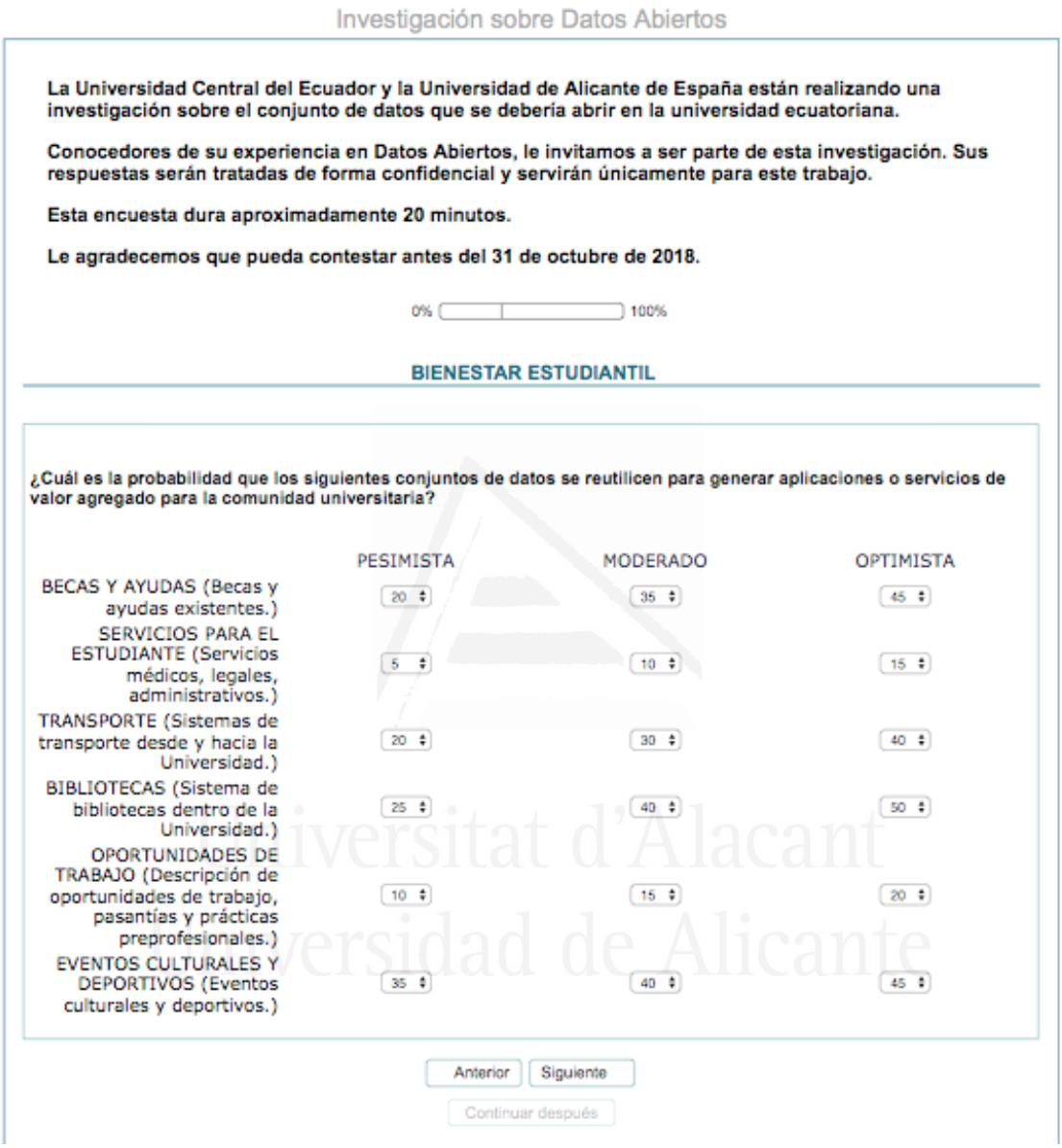
Anterior Siguiente

Continuar después

Pantalla con la segunda clasificación de conjuntos de datos a seleccionar.



Pantalla con la tercera clasificación de conjuntos de datos a seleccionar.



Pantalla con la cuarta clasificación de conjuntos de datos a seleccionar.

Investigación sobre Datos Abiertos

La Universidad Central del Ecuador y la Universidad de Alicante de España están realizando una investigación sobre el conjunto de datos que se debería abrir en la universidad ecuatoriana.

Conocedores de su experiencia en Datos Abiertos, le invitamos a ser parte de esta investigación. Sus respuestas serán tratadas de forma confidencial y servirán únicamente para este trabajo.

Esta encuesta dura aproximadamente 20 minutos.

Le agradecemos que pueda contestar antes del 31 de octubre de 2018.

