UNIVERSIDAD DE VALENCIA

FACULTAD DE FILOSOFÍA Y CIENCIAS DE LA EDUCACIÓN

Departamento de Educación Comparada e Historia de la Educación

PROGRAMA DE DOCTORADO EN EDUCACIÓN

LA INVESTIGACIÓN BIBLIOMÉTRICA EN HISTORIA DE LA EDUCACIÓN. SITUACIÓN ACTUAL, DESARROLLO DE BASES DE DATOS ESPECÍFICAS Y PROPUESTAS DESDE LA INTELIGENCIA ARTIFICIAL

TESIS DOCTORAL

Presentada por:

Jacobo Roda Segarra

Dirigida por:

José Luis Hernández Huerta

Santiago Mengual Andrés

Andrés Payà Rico

Valencia

Noviembre de 2022

Agradecimientos

Todo cambió cuando recibí, hace ya unos cuantos años, un email de un tal Santiago Mengual, cuyo nombre me sonaba vagamente por sus críticas tenaces a mi Trabajo Fin de Máster sobre *Actividades procedurales y aprendizaje guiado por software*, del que había sido parte del tribunal. En este caso, su mensaje era de un tono radicalmente distinto, dándome la enhorabuena porque mi extraño trabajo había llamado su atención, e invitándome a continuar con el doctorado bajo su dirección. Las implicaciones que, años después, tuvo aquello en mi carrera profesional y en mi vida personal y familiar fueron tan grandes, que solo pensar en ello me da vértigo. Nunca podré agradecerte suficientemente tu inestimable ayuda todos estos años, y por permitirme acompañarte en tus apasionantes proyectos.

José Luis Hernández y Andrés Payà, para los cuales era un perfecto desconocido, confiaron ciegamente en mí desde el primer momento en el que me uní al ambicioso proyecto que tenían entre manos. Las expectativas tan altas que tenían en mí me abrumaron y aterrorizaron a partes iguales. Espero haber estado a la altura de las circunstancias, y seguir estándolo con todos aquellos retos que me seguís proponiendo día a día. Gracias por vuestros sabios consejos, por vuestra ayuda y por todas esas agradables charlas que hemos tenido, y que seguiremos teniendo.

Como tampoco me cansaré de agradecer a Ana, Álex y Emma, cada uno a su manera, el estoicismo con el que soportan las implicaciones que tienen en sus vidas todos y cada uno de los proyectos en los que me he metido (y me seguiré metiendo, lo siento ya por adelantado). Gracias, Ana, por apoyar ciegamente este cambio de rumbo a pesar de saber que lo iba a poner todo del revés.

Por último, reservo un agradecimiento muy especial para mis padres, Vicente y Elena. Espero que con este trabajo, y con sus implicaciones futuras, se puedan quitar una espinita que han tenido clavada durante más de veinte años. Quizá me lo he tomado con más paciencia de lo que os hubiera gustado, pero estad tranquilos porque he dado un largo e interesante paseo por la vida antes de volver a la casilla de salida.

Índice

Índice de tablas

Índice de figuras

Índice de siglas

1FN - 1ª Forma Normal de Boyce-Codd

2FN - 2ª Forma Normal de Boyce-Codd

3FN - 3ª Forma Normal de Boyce-Codd

DBR - Design-Based Research

DT - Decision Trees

EAV - Entidad-Atributo-Valor

ER - Entidad-Relación

FN - Falsos negativos

FP - Falsos positivos

HE - Historia de la Educación

IA - Inteligencia Artificial

LR - Logistic Regression

MLP - Multilayer Perceptron

MVC - Modelo-Vista-Controlador

NB - Naïve Bayes

ORM - Object Relational Mapping

PHP - Pre-Procesador de Hipertexto

PMV - Producto Mínimo Viable

RF - Random Forest

SQL - Structured Query Language

TIC - Tecnologías de la Información y la Comunicación

TN - Negativos verdaderos

TP - Positivos verdaderos

W3C - World Wide Web Consortium

Capítulo 1: Introducción

1.1 Introducción

—He dicho, y lo repito, que Tránor quedará convertido en ruinas dentro de cinco siglos.

—¿No considera que su declaración es desleal?

—No, señor. La verdad científica está más allá de toda lealtad y deslealtad.

—¿Está seguro de que su declaración representa la verdad científica?

—Lo estoy.

—¿En qué se basa?

—En las matemáticas de la psicohistoria.

—¿Puede demostrar que estas matemáticas son válidas?

—Solo a otro matemático.

Isaac Asimov, *Fundación*

La presente tesis dista mucho de tratar de predecir ningún debacle, tal como auguró Asimov respecto a la ficticia caída de Tránor con sus métodos psicohistóricos en la cita con la que abre esta introducción. Más bien todo lo contrario ya que, tal como se mostrará a lo largo de la investigación, la historia de la educación (en adelante, HE) goza de una envidiable salud, con un crecimiento enorme en cuanto a la producción científica durante los últimos años. Sin embargo, el desarrollo del concepto de la psicohistoria en la ficción de Asimov encierra una idea atractiva para cualquier investigador. Si, tal como la definió en su *Fundación*, la psicohistoria era la "rama de las matemáticas que trata sobre las reacciones de conglomeraciones humanas ante determinados estímulos sociales y económicos" (Asimov, 2006, p. 27), ¿podrían las matemáticas en la realidad ser capaces de analizar, sintetizar y predecir la complejidad de las relaciones humanas?

La presente tesis no trata de responder a dicha pregunta, que se aleja del campo de estudio que nos ocupa, y entra de lleno en la estadística y en las modernas técnicas de inteligencia artificial (IA) de las que, no obstante, sí que se hablará a lo largo de la tesis. Sin embargo, sí que podemos adaptar y concretar la cuestión al campo de la investigación en HE. La pregunta, en este caso, se podría describir de la siguiente forma: ¿se pueden analizar cuantitativamente las investigaciones en HE? La respuesta inmediata es afirmativa, y se desprende de la cienciometría y, en concreto, de sus técnicas bibliométricas, que han dedicado las últimas décadas a los análisis cuantitativos de la producción científica con diversos fines, algunos de ellos no exentos de críticas. Sin embargo, cuando nos centramos en la HE específicamente, descubrimos que las investigaciones encierran complejidades que pasan desapercibidas para los actuales estudios cienciométricos. Una riqueza propia de la disciplina que, en muchos casos, desaparece frente a los reduccionismos cuantitativos propios de la cienciometría.

Así pues, la cuestión se podría afinar más: ¿se puede analizar cuantitativamente la complejidad contenida en las investigaciones en HE? En este caso, la respuesta no es tan obvia, al menos con las técnicas con las que se realizan en la actualidad este tipo de investigaciones. Explorar este camino va a exigir, tal como se mostrará, una serie de etapas intermedias, que en algunos casos, deberán acercarse a las ciencias de la computación, para poder desarrollar soluciones tecnológicas que permitan profundizar en la problemática del análisis de la investigación en HE desde las particularidades de las mismas.

Respecto a la pertinencia de analizar la disciplina de la HE desde una perspectiva cuantitativa teniendo en cuenta sus particularidades, la presente tesis, desarrollada en el marco de la ayuda predoctoral PRE2020-093276 (financiada por MCIN/AEI/10.13039/501100011033 y por FSE invierte en tu futuro), es una pieza más del proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global" (referencia de ayuda PID2019-105328GB-I00) que, entre otros, persigue los siguientes objetivos generales dentro de los cuales se enmarca el presente trabajo de investigación:

- OG1: Cartografiar globalmente los espacios de socialización, las redes de comunicación y la producción científica internacional en HE.
- OG2: Estudiar la producción y evolución historiográfica, durante los últimos veinticinco años, de temas histórico-educativos clave que puedan proporcionar elementos de reflexión y análisis para los retos educativos actuales.
- OG4: Desarrollar herramientas para la investigación en red adaptadas a las exigencias y necesidades de la comunidad científica global de historiadores de la educación y posibilidades ofrecidas por las tecnologías de la información y la comunicación.

Cada una de las etapas que se ha seguido durante el desarrollo de esta tesis ha implicado unas necesidades diferentes, lo que ha exigido una flexibilidad metodológica en función de los objetivos planteados en cada una de las fases. Cabe destacar que en la presente tesis se lleva a cabo una revisión narrativa, el desarrollo específico de herramientas *software* para el estudio de la HE, propuestas desde la IA para el análisis de datos de HE y un estudio bibliométrico. A pesar de esta diversidad, cada una de estas etapas sigue un hilo conductor preciso que responde a los objetivos del proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global " anteriormente expuesto. De esta forma, el desarrollo de herramientas *software* específicas responde al OG4, la investigación en IA es una propuesta de automatización para el OG1 y el estudio bibliométrico forma parte del OG2. Todo ello enmarcado en la revisión narrativa previa, que reflexiona sobre la cienciometría y, en concreto, sobre los estudios bibliométricos, así como sus dificultades de aplicación a la HE.

Así pues, cada una de estas etapas se articula en un capítulo diferente, que son abordados con diferentes metodologías adecuadas a los objetivos planteados. Esto, lejos de restarle un ápice de rigor a la investigación, permite afrontar cada una de las etapas intermedias con las herramientas metodológicas más adecuadas y adaptadas a la diversidad de objetivos planteados en la tesis.

1.2 Estructura del trabajo

La presente tesis recurre a la complementariedad metodológica para dar respuesta a cada una de las necesidades los problemas de investigación planteados. Así pues, se han empleado diferentes estrategias y procesos para abordar el problema objeto de estudio. Blanco y Pirela (2016) afirman que este planteamiento, lejos de suponer una confrontación entre técnicas, aporta una alternativa integradora en la producción de conocimiento. En esta línea continúa Martínez (2005) al afirmar que usar diferentes enfoques permite integrar en un todo coherente el objeto de estudio, superando la fragmentación a la que tiende el método científico.

La primera investigación realizada, que cubre la totalidad del capítulo 2 y que se titula "Principios cienciométricos como soporte a la investigación en historia de la educación", parte de una revisión narrativa, de corte cualitativo, para analizar la evolución de la cienciometría, los diferentes enfoques epistemológicos y las críticas que ha recibido (apartado 2.2). Desde la aproximación a la cienciometría se expondrá una de sus técnicas, los estudios bibliométricos, su fundamentación y limitaciones cuando se trata de analizar datos más allá de los que nos ofrecen los registros bibliográficos (apartado 2.3). Relacionado con esta cuestión, y estableciendo el núcleo de la tesis, se destacará la información específica de HE y su extracción (apartado 2.4), lo que nos obligará a detenernos en la biblioteconomía y la hermenéutica, así como diversos planteamientos para enriquecer esta información bibliográfica con datos obtenidos del contenido de las investigaciones (apartado 2.5).

A continuación, basándonos en las necesidades y limitaciones detectadas en el capítulo previo y respondiendo al OG4 del proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global", el capítulo 3, "Desarrollo de una base de datos específica para historia de la educación", describe la implementación de una herramienta propia para HE, denominada Hecumen, desde la metodología del Design-Based Research (DBR), cuyos principios persiguen la mejora o innovación a través de un esfuerzo colaborativo entre investigadores y profesionales a través de una serie de ciclos recursivos (Design-based Research Collective, 2003). En

estas iteraciones se propondrá un diseño (apartado 3.2.2) que será puesto a prueba (*tested*) por investigadores y, tras una reflexión (apartado 3.2.3) y un refinamiento (apartado 3.2.4), se iniciará una nueva iteración en el ciclo hasta llegar al objetivo final, que no es otro que la elaboración de unos principios de diseño que puedan ser un punto de partida para futuras investigaciones (apartado 3.2.5).

El capítulo 4, "Inteligencia artificial para automatizar la clasificación de artículos en historia de la educación", aborda la problemática a nivel de recursos que supone la extracción de temáticas y épocas estudiadas en cada uno de las investigaciones de HE publicados, proponiendo una solución desde las ciencias de la computación en su especialidad de la IA para la clasificación y catalogación automática de textos, lo que da respuesta al OG1 del proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global". Para tal efecto se realiza una conceptualización de la IA, así como de los algoritmos utilizados (apartados 4.1 y 4.2), se propone un diseño de clasificador para los textos de HE (apartado 4.3) y se implementa, realizando las modificaciones pertinentes en la base de datos Hecumen desarrollada en el capítulo 3 (apartado 4.4). Los resultados del sistema se describen en el apartado 4.5.

Una vez desarrolladas y fundamentadas estas herramientas, el capítulo 5, " La historia de la educación a través de las revistas especializadas. Temáticas, producción científica y bibliometría (1961-2022)", analiza la aplicación práctica de la base de datos Hecumen en la realización de un estudio bibliométrico en el que se han descrito aquellas cuestiones específicas de HE, como son las temáticas y las épocas estudiadas, que no aparecen en el resto de estudios bibliométricos. Respecto a la discusión de cada uno de los capítulos, esta se realizará en los diferentes apartados del capítulo 6.

Para finalizar esta introducción y volviendo a la cita de Asimov con el que se abría el capítulo, a pesar del evidente acercamiento a las ciencias de la computación y a las matemáticas de la presente tesis, no se pretende desvirtuar la investigación en HE, sino enriquecerla con un acercamiento a otros campos de investigación que, aunque a priori pueden parecer en un lugar diametralmente opuesto a la HE, pueden aportar nuevas soluciones desde diferentes ópticas.

Capítulo 2: Principios cienciométricos como soporte a la investigación en historia de la educación

2.1 Introducción

El presente capítulo repasa de una manera panorámica los orígenes de la cienciometría, disciplina en la que se basan, eminentemente, los actuales índices e indicadores de producción científica a nivel nacional e internacional. También se muestran los diferentes retos epistemológicos a los que se enfrentó desde sus orígenes, así como las dudas y críticas que suscitó esta óptica de medición y valoración cuantitativa de la producción científica, cuestiones y discursos que hoy en día perduran.

Esta revisión narrativa del nacimiento y evolución de la cienciometría permite establecer un marco conceptual en el que enmarcar los estudios bibliométricos, como una de las técnicas dentro de la cienciometría, y sus limitaciones a la hora de realizar investigaciones en el campo de la HE. Se expondrá razonadamente una serie de datos clave y propios de la HE que no quedan evidenciados en los registros bibliográficos, y la limitación que esto supone a la hora de trazar un mapa de la disciplina en términos globales.

Por último, se propondrá un acercamiento a la biblioteconomía y la hermenéutica con la finalidad de posibilitar la obtención de información desde los mismos contenidos de los artículos, y no solamente desde sus registros bibliográficos; la finalidad de este planteamiento es enriquecer los datos descriptivos de la producción científica en HE, con el objetivo de analizar la disciplina desde una perspectiva global.

Asimismo, se expondrán las dificultades asociadas a este enfoque, dejando el inicio del camino trazado para diversas propuestas que superen estas problemáticas, y que serán desarrolladas en sucesivos capítulos.

2.2 La ciencia de la ciencia

En 1970, Eugene Garfield puso el foco de atención en la aparición de una nueva generación de sociometristas preocupados por el estudio de la ciencia y de los científicos desde una perspectiva histórica, sociológica, económica y conductual (Garfield, 1970). Los denominó, parafraseando la charla que dio Derek J. de Solla Price en la primavera de cinco años atrás en la Universidad de Chicago, científicos de la ciencia (Price, 1965).

La connivencia de estos dos autores no es casual, puesto que ambos tuvieron un papel fundamental en el desarrollo de la ciencia que estudiaba "la ciencia". Garfield fundó en 1960 el Institute for Scientific Information (ISI), conocido por ser la empresa que desarrolló originalmente las bases de datos que componen la actual Web of Science (WOS) y sus índices de citas. Chernyi (2009) describe el camino seguido por ISI en el que fue adquirida por Thompson Corporation en 1992, adoptando los sucesivos nombres de ISI-Thompson y Thompson Scientific para, finalmente, independizar su filial científica en 2016 bajo el nombre de Clarivate Analytics.

Garfield planteó identificar de una forma objetiva el impacto de las investigaciones de autoras y autores en sus respectivos campos. Para ello, el ISI elaboró el Science Citation Index (SCI), que consistía en una lista ordenada de artículos citados (denominados referencias), acompañados de aquellos artículos que los habían citado (conocidos como fuentes). El listado no incluía todos los contenidos de los artículos, sino que se limitaba a una serie de datos bibliográficos. Tal como explica Garfield (1970), en el caso de las referencias, los datos bibliográficos que se almacenaban eran los autores, la publicación donde había aparecido, el tipo de documento, el año de citación, el volumen y las páginas. En lo que se refiere a las fuentes, por cada uno de los autores citados se almacenaba el título del artículo que lo citaba, los autores, la publicación, el tipo de documento, el año de citación, el volumen y las páginas.

Por su parte, Price fue considerado por Garfield como el padre de la "ciencia de la ciencia" (tal como la denominaron originalmente), debido a su enorme impacto por el uso de indicadores cuantitativos a la hora de establecer los principios para valorar la

influencia de la producción científica (Garfield, 2009). Sin embargo, el nombre mediante el cual conocemos hoy en día a la ciencia de la ciencia, esto es, la cienciometría, fue acuñado por el polifacético investigador ruso Vasily Vasilyevich Nalimov con el vocablo original *naukometria*. Además de bautizar la nueva ciencia, también aportó la idea de utilizar las citaciones como indicadores (Wouters, 1999). El nombre actual quedó establecido en la comunidad científica internacional con la aparición de la revista Scientometrics en 1978, fundada por Tiber Braun en Hungría (Brindha y Murugesapandian, 2016).

En cuanto a su definición y a la concreción de sus objetivos, Koenig y Bookstein afirmaron que la cienciometría era la ciencia de medir "ciencia" (1995), cuestión que conecta con la temprana definición que realizaron Garfield y Price. Por su parte, Brindha y Murugesapandian (2016), de una manera más genérica, acotan sus objetivos al proceso de proveer información sobre la estructura del conocimiento y la forma en que es comunicado. Siguen diciendo que esto se consigue a través de la medida de patrones en las publicaciones en cualquier forma de comunicación escrita, aportando medidas evaluativas que reflejen la producción de un país y mostrando los patrones de citaciones en la literatura científica. Beck, en la editorial publicada en el número inaugural de la revista Scientometrics, dice que la cienciometría es el estudio de la evaluación cuantitativa y comparada de la actividad científica, productividad y progreso (Beck, 1978).

En la misma editorial del primer número de la revista Scientometrics, Beck (1978) apunta tres cuestiones de suma relevancia en la cienciometría:

1. La cienciometría puede contribuir a un uso más adecuado de los fondos en la investigación científica.
2. La cienciometría se fundamenta en el desarrollo de sistemas computerizados para catalogar y procesar la información, ya que permiten acumular una cantidad importante de datos y realizar evaluaciones estadísticas significativas.
3. La proliferación de revistas científicas es alarmante.