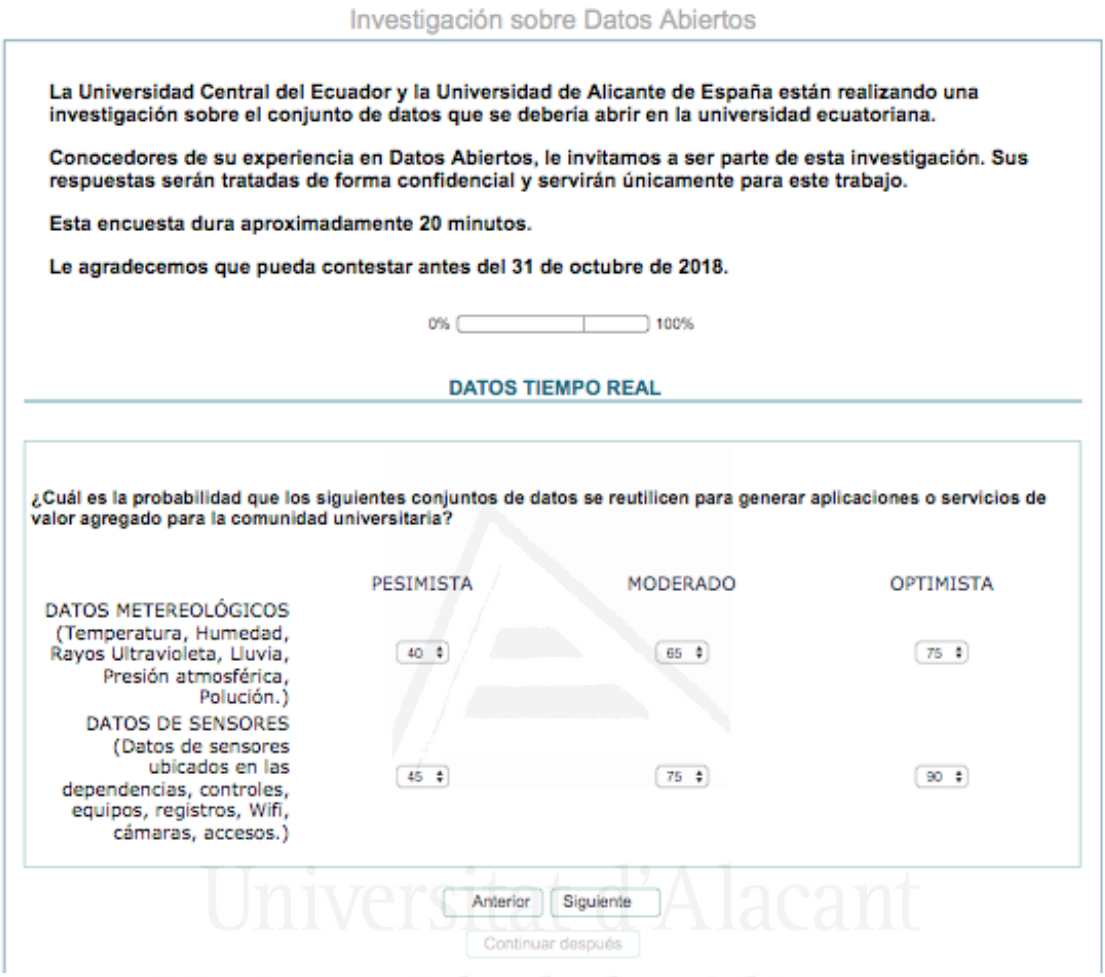
0% 100%

ORGANIZACIÓN E INFRAESTRUCTURA

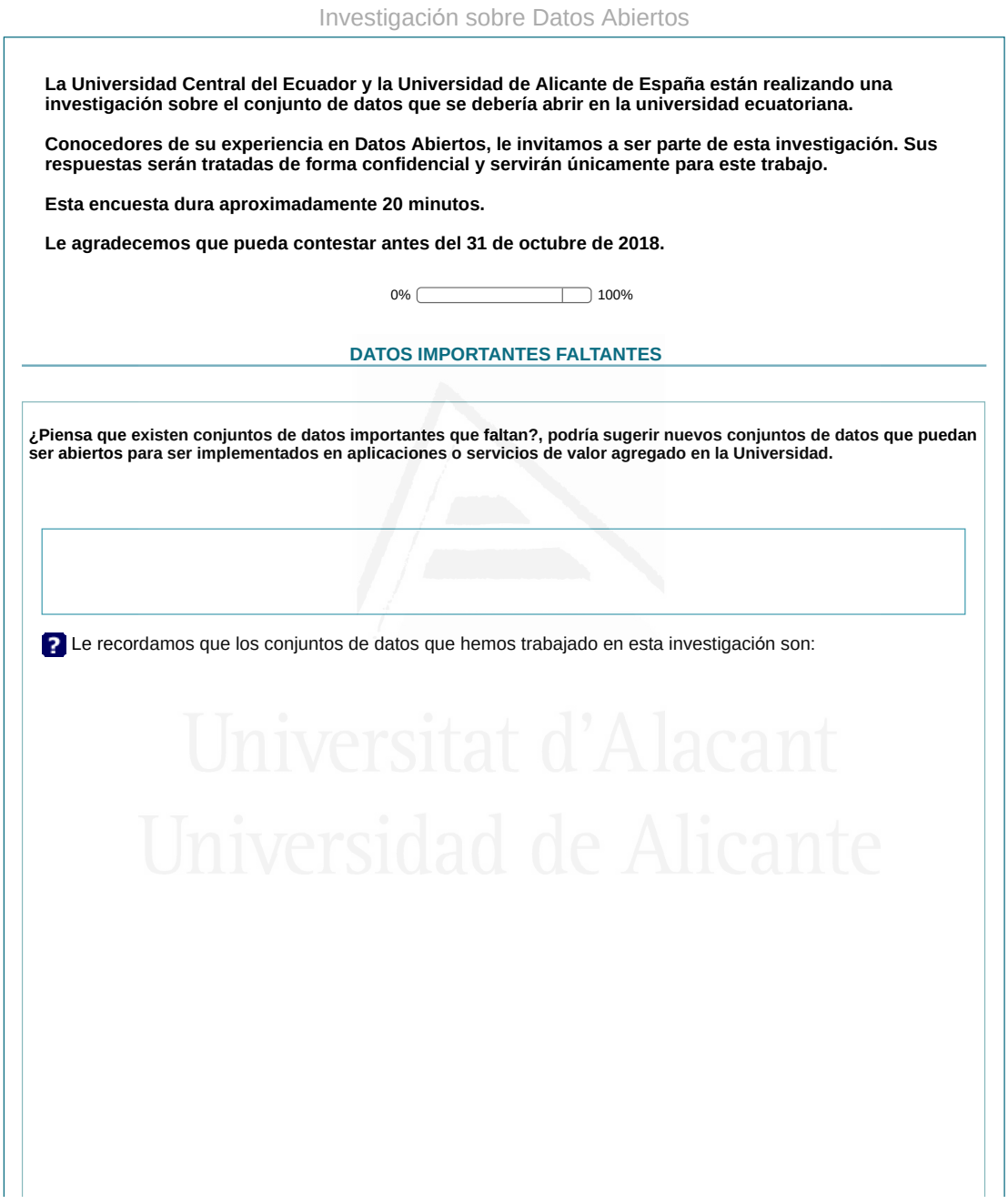
¿Cuál es la probabilidad que los siguientes conjuntos de datos se reutilicen para generar aplicaciones o servicios de valor agregado para la comunidad universitaria?

	PESIMISTA	MODERADO	OPTIMISTA
DATOS GEOPOSICIONADOS (Datos georeferenciales de todos los edificios, dependencias, equipos de comunicaciones, máquinas expendedoras, zonas de jardinería y otros.)	55	70	80
ORGANIGRAMA DE DIRECCIÓN (Datos de personal que dirige las unidades administrativas , académicas e investigación.)	30	40	50
RESTAURANTES Y PUESTOS DE COMIDA (Descripción de restaurantes, puestos de comida y bares.)	50	60	65
ENERGÍA ELÉCTRICA (Consumo de energía eléctrica tomado en forma física y digital.)	70	80	90
AGUA (Consumo de agua tomado en forma física y digital.)	40	50	60
TELEFONÍA (Consumo telefónico tomado en forma física y digital.)	45	50	55
COLECTORES DE BASURA (Sistema de recolección interna de basura, tachos de basura, horarios de recolección.)	60	65	70