Los dos últimos puntos serán abordados en profundidad más adelante, ya que constituirán el eje vertebrador del capítulo 3. Por lo que respecta al primer punto, el planteamiento de la cienciometría el servicio de la evaluación y la medición de la producción científica se entiende como la base de una de las críticas que se le hace. Millán et al. (2017) sostienen que, desde esta perspectiva, la cienciometría no será más que un instrumento para el sometimiento y control ideológico. Esto es debido, principalmente, a su indeterminación epistemológica y su falta de fundamentación filosófica. En otros términos, su problemática radica en la incapacidad para definir un objeto de estudio estable. Millán et al. (2017) continúan evidenciando que, a pesar de que Garfield y Price sean considerados los padres de la cienciometría, tenían una concepción diferente de lo que debía ser, lo que redunda en la falta de fundamentación filosófica a la que se ha aludido. Si para Garfield el objeto de estudio era, fundamentalmente, el análisis cuantitativo de las citas, Price amplió el objeto para enfocar el problema desde una perspectiva en la que se pudiese analizar la lógica existente tras la evolución del conocimiento científico, lo que influyó en una primera etapa sociológica de la cienciometría.

Entender la cienciometría desde una perspectiva sociológica nos obliga, por tanto, a analizar las aportaciones de Robert K. Merton y su visión de la ciencia como una institución autónoma pero fundamentalmente social. Esta concepción pone en relevancia la circulación de los productos (investigaciones) entre los agentes (investigadores) y regulada por un ethos propio, entendido como un conjunto de valores y normas aceptadas por la propia comunidad científica (Ávila-Toscano, 2018). Esta difusión de la producción científica entre los científicos remarca la conexión de las investigaciones con el momento histórico en el que se producen, lo que cambia la manera de entender la actividad científica: ya no se trata del descubrimiento progresivo de una realidad, sino que es producto mismo de la actividad científica y del momento en el que se realiza.

Esta visión, junto con la línea establecida por Price, no puede desligarse de las coordenadas establecidas por Kuhn y su planteamiento fundamentalmente social en el que se liga la actividad científica al contexto social e histórico. Tal como plantea Kuhn

(2019), la comunidad científica se basa en los paradigmas, que son el conjunto de verdades y creencias aceptados en una época determinada por el conjunto de investigadoras e investigadores. Cuando se acumulan nuevas evidencias científicas que no encajan con el paradigma vigente, se produce una revolución científica que reemplaza el paradigma actual por uno nuevo que responda a estas nuevas evidencias. Así pues, la ciencia no evoluciona de una manera lineal por la simple aplicación del método científico, sino que lo hace por la dimensión social e histórica de sus agentes, que son los que aceptan un paradigma u otro en función de las evidencias.

El carácter cultural y contextualizado en el momento histórico de la ciencia pone el acento en la valoración crítica que pueden realizar las ciencias sociales a la hora de estudiar la ciencia, lo que trasciende a los análisis reduccionistas numéricos (Ávila-Toscano, 2018) propuestos originalmente por Garfield. En esta misma línea, Valle (2018) afirma que las ciencias sociales pueden realizar aportaciones significativas en lo que respecta a las cuestiones comportamentales de los agentes implicados en la producción científica, lo que supera las limitaciones del cómputo bibliométrico de citas.

Al respecto de los procesos bibliométricos dentro de la cienciometría dedicaremos el siguiente apartado.

2.3 Los estudios bibliométricos dentro de la cienciometría

La bibliometría es una de las disciplinas de la cienciometría cuyo objeto de estudio es la producción escrita (artículos, editoriales, documentos de conferencias...), analizada mediante procesos matemáticos con objeto de extraer indicadores bibliométricos en base a unos datos bibliográficos (Colorado y Anaya, 2018). Es decir, los estudios bibliométricos realizan un análisis cuantitativo y estadístico de los datos bibliográficos contenidos en las publicaciones científicas, sintetizados en unos indicadores bibliométricos. Cuando se plantea un estudio bibliométrico, las fuentes a las que se suele recurrir para la obtención de los documentos que formarán parte de la muestra suelen ser directorios o catálogos colectivos de revistas, así como servicios de indexación y resúmenes.

Los estudios bibliométricos forman parte de la cienciometría y, aunque su nombre fue propuesto por primera vez en 1969 por Alan Pritchard, podemos encontrar precursores de investigaciones bibliométricas en los análisis estadísticos realizados por Cole y Eales en 1917 sobre las investigaciones publicadas en el campo de la anatomía comparada (Araújo y Arencibia, 2002). Previamente al término bibliometría, este tipo de estudios se conocían como biografías estadísticas.

Los indicadores que se obtienen de los estudios bibliométricos sintetizan la información bibliográfica analizada en las muestras y permiten comparar cuantitativamente diferentes estudios bibliométricos. Según el autor al que recurramos, estos se clasifican en diferentes grupos. Por ejemplo, Ardanuy (2012) realiza la siguiente clasificación:

- Indicadores personales, que describen determinadas variables de las autoras y autores de los estudios como los países de procedencia, el sexo o las afiliaciones institucionales.
- Indicadores de producción, que se centran en contabilizar el número de publicaciones, tanto del autor (lo que permite obtener las autoras y autores más productivos) como de la institución de afiliación, país, etcétera.

Permiten, por tanto, conocer el número total de autores de una muestra y el número total de artículos y se pueden extraer indicadores como la media de autores por artículo o, al revés, la media de artículos por autor.

- Indicadores de dispersión, con los que se identifican aquellos artículos que han recibido el mayor número de citas y que, por lo tanto, han sido más relevantes e influyentes en el campo de estudio.
- Indicadores de impacto, fundamentado en lo hemos expuesto en el apartado 2.2 sobre Garfield en su intento de analizar objetivamente el impacto de la producción científica.
- Indicadores de colaboración, en muchas ocasiones representados por redes de colaboración entre autores o instituciones, o como aquellas parejas de autores que publican con mayor frecuencia. Dentro de los indicadores de colaboración solemos encontrar los análisis de co-citaciones, que reflejan la frecuencia con la que un par de artículos se han citado conjuntamente, lo que establecería similitudes temáticas entre ambos (Garfield, 1979). También se realizan los análisis de acoplamiento bibliográfico (*bibliographic coupling),* concepto acuñado por Kessler (1963) que, al igual que las citaciones, detecta las parejas de artículos con mayor frecuencia, pero si en el caso anterior era sobre los más citados, en el acoplamiento bibliográfico son los que más se citan. Se trataría de las parejas de investigaciones cuyas citaciones se repiten con mayor frecuencia.
- Indicadores de obsolescencia, que reflejan la antigüedad de los artículos contenidos en las muestras estudiadas.
- Indicadores de forma y contenido, esto es, la tipología de los textos según se clasifican en las bases de datos de las que se han extraído. Según estos indicadores, se puede identificar el número de artículos, capítulos de libros, comunicaciones en congresos, etcétera.

Moreno (1997), por su parte, propone una clasificación en función del número características de los documentos que se estudien:

- Indicadores unidimensionales, cuando solo estudian una característica sin tener en cuenta los aspectos comunes que pueden tener los documentos. Estos indicadores se pueden obtener de diversas maneras:
- Obtenidos según el uso de los documentos en los centros donde hay depósitos documentales. Un ejemplo de este indicador lo encontramos, por ejemplo, en los libros más prestados en una biblioteca.
- Obtenidos de los mismos documentos publicados, de los que se puede obtener su grado de actualidad o las tipologías. Cabe destacar que estos indicadores, con diferente nomenclatura, son similares a los de obsolescencia y forma y contenido, respectivamente, descritos previamente. En base a los documentos publicados también se pueden obtener las temáticas de los documentos, el grado de colaboración en su redacción y su visibilidad (esta última referida al factor de impacto de un documento).
- Indicadores multidimensionales, en los que se tienen en cuenta diversas variables de los documentos al mismo tiempo. Para ello se recurre a técnicas matemáticas como el análisis multivariante, mediante el cual se obtienen *clusters* y mapas. Según las variables analizadas, los indicadores multidimensionales permiten la elaboración de mapas de citaciones o mapas de las co-palabras entre otros.

Con la realización de un número elevado de estudios bibliométricos, determinados resultados comenzaron a repetirse de tal forma que emergieron patrones y regularidades. Estos patrones sirvieron para que diversos autores establecieran una serie de leyes bibliométricas. Ardanuy (2012) identifica las siguientes:

- Ley de productividad de los autores: evoluciona de la ley de Lotka, propuesta en 1926, y establece la relación desigual entre los autores y su producción a lo largo del tiempo, es decir, que un conjunto reducido de autores produce una gran parte de la producción científica.
- Ley de dispersión de la bibliografía: de una manera similar a la Ley de productividad de los autores, en el caso de la Ley de dispersión de la bibliografía solo un número reducido de revistas es al que se recurre a la hora

de confeccionar las bibliografías de los artículos. Esto fue estimado por Bradford en 1934.

- Ley de crecimiento exponencial: expuesta por Price en 1956, afirma que la producción científica crece de una manera exponencial, de tal forma que cuando una disciplina llega a esta fase, su producción se duplica cada 10-15 años.
- Ley de obsolescencia, también afirmada por Price, refleja que la obsolescencia de la bibliografía no es igual para todas las disciplinas, siendo esta más acusada en las ciencias experimentales y más lenta en las humanidades.

Sin embargo, los estudios bibliométricos tampoco están exentos de críticas. Por una parte, en lo que se refiere a los indicadores de impacto, se está correlacionando el número de citas que recibe un artículo con su importancia en el campo, invisibilizando a una gran parte de la producción científica que, aunque pueda haber sido relevante e influyente, no ha recibido el mismo volumen de citas que otros trabajos (Ardanuy, 2012). Por otra parte, estos indicadores de impacto pueden ser manipulados con prácticas como las de las autocitas, o citas que realizan las autoras y autores en sus trabajos al citar sus propios trabajos anteriores de una manera injustificada. Estas prácticas incrementan de forma artificial su factor de impacto y, por lo tanto, aumenta erróneamente la percepción de calidad en la investigación (Masic y Jankovic, 2021).

Por lo que respecta al objeto de la presente investigación, los análisis bibliométricos adolecen de una serie de información crucial para la investigación en HE ya que se limitan a un procesamiento de los datos bibliográficos, sobre lo cual profundizaremos en la siguiente sección.

2.4 Datos específicos de la producción científica en historia de la educación

Si el objeto de la HE es el estudio de una realidad educativa y de sus prácticas, enmarcadas en un todo histórico que, de una manera holística, incluye aspectos políticos, sociales, económicos o culturales (Guichot, 2006), el factor cronológico que contextualiza el objeto de estudio en la investigación científica en HE resulta angular. Esta variable cronológica de las investigaciones, a menudo reforzada por una tendencia historicista que pone el acento en el contexto de la realidad educativa estudiada, resulta de vital importancia cuando se trata de llevar a cabo una investigación en HE. Así pues, una investigación en HE pone su foco de atención, indirectamente, en un momento histórico que se interrelaciona con la realidad educativa estudiada, cuestión que también ha sido aprovechada desde un punto de vista pragmático como fundamentación teórica de las prácticas pedagógicas presentes en un afán de su propia legitimación (Cucuzza, 1996), aunque la importancia de su conocimiento y estudio radique en que la educación es un fenómeno histórico propio.

A este respecto hay que destacar una serie de limitaciones, desde el punto de vista de la HE, en la información obtenida en los estudios cienciométricos. Tal como describimos en el apartado 2.3, los indicadores bibliométricos se basan en la información bibliográfica. Uno de estos datos que se pueden encontrar en los diferentes directorios bibliográficos es el año de publicación de la investigación. Sin embargo, es baladí mencionar que el dato de la publicación de una investigación, o incluso el año en el que se realizó, no tiene nada que ver con la época estudiada en la propia investigación. El hecho de que la HE ponga su foco en el pasado, además de la cuestión crucial de que los procesos educativos están insertos en unas coordenadas espacio-temporales que los configuran (Guichot, 2006), implica que tenemos una primera discrepancia entre la información bibliográfica relativa a la fecha de publicación y el periodo estudiado en la propia investigación. Es decir, en una investigación de HE hay, al menos, dos juegos de fechas:

- Fechas externas: accesibles desde los propios datos bibliográficos del documento, y que ubican temporalmente el documento en el momento en el que fue publicado.
- Fechas internas: insertas dentro del contenido del documento, y que sitúan la investigación en la época en la que se ubica el fenómeno educativo estudiado.

Es por esto que, volviendo a los análisis cienciométricos y bibliométricos expuestos en los apartados 2.2 y 2.3 respectivamente, el análisis de las fechas contenidas en los datos bibliográficos de una serie de publicaciones de HE se refiere a fechas externas y no van a ofrecer ningún tipo de información respecto a las épocas estudiadas, lo que dificulta cualquier tipo de investigación sobre la producción científica en el campo al poner el foco exclusivamente en *cuándo* se ha estudiado (que también ofrece una información muy relevante), pero no en *qué épocas* se han estudiado. Las investigaciones cienciométricas en HE no deberían estar constreñidas por este tipo de limitaciones, no solo porque estamos dejando de lado la dimensión historicista de la disciplina, reduciendo la producción a unos meros indicadores cuantitativos de productividad, en muchas ocasiones ligados a la crítica que mencionamos en el apartado 2.2 de Millán et al. (2017) al respecto del instrumento de sometimiento y control que suponen, sino porque, desde el enfoque culturalista, la educación no solo tiene la capacidad de transformar la cultura, sino que también puede conservarla (Guichot, 2006). Así pues, la HE, en tanto que historia de los diversos planteamientos educativos a lo largo del tiempo, también es testigo en cierta manera la cultura que encierra cada momento educativo. Soslayar esta información histórica en las investigaciones cienciométricas y bibliométricas en HE implica dejar de mostrar una dimensión fundamental de la disciplina, reduciendo una parte importante de la riqueza de las investigaciones de HE. La obtención de las fechas internas del documento resulta, en la mayoría de los casos, una tarea trivial simplemente accediendo a datos bibliográficos del documento, como por ejemplo el título. En la tabla 1 se muestran los resultados del análisis de los datos bibliográficos de 11 revistas especializadas en HE de prestigio internacional, seleccionadas según el criterio planteado por Hernández Huerta, Payá y Sanchidrián (2019), referido a la profesionalización de su gestión editorial para estar

acreditadas internacionalmente. En estas revistas se han identificado el número de artículos indexados que contaban con alguna referencia temporal de la época estudiada en el título.

Tabla 1. Porcentaje de artículos con referencias temporales.

Revista	**Artículos totales**	**Con referencia temporal**	**Porcentaje**
Childhood in the Past: An International Journal	90	13	14,44%
Espacio, Tiempo y Educación	106	46	43,40%
Histoire de l'éducation	253	168	66,40%
História da Educação	470	220	46,81%
Historia Social y de la Educación/ Social and Education History	89	20	22,47%
Historia y Memoria de la Educación	100	51	51,00%
History of Education & Children's Literature	764	297	38,87%
History of Educacion. Journal of the History of Education Society	996	479	48,09%
History of Education Quarterly	336	101	30,06%

History of Education Review	169	67	39,64%
Paedagogica Historica: International Journal of the History of Education	1576	640	40,61%

Fuente: elaboración propia.

De media, un 40,16% de los artículos a los que da cobertura Scopus de estas 11 revistas referidas explicitan el año o el intervalo de años objeto de estudio en el título del artículo. Esto supone una base adecuada de la que partir a la hora de identificar el periodo histórico analizado en las investigaciones, tanto de manera manual para una muestra reducida, como mediante el procesado de esta información bibliográfica de una manera computerizada para muestras más grandes.

Por otra parte, tampoco se puede dejar de lado la cuestión de que en la investigación en HE se solapan, simultáneamente, las dimensiones sincrónicas y diacrónicas (Cucuzza, 1996). Dicho de otra forma, el fenómeno educativo estudiado puede estar enmarcado en diferentes contextos cuyos planteamientos no tienen por qué coincidir. Se puede presuponer una corriente de pensamiento mayoritaria en la época estudiada, pero encontrarnos frente a un planteamiento pedagógico rupturista puntual y concreto. Es por esto que la información de la época estudiada es insuficiente a la hora de reflejar la riqueza y variedad de las investigaciones en HE, con lo que debemos plantearnos un análisis temático de los contenidos, que concretará de una manera mucho más precisa el fenómeno educativo estudiado.

Sin embargo, la identificación temática de las investigaciones no resulta un proceso tan simple como en el caso de la época estudiada. A continuación, se expone el acercamiento a la ciencia de la biblioteconomía y documentación y se introduce la problemática de la interpretación hermenéutica de los textos para la clasificación temática.

2.5 Clasificación de la producción científica: biblioteconomía y hermenéutica

La biblioteconomía tiene su origen en la Antigüedad del mundo oriental y, posteriormente, se extenderá a occidente (Garrido, 1990). A pesar de la evolución de la disciplina a lo largo de los siglos y los retos a los que se ha visto enfrentada, ya desde sus inicios, en su periodo precientífico, uno de sus objetivos fue la clasificación de los materiales atesorados en las primeras bibliotecas (Orera, 1995). Sin embargo, el procedimiento era radicalmente distinto al que podemos encontrar hoy en día, puesto que en la denominada etapa pretécnica no existían normas para la catalogación y los criterios dependían individualmente de los responsables de los mismos (Garrido, 1990). Ello hacía que cada biblioteca siguiera un planteamiento distinto. No obstante, esto no implicaba un problema mayor ya que los ejemplares que podían almacenarse previamente a la invención de la imprenta también tenían, en la mayoría de los casos, un carácter único.

Hesse (2010) apuntó a una separación entre biblioteconomía y bibliografía al enumerar ciertos retos a los que se enfrentaba la primera, sobre todo referidos a las colecciones, los usuarios y la biblioteca como servicio organizado (Orera, 1995). A pesar de esto, la catalogación de materiales escritos se presenta como un puente entre ambas disciplinas en tanto que es un objetivo común para biblioteconomía y bibliografía. Además, al acercarnos a la biblioteconomía abrimos la puerta a la clasificación de los textos y, con ello, a la posibilidad de poder enriquecer los datos tradicionales bibliográficos con información extraída del contenido de los textos, y no simplemente de su información descriptiva externa. Si, tal como se ha explicado en el apartado 2.3, la información bibliográfica es la base de los indicadores bibliométricos, el hecho de integrar los contenidos de los textos desde la perspectiva de la biblioteconomía permitiría plantear nuevos indicadores bibliométricos. Estos podrían reflejar aspectos particulares de la HE que no están cubiertos por los datos bibliográficos tradicionales.

Sin embargo, este enfoque implica mezclar cuestiones que, desde la biblioteconomía, deben estar separadas. Un ejemplo de esto son los aspectos descriptivos de las obras (lo que correspondería con la información bibliográfica actual)

y los contenidos de las mismas. Esta separación sería similar a la que existe entre lo que se ha denominado previamente como fechas externas (provenientes de los datos bibliográficos de los trabajos) y fechas internas (deducidas de los contenidos de las investigaciones). Sin embargo, esto ha sucedido en diferentes momentos de la historia de la biblioteconomía, tal como afirma Garrido (1990) al describir que dicha mezcla fue habitual en la época pretécnica. No obstante, plantear esta mezcla no implicaría alejarnos de un método riguroso para añadir información extraída de los contenidos de los textos en los datos bibliográficos descriptivos.

Esta combinación de ambos tipos de datos (por ejemplo, fechas internas y externas) es fundamental si, tal como se ha expuesto previamente, se pretende analizar desde una perspectiva cuantitativa las cuestiones propias insertas en las investigaciones de HE, como son las épocas investigadas y las temáticas. Así pues, se debe establecer un puente entre la información contenida dentro de los materiales publicados y las variables descriptivas que podemos encontrar en los registros bibliográficos.

El hecho de acceder a los contenidos de una investigación para extraer la temática estudiada dentro de la HE se enfrenta a un problema clásico: la interpretación de los textos o hermenéutica. Cuando se hace referencia aquí a hermenéutica se limita a su acepción más tradicional, al margen de las transformaciones que sufrió el término en el siglo XX con la hermenéutica de la factilidad de Heidegger, o de la ontología de la comprensión de Gadamer (Sánchez, 2019). Según las tres acepciones posibles para definir la hermenéutica clásica de acuerdo con Grondin (2014), aquella que viene a colación del problema planteado para clasificar textos es la que se refiere al "arte" de interpretación de textos en sus tres variantes clásicas: *hermeneutica sacra*, *hermeneutica juris* y *hermeneutica profana* o, incluso, desde la perspectiva del idealismo romántico, una única hermenéutica general, integradora de las tres anteriores, tal como pretendió Schleiermacher (2002). Dejando al margen este planteamiento unificador, sería la hermenéutica profana la que se encargaría de la interpretación de textos con carácter general, al margen de aquellos sagrados (*hermeneutica sacra*) o jurídicos (*hermeneutica juris*).

A pesar de la necesidad de acercamiento al enfoque hermenéutico en tanto que la extracción de determinados datos del contenido de los artículos implica una lectura y comprensión del mismo, los objetivos se limitan a la identificación de la época estudiada en cada una de las investigaciones, así como la clasificación temática. Tal como se ha expuesto en el apartado 2.4, la primera es obvia en un gran porcentaje de trabajos simplemente accediendo a dato bibliográfico del título. La segunda sí que requiere un acercamiento a los contenidos de la obra, lo que lleva parejo una problemática respecto al volumen de producción científica.

Si volvemos a las palabras que dejó escritas Beck en la primera editorial de Scientometrics (1978), tildó de alarmante la proliferación de revistas científicas. Tal como se expondrá en el estudio bibliométrico del capítulo 5, esta afirmación adquiere un cariz premonitorio a raíz de la eclosión en cuanto a producción científica que se produjo en la HE en las siguientes décadas, y que deja en unas cifras mínimas, comparativamente, la producción de la década de los 70 del pasado siglo. Si lo que se pretende es poder clasificar tal volumen de producción científica según las variables propias de la HE, es inviable realizar una interpretación en profundidad de cada uno de los trabajos para extraer estos datos del contenido y añadirlos a los datos descriptivos bibliográficos de las obras, a no ser que esta labor fuera realizada desde dentro de los propios equipos editoriales de las diferentes revistas. Pero, aún así, requeriría cierto consenso entre las revistas especializadas a la hora de seleccionar los temas a los que puede pertenecer un artículo.

A pesar de las dificultades identificadas, en esta selección de categorías encontramos una simplificación del problema. Si la extracción del tema estudiado dentro de los contenidos de una investigación se reduce a identificar la temática de entre un grupo finito y reducido de posibles categorías, la aproximación al enfoque hermenéutico es sutil, ya que no es necesaria la comprensión en profundidad el texto, ni mucho menos el desciframiento de los contenidos psíquicos o vitales del autor cuando produjo el texto, tal como propuso Dilthey (Sánchez, 2019). Para poder abordar este problema, la clave radica en el hecho de que las categorías en las que se debe clasificar el texto ya estén dadas previamente, con lo que podemos transitar de un problema de

comprensión e interpretación hermenéutica a otro más abordable de clasificación. Por suerte, los problemas de clasificación no son específicos de la HE y han sido ampliamente estudiados en otras disciplinas, como son las ciencias de la computación.

Este camino de acercamiento a la disciplina de las ciencias de la computación es el que se seguirá en los sucesivos capítulos de la presente tesis, con la finalidad de plantear soluciones desde unas ópticas diferentes, con objeto de abordar la problemática de cartografiar la HE de una manera eficaz.

2.6 Conclusiones

El presente capítulo ha pretendido revisar la evolución histórica de la cienciometría, como la ciencia que estudia la ciencia, y uno de sus objetivos iniciales, que era el poder medir de una manera cuantitativa la influencia de las investigaciones en sus respectivos campos. Este objetivo sentó las bases de los actuales indicadores de citaciones, que condicionan la producción científica a nivel mundial.

En este último sentido también se han mostrado las críticas referidas, precisamente, a su falta de fundamentación epistemológica y al hecho de ser una ciencia al servicio de las agencias de acreditación. También se han investigado las diferentes visiones que tuvieron Garfield y Price al respecto de la cienciometría. La visión más cercana del último a la sociología acercó la investigación hacia Merton, que entendía la ciencia como un producto de cada época, e imposible de desligar de su contexto histórico y hacia Kuhn y su visión eminentemente social de la ciencia, ligada al contexto social e histórico.

Posteriormente se ha profundizado en los estudios bibliométricos, como una de las técnicas de la cienciometría, y su vinculación a la información bibliográfica de los registros que proveen las bases de datos. Es precisamente esta vinculación la que hace que, cuando se trata de estudios bibliométricos en HE, parte de la información específica de las investigaciones en HE, no quede reflejada. Dicha información son las temáticas específicas y las épocas estudiadas, que no aparecen en los registros bibliográficos, y que hace que se invisibilice parte de la riqueza de las investigaciones en HE cuando se trata de cartografiar y estudiar la disciplina desde una perspectiva global. La obtención de estos datos ha llevado la presente investigación al terreno de la biblioteconomía y la hermenéutica, en tanto que hay que acceder a los contenidos de los artículos para realizar una comprensión e interpretación de los mismos.

En este sentido, si la aparición de la imprenta, ideada por Gutenberg en el siglo XV, influyó en la evolución de la biblioteconomía (Orera, 1995) y, por tanto, en las técnicas que se utilizaban en la clasificación de los materiales escritos, la progresiva

digitalización de materiales escritos que estamos viviendo desde las últimas décadas del siglo XX también debería replantear las coordenadas en las que se realizan las clasificaciones temáticas. En lo que se refiere al campo de la HE, la eclosión de publicaciones especializadas en los últimos años motivado, entre otras cuestiones, por la facilidad de la publicación digital, tal como se mostrará en el capítulo 5, abre la puerta a explorar otras formas automatizadas de clasificación que, aunque provenientes de otras disciplinas, pueden resultar de ayuda en la tarea de la clasificación temática en HE. Una propuesta a este respecto desde la disciplina de la IA será expuesta en el capítulo 4. Pero, como paso previo, es necesario el diseño e implementación de una base de datos específica que permita combinar en un único lugar, tanto la información bibliográfica cuantitativa de los artículos, como los datos específicos de las investigaciones en HE. Este, por lo tanto, será desarrollado a continuación.

Capítulo 3: Desarrollo de una base de datos específica para historia de la educación

3.1 Introducción

Tal como se ha expuesto en el capítulo anterior, los estudios de HE poseen una serie de características propias, como son las épocas y las temáticas estudiadas, que no forman parte de la información bibliométrica de las bases de datos en las que aparecen indexadas -dado su carácter generalista-. Por ello, la ausencia de esta información imposibilita llevar a cabo una cartografía de la disciplina atendiendo a dimensiones propias de HE, así como la realización de estudios bibliométricos que reflejen la riqueza de las investigaciones en el área.

Esto exige, como paso intermedio para el análisis de la disciplina desde una perspectiva global, el desarrollo de una herramienta informática que permita aunar tanto los datos de los registros bibliográficos existentes como los nuevos datos cualitativos referidos a temáticas y épocas estudiadas.

El presente capítulo, por tanto, cubrirá esta parte de la investigación que llevamos a cabo, donde se diseñará, desarrollará e implementará una base de datos específica (Hecumen), que permita combinar tanto la información bibliográfica cuantitativa de los artículos como los datos específicos de las investigaciones en HE. En cada una de las fases de desarrollo se valorarán diferentes alternativas en cuanto a tecnologías, patrones de diseño o tipos de bases de datos, optando por la más adecuada dada las características y requisitos de la investigación.

Tal como ya citamos en el capítulo 2, Beck (1978) apuntó en la editorial del primer número de Scientometrics que la proliferación de revistas científicas era alarmante, lo que traía consigo la problemática de abordar este conocimiento histórico. En este sentido, desde mediados de los años 90 nos encontramos con un panorama de eclosión de publicaciones en el campo de la HE, debido a factores como la maduración de la HE durante los años 80 y 90, la reducción de los costes de publicación y distribución con la edición electrónica y la extensión de los sistemas de acreditación de profesores universitarios que les exigen publicar (Huerta et al., 2019).

Teniendo en cuenta que, tradicionalmente, uno de los problemas de la HE ha sido la complejidad en la búsqueda de fuentes (Salas, 2019), este incremento de publicaciones electrónicas vendría parejo a la necesidad de organizarlas y clasificarlas, de tal forma que la producción científica pueda ser analizada desde una perspectiva global, ayudando a establecer procesos que faciliten su difusión global, y ofreciendo un punto de entrada a la labor de futuros investigadores del campo de la HE. Las Tecnologías de la Información y Comunicación (TIC) permitirían abordar esta problemática desde una perspectiva diferente, lo que no estaría reñido con la génesis de la HE, equidistante a diferentes campos del conocimiento como la historia, la educación o las ciencias sociales (McCulloch, 2011). En esta línea, la incorporación de los procesos técnicos que aportan las TIC ampliaría las posibilidades de estudio y análisis de la HE, lejos de desvirtuarla. De hecho, la comunidad científica de HE no se ha mantenido al margen de los cambios y las posibilidades que ofrecen las TIC en sus investigaciones (Rico y Motilla, 2016). Siguiendo la misma línea, Beck (1978) también remarcó que la cienciometría se fundamenta en el desarrollo de sistemas computerizados para catalogar y procesar la información, ya que permiten acumular una cantidad importante de datos y realizar evaluaciones estadísticas significativas.

Un análisis de la HE desde una perspectiva global requiere ampliar la perspectiva desde la que analizamos el área, transitando desde lo particular hasta lo general, sin que esto implique necesariamente la pérdida del detalle de lo particular. En este sentido, la facilidad en la captura de datos y el abaratamiento de su almacenaje (Liao et al., 2012) permite procesar grandes volúmenes de información que, con un adecuado diseño, permitiría tener una visión global de la HE, al mismo tiempo que se mantendría la particularidad de los datos individuales, permitiendo a los investigadores centrarse, tanto en un nivel particular como en uno general, en función de las necesidades de la investigación. Estos datos individuales a los que nos referimos son los llamados metadatos, que es la información bibliográfica que complementa los contenidos textuales de un artículo científico, ofrecidos por los proveedores de datos bibliográficos, y que permiten localizar y catalogar la información contenida en el mismo (Kammerer et al., 2021). Los metadatos, pues, resultan clave para el proceso de catalogación y clasificación de grandes volúmenes de producción científica en el campo de la HE.

A pesar de que gran parte de la investigación producida en el campo de la HE está clasificada en bases de datos existentes, como Scopus, Web of Science, Dialnet o Redalyc -entre otras-, no hay un lugar centralizado, por lo general, en el que poder consultar toda la producción científica de HE. Así pues, un investigador tiene que acceder a diversas bases de datos para analizar un tema concreto dentro de la HE, con lo que una unificación de la producción albergada en estas bases de datos implicaría un salto cualitativo en lo que se refiere a la futura investigación en HE.

Por otra parte, como ya se ha mencionado con anterioridad, las citadas bases de datos tienen un carácter generalista ya que cubren diversas disciplinas, con lo que la información que registran es aquella común entre todas ellas. Es por esto que no hay un etiquetado de la producción científica según los temas y criterios específicos de HE. En este sentido, añadir una capa de información cualitativa a los metadatos de los artículos del campo de HE permitiría realizar búsquedas más precisas, de tal forma que se pudieran cribar periodos históricos y épocas estudiadas, así como categorías específicas de HE. Por citar un ejemplo, en las bases de datos generalistas se puede realizar una búsqueda de artículos que fueron publicados en determinados años, pero no se puede buscar aquellos artículos de HE que han centrado su estudio en una época determinada, ya que sería una información cualitativa. Con esta capa de información cualitativa, las investigadoras e investigadores de HE podrían obtener de una manera precisa la producción científica cuyo objeto de estudio ha sido una época o años determinados.

En tercer lugar, esta base de datos unificada podría establecerse como un punto de encuentro virtual entre investigadores del campo de HE, de tal forma que permitiera acceder o mantenerse informado de las últimas investigaciones publicadas en el campo, lo que puede propiciar y facilitar la generación de redes transnacionales de investigación a las que se ha hecho alusión anteriormente. Esto conecta con la realidad TIC contemporánea, marcada por las redes sociales y las múltiples oportunidades que ofrecen para aprender, lo que aplicado al área de estudio es denominado por Rico et al. (2016) como “historia de la educación 2.0”.

Dentro de este contexto, la presente fase de la tesis se centra en aportar una serie de soluciones desde las TIC para el campo de HE. El capítulo describe el desarrollo de la herramienta Hecumen, cuyo objetivo es la centralización en un único conjunto de datos de la información bibliográfica de diversas revistas del campo de la HE, y que permite a diversos investigadores añadir información cualitativa a los artículos, etiquetando y catalogando datos referentes a la categoría de los temas en cada uno de los artículos, los periodos históricos o las épocas estudiadas.

El capítulo, por tanto, se organizará de la siguiente manera:

- El apartado 3.2 se centra en el proceso del desarrollo de Hecumen, describiendo la metodología utilizada (3.2.1), llevando a cabo las fases de diseño (3.2.2), análisis de los resultados (3.2.3), rediseño de la aplicación en base a la información del análisis (3.2.4) y, finalmente, concluyendo con un análisis final y el establecimiento de unos principios de diseño (3.2.5).
- El apartado 3.3 establece las conclusiones del proceso.

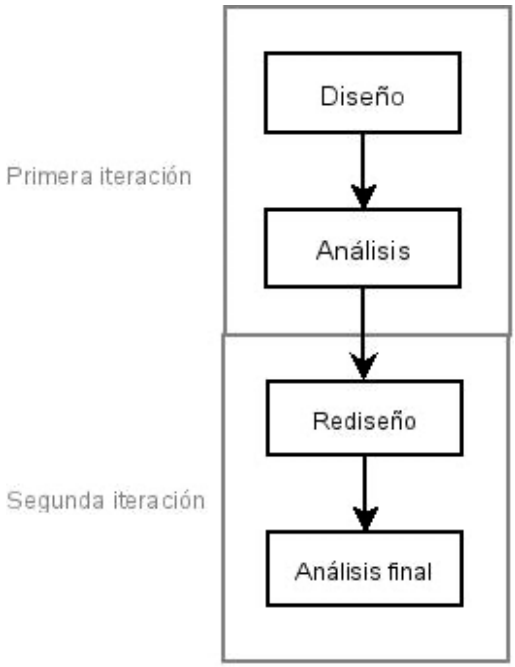
3.2 Desarrollo de la herramienta Hecumen

3.2.1 Metodología

A nivel metodológico, el desarrollo de la herramienta Hecumen se ha basado en la metodología DBR, cuyos principios persiguen la mejora o innovación a través de un esfuerzo colaborativo entre investigadores y profesionales a través de una serie de ciclos recursivos (Design-based Research Collective, 2003). En estas iteraciones se propone un diseño que es probado (testeado) por los profesionales y, tras un debate y un refinamiento, se vuelve a iniciar el ciclo hasta llegar al objetivo final, que no es otro que la elaboración de unos principios de diseño que puedan ser un punto de partida para futuras investigaciones.

Esta metodología pone el acento en la conexión del investigador con los problemas del mundo real a través de la estrecha colaboración entre investigador y profesional (Amiel y Reeves, 2008). La cercanía entre ambos perfiles persigue que el resultado de la investigación, que en este caso se trata de una herramienta *software*, esté adaptada en un alto grado a las necesidades del profesional, ya que el producto ha sido el resultado de esta interacción constante entre investigador y profesional. Dentro de este marco metodológico, se adoptó el enfoque de 4 fases para el desarrollo de la aplicación Hecumen; está compuesto por dos iteraciones en la que cada una de ellas incluye un diseño anterior y un análisis posterior (Design-based Research Collective, 2003). Tal como se puede observar en el diagrama de flujo de la figura 1, la entrada de cada una de las etapas depende de la salida o información obtenida en la fase previa. Así pues, debido al carácter iterativo de la metodología y por haber escogido el enfoque de cuatro fases, el ciclo completo ocurre en dos ocasiones.

Figura 1. DBR a cuatro fases.



Fuente: elaboración propia.

3.2.2 Diseño

La fase de diseño es la primera de las cuatro etapas de la metodología DBR, dentro de la primera iteración, y se ocupará de recabar los requisitos que debe tener la aplicación, realizar una investigación para dotar de sustento teórico a las decisiones de diseño y definir la estructura de la base de datos e implementar la solución.