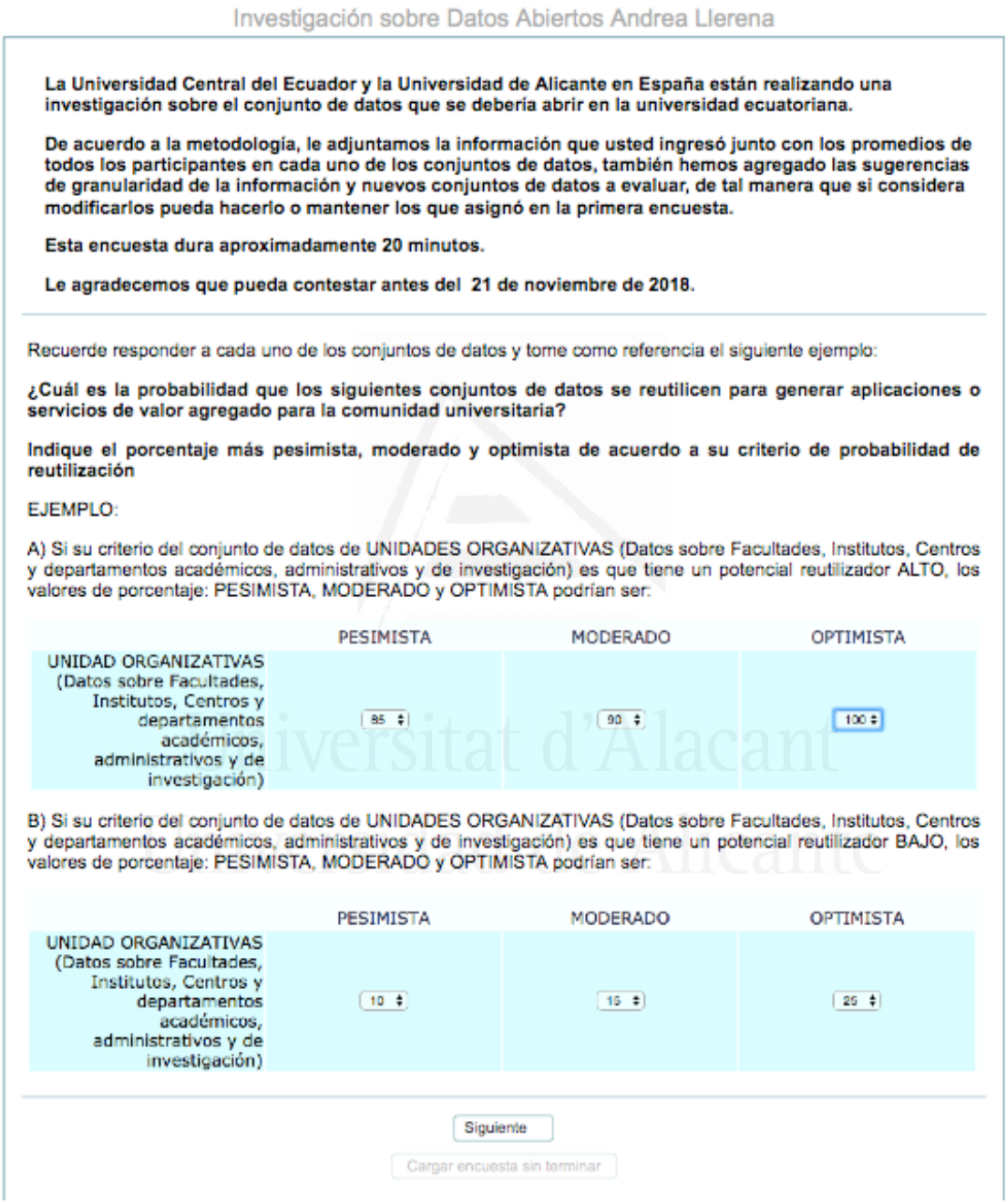
Pantalla con la quinta clasificación de conjuntos de datos a seleccionar.



Pantalla donde se presenta la segunda pregunta de investigación, es un espacio libre para aumentar conjuntos de datos que no se encontraban en la línea base.



Pantalla inicial con la segunda iteración detallada por experto, en el texto se puede observar que se comunica al experto de los resultados de la primera iteración.



Investigación sobre Datos Abiertos Andrea Llerena

La Universidad Central del Ecuador y la Universidad de Alicante en España están realizando una investigación sobre el conjunto de datos que se debería abrir en la universidad ecuatoriana.

De acuerdo a la metodología, le adjuntamos la información que usted ingresó junto con los promedios de todos los participantes en cada uno de los conjuntos de datos, también hemos agregado las sugerencias de granularidad de la información y nuevos conjuntos de datos a evaluar, de tal manera que si considera modificarlos pueda hacerlo o mantener los que asignó en la primera encuesta.

Esta encuesta dura aproximadamente 20 minutos.

Le agradecemos que pueda contestar antes del 21 de noviembre de 2018.

Recuerde responder a cada uno de los conjuntos de datos y tome como referencia el siguiente ejemplo:

¿Cuál es la probabilidad que los siguientes conjuntos de datos se reutilicen para generar aplicaciones o servicios de valor agregado para la comunidad universitaria?

Indique el porcentaje más pesimista, moderado y optimista de acuerdo a su criterio de probabilidad de reutilización

EJEMPLO:

A) Si su criterio del conjunto de datos de UNIDADES ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y departamentos académicos, administrativos y de investigación) es que tiene un potencial reutilizador ALTO, los valores de porcentaje: PESIMISTA, MODERADO y OPTIMISTA podrían ser:

	PESIMISTA	MODERADO	OPTIMISTA
UNIDAD ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y departamentos académicos, administrativos y de investigación)	85	90	100

B) Si su criterio del conjunto de datos de UNIDADES ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y departamentos académicos, administrativos y de investigación) es que tiene un potencial reutilizador BAJO, los valores de porcentaje: PESIMISTA, MODERADO y OPTIMISTA podrían ser:

	PESIMISTA	MODERADO	OPTIMISTA
UNIDAD ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y departamentos académicos, administrativos y de investigación)	10	15	25

Siguiente

Cargar encuesta sin terminar

Pantalla con la segunda iteración detallada por experto, en el texto se puede observar que se entregan los promedios de cada valor del número difuso que ingresaran, así como lo valores que ingresó en la primera iteración.