3.2.2.1 Requisitos

Los requisitos que debía cumplir la herramienta Hecumen fueron extraídos de los datos cualitativos en la entrevista exploratoria semiestructurada con los investigadores principales del proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global". El carácter inicial de este tipo de entrevista permite obtener orientaciones para concretar el diseño de la herramienta o instrumento (Meneses y Rodríguez-Gómez, 2011), lo que permitió identificar los siguientes requisitos:

- R1: La base de datos debía almacenar los datos bibliográficos de una serie de artículos de revistas de HE, extraídas de diversas bases de datos. A pesar de que diferentes artículos de revistas compartirían una serie de metadatos (por ejemplo título, resumen, autores y palabras clave entre otros), otros podían ser específicos de un artículo concreto y no darse en otros.
- R2: La herramienta debía ser accesible por diferentes investigadores ubicados en diversos lugares geográficos.
- R3: El acceso de los investigadores se basaría en un sistema de permisos diferenciados, de tal forma que solo algunos investigadores pudieran realizar tareas administrativas en el sistema.
- R4: En lo que respecta a la interacción de los usuarios con la plataforma, la función básica de los investigadores sería el catalogar y etiquetar cada uno de los artículos asignados en función de la categoría de los temas en los artículos (según la clasificación establecida en el proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global"), los

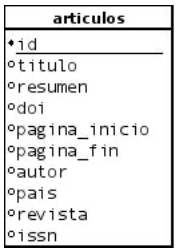
periodos históricos o las épocas estudiadas. Para tal cometido, los investigadores deberían poder visualizar los datos bibliográficos contenidos en la base de datos referidos a cada uno de los artículos asignados.

En base a estos requisitos, se procedió a realizar una investigación teórica que permitiera fundamentar las decisiones referidas al diseño de la aplicación, analizando los aciertos y las dificultades de proyectos similares.

3.2.2.2 Diseño de la base de datos

Una base de datos es un conjunto de información relacionada que contiene, además de los datos operacionales que se desean almacenar, la estructura misma de los datos (Karamcheti, 2007). Esta estructura de los datos se denomina *schema*, y define la organización de la base de datos en tablas, los campos que contiene cada una de esas tablas, las características de los campos y las relaciones que se establecen entre unos campos y otros. Los datos operacionales que pretendemos almacenar en la base de datos deben ajustarse al *schema*, en concreto a las características de las tablas donde se guardará la información, puesto que cada una de estas tablas no es más que una n-tupla que contiene de una manera secuencial cada uno de los registros, todos ellos compartiendo las mismas características de la tabla. Así pues, cada tabla define a cada una de las entidades que poblarán la base de datos. A modo de ejemplo de estos dos conceptos, aplicado a la información bibliográfica que nos atañe, necesitaríamos una entidad-tabla para almacenar la información relativa a cada uno de los artículos. Su *schema* simplemente enumeraría los campos que debería contener la tabla, es decir, los datos que debería almacenar esta tabla y que definen a un artículo, tal como se puede observar en la figura 2.

Figura 2. Definición de la tabla de artículos.



Fuente: elaboración propia.

Esta tabla almacena entidades de tipo artículo, que se caracterizan por contener una serie de datos comunes como son el título, el resumen, el DOI, las páginas de inicio y fin, así como los autores y la revista donde fue publicado. El campo *id* es un campo necesario en toda tabla, que almacena una serie de números enteros correlativos que aseguran que cada registro sea único, lo que se denomina la clave primaria. Si este *schema* es la estructura de la tabla, los datos operacionales serían cada uno de los registros que se insertan (es decir, la información de cada uno de los artículos que queremos almacenar).

Sin embargo, esta aproximación no sería la más adecuada, porque almacenar todos los artículos de una misma revista implicaría una redundancia importante de datos (el campo que almacena el nombre de la revista siempre sería el mismo); lo mismo sucedería con el campo de autores, que podría verse repetido en diversas ocasiones. Además, esta forma de almacenar la información dificultaría el posterior procesamiento de los datos registrados. La solución viene dada de una característica que se ha apuntado con anterioridad, la relación entre campos, en combinación con las formas normales de Boyce-Codd.

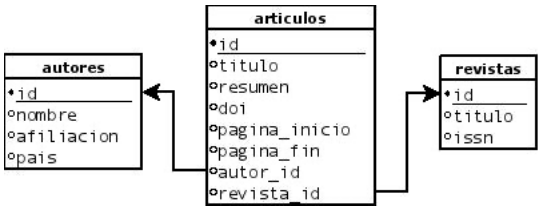
El modelo entidad relación (ER) permite definir los datos del sistema junto con las relaciones que existan entre ellos (Cerrada et al., 2000). Por su parte, la redundancia en los datos se puede eliminar recurriendo a la normalización de la base de datos mediante las formas normales de Boyce-Codd, que sigue tres fases: 1ª forma normal (1FN), 2ª forma normal (2FN) y 3ª forma normal (3FN) (Arini et al., 2019). En el proceso

de normalización de una base de datos, se deben seguir los requisitos exigidos en cada una de las formas normales, siendo:

- 1FN: Una base de datos está en 1FN si en cada uno de los campos de cada registro hay solo un dato.
- 2FN: Nos encontramos ante 2FN si se cumplen los criterios de 1FN y, además, hay una clave primaria que distingue cada registro y del que todos los demás campos que no sean clave primaria dependen de ella.
- 3FN: Los requisitos de la 3FN son los de la 2FN junto con el hecho de que no haya ninguna columna que dependa de otra que no sea la clave primaria.

En la figura 2 se puede observar que dicha base de datos cumple 1FN (cada registro almacenaría un único artículo) y 2FN (ya que, tal como apuntamos, se ha definido una clave primaria llamada *id*, que identifica a todo el registro). Sin embargo, no cumpliría 3FN, ya que hay una columna que depende de otras que no son la clave primaria. Esta es el ISSN, cuyos códigos son dependientes de la revista (un ISSN determinado siempre será de una única revista). Puesto que esta pareja de datos, revista-ISSN, configura una entidad independiente, la manera más simple de alcanzar 3FN consiste en extraerla de la tabla artículos, de tal forma que configure una entidad nueva. El resultado de esta normalización cumpliendo 3FN se puede observar en la figura 3.

Figura 3. Resultado de la normalización.



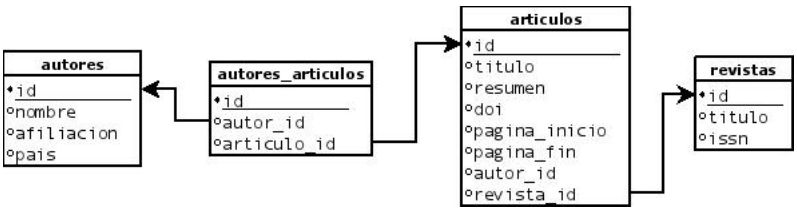
Fuente: elaboración propia.

Hasta este punto, la cardinalidad con la que nos encontramos en el diseño de la base de datos siempre es uno a uno (1-1) o uno a mucho (1-N). La primera es la que se da entre los campos dentro de una misma tabla (a un título le corresponde solo un

resumen, solo un DOI…) con lo que está integrada en la misma estructura. Por su parte, la cardinalidad 1-N es la que se da cuando la cardinalidad de un lado de la relación se reduce a 1, pero puede ser muchos del otro lado. Esta es la que encontramos en la relación entre artículos y revistas (un artículo solo puede pertenecer a una revista, pero una revista puede tener múltiples artículos). Al trasladar esta cardinalidad 1-N a la relación entre artículos y autores, podemos observar que se trata de algo deficitario. Un artículo debería poder tener más de un autor y, a su vez, un autor podría haber escrito más de un artículo, con lo que debemos realizar otra iteración en el proceso de diseño de la base de datos para integrar las relaciones muchos a muchos (N-N).

Una cardinalidad N-N exige la creación de tablas intermedias que relacionen diversos registros de una primera tabla con diversos registros de una segunda tabla. La modificación necesaria queda plasmada en la figura 4, en la que se añade la tabla *autores_articulos* (el nombre de esta tabla, en castellano, es a modo de ejemplo, por lo que no sigue la nomenclatura que se utilizará posteriormente en el diseño final). En esta tabla, en cada registro se almacena el par de claves primarias de los registros correspondientes de las tablas de autores y artículos. Así pues, esta solución permite registrar a múltiples autores por cada uno de los artículos, siendo que anteriormente esto no podía suceder sin incumplir 1FN, cuya exigencia era que no aparecieran varios datos en cualquier campo de un registro.

Figura 4. Tablas intermedias para relación N-N.



Fuente: elaboración propia.

No obstante, este proceso de normalización provoca una dispersión de datos relacionados (Green, 1996), evidencia que se refleja en la extensión de la figura 4 en comparación con la figura 2, lo que implica mayor coste computacional para acceder unos datos que están relacionados entre sí. Sin embargo, la claridad de esta forma de

diseñar una base de datos resulta evidente, lo que puede simplificar el mantenimiento de la herramienta, así como las posibles ampliaciones futuras.

Atendiendo al primero de los requisitos de la aplicación Hecumen (R1) se indicó que, a pesar de que los artículos compartirían una serie de metadatos comunes, cabía la posibilidad de que algunos datos fueran específicos de determinados artículos, pero que no aparecieran en otros (como por ejemplo número de páginas, detalles de la financiación del artículo, así como detalles referentes a la conferencia en la que fue presentado en el caso de comunicaciones). Esta situación de contingencia y dispersión de determinados datos plantea la problemática de su almacenamiento.

Una primera solución podría consistir en añadir tantos campos como metadatos puedan existir en cualquier revista. Sin embargo, este planteamiento conlleva el hecho de que muchos registros tendrían campos vacíos (aquellos en los que el registro no contuviera tal metadato) y, no menos importante, habría que conocer de antemano todos los campos posibles de todas las revistas. Esto se podría conocer para un listado determinado de revistas, pero se perdería la flexibilidad de añadir futuras revistas no previstas inicialmente, ya que exigiría cambiar el modelo de datos y, en consecuencia, reescribir la aplicación (Xie et al., 2013).

Otra alternativa consiste en optar por relaciones de herencia en las entidades, en las que hay una entidad base con unos campos comunes a todos los registros, mientras que otras subentidades heredan estos metadatos comunes y, además, almacenan los datos específicos, tal como describen Cerrada et al. (2000). Este enfoque, adoptado en proyectos como el descrito por Rashid y Chitchyan (2003), aunque eliminaría la problemática de los campos vacíos de la primera solución, no soslaya el hecho de que habría que conocer todos los metadatos posibles de las revistas presentes y futuras, puesto que en el caso de tener que añadir una que tuviera algún metadato no previsto, habría que modificar el modelo de datos y, tal como se ha indicado previamente, reescribir la aplicación.

La tercera opción consiste en recurrir a un modelo de bases de datos no *Structured Query Language* (SQL). Todo lo expuesto hasta el momento se ha basado en el modelo declarativo SQL, que es el lenguaje de programación para la obtención de datos desarrollado por IBM en la década de 1970 (Feng, 2006). Sin embargo, existe la alternativa de los modelos de bases de datos no-SQL orientado a documentos (Piroska et al., 2012) en los que no hay una estructura predefinida para los registros, ya que estos se almacenan como documentos de contenido variable. A pesar de que el planteamiento puede ser adecuado para la problemática de los datos heterogéneos, Feng (2006) sostiene que hay una serie de dificultades relacionadas con el uso de bases de datos no-SQL, como por ejemplo a la hora de la combinación de datos de diferentes entidades, mucho más complejo que con su equivalente en bases de datos relacionales SQL. Teniendo en cuenta que el grueso de la aplicación va a tener que combinar datos bibliográficos de diferentes tablas, tampoco resulta la opción más adecuada.

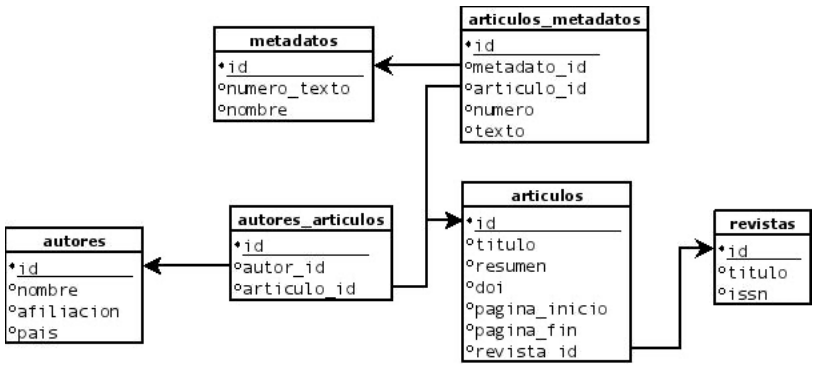
La última vía, que es por la que se ha optado, se basa en lo expuesto anteriormente de las relaciones N-N, y es el denominado modelo Entidad-Atributo-Valor (EAV). En este modelo, los datos son almacenados en tripletas de campos que relacionan una entidad con un atributo y con su respectivo valor. Tanto las entidades como los atributos están en tablas independientes, con lo que la tabla intermedia que permite implementar el modelo EAV es similar a la que se utilizaba en las relaciones N-N ya expuestas. Este enfoque permite modificar las estructuras de los datos simplemente cambiando los valores en las tablas sin la necesidad de tener que cambiar la estructura relacional de la base de datos, tal como se exigiría en un modelo íntegramente relacional (Ganslandt et al., 1999). Además, el modelo EAV es ampliamente utilizado en bases de datos relacionadas con la atención sanitaria (Batra et al., 2018), precisamente por la misma problemática que se ha expuesto con la información bibliográfica: la contingencia y dispersión de determinados datos, difícilmente previsibles a la hora del diseño de una base de datos íntegramente relacional.

Sin embargo, el modelo EAV también adolece de una serie de problemas relacionados con una falta de eficiencia a la hora de realizar búsquedas (Batra et al., 2018), con lo que no se utilizará íntegramente en el diseño de la base de datos Hecumen.

Puesto que gran parte de los datos son comunes a todos los registros y conocidos de antemano, se opta por recurrir a un diseño principalmente relacional, en combinación con un modelo EAV para poder registrar metadatos contingentes y dispersos en caso de encontrarlos en el momento de la ingesta de datos en el sistema.

En la figura 5 se modifica el diseño expuesto hasta el momento para incorporar, tanto la agrupación de artículos por números (R1), como las tablas necesarias para poder almacenar datos variables según el modelo EAV.

Figura 5. Tablas necesarias para el modelo EAV.



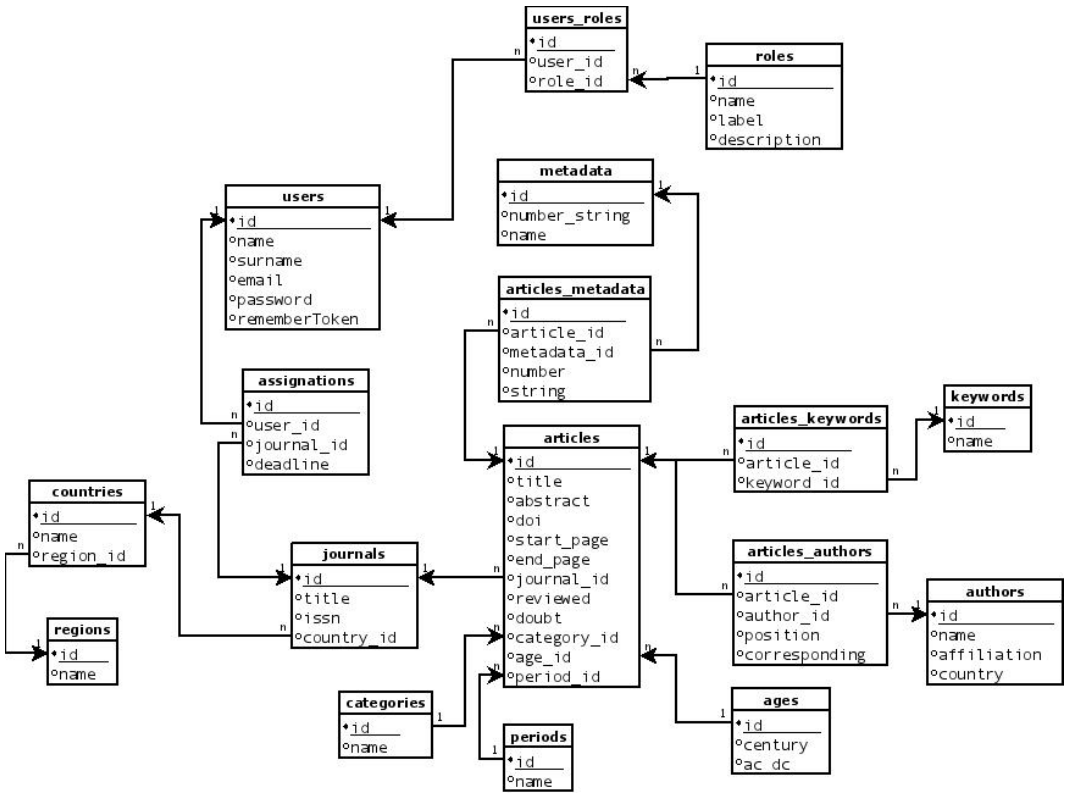
Fuente: elaboración propia.

Las tablas *articulos_metadatos* y *metadatos* resuelven la problemática de la dispersión y contingencia de los datos mediante el modelo EAV, de tal forma que los metadatos que no puedan ser almacenados en el esquema relacional base, quedarán registrados con estas dos tablas. Cuando, en la ingesta de datos, nos encontremos ante un metadato no previsto en el esquema relacional base, dicho metadato se insertará en la tabla metadatos, indicando si se trata de un dato de tipo numérico o de texto (campo *numero_texto*, necesario para el posterior procesamiento de los datos), y relacionándolo con el valor correspondiente en la tabla *articulos_metadatos*. Esta última, adicionalmente, queda vinculada con el artículo correspondiente de la tabla de artículos.

El modelo ER puede plasmarse en un diagrama ER, que proporciona una descripción básica de la base de datos relacional, mostrando los componentes (las

tablas-entidad), así como la cardinalidad y relación entre las entidades (Arini et al., 2019). Así pues, en base al desarrollo aquí expuesto, se presenta en la figura 6 el diagrama ER de la aplicación Hecumen:

Figura 6. Diagrama ER.



Fuente: elaboración propia.

Los aspectos a destacar de este diagrama ER son los siguientes:

- La denominación de las tablas-entidad, así como de los campos contenidos en cada una de ellas, se ha traducido al inglés por coherencia con el siguiente punto del diseño, en el que se implementará la herramienta íntegramente en inglés.
- El modelo EAV para los metadatos contingentes de los artículos queda plasmado en la tabla intermedia *articles_metadata*, y su catalogación como dato numérico o textual quedan almacenados en una única tabla denominada *metadata*. Esto cubre íntegramente R1.
- Para cumplir los requisitos R2 y R3 se debe almacenar la información de diferentes usuarios que puedan acceder simultáneamente con diferentes

permisos de acceso. Esto se resuelve con las tablas *users*, *users_roles* y *roles*, con el esquema ya visto de cardinalidad N-N con tres tablas.