Investigación sobre Datos Abiertos Andrea Llerena

La Universidad Central del Ecuador y la Universidad de Alicante en España están realizando una investigación sobre el conjunto de datos que se debería abrir en la universidad ecuatoriana.

De acuerdo a la metodología, le adjuntamos la información que usted ingresó junto con los promedios de todos los participantes en cada uno de los conjuntos de datos, también hemos agregado las sugerencias de granularidad de la información y nuevos conjuntos de datos a evaluar, de tal manera que si considera modificarlos pueda hacerlo o mantener los que asignó en la primera encuesta.

Esta encuesta dura aproximadamente 20 minutos.

Le agradecemos que pueda contestar antes del 21 de noviembre de 2018.

0% 100%

ACADÉMICO

¿Cuál es la probabilidad que los siguientes conjuntos de datos se reutilicen para generar aplicaciones o servicios de valor agregado para la comunidad universitaria?

La media obtenida en cada uno de los siguientes conjuntos de datos fueron:

	MEDIA		
	PESIMISTA	MODERADO	OPTIMISTA
UNIDADES ORGANIZATIVAS	46	59	76
TITULACIONES	51	64	81
ESTUDIANTES	51	69	81
SEGUIMIENTO DE EGRESADOS	38	54	71
DOCENTES	41	51	64
ADMINISTRATIVOS	28	46	54

Sus respuestas fueron:

	SUS RESPUESTAS		
	PESIMISTA	MODERADO	OPTIMISTA
UNIDADES ORGANIZATIVAS	60	70	80
TITULACIONES	75	85	90
ESTUDIANTES	75	85	90
SEGUIMIENTO DE EGRESADOS	75	85	90
DOCENTES	60	70	80
ADMINISTRATIVOS	60	70	80

	PESIMISTA	MODERADO	OPTIMISTA
UNIDAD ORGANIZATIVAS (Datos sobre Facultades, Institutos, Centros y Departamentos Académicos, Administrativos y de Investigación, Procesos y servicios)	...	...	...

Bibliografía

[1] J. Gantz and D. Reinsel, "Big Data , Bigger Digital Shadows , and Biggest Growth in the Far East," *Idc*, vol. 2007, no. December 2012, pp. 1–16, 2012.

[2] G. Vickery, "Review of Recent Studies on Psi Re-Use and Related Market Developments Review of Recent Studies on Psi Re-Use and Related Market Developments 1 Executive Summary / Key Findings 3 Task Description 6."

[3] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, Adoption Barriers and Myths of Open Data and Open Government," *Inf. Syst. Manag.*, vol. 29, no. 4, pp. 258–268, 2012.

[4] S. Martin, M. Foulonneau, S. Turki, M. Ihadjadene, U. Paris, and P. Tudor, "Risk analysis to overcome barriers to open data," *Electron. J. e-Government*, vol. 11, no. 1, pp. 348–359, 2013.

[5] A. Ramirez-Alujas, "Open Government and Modernization of Public Management: Current Trends and the (inevitable) way forward. Seminal Reflections," *Rev. Enfoques*, vol. Vol. IX, no. 15, pp. 99–115, 2011.

[6] J. J. Zubcoff *et al.*, "The university as an open data ecosystem," *Int. J. Des. Nat. Ecodynamics*, vol. 11, no. 3, pp. 250–257, 2016.

[7] B. Obama, "Open government directive," *White House*, 2009. .

[8] I. Susha, A. Zuiderwijk, M. Janssen, and Å. Grönlund, "Benchmarks for Evaluating the Progress of Open Data Adoption: Usage, Limitations, and Lessons Learned," *Soc. Sci. Comput. Rev.*, vol. 33, no. 5, pp. 613–630, 2015.

[9] J. Attard, F. Orlandi, S. Scerri, and Soren Auer, "A systematic review of open government data initiatives," *Gov. Inf. Q.*, vol. 32, no. 4, pp. 399–418, 2015.

[10] M. Lee, E. Almirall, and J. Wareham, "Open data and civic apps," *Commun. ACM*, vol. 59, no. 1, pp. 82–89, 2015.

[11] J. C. Bertot, P. T. Jaeger, and J. M. Grimes, "Using ICTs to create a culture of

transparency: E-government and social media as openness and anti-corruption tools for societies," *Gov. Inf. Q.*, vol. 27, no. 3, pp. 264–271, 2010.

[12] S. A. Chun, S. Shulman, R. Sandoval, and E. Hovy, "Government 2.0: Making connections between citizens, data and government," *Inf. Polity*, vol. 15, no. 1–2, pp. 1–9, 2010.