- R4 exige que la herramienta permita a los investigadores catalogar cada uno de los artículos en función de la categoría, los periodos históricos o las épocas estudiadas. Por este motivo se han añadido los campos *category_id*, *age_id* y *period_id* en la tabla *articles* que referencian a sus respectivas tablas de categorías, épocas y periodos históricos. En este caso la cardinalidad es 1-N (un artículo tiene una categoría, un periodo y una época, pero tanto una categoría como un periodo o una época puede pertenecer a diferentes artículos).
- Adicionalmente se han añadido las tablas necesarias para almacenar las palabras clave con una cardinalidad N-N (un artículo puede tener varias palabras clave, y una palabra clave puede pertenecer a varios artículos). Estas tablas son *articles_keywords* y *keywords*. Además, se ha añadido una tabla de regiones (*regions*) que permite identificar tanto la región a la que pertenece un país para posibles análisis geográficos en términos globales.

Una vez realizado el primer diseño de la base de datos, se procede a la implementación de la herramienta, tal como se describe en la siguiente sección.

3.2.2.3 Implementación del Producto Mínimo Viable

Los requisitos de la aplicación Hecumen, extraídos de la entrevista exploratoria, ponen el acento en el acceso de varios investigadores simultáneamente a la aplicación, desde diferentes lugares geográficos, pero todos trabajando alrededor de una misma base de datos común. Implementar la herramienta como una aplicación *web*, accesible a través de internet, se adapta a estas cuestiones.

Una aplicación *web* es un sistema de *software* basado en las tecnologías y estándares del World Wide Web Consortium (W3C) que provee recursos específicos, tales como servicios y contenidos, a través de una interfaz de usuario en el navegador *web* (Kappel et al., 2006). Además, tal como manifiestan Bruno et al. (2005), permite funciones de procesado de información de una forma remota, desde un navegador de

internet, ejecutándose parcialmente en un servidor *web*, un servidor de aplicaciones o un servidor de bases de datos.

Uno de los múltiples lenguajes de programación que se utiliza para desarrollar aplicaciones *web* es el Pre-Procesador de Hipertexto (PHP). Se tata de un lenguaje de *script* (en términos generales, estos lenguajes no son compilados, aunque hay excepciones como el lenguaje Falcon) que se ejecuta del lado del servidor, diseñado específicamente para aplicaciones *web* y que, entre otras, cuenta con una serie de ventajas, como por ejemplo su rendimiento, escalabilidad, portabilidad y el hecho de ser de código abierto (Supaartagorn, 2011).

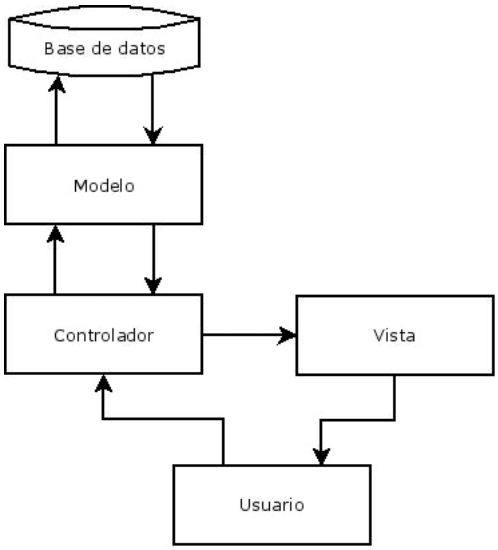
Sin embargo, no es habitual iniciar un desarrollo desde cero con el lenguaje escogido, sino que se suele recurrir a los *web frameworks*. Un *framework*, dentro del contexto del desarrollo de aplicaciones *web*, es un conjunto de librerías de desarrollo que facilitan el diseño de aplicaciones *web*, añadiendo rigor al desarrollo, garantizando una arquitectura coherente y permitiendo la automatización de tareas debido a una serie de rutinas que ya han sido implementadas nativamente (Laaziri et al., 2019). Uno de estos *frameworks* es Laravel, desarrollado por Taylor Otwell y liberado como proyecto de código abierto. Otro de los frameworks ampliamente utilizados es Symfony y, a pesar de que ambos están desarrollados sobre un conjunto de librerías comunes llamadas "The Symfony Components" (Laaziri et al., 2019), Laravel cuenta con un potente sistema de desarrollo por capas (llamado *middleware*), que permite añadir capas de funcionalidades a un programa según las necesidades. Además, estas capas se pueden activar o desactivar en función de las partes de la aplicación en la que el usuario se encuentre, de tal forma que aporta mucha flexibilidad si cambian los requisitos de la aplicación, lo que puede suceder cuando se trabaja con un enfoque de sucesivas iteraciones, tal como ha sido el desarrollo de Hecumen. Así pues, Laravel ha sido el *framework* escogido para el desarrollo de Hecumen.

Laravel está dirigido al desarrollo de aplicaciones *web* siguiendo el patrón de diseño Modelo-Vista-Controlador (MVC) (Yu, 2015). Los patrones de diseño son modos y soluciones a problemas de programación que suceden con frecuencia, en muchas

ocasiones similares, emergiendo al final como patrones (Lott y Patterson, 2007). El patrón de diseño MVC fue descrito originalmente por Trygve Reenskaug en 1979, posteriormente implementado por Jim Althoff, y consiste en un desacoplamiento del acceso a los datos (modelo), la lógica del procesado (controlador) y la representación de los datos (vista) (Curry y Grace, 2008).

Las acciones del usuario se envían al controlador, que las procesa y, en caso de ser necesario, se comunica con el modelo (por ejemplo, para solicitar información a la base de datos o para alterar un registro), para finalmente enviar los resultados a la vista, que será la encargada de reflejar los cambios en la interfaz de usuario. Este proceso queda ilustrado en la figura 7.

Figura 7. Esquema MVC.



Fuente: elaboración propia.

En base a la serie de decisiones tomadas durante el proceso de diseño que se han descrito hasta el momento, tanto de la estructura de la base de datos como del uso de un lenguaje y *framework* específico, se procedió a implementar un Producto Mínimo Viable (PMV). El PMV, tal como lo definió Ries (2011), consiste en un nuevo producto funcional en el que se ha omitido todo lo innecesario, de tal forma que su desarrollo consuma una menor cantidad de recursos, y cuya finalidad es poder obtener información sobre su uso para ir mejorándolo y ampliándolo en sucesivas iteraciones.

Se estableció que las características que tendría este PMV sería el acceso con diferentes cuentas de usuario, cada una con sus roles (R2 y R3), así como la visualización de los datos bibliográficos, el etiquetado y la catalogación de los artículos (R4). R1 queda implícito en el PMV, ya que el diseño de la base de datos es requisito necesario para cualquiera de las características del PMV. Las implicaciones de estos requisitos en cada una de las capas del modelo MVC se refieren a continuación:

- En lo que respecta al modelo, se tradujo el diagrama ER previamente diseñado (figura 6) a Eloquent, el Object Relational Mapping (ORM) de Laravel. Un ORM es un método de programación que permite traducir conceptos de la programación orientada a objetos a aquellos relativos a las bases de datos relacionales (Budiman et al., 2017). Así pues, la base de datos era accesible desde objetos PHP, tanto para operaciones de inserción y modificación, como de obtención de registros. Además, se implementaron los métodos en estos objetos que permitían acceder a otros objetos de las bases de datos a través de sus diferentes relaciones de cardinalidad (1-1, 1-N y N-N).
- Las implicaciones en el controlador consistían en la implementación de tres objetos controladores. El primero se especializaría en la lógica del proceso de autentificación mediante roles, el segundo controlaría el panel de control (la primera página que se muestra tras la autentificación) mientras que el tercero se centraría en la obtención de los datos de las revistas asignadas, así como las rutinas necesarias para el etiquetado y catalogado de los artículos.

  El controlador dedicado a la lógica del proceso de autentificación mediante roles (*LoginController*) implementó dos métodos: *login*, que interactuaba con la entidad *User* del modelo y, tras un proceso de verificación, incorporaba la información correspondiente a la sesión para que el usuario se mantuviera dentro del sistema sin que tuviera que solicitarle de nuevo los datos. Por su parte, el método *logout* eliminaba la información de la sesión y redirigía a la página de credenciales.

  El controlador que se encargaba del panel de control (*DashboardController*) era el que menos complejidad implicaba, puesto que simplemente daba información

sobre las revistas asignadas al usuario, así como el número de artículos pendientes de catalogación. Es por esto que interactuaba con las entidades *User*, *Article* y *Journal* del modelo. Contaba con un único método (*getDashboard*).

El controlador *JournalController* resultó el más complejo de implementar, ya que implicó implementar la lógica de creación, lectura, actualización y borrado (CRUD, por sus siglas en inglés) de las categorías, años y épocas estudiadas de los artículos, de tal forma que los usuarios pudieran etiquetar los artículos. Es por esto que se implementaron los métodos *filterJournals* (mostraba solo los artículos de una revista determinada) y *editArticle* (unificaba la creación, actualización y borrado de los datos sobre categorías, años y épocas estudiadas en función de la información suministrada por el formulario). Este controlador es el que interactuaba con más entidades del modelo debido a que los datos con los que trabajaba interrelacionaban muchos datos. Así pues, trabajaba con las entidades *User*, *Journal*, *Article*, *Category*, *ArticleAuthor*, *ArticleCategory*, *Period*, *ArticlePeriod*, *Age*, *Author*, *ArticleAge*, *Keyword*, *ArticleKeyword* y *Assignation*.

Un resumen de la capa de controlador puede observarse en la tabla 2.

Tabla 2. Controladores, métodos y relaciones con el modelo.

Controlador	**Métodos**	**Relación con el modelo**
LoginController	*login* *logout*	*User*
DashboardController	*getDashboard*	*User* *Article* *Journal*
JournalController	*filterJournals*	*User*

	editArticle	*Journal*
		Article
		Category
		ArticleAuthor
		ArticleCategory
		Period
		ArticlePeriod
		Age
		Author
		ArticleAge
		Keyword
		ArticleKeyword
		Assignation

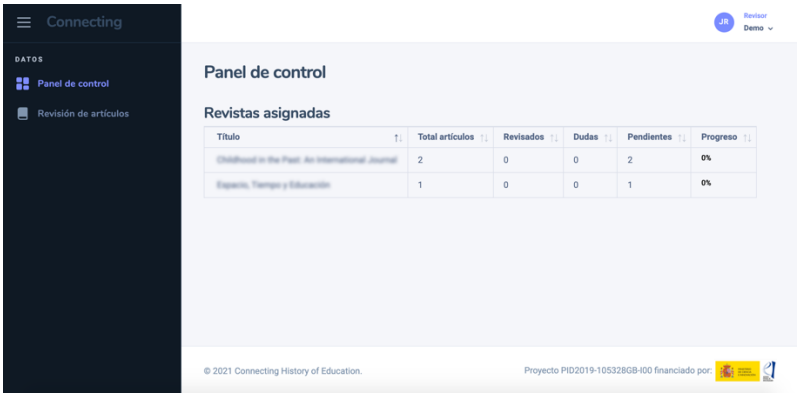
Fuente: elaboración propia.

- El trabajo en la vista se centró en el diseño estático de las plantillas HTML de las diferentes secciones, así como la lógica para mostrar los datos dinámicos en función de las interacciones del usuario. El resultado se puede observar en las figuras 8, 9, 10 y 11.

Figura 8. Pantalla de acceso.

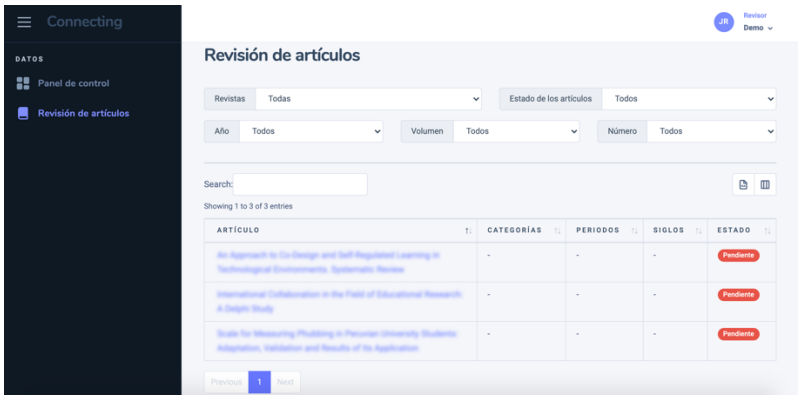
Fuente: elaboración propia.

Figura 9. Panel de control.



Fuente: elaboración propia.

Figura 10. Listado de artículos.



Fuente: elaboración propia.

Figura 11. Revisión de artículo.

Fuente: elaboración propia.

3.2.3 Análisis

Una vez completada la fase de diseño se procede al análisis, ambas dentro de la primera iteración de la metodología DBR a cuatro fases. Para esta etapa, el PMV resultante de la fase de diseño anterior es puesto a prueba por los profesionales que utilizarán la herramienta Hecumen una vez finalizada. En este caso, se trata de los investigadores principales del proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global", que refieren una serie de cuestiones en relación a las funcionalidades de la herramienta.

En primer lugar, se muestran conformes con el diseño de la herramienta a nivel general, tanto en lo referente a su forma de acceso a través de internet, como en la sencillez de uso gracias a su interfaz. A un nivel más pormenorizado detallan que sería pertinente que los investigadores, a la hora de etiquetar un artículo, pudieran contar con un campo de observaciones en el que dejar anotadas cuestiones respecto a sus decisiones de etiquetado y catalogación. Además, constatan que hay un error de diseño en cuanto a dicho etiquetado: cada uno de los artículos puede estar incluido en más de una categoría, periodo histórico o haber estudiado más de una época, al contrario de lo que permite la herramienta en este punto al haber escogido una cardinalidad 1-1. Por último, proponen dos secciones accesibles exclusivamente por los roles de administrador del sistema; la primera debería permitir la gestión de los usuarios (altas, bajas, datos de acceso y roles), mientras que la segunda mostraría información estadística, tanto del trabajo que se está realizando, como de los resultados ya obtenidos.

En base a esta información proporcionada por los investigadores principales, se enumeran los siguientes puntos del análisis:

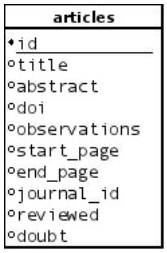
- A1: Añadir un campo de observaciones para cada artículo.
- A2: Permitir catalogar cada artículo en múltiples categorías, periodos históricos o épocas estudiadas.
- A3: Zona de administrador para la gestión de los usuarios.
- A4: Zona de administrador con información estadística.

3.2.4 Rediseño

Tras haber completado la fase de análisis previa se da por finalizada la primera iteración, comenzando la segunda y última iteración, que está compuesta por dos fases: rediseño y análisis final. En lo que respecta al rediseño, se parte de la enumeración previa que determina los requisitos de rediseño. Así pues, se procede a implementar los cambios tal como se detalla a continuación.

Los cambios que exige A1 son mínimos y afectan a la base de datos, al modelo, al controlador y a la vista correspondiente. En lo que respecta a base de datos, se actualiza el diagrama ER para incluir el campo observaciones que queda tal como se puede observar en la figura 12.

Figura 12. Modificaciones en la tabla de artículos.



Fuente: elaboración propia.

Se realizan los cambios necesarios en el modelo para poder trabajar con este nuevo campo, así como en el controlador, donde se añade la lógica de obtención, modificación y guardado del nuevo campo. Por último, se modifica la vista, añadiendo a la interfaz el campo necesario para que el usuario pueda registrar este nuevo dato, resultando como se muestra en la figura 13.

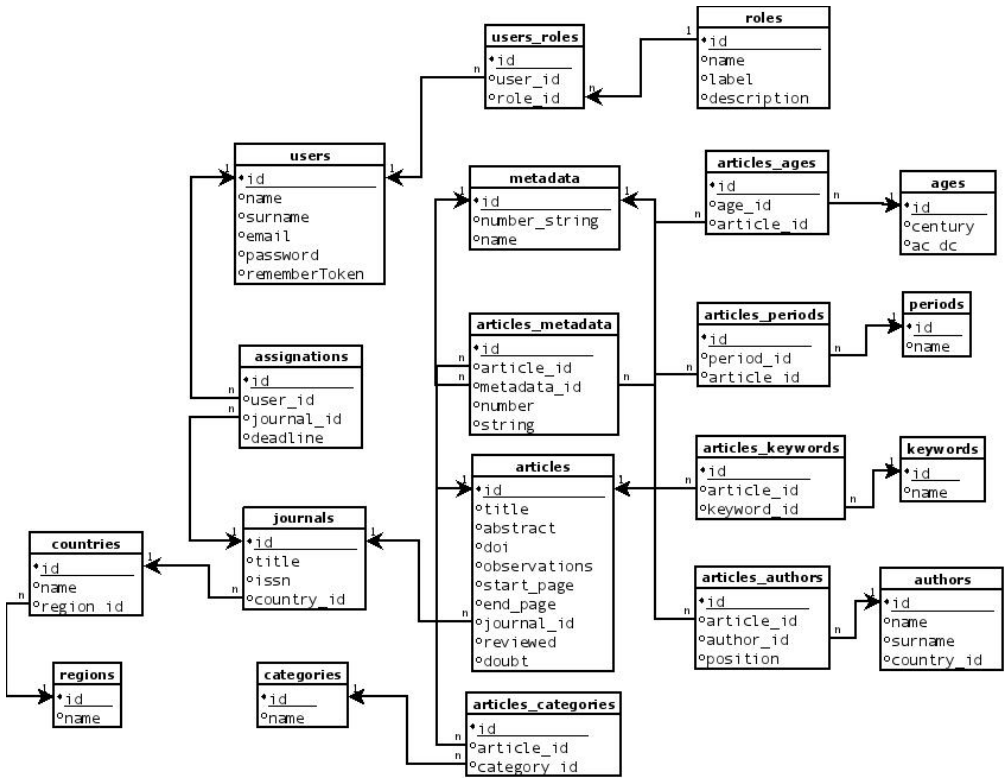
Figura 13. Cambios en la interfaz.

Fuente: elaboración propia.

A2 exige un número más elevado de cambios, ya que implica cambiar la cardinalidad de tres campos, pasando de 1-N a N-N. Así, un artículo podrá tener múltiples categorías y, al mismo tiempo, una categoría puede estar presente en más de un artículo. Lo mismo sucedería con los periodos históricos y las épocas estudiadas.

Como se ha expuesto anteriormente, una cardinalidad N-N implica que son necesarias unas tablas intermedias que permitan relacionar los registros de una primera tabla (artículos) con una segunda tabla (categorías, periodos históricos y épocas estudiadas). Así pues, se añaden las tablas *articles_categories*, *articles_periods* y *articles_ages*. Estos cambios en el diagrama ER exigieron cambios en el modelo, definiendo estas nuevas relaciones en el ORM Eloquent, así como en el controlador a la hora de acceder, modificar y guardar dichos datos. Como consecuencia, el diagrama ER completo resulta como se puede ver en la figura 14.

Figura 14. Rediseño del diagrama ER.



Fuente: elaboración propia.

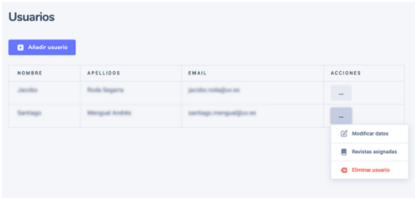
También se tuvieron que realizar cambios en la vista a nivel de interfaz para que el usuario pudiera realizar una selección múltiple, que se reflejó en los cambios de la figura 15.

Figura 15. Cambios en la interfaz.

Fuente: elaboración propia.

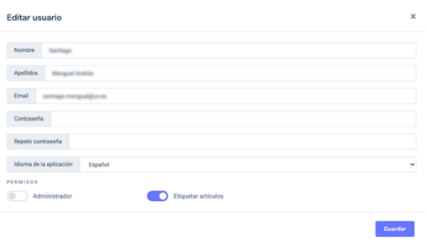
Tanto A3 como A4 no implicaron cambios en el diseño de la base de datos ni en el modelo. En lo referente a A3, esta ya estaba preparada para registrar la información de acceso de diferentes usuarios, cada uno de ellos con la posibilidad de tener asignados varios roles. A4 consiste en el análisis de los datos que ya hay almacenados. Así pues, ambos resultados del análisis se limitaron a modificaciones o creaciones de nuevas vistas y la implementación de la lógica correspondiente en el controlador. Estos cambios en la interfaz se pueden identificar en las figuras 16, 17 y 18.

Figura 16. Gestión de usuarios.



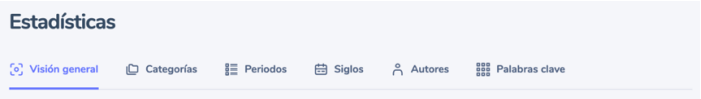
Fuente: elaboración propia.

Figura 17. Alta de usuarios.



Fuente: elaboración propia.

Figura 18. Estadísticas.



Fuente: elaboración propia.

3.2.5 Análisis final

La última fase de la metodología del DBR de cuatro fases consiste en un análisis final en el que se establecen unos principios de diseño que, a modo de conclusión tras el desarrollo de la investigación realizada, permiten sentar las bases de futuras investigaciones (Design-based Research Collective, 2003).

Los principios de diseño de la presente investigación son los siguientes:

- Las bases de datos cuyo objetivo sea almacenar información bibliográfica deben tener prevista la particularidad de la contingencia y dispersión de determinados metadatos, así como la imposibilidad de poder conocer de antemano los nuevos metadatos que se creen en futuras publicaciones. A pesar de que existen diferentes enfoques para solucionar esta problemática (como, por ejemplo, los campos vacíos, la relaciones de herencia en las entidades o las bases de datos no-SQL), se ha considerado que la combinación de un diseño relacional clásico junto con un modelo EAV permite tener al mismo tiempo la velocidad de acceso a los datos del modelo relacional, junto con la flexibilidad del modelo EAV. Además, permite añadir metadatos nuevos sin tener que reescribir la aplicación.
- Cuando los requisitos de una herramienta son que debe ser accesible por diferentes usuarios simultáneamente, desde diferentes lugares geográficos, trabajando con una base de datos común, la decisión de implementar dicho *software* como una aplicación *web* resuelve estas problemáticas frente al hecho de escribir una herramienta de escritorio. Además, simplifica el despliegue de las sucesivas versiones y correcciones de errores, ya que la herramienta está almacenada y se ejecuta parcialmente desde un servidor *web* centralizado (Bruno et al., 2005).

- El uso de *web frameworks* para el desarrollo de una aplicación *web* facilita el diseño de la herramienta, añadiendo rigor, garantizando una arquitectura coherente y automatizando una serie de tareas (Laaziri et al., 2019). Esto, además de optimizar el tiempo de desarrollo, permite centrarse en las funcionalidades específicas, obviando las cuestiones comunes a la mayoría de las aplicaciones *web*, cuya implementación consume una cantidad de tiempo considerable. Además, se basan en patrones de diseño ampliamente utilizados y refinados, como el MVC que utiliza Laravel, descrito por primera vez en 1979, lo que aporta una garantía de estabilidad a la herramienta desarrollada.
- La opción de centrarse en un PMV (Ries, 2011), en lugar de abordar el desarrollo de la aplicación completa desde el primer momento, permite tener un producto funcional en menor tiempo en el que se ha omitido todo lo innecesario. Esto, además del evidente ahorro de recursos, facilita la obtención de información por parte de los usuarios que van a utilizar la herramienta. Esta información se podrá utilizar en sucesivas iteraciones para añadir detalle y mejorar la aplicación.

3.3 Conclusiones

El significativo aumento de la producción científica en el campo de la HE lleva pareja la problemática del acceso a estas investigaciones que, en la actualidad, están dispersas en diferentes bases de datos generalistas (Scopus, Web of Science, Dialnet o Redalyc). Además, estas bases de datos se limitan a los metadatos bibliográficos, y no catalogan la producción según los criterios de la historia de la educación, como temas específicos del campo o periodos histórico, lo que dificulta el análisis de la investigación producida y la cartografía del campo.

El presente capítulo se ha centrado en el diseño e implementación de una base de datos *online* para el almacenamiento centralizado de información bibliográfica de artículos publicados en HE que, además, permite el catalogado y etiquetado con información cualitativa al respecto de las temáticas y las épocas estudiadas.

Para ello se ha recurrido a la metodología DBR en cuatro fases, que se articula alrededor de dos iteraciones, con un diseño anterior y un análisis posterior en cada una, lo que le da un carácter iterativo y permite mejorar la herramienta en función de la información obtenida en cada una de las fases.

Las ventajas que aporta esta herramienta a la comunidad de investigadores de HE son las siguientes:

- Ofrece un lugar centralizado en el que consultar la producción científica en HE, sin tener que buscar en diferentes bases de datos generalistas a la hora de identificar, por ejemplo, los estudios que se han realizado sobre una temática concreta.
- Aporta la información cualitativa específica del campo, como son las temáticas y las épocas estudiadas, lo que es un primer paso de cara a cartografiar y estudiar la disciplina en términos globales según sus parámetros específicos.

- Es un punto de partida para establecer una comunidad de investigadores de HE, que podría establecerse como un punto de encuentro virtual, de tal forma que se pudiera acceder a las últimas investigaciones publicadas.

Sin embargo, el ofrecer la posibilidad de catalogar la producción científica con información cualitativa, como son las temáticas y las épocas estudiadas en los artículos, plantea una nueva problemática, en este caso de índole práctico: esta catalogación se debe hacer manualmente ya que, tal como se indicó en el capítulo 2, requiere acceder a los contenidos del artículo (no solo a su información bibliográfica externa) para interpretar e identificar dichos datos y registrarlos en Hecumen.

Dada la enorme producción en la disciplina, esta tarea puede ser ingente a nivel de recursos humanos, o del tiempo de gran cantidad de investigadores. Es por esto que, de nuevo, acercándonos a las ciencias de la computación, se han explorado otro tipo de soluciones de la mano de la IA y la clasificación automática de textos. El siguiente capítulo desarrollará este punto en profundidad.

Capítulo 4: Inteligencia artificial para automatizar la clasificación de artículos en historia de la educación

4.1 Introducción

El diseño de la base de datos Hecumen, descrita en el capítulo previo, permite almacenar en un único lugar tanto información proveniente de los registros bibliográficos como información cualitativa específica de HE, como son las temáticas y las épocas estudiadas, de tal forma que los artículos quedarían catalogados según los parámetros de la HE.

El problema radica en que, al tratarse de información cualitativa, este catalogado debería ser realizado manualmente por diferentes investigadores en función de la información bibliográfica del artículo (título, palabras clave y resumen). Esto podría convertirse en una tarea enorme con una importante dedicación de tiempo, ya que los investigadores deberían revisar los artículos manualmente uno a uno para etiquetar la información. Sin embargo, el empleo de tecnologías basadas en la IA puede ayudar en este proceso automatizándolo.

Como marco filosófico en esta aproximación a la IA, Wittgenstein, en su primera etapa dentro de la filosofía analítica, le dio una preponderancia al lenguaje de la lógica, estableciendo un isomorfismo entre el lenguaje de la lógica y el mundo, de tal forma que el lenguaje de la lógica permitía representar el mundo en su totalidad, a excepción de esta misma capacidad de representación, que solo podía ser mostrada (Wittgenstein, 1989). En consecuencia, este lenguaje de la lógica permitía describir los dos únicos tipos de enunciados con sentido: los abstractos de la matemática, y los concretos y empíricos del mundo, lo que establece un temprano paralelismo que podemos observar en la actualidad en los sistemas de IA: una lógica subyacente que se utiliza tanto para describir los procesos matemáticos y computacionales, como los fenómenos del mundo que trata de clasificar y predecir. De hecho, si añadimos a este contexto tanto el trabajo de Boole como de Shannon al mundo de la computación, el primero en su búsqueda de un lenguaje simbólico y el segundo en cuando a la realización física de los operadores lógicos, todo el desarrollo de la IA se fundamenta en el lenguaje de la lógica formal que todavía no ha podido ser abandonado (Mira et al., 2003).

En cuanto a su definición, Dobrev (2012) afirma que la IA sería aquel programa informático que en un mundo arbitrario (una representación informática simplificada de determinados aspectos del mundo) no daría unas respuestas peores que las que podría dar un ser humano. Esta definición, a pesar de ser muy genérica, resulta interesante ya que introduce el concepto de comparación con las respuestas humanas, lo que refiere al concepto de imitación que más adelante se expondrá con el planteamiento de Alan Turing.

Pero antes de hablar de la aportación concreta de Turing al campo de la IA, debemos exponer sus planteamientos, que sentaron las bases de la ciencia de la computación. Sus aportaciones fueron fundamentales al desarrollar un sistema que permitiera modelizar cualquier tipo de cálculo: las máquinas de Turing, propuestas en 1936, consisten en un modelo teórico compuesto por una cinta en la que hay instrucciones y datos indistintamente, un cabezal que puede desplazarse por la cinta leyendo y escribiendo datos, así como un mecanismo de control que en cada momento puede encontrarse en una serie finita de estados (Brookshear, 1993). Más aún, en base a este modelo conceptual, teorizó su idea de máquina de Turing universal, que consistía en una máquina de Turing que pudiera ser capaz de ejecutar de una manera genérica otras máquinas de Turing, codificadas en la cinta, y que estuvieran especializadas en la resolución de determinados problemas concretos. Este planteamiento es la base del funcionamiento de cualquier microprocesador moderno que, de una manera genérica, ejecuta las instrucciones contenidas en un programa que está especializado en una tarea concreta.

En cualquier caso, su planteamiento teórico fue independiente de la posible implementación física; de hecho, su desarrollo fue anterior al diseño de los ordenadores, pero tal como continúa Brookshear (1993), estableció un límite superior en cuanto a la potencia computacional de cualquier sistema computacional presente o futuro. Esto es la llamada tesis de Turing, que afirma que el poder computacional de una máquina de Turing es tan grande como el de cualquier sistema computacional posible. En este sentido, un algoritmo de IA, puesto que es un tipo de proceso computacional,

tendrá como límite superior en cuanto a potencia de cálculo el modelo de la máquina de Turing.

Aunque este planteamiento de Turing aplica a las bases comunes de la computación (independientemente de que se trate de IA o no), no es de extrañar que también realizara una aportación explícita al campo de la IA: el test de Turing. Esta prueba reformula la cuestión filosófica sobre si las máquinas podrían pensar algún día a algo mucho más empírico y analizable, esto es, si parece que piensen; con este enfoque, Turing realiza una reducción conductista de la IA asemejándola a un juego de imitaciones (Mira et al., 2003), lo que conecta con la definición de IA propuesta anteriormente de Dobrev.

Si, tal como se ha expuesto con la tesis de Turing, la IA y los procesos computacionales tradicionales tienen una misma cota superior a nivel de potencia computacional, pero la primera puede resolver problemas que los segundos no, ambos enfoques deben tener una serie de diferencias, a pesar de su fundamento computacional común. Una de estas diferencias radica en la posibilidad de que pueda existir una solución algorítmica o no. La IA puede abordar problemas para los que no hay una solución algorítmica desde la computación tradicional y, en caso de haberla, sea ineficiente por su explosión combinatoria, tal como afirman Rich y Knight (1991). En este enfoque, el concepto autoprogramable por aprendizaje es uno de los elementos clave que establece la diferencia entre un problema de IA o uno computacional resoluble mediante programación tradicional. De hecho, Mira et al. (2003) establecen que la condición para que podamos hablar de IA es que tiene que haber algún tipo de aprendizaje, ya sea simbólico o conexionista, y este se produce en base a unos algoritmos establecidos.

4.2 Algoritmos para clasificación y predicción

En términos generales, la IA se puede clasificar en dos grandes ramas: la simbólica, que tuvo su auge en los años 70 del siglo XX y que se caracterizó por los sistemas basados en conocimientos y los sistemas expertos, y la conexionista, auto-programable por aprendizaje y en la que el conocimiento está en la propia estructura de la red. La IA conexionista trata de superar las dificultades que se encuentra la IA simbólica en tareas como el reconocimiento de caracteres, la percepción o la memoria asociativa entre otras (Mira et al., 2003).

Un algoritmo suele ser definido como una máquina abstracta (Moschovakis, 2001) que, tras una serie de pasos secuenciales, ofrece unos resultados que pueden estar en función de unas variables de entrada, lo que enlaza con la noción del modelo conceptual de la máquina de Turing expuesto previamente. De una manera más genérica, Knuth (1997) lo define como una serie finita de reglas y operaciones que resuelven un problema específico.

Cuando el problema a resolver es la clasificación de un conjunto de datos de entrada en diferentes grupos, encontramos una serie de algoritmos específicos utilizados ampliamente en estudios de índole educativo para tareas como la predicción de abandono de un curso o la clasificación del rendimiento académico entre otros. Entre los algoritmos utilizados con más frecuencia en este tipo de estudios encontramos el J48, Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), ANN – Multilayer Perceptron (MLP) o Naïve Bayes (NB) (Landa et al., 2021; Castrillón et al., 2020; Gil et al., 2018; Jokhan et al., 2022; Mourdi et al., 2020).

Los algoritmos DT, RF y J48 se engloban dentro del grupo de algoritmos basados en árboles de decisión que, tal como afirman Gil et al. (2018), se basan en una serie de reglas obvias e identificables, como por ejemplo aquellas que permiten realizar una predicción meteorológica; así, las entradas del algoritmo son siempre el mismo tipo de variable (por ejemplo, entrada 1: presión atmosférica, entrada 2: temperatura…). A pesar de que, en el caso de la clasificación de textos, las reglas del sistema puede que

no resulten tan obvias e identificables, este tipo de algoritmos se utilizó de una manera pionera en el estudio de Apté et al. (1994).

El caso del clasificador NB fue utilizado en la investigación de Kamruzzaman (2010), precisamente para la clasificación de textos. Tal como explica el autor, este tipo de algoritmo se basa en el teorema de Bayes, y se fundamenta en la asunción de que el efecto del valor de un atributo (por ejemplo, la aparición de una palabra determinada) es independiente de los valores de otros atributos (la aparición de otras palabras distintas), lo que se conoce como la hipótesis de independencia.

Por su parte, en el algoritmo LR se asume que la salida sea una combinación lineal de las entradas, modeladas a través de una función logística o sigmoide (Gutiérrez et al., 2020), lo que hace que la transición entre 0 (no pasa el filtro) y 1 (sí que lo hace) sea suave, sin saltos abruptos, tal como queda descrita por su expresión:

$$\phi(x) = \frac{1}{1 + e^{-x}}$$

En el estudio realizado por Prabhat y Khullar (2017) se comparó la precisión de los modelos LR y NB a la hora de clasificar los sentimientos expresados en textos *online*, dando como resultado un 10,1% más de precisión con el modelo de LR. Sin embargo, en el trabajo de Aborisade y Anwar (2018), orientado a la detección de la autoría de mensajes publicados en la red social Twitter, la mejora en la precisión de LR respecto a NB se reducía a un 1,3% de diferencia.

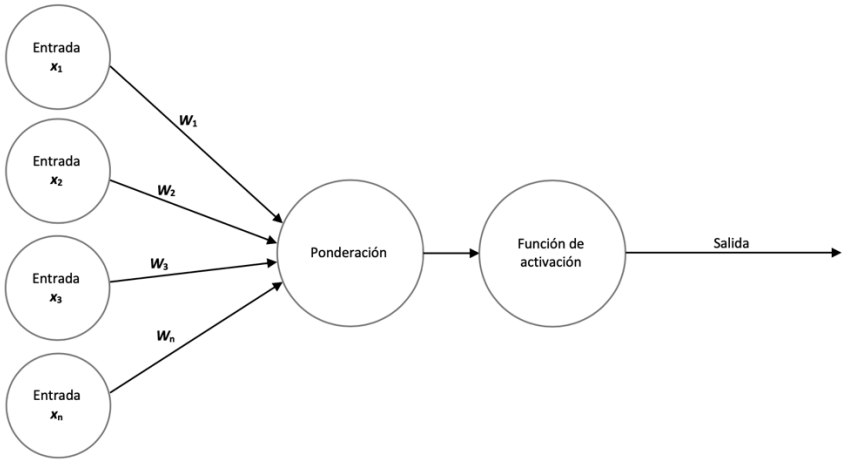
El caso del algoritmo MLP consiste en una simplificación del funcionamiento de las redes neuronales biológicas, cuyo funcionamiento básico fue sistematizado desde el punto de vista de la lógica proposicional por McCulloch y Pitts en 1943. Si una neurona, desde el punto de vista biológico, está compuesta por un cuerpo central (soma), una serie de ramificaciones de entrada (dendritas) y una larga propagación llamada axón que conecta o no de una manera química, eléctrica o mixta con las dendritas de otras neuronas (sinapsis) en función de la información que llega desde las dendritas (Mira et

al., 2003), una neurona artificial imita este modo de funcionamiento de la siguiente forma:

- Cada una de las dendritas pasa a ser un valor de entrada.
- El soma, a través de una función matemática, pondera la información de entrada en función de unos pesos.
- El axón se convierte en valor de salida, que dependerá de una función de activación que se aplica al resultado de la función del soma.

El modelo referido, llamado *perceptron* e ideado por Rosenblatt (1958), se puede observar en la figura 19.