[13] A. Zuiderwijk, M. Janssen, S. Choenni, R. Meijer, R. Sheikh_Alibaks, and R. S. Alibaks, "Socio-technical impediments of open data," *Electron. J. eGovernment*, vol. 10, no. 2, pp. 156–172, 2012.

[14] J. J. Aparicio, Juan Manuel; Fuster, Andrés; Garrigós, Irene; Maciá, Francisco; Mazon, Jose Norberto;VAquer Llorenç; Zubcoff, *ECOSISTEMA DE DATOS ABIERTOS De la Universidad de Alicante*. Alicante.

[15] G. Concha and A. Naser, "Panorama de gobierno Electrónico en la región: resultados e impactos," *El desafío hacia el Gob. abierto en la hora la Igual.*, p. 134, 2012.

[16] A. Naser and D. Rosales, "Panorama regional," pp. 23–67, 2016.

[17] Alon Peled, "When Transparency and Collaboration Collide: The USA Open Data Program," *J. Am. Soc. Inf. Sci. Technol.*, vol. 11, no. 62, pp. 2085–2094, 2011.

[18] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits , Adoption Barriers and Myths of Open Data and Open Government Benefits , Adoption Barriers and Myths of Open Data and Open Government," vol. 0530, 2012.

[19] P. Conradie and S. Choenni, "On the barriers for local government releasing open data," *Gov. Inf. Q.*, vol. 31, pp. S10–S17, 2014.

[20] B. Kitchenham *et al.*, "Systematic literature reviews in software engineering-A tertiary study," *Inf. Softw. Technol.*, vol. 52, no. 8, pp. 792–805, 2010.

[21] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic Mapping Studies in Software Engineering," *12Th Int. Conf. Eval. Assess. Softw. Eng.*, vol. 17, p. 10, 2008.

[22] S. Casteleyn, I. Garrigós, and J.-N. Mazón, "Ten years of Rich Internet applications: A

systematic mapping study, and beyond," *ACM Trans. Web*, vol. 8, no. 3, 2014.

[23] S. Barney, K. Petersen, M. Svahnberg, A. Aurum, and H. Barney, "Software quality trade-offs: A systematic map," *Inf. Softw. Technol.*, vol. 54, no. 7, pp. 651–662, 2012.

[24] E. R. P. M. Engström, "A systematic review on regression test selection techniques," *ACM Int. Conf. Proceeding Ser.*, vol. 53, no. 1, pp. 14–40, 2010.

[25] D. Garson, *The Delphi Method in Quantitative Research*, 2014th ed. Asheboro, NC 27205 USA: Publishing, Statistical Associates, 2014.

[26] H. Danladi, M. Rusli, and A. Makmom, "Delphi method of developing environmental well-being indicators for the evaluation of urban sustainability in Malaysia," *Procedia Environ. Sci.*, vol. 30, pp. 244–249, 2015.

[27] C. Wu and W. Fang, "Combining the Fuzzy Analytic Hierarchy Process and the fuzzy Delphi method for developing critical," pp. 751–768, 2011.

[28] M. J. dos Santos and E. de Mello Fagotto, "Cloud Computing Management Using Fuzzy Logic," *IEEE Lat. Am. Trans.*, vol. 13, no. 10, pp. 3392–3397, 2015.

[29] M. Mendonca, I. Rossato Chrun, M. Antonio Ferreira Finocchio, and E. Eire De Mello, "Fuzzy cognitive maps applied to student satisfaction level in an university," *IEEE Lat. Am. Trans.*, vol. 13, no. 12, pp. 3922–3927, 2015.

[30] C. Lin and L. Z. Chuang, "Using Fuzzy Delphi Method and Fuzzy AHP for Evaluation Structure of the Appeal of Taiwan ' s Coastal Wetlands Ecotourism," pp. 347–358, 2012.

[31] W. Liu, "Application of the Fuzzy Delphi Method and the Fuzzy Analytic Hierarchy Process for the Managerial Competence of Multinational Corporation Executives," vol. 3, no. 4, pp. 313–317, 2013.

[32] P. Chang, C. Hsu, and P. Chang, "Fuzzy Delphi method for evaluating hydrogen production technologies," *Int. J. Hydrogen Energy*, vol. 36, no. 21, pp. 14172–14179, 2016.

[33] Y. L. Hsu, C. H. Lee, and V. B. Kreng, "The application of Fuzzy Delphi Method and Fuzzy AHP in lubricant regenerative technology selection," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 419–425, 2010.

[34] S. A. Publishing, *2014 Edition Single User License . Do not copy or post . 2014 Edition ISBN : 978-1-62638-018-9 Single User License . Do not copy or post .* 2014.

[35] Y. and others George, J Klir and Bo, *Fuzzy sets and fuzzy logic: Theory and applications*. 1995.

[36] S. Hsueh, "Assessing the effectiveness of community-promoted environmental protection policy by using a Delphi-fuzzy method : A case study on solar power and plain afforestation in Taiwan," *Renew. Sustain. Energy Rev.*, vol. 49, pp. 1286–1295, 2015.