Figura 19. Esquema de perceptron. Fuente: elaboración propia



La ponderación no es más que el sumatorio de cada una de las entradas (x_1, x_2, x_3… x_n) multiplicadas por sus respectivos pesos (w_1, w_2, w_3…w_n), tal como refiere la siguiente expresión:

$$\sum_{i=1}^{n} x_i w_i$$

El resultado de esta función de ponderación es pasado por la función de activación, que es la que determinará si hay salida o no que propagar a la siguiente

neurona artificial. Aunque hay diversas funciones de activación que se pueden utilizar para permitir o no el paso de la información de ponderación, según Glorot et al. (2011) las más frecuentes son la función sigmoidea y la tangente hiperbólica. Estos modelos se basaban en un funcionamiento de "todo o nada", en el sentido de que la salida de cada neurona artificial finalmente sería 1 (en caso de activarse) o 0 (en caso de permanecer inactiva), lo que determinaba un umbral de activación, y permitiría su tratamiento desde el punto de vista de la lógica proposicional (McCulloch y Pitts, 1943). Sin embargo, una neurona artificial aislada no tiene demasiada potencia computacional, lo que nos lleva al funcionamiento del algoritmo MLP, que consistiría en la interconexión en paralelo de un gran número de neuronas simples organizadas en una arquitectura multicapa (Aldabas-Rubira, 2002; Mira et al., 2003). En términos generales nos encontramos ante tres capas de neuronas: capa de entrada, capa(s) oculta(s) y capa de salida. El flujo básico de funcionamiento comenzaría con los datos que se le dan a la capa de entrada, que pasan a la capa o capas ocultas, que son las que procesan estos datos, para finalmente dar una respuesta en la capa de salida.

Si volvemos a la afirmación expuesta con anterioridad por Mira et al. (2003) en la que se afirmaba que la condición para que podamos hablar de IA es que tiene que haber algún tipo de aprendizaje, los algoritmos presentados no deben dar siempre una misma respuesta, sino que esta respuesta cambiará en función del aprendizaje al que sea expuesto, lo que es conocido como entrenamiento. En este proceso se parte de un conjunto de valores de entrada, para los cuales ya se conoce cuál debe ser su salida. Al algoritmo se le presenta esta información y, en base a su diseño, adaptará su estructura interna para establecer una relación entre dichas entradas y salidas. La finalidad es que, una vez ha configurado su estructura interna según este proceso específico de entrenamiento, sea capaz de dar unos datos de salida con el mayor grado de precisión posible para una nueva muestra de valores de entrada, de los cuales no se conoce la salida esperada. Este es el motivo por el cual un algoritmo de IA necesita de una cantidad importante de datos de entrada, con sus respectivos valores de salida esperados, para realizar este proceso de entrenamiento, y ser capaz de dar salidas adecuadas para futuros valores no entrenados, lo que es conocido como generalización (Aldabas-Rubira, 2002).

4.3 Diseño del clasificador

Para el problema que nos ocupa de clasificación de textos de HE en diferentes categorías específicas, los datos de entrada que son necesarios para poder entrenar a una IA es una cantidad de artículos previamente clasificados por investigadores, de tal forma que la IA pueda aprender en base a estos resultados conocidos de antemano, para poder generalizar su clasificación a futuros artículos no catalogados. Desde este punto de vista, y en base a todo el modelo propuesto hasta el momento, el problema de la clasificación de un artículo en una categoría determinada consistiría en definir un conjunto de variables de entrada, que serían la entrada a la IA y que, tras los mecanismos de aprendizaje del algoritmo, darían como salida la categoría a la que pertenece el artículo.

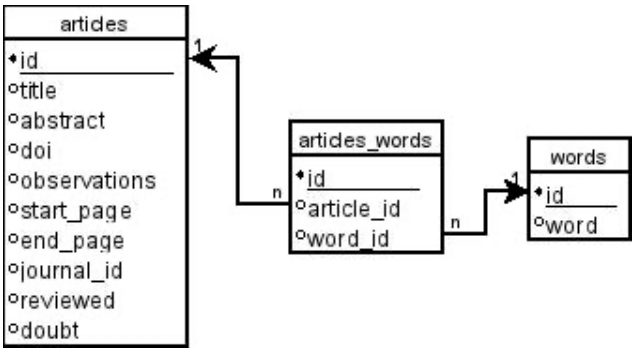
Las variables de entrada deberían ser las mismas que consideraría un investigador a la hora de clasificar un artículo determinado, a saber, el título, el resumen y las palabras clave. Sin embargo, a diferencia de un investigador, la IA no va a hacer una valoración cualitativa del significado de estas palabras, sino que se va a centrar en aspectos cuantitativos, con lo que una primera fase del proceso debe consistir en traducir la información bibliográfica de cada artículo en una serie de datos cuantificables. Para ello se proponen los siguientes pasos:

- Obtención de los campos título, resumen y palabras clave de cada uno de los artículos. Puesto que la capacidad de predicción puede variar en función de las diferentes permutaciones de estos tres datos, se almacenarán independientemente.
- Eliminación de los signos de puntuación de la cadena de texto resultante: punto, coma, punto y coma, signos de admiración y de interrogación, guiones…
- Eliminación de las palabras vacías de significado (artículos, pronombres, preposiciones…), así como las genéricas propias del campo que aparecen en la mayoría de los artículos y que no aportan valor a la discriminación (historia, educación, historia de la educación…)

- Dividir la cadena resultante según los espacios. Esto daría como resultado una secuencia de palabras.
- Almacenar estas palabras en una base de datos relacional, de tal forma que no se repitan palabras que ya se hayan insertado previamente, relacionándolas con los artículos en los que aparecen. Cada una de las palabras pasa a tener un valor numérico único.

Para este último punto podemos elaborar un diagrama ER con las tablas necesarias para almacenar la frecuencia de aparición de palabras por artículo. Nos encontramos ante una cardinalidad N-N, puesto que cada artículo contiene varias palabras y, al mismo tiempo, cada palabra puede aparecer en más de un artículo. Así pues, sería necesaria una tabla intermedia que relacionara cada artículo con sus palabras, tal como se puede observar en la figura 20.

Figura 20. Esquema ER para almacenamiento de palabras.



Fuente: elaboración propia.

Si aplicamos este proceso de codificación a todos los artículos que hayan sido catalogados previamente por un investigador dentro de una categoría determinada, obtendríamos un listado de todas las palabras que aparecen en la información bibliográfica de cada uno de los artículos de una misma categoría, lo que nos permitiría ordenarlas de mayor a menor frecuencia de aparición. Este listado de palabras de mayor frecuencia de aparición podría considerarse la entrada de IA, pero el número de entradas de la IA debería ser fija y predefinida, con lo que habría que seleccionar las *n* palabras cuya frecuencia de aparición es mayor cada uno de los artículos de una

categoría. Este valor de *n* podría variar la capacidad de predicción del sistema, al igual que, tal como se ha mencionado previamente, el uso de diferentes permutaciones de título, resumen y palabras clave.

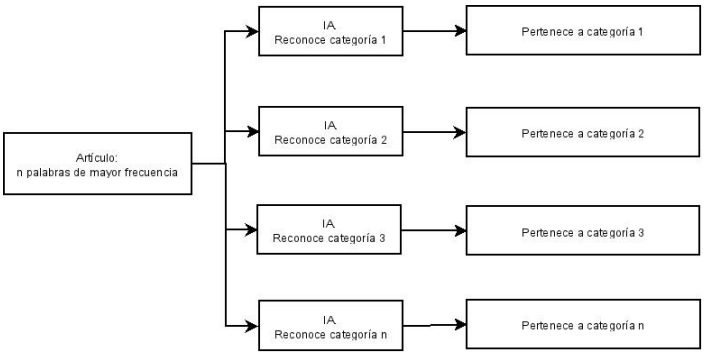
Este planteamiento se basa en la ley de Zipf, o ley del mínimo esfuerzo, que describe la relación entre las palabras de un texto y el orden de serie de estas palabras. Si, tal como se plantea en esta ley, en una lista de frecuencias de aparición en orden decreciente se multiplica la posición de una palabra determinada por su frecuencia, el resultado es una constante (Araujo, 2006). Esto establecería que en un artículo se tiende a utilizar el mínimo número de palabras diferentes, lo que refuerza el uso de estas palabras más frecuentes para clasificar el artículo.

El proceso de entrenamiento consistiría en suministrar a la IA el vector de las *n* palabras más frecuentes por cada uno de los artículos, aportando también la categoría previamente escogida por los investigadores. El proceso de aprendizaje automático establecería relaciones entre las entradas y la salida lo que configuraría un patrón de importancia de estas palabras de mayor frecuencia de aparición en la categoría, pero con la salvedad que este patrón no es dado por el investigador, sino que es un producto de la IA, lo que permitiría establecer relaciones entre términos que hacen que el artículo pueda categorizarse en una u otra categoría.

En lo que respecta a la clasificación en sí misma, este enfoque nos plantea una primera decisión de diseño, que estribaría entre crear una única IA que fuera capaz de clasificar un artículo entre una o varias categorías, o bien tener diferentes IAs, cada una de ellas especializada en una categorización concreta. El hecho de que las variables independientes vayan a ser las mismas (las *n* palabras de mayor frecuencia de aparición en los metadatos del artículo), hace que no haya una diferencia cualitativa entre dos artículos; adicionalmente existe el condicionante de que cada artículo puede estar incluido en diferentes categorías, lo que abogaría por un planteamiento de múltiples IA, cada una especializada en la clasificación de los artículos en una categoría determinada. En este caso el diseño de la IA sería exactamente el mismo, independientemente de la categoría a la que fuera a estar dirigida, lo que cambiaría sería el proceso de

entrenamiento, de tal forma que a cada IA se le suministraría un conjunto de las *n* palabras de mayor frecuencia en los artículos de la misma categoría que pretende identificar. Como resultado, de cada IA emergería un patrón de relevancia de palabras relacionadas que son las que definirían la categorización del artículo. En la figura 21 puede observarse el flujo de este procedimiento.

Figura 21. Procedimiento de clasificación.



Fuente: elaboración propia.

4.4 Implementación

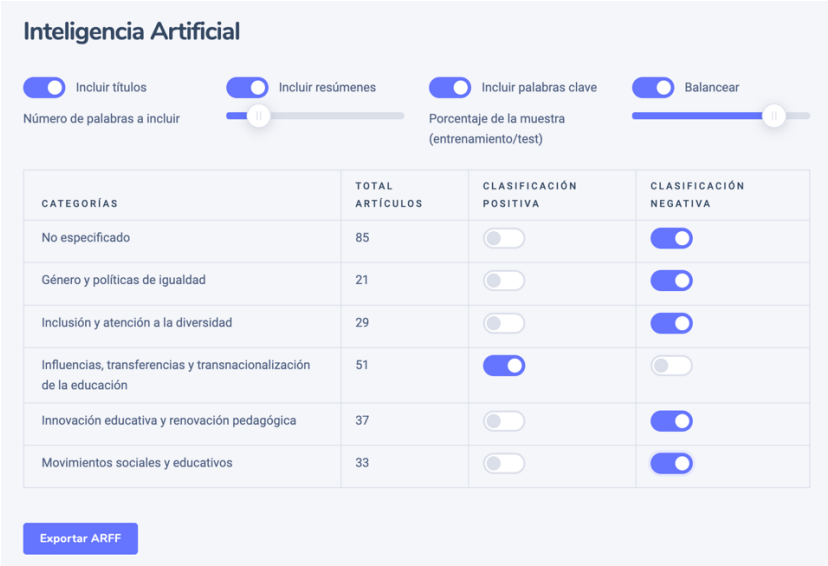
Para la aplicación práctica del diseño descrito se ha recurrido a diferentes herramientas. Por una parte, los datos para el entrenamiento de la red neuronal han sido extraídos de la base de datos Hecumen, cuyo diseño e implementación han sido descritos en el capítulo 3. Esta herramienta tiene la capacidad de centralizar la información bibliográfica de los artículos de diferentes revistas del campo de HE, etiquetados manualmente por investigadores del campo según la categoría a la que pertenecen. Para el desarrollo del presente capítulo se han añadido unas funcionalidades a la base de datos Hecumen para poder extraer los datos según el formato exigido por la siguiente herramienta, Weka, desarrollada en la Universidad de Waikato (Nueva Zelanda), y liberada como código abierto (https://www.cs.waikato.ac.nz/ml/weka/). Esta aplicación permite aplicar una serie de algoritmos de clasificación y predicción inteligente a una muestra dada (Frank et al., 2016).

4.4.1 Extracción de datos de Hecumen

Para poder extraer la información para el entrenamiento de la IA, se han realizado una serie de modificaciones en la base de datos Hecumen consistentes en el procesado de los metadatos de los artículos, así como en el exportado de un tipo de archivo de datos llamado ARFF, requerido por Weka. Para tal efecto se ha programado una interfaz específica que permite seleccionar qué categorías incluir para la clasificación positiva (pertenece a una categoría) y para la clasificación negativa (no pertenece a una categoría). El procedimiento consiste en seleccionar como clasificación positiva la categoría para la que queremos entrenar la IA, y como clasificación negativa, al menos, una de las categorías restantes. El poder seleccionar diferentes categorías que formarán parte de la clasificación negativa permitirá analizar si el rendimiento de la IA mejora en función de esta selección. Por otra parte, se ha implementado una opción de balancear las categorías resultantes El problema de los datos desbalanceados es uno de los mayores retos a la hora de la clasificación de categorías por parte de una IA (Su et al., 2006), ya que se produce un sesgo en el que se la efectividad del algoritmo es mayor

para la categoría mayoritaria, mientras que decrece para la minoritaria (Jiang et al., 2012). En caso de encontrarse frente a datos desbalanceados, se ha implementado la técnica del *undersampling* (Lemaître et al., 2017), de tal forma que se eliminan aleatoriamente elementos de la categoría mayoritaria para acercarla al número de elementos de la minoritaria. Además, la interfaz permite seleccionar la inclusión o no de los metadatos título, resumen y palabras clave, lo que permite realizar diferentes permutaciones a la hora de diseñar los experimentos de clasificación. También permite extraer los datos en dos archivos, uno para el entrenamiento y otro para el test, asignando un porcentaje de artículos a cada archivo, junto con el número de *n* palabras con mayor frecuencia a tener en cuenta. El resultado dentro de la misma herramienta Hecumen puede observarse en la figura 22.

Figura 22. Interfaz de Hecumen para la extracción del archivo ARFF.



Fuente: elaboración propia.

Tras el proceso descrito en el apartado 4.3, el resultado es un archivo ARFF en el que, tras un bloque de cabecera, se registran, en cada línea, los datos de cada uno de los artículos. Por cada artículo hay tantos números separados por comas como las *n* palabras de mayor frecuencia en los metadatos seleccionados. Así, cada número identifica a una palabra única. Al final de la línea aparece la categorización *yes* o *no*, que corresponde a las categorías que hayamos escogido como clasificación positiva y

negativa respectivamente. El resultado de un archivo ARFF de ejemplo generado, para n = 10, se puede encontrar a continuación:

177,315,286,360,67,328,297,27,362,300,no

104,171,17,45,554,565,273,566,115,567,no

647,513,645,654,87,183,357,680,537,104,no

718,108,717,705,728,580,721,719,478,49,no

1072,198,328,269,407,1122,1112,97,1139,556, no

113,1212,182,485,20,738,653,1238,38,198,no

796,444,1211,1458,198,97,1454,308,470,1187,no

1591,1593,118,1594,1595,1006,104,924,702,158,no

17,1591,188,1889,692,549,1890,1916,1929,923,no

337,477,177,955,845,2005,1498,982,1066,1107,no

En lo que respecta a su equivalente con las palabras originales en lugar de sus códigos numéricos el resultado sería el siguiente:

state,schools,primary,politicians,after,funding,financial,support,private,Dutch,no

war,peace,education,have,Introduction,aspects,general,theme,developed,detail,no

textbook,narratives,textbooks,multidirectional,future,past,new,interact,authors,war,no

priority,pedagogy,over,Priority,Part,schooling,monitorial,discovery,literature,two,no

administration,educational,funding,had,histories,no,Funding,school,global,special,no

child,protection,political,ideas,social,more,concept,societal,international,educational,no

writing,pedagogic,evolution,inspection,educational,school,configuration,century,within,forms,no

art,exhibitions,War,refugee,childrens,cultural,war,German,World,re,no

education,art,culture,critical,case,studies,possibilities,Read,consciousness,about,no

national,language,state,curricula,identity,states,Swiss,nation,Switzerland,became,no

4.4.2 Entrenamiento de la inteligencia artificial

En el momento de la extracción del archivo ARFF para el entrenamiento (1 de junio de 2022), el número de artículos catalogados temáticamente en Hecumen era de 256, con el desglose de categorías que puede analizarse en la tabla 3.

Tabla 3. Número de artículos clasificados en Hecumen por categoría.

Categoría	**Artículos**
No especificado	85
Género y políticas de igualdad	21
Inclusión y atención a la diversidad	29
Influencias, transferencias y transnacionalización de la educación	51
Innovación educativa y renovación pedagógica	37
Movimientos sociales y educativos	33

Fuente: elaboración propia.

Cabe mencionar que la categoría “no especificado” incluye a todos aquellos artículos que los investigadores no han podido clasificar en ninguna de las otras, y a todos los efectos se considera una categoría en sí misma.

Para la generación de los archivos ARFF se han escogido las 10 palabras más frecuentes, que han sido extraídas de todos los metadatos (título, resúmenes y palabras clave). Las unidades de significado que implican más de una palabra (por ejemplo, "política educativa", puesto que están formadas por palabras que siempre aparecen juntas, implica una frecuencia de aparición similar, con lo que este enfoque tiene en cuenta de una manera implícita dichas unidades de significado. Con esos parámetros se han extraído 6 archivos ARFF, uno por cada una de las categorías. La categoría correspondiente ha sido marcada como clasificación positiva, mientras que las 5 restantes han sido consideradas como clasificación negativa. Esta estrategia daba como resultado una muestra desbalanceada, tal como se puede analizar en la tabla 4. La categoría que más desbalanceada resultaba era “Género y políticas de igualdad” (un 8,20% de artículos de la categoría frente al 91,8% de artículos con otras categorías), mientras que la que menos apenas pasaba el 33%. Así pues, se decidió activar la opción de balancear la muestra que, como se ha descrito previamente, aplicaba un algoritmo de *undersampling* para eliminar aleatoriamente elementos de la categoría mayoritaria para equiparala a la minoritaria, de tal forma que la muestra final se aproximara al balance ideal del 50%.

Tabla 4. Balance de los artículos de Hecumen por categoría.

Categoría	**Artículos clasificación positiva**	**Artículos clasificación negativa**	**Balance**
No especificado	85	171	33,20%
Género y políticas de igualdad	21	235	8,20%
Inclusión y atención a la diversidad	29	227	11,33%
Influencias, transferencias y transnacionalización de la educación	51	205	19,92%
Innovación educativa y renovación pedagógica	37	219	14,45%
Movimientos sociales y educativos	33	223	12,89%

Fuente: elaboración propia.