[37] Y. Wang, G. Yeo, and A. K. Y. Ng, "Choosing optimal bunkering ports for liner shipping companies : A hybrid Fuzzy-Delphi – TOPSIS approach," *Transp. Policy*, vol. 35, pp. 358–365, 2014.

[38] W.-K. Liu, "Application of the Fuzzy Delphi Method and the Fuzzy Analytic Hierarchy Process for the Managerial Competence of Multinational Corporation Executives," *Int. J. e-Education, e-Business, e-Management e-Learning*, vol. 3, no. 4, pp. 313–317, 2013.

[39] F. Herrera and L. Martónez, "A 2-Tuple Fuzzy Linguistic Representation Model for Computing with Words," *IEEE Trans. Fuzzy Syst.*, vol. 8, no. 6, pp. 746–752, 2000.

[40] L. O. Seman, G. Gomes, R. Hausmann, and E. A. Bezerra, "A quadratic fuzzy regression approach for handling uncertainties in Partial Least Squares Path Modeling," *IEEE Lat. Am. Trans.*, vol. 16, no. 1, pp. 192–201, 2018.

[41] https://opendatacharter.net, "G8 Open Data Charter," *G8 Lough Erne 2013*, no. June, pp. 1–10, 2013.

[42] J. Hagel III, "'The Coming Battle for Customer Information'.," *Harvard Business Review.*, 1997.

[43] Open Data Charter, "International Open Data Charter," no. September, p. 8, 2015.

[44] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: A proposal and a discussion," *Requir. Eng.*, vol. 11, no. 1, pp. 102–107, 2006.

[45] A. Latif, A. Scherp, and K. Tochtermann, "LOD for Library Science: Benefits of Applying Linked Open Data in the Digital Library Setting," *KI - Künstliche Intelligenz*, vol. 30, no. 2, pp. 149–157, 2015.

[46] A. Ramos-Soto, A. Bugarín, S. Barro, and F. Díaz-Hermida, "Automatic linguistic descriptions of meteorological data," *Proc. Cist.*, pp. 1–6, 2013.

[47] S. Chakraborty, M. H. H. Rahman, and M. H. Seddiqui, "Linked open data representation of historical heritage of Bangladesh," *16th Int'l Conf. Comput. Inf. Technol. ICCIT 2013*, no. March, pp. 242–248, 2014.

[48] E. Piedra, N.; Chicaiza, J.; Lopez, J.; Tovar Caro, "Towards a Learning Analytics Approach for Supporting discovery and reuse of OER," no. March, pp. 978–988, 2015.

[49] C. Millette and P. Hosein, "A consumer focused open data platform," *2016 3rd MEC Int. Conf. Big Data Smart City, ICBDSC 2016*, pp. 101–106, 2016.

[50] J. N. Rouder, "The what, why, and how of born-open data," *Behav. Res. Methods*, vol. 48, no. 3, pp. 1062–1069, 2016.

[51] a O. Erkimbaev, V. Y. Zitserman, G. a Kobzev, V. a Serebrjakov, and K. B. Teymurazov, "Publishing scientific data as linked open data," *Sci. Tech. Inf. Process.*, vol. 40, no. 4, pp. 253–263, 2013.

[52] A. Callahan, J. Cruz-Toledo, and M. Dumontier, "Ontology-Based Querying with Bio2RDF's Linked Open Data.," *J. Biomed. Semantics*, vol. 4 Suppl 1, no. Suppl 1, p. S1, 2013.

[53] S. O'Riain, E. Curry, and A. Harth, "XBRL and open data for global financial ecosystems: A linked data approach," *Int. J. Account. Inf. Syst.*, vol. 13, no. 2, pp. 141–162, 2012.

[54] D. S. Sayogo and T. A. Pardo, "Exploring the motive for data publication in open data initiative: Linking intention to action," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, no. 2011, pp. 2623–2632, 2011.

[55] P. Ciancarini, F. Poggi, and D. Russo, "Big Data Quality: A Roadmap for Open Data," *2016 IEEE Second Int. Conf. Big Data Comput. Serv. Appl.*, pp. 210–215, 2016.

[56] P. Doshi *et al.*, "Open data 5 years on: a case series of 12 freedom of information requests for regulatory data to the European Medicines Agency," *Trials*, vol. 17, no. 1, p. 78, 2016.

[57] S. Oyama, Y. Baba, I. Ohmukai, H. Dokoshi, and H. Kashima, "Crowdsourcing chart digitizer: task design and quality control for making legacy open data machine-readable," *Int. J. Data Sci. Anal.*, 2016.

[58] N. B. Hounsell, B. P. Shrestha, M. McDonald, and A. Wong, "Open Data and the Needs of Older People for Public Transport Information," *Transp. Res. Procedia*, vol. 14, pp. 4334–4343, 2016.