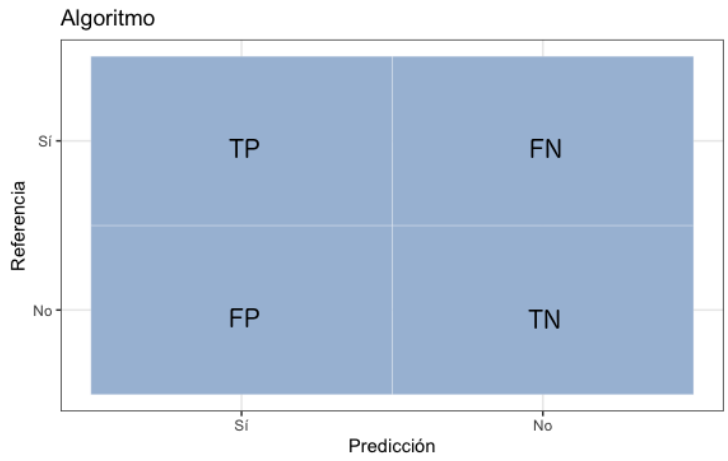
A continuación, se han seleccionado los algoritmos con los que se iban a entrenar cada una de las IA. Tal como se ha expuesto en el apartado 4.2, los algoritmos RF, LR, MLP, NB, J48 y DT han sido ampliamente utilizados en literatura científica del campo de la educación (Landa et al., 2021; Castrillón et al., 2020; Gil et al., 2018; Jokhan et al., 2022; Mourdi et al., 2020), con lo que se va a comparar su rendimiento para el problema de la clasificación de artículos de HE.

Para analizar el rendimiento de cada uno de los algoritmos se han cuantificado las siguientes variables, resultantes de aplicar los algoritmos a la muestra de cada una de las categorías:

- Positivos verdaderos (TP): tanto investigador como algoritmo coincidieron en que el artículo formaba parte de la categoría.
- Negativos verdaderos (TN): el investigador y el algoritmo coincidieron en que el artículo no formaba parte de la categoría.
- Falsos positivos (FP): el algoritmo identificó al artículo dentro de la categoría, mientras que el investigador no.
- Falsos negativos (FN): el investigador clasificó el artículo dentro de la categoría, pero el algoritmo no.

Estos valores se han representado en matrices de confusión, en las que se representan estos 4 datos según el siguiente esquema representado en la figura 23.

Figura 23. Esquema de la matriz de confusión.



Fuente: elaboración propia.

Para la validación de los datos se ha escogido el procedimiento de valoración cruzada con 10 conjuntos. En el mismo, la muestra se divide en 10 subconjuntos más pequeños y, en cada iteración, se utiliza uno de los subconjuntos como test y el resto como entrenamiento (Alshaikh et al., 2021).

En base a estos datos se han calculado, por cada uno de los algoritmos y categorías, las siguientes métricas:

- Exactitud (*accuracy*): permite obtener la tasa de aciertos (tanto positivos como negativos) que ha logrado el algoritmo. Esta métrica funciona adecuadamente cuando el número de muestras positivas y negativas para el entrenamiento están balanceadas (Ruiz y Srinivasan, 2002), lo que se ha logrado, tal como se ha descrito previamente, mediante la técnica del *undersampling*. La expresión que describe esta métrica es la siguiente:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precisión (*precision*): esta métrica hace referencia al número de clasificaciones positivas reales de la clase respecto a todas las realizadas por el algoritmo, es decir, a la capacidad del algoritmo para identificar positivamente la categoría. Su fórmula es:

$$Precision = \frac{TP}{TP + FP}$$

- Exhaustividad (*recall*): mide la capacidad del algoritmo para la predicción correcta de la categoría, y se representa de la siguiente manera:

$$Recall = \frac{TP}{TP + FN}$$

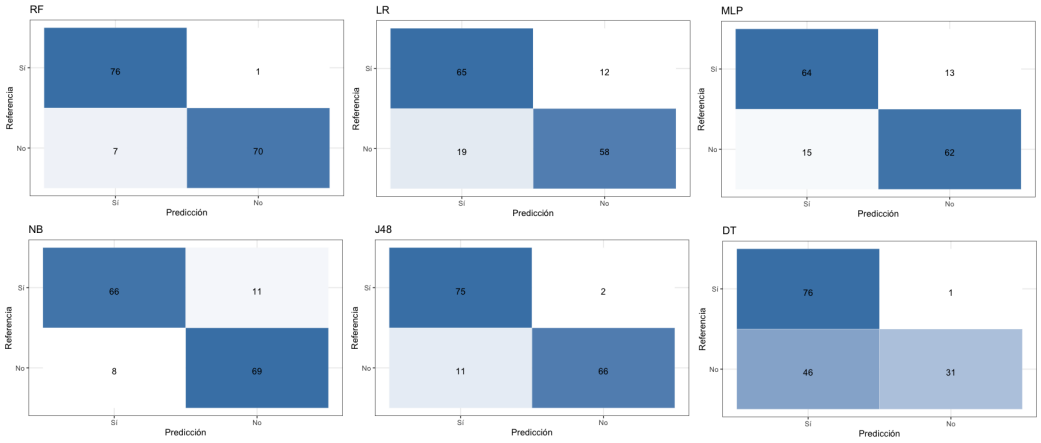
- Valor-F (*f1-score*): es el resultado de la media armónica entre las métricas de *precision* y *recall*, calculada según la siguiente expresión:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.5 Resultados

Tal como se ha descrito previamente, se ha generado un archivo ARFF por cada una de las categorías. Para los parámetros de exportado se han seleccionado las 10 palabras con mayor frecuencia de aparición entre título, resumen y palabras clave, así como un balanceo de la muestra mediante la técnica del *undersampling*. Cada uno de estos archivos se ha introducido en Weka para obtener las variables de predicción con cada uno de los 6 algoritmos escogidos (RF, LR, MLP, NB, J48 y DT), lo que se ha representado en diferentes matrices de confusión. Las matrices correspondientes a cada una de las categorías pueden observarse en las figuras 24, 25, 26, 27, 28 y 29. En base a estos datos se han calculado las métricas de *accuracy*, *precision*, *recall* y *f1-score* con las fórmulas descritas previamente por cada uno de los algoritmos. Los resultados pueden analizarse en las tablas 5, 6, 7, 8, 9 y 10.

Figura 24. Matrices de confusión para la categoría "No especificado".



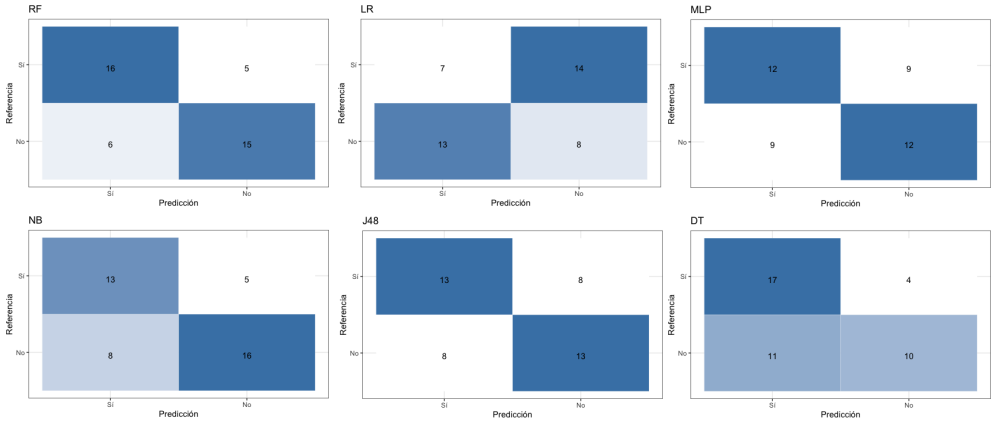
Fuente: elaboración propia.

Tabla 5. Métricas de los algoritmos para la categoría "No especificado".

Algoritmo	Accuracy	Precision	Recall	F1-score
RF	0,95	0,92	0,99	0,95
LR	0,80	0,77	0,84	0,80
MLP	0,82	0,81	0,83	0,82
NB	0,88	0,89	0,86	0,87
J48	0,92	0,87	0,97	0,92
DT	0,69	0,62	0,99	0,76

Fuente: elaboración propia.

Figura 25. Matrices de confusión para la categoría "Género y políticas de igualdad".



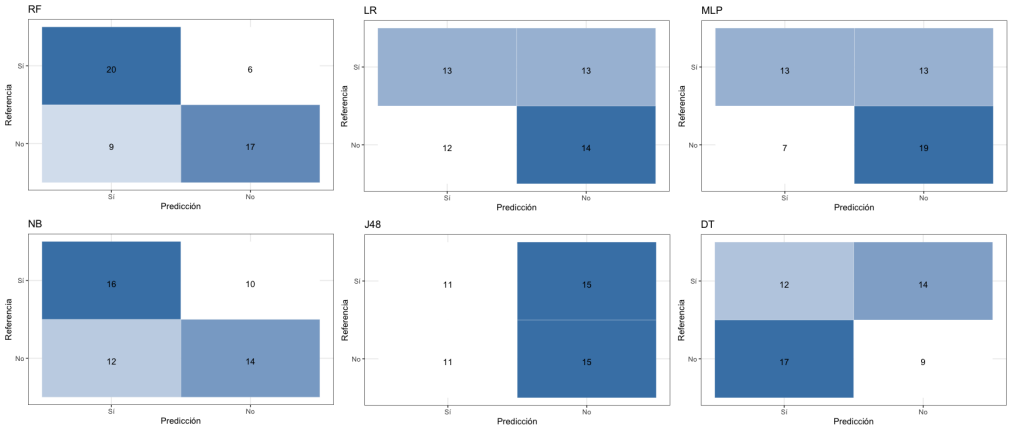
Fuente: elaboración propia.

Tabla 6. Métricas de los algoritmos para la categoría "Género y políticas de igualdad".

Algoritmo	Accuracy	Precision	Recall	F1-score
RF	0,74	0,73	0,76	0,74
LR	0,36	0,35	0,33	0,34
MLP	0,57	0,57	0,57	0,57
NB	0,69	0,62	0,72	0,67
J48	0,62	0,62	0,62	0,62
DT	0,64	0,61	0,81	0,70

Fuente: elaboración propia.

Figura 26. Matrices de confusión para la categoría "Inclusión y atención a la diversidad".



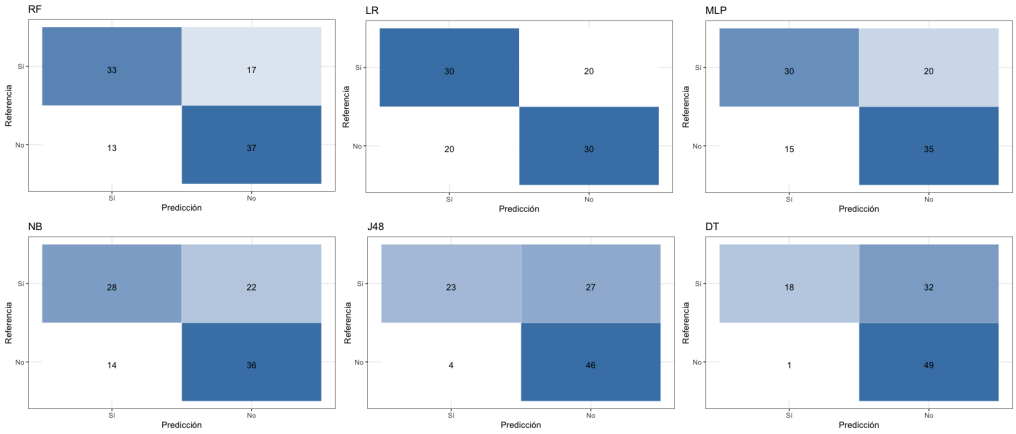
Fuente: elaboración propia.

Tabla 7. Métricas de los algoritmos para la categoría "Inclusión y atención a la diversidad".

Algoritmo	**Accuracy**	**Precision**	**Recall**	**F1-score**
RF	0,71	0,69	0,77	0,73
LR	0,52	0,52	0,50	0,51
MLP	0,62	0,65	0,50	0,57
NB	0,58	0,57	0,62	0,59
J48	0,50	0,50	0,42	0,46
DT	0,40	0,41	0,46	0,43

Fuente: elaboración propia.

Figura 27. Matrices de confusión para la categoría "Influencias, transferencias y transnacionalización de la educación".



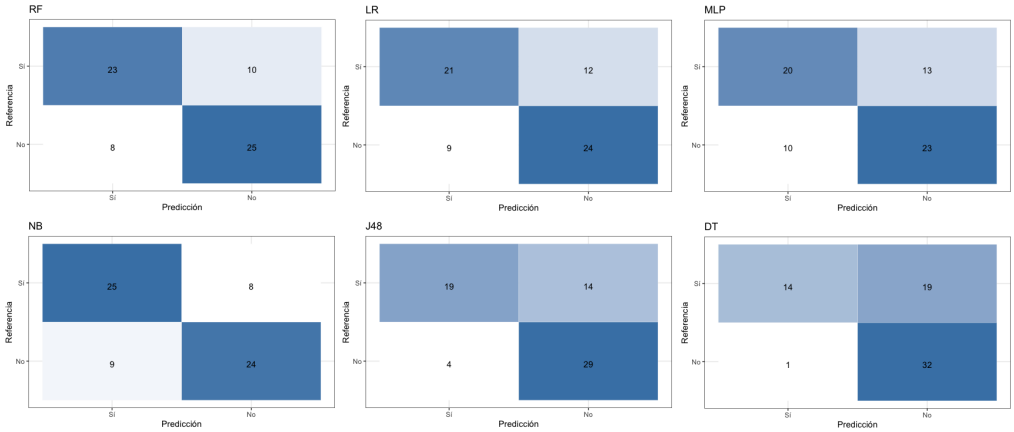
Fuente: elaboración propia.

Tabla 8. Métricas de los algoritmos para la categoría "Influencias, transferencias y transnacionalización de la educación".

Algoritmo	**Accuracy**	**Precision**	**Recall**	**F1-score**
RF	0,70	0,72	0,66	0,69
LR	0,60	0,60	0,60	0,60
MLP	0,65	0,67	0,60	0,63
NB	0,64	0,67	0,56	0,61
J48	0,69	0,85	0,46	0,60
DT	0,67	0,95	0,36	0,52

Fuente: elaboración propia.

Figura 28. Matrices de confusión para la categoría "Innovación educativa y renovación pedagógica".



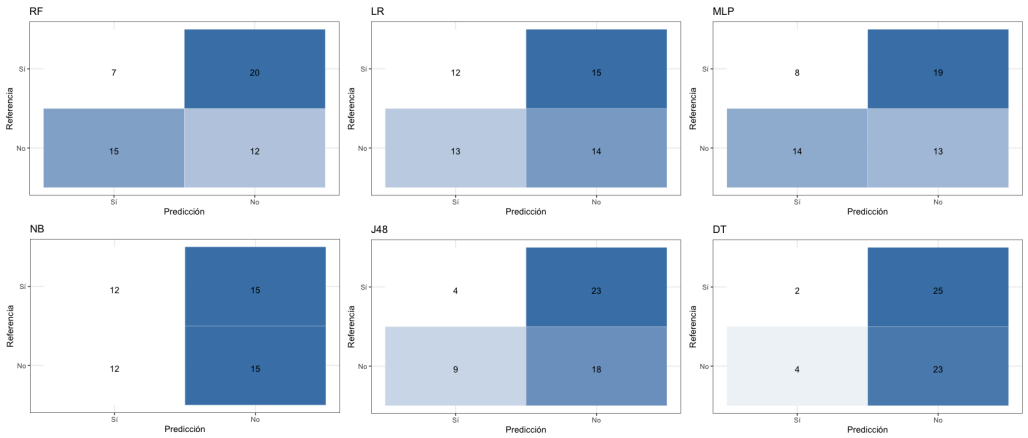
Fuente: elaboración propia.

Tabla 9. Métricas de los algoritmos para la categoría "Innovación educativa y renovación pedagógica".

Algoritmo	**Accuracy**	**Precision**	**Recall**	**F1-score**
RF	0,73	0,74	0,70	0,72
LR	0,68	0,70	0,64	0,67
MLP	0,65	0,67	0,61	0,64
NB	0,74	0,74	0,76	0,75
J48	0,73	0,83	0,58	0,68
DT	0,70	0,93	0,42	0,58

Fuente: elaboración propia.

Figura 29. Matrices de confusión para la categoría "Movimientos sociales y educativos".



Fuente: elaboración propia.

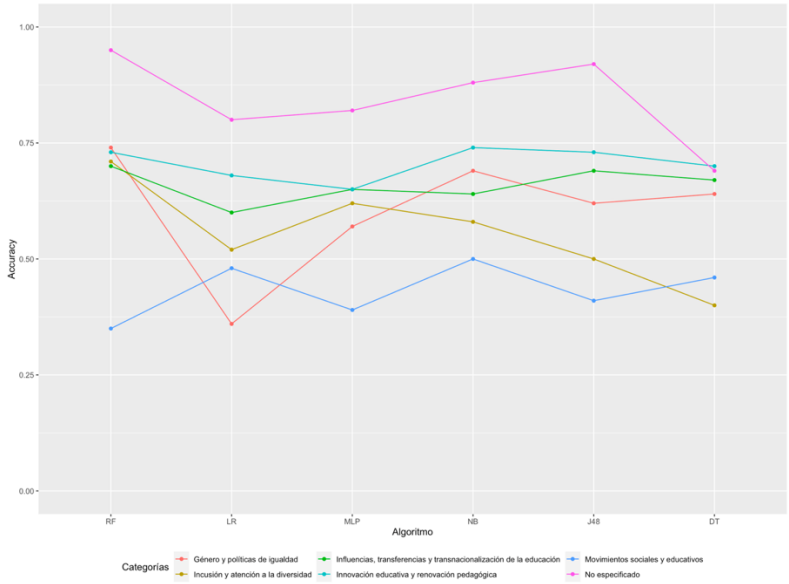
Tabla 10. Métricas de los algoritmos para la categoría "Movimientos sociales y educativos".

Algoritmo	**Accuracy**	**Precision**	**Recall**	**F1-score**
RF	0,35	0,32	0,26	0,29
LR	0,48	0,48	0,44	0,46
MLP	0,39	0,36	0,30	0,33
NB	0,50	0,50	0,44	0,47
J48	0,41	0,31	0,15	0,20
DT	0,46	0,33	0,07	0,12

Fuente: elaboración propia.

Si atendemos exclusivamente a la métrica *accuracy* que, al tratarse siempre de muestras balanceadas refleja de una manera global el funcionamiento del algoritmo, en la figura 30 se puede observar comparativamente la exactitud