[59] A. Lausch, A. Schmidt, and L. Tischendorf, "Data mining and linked open data - New perspectives for data analysis in environmental research," *Ecol. Modell.*, vol. 295, pp. 5–17, 2015.

[60] H. Demski, S. Garde, and C. Hildebrand, "Open data models for smart health interconnected applications: the example of openEHR," *BMC Med. Inform. Decis. Mak.*, vol. 16, no. 1, p. 137, 2016.

[61] M. Kassen, "A promising phenomenon of open data: A case study of the Chicago open data project," *Gov. Inf. Q.*, vol. 30, no. 4, pp. 508–513, 2013.

[62] W. Brunette *et al.*, "Open data kit sensors: a sensor integration framework for android at the application-level," *Proc. 10th Int. Conf. Mob. Syst. Appl. Serv. - MobiSys '12*, p. 351, 2012.

[63] T. Silva, V. Wuwongse, and H. N. Sharma, "Disaster mitigation and preparedness using linked open data," *J. Ambient Intell. Humaniz. Comput.*, vol. 4, no. 5, pp. 591–602, 2013.

[64] F. G. De Andrade and R. José, "Semantic Annotation of Geodata Based on Linked-Open Data," vol. 2, pp. 9–16, 2015.

[65] Y.-A. Lai, Y.-Z. Ou, J. Su, S.-H. Tsai, C.-W. Yu, and D. Cheng, "Virtual disaster management information repository and applications based on linked open data," *2012 Fifth IEEE Int. Conf. Serv. Comput. Appl.*, pp. 1–5, 2012.

[66] N. Kobayashi and T. Toyoda, "BioSPARQL: Ontology-based smart building of SPARQL queries for biological Linked Open Data," *ACM Int. Conf. Proceeding Ser.*, no. 1, pp. 47–49, 2012.

[67] P. Colpaert, J. Sarah, P. Mechant, E. Mannens, and R. Van de Walle, "The 5 stars of open data portals," *7th Int. Conf. Methodol. Technol. tools enabling e-Government*, 2013.

[68] X. Masip-Bruin, G.-J. Ren, R. Serral-Gracia, and M. Yannuzzi, "Unlocking the Value of Open Data with a Process-Based Information Platform," *IEEE 15th Conf. Bus. Informatics*, 2013.

[69] B. Isaac, Antoine |Haslhofer, "Europeana Linked Open Data – data.europeana.eu," *Semant. Web*, vol. 4, no. 3, pp. 291–297, 2013.

[70] E. Rozell, J. Erickson, and J. Hendler, "From international open government dataset search to discovery: a semantic web service approach," *ICEGOV '12 Proc. 6th Int. Conf. Theory Pract. Electron. Gov.*, pp. 480–481, 2012.

[71] D. Maier, V. M. Megler, and K. Tufte, "Challenges for Dataset Search," in *Database Systems for Advanced Applications*, 2014, pp. 1–15.

[72] L. M. Koesten, E. Kacprzak, J. F. A. Tennison, and E. Simperl, "The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 1277–1289.

[73] E. Kacprzak, L. Koesten, L. D. Ibáñez, T. Blount, J. Tennison, and E. Simperl, "Characterising dataset search—An analysis of search logs and data requests," *J. Web*

Semant., no. xxxx, 2018.

[74] M. de Rijke, "Learning to Search for Datasets," in *Companion Proceedings of the The Web Conference 2018*, 2018, p. 1483.

[75] L. Mlynárová, J. P. Nap, and T. Bisseling, "The SWI/SNF chromatin-remodeling gene AtCHR12 mediates temporary growth arrest in Arabidopsis thaliana upon perceiving environmental stress," *Plant J.*, vol. 51, no. 5, pp. 874–885, 2007.

[76] T. Jetzek, M. Avital, and N. Bjorn-Andersen, "Data-driven innovation through open government data," *J. Theor. Appl. Electron. Commer. Res.*, vol. 9, no. 2, pp. 100–120, 2014.

[77] E. Lakomaa and J. Kallberg, "Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs," *IEEE Access*, vol. 1, pp. 558–563, 2013.

[78] G. Magalhaes, C. Roseira, and L. Manley, "Business Models for Open Government Data," in *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, 2014, pp. 365–370.

[79] A. Zuiderwijk, M. Janssen, K. Poulis, and G. van de Kaa, "Open Data for Competitive Advantage: Insights from Open Data Use by Companies," in *Proceedings of the 16th Annual International Conference on Digital Government Research*, 2015, pp. 79–88.

[80] IDC, "Impact Assessment of Odine Programme," 2017.

[81] A. E. Prieto, J. N. Mazón, A. Lozano-Tello, and L. D. Ibáñez, "Supporting open dataset publication decisions based on Open Source Software reuse," *CEUR Workshop Proc.*, vol. 2062, 2018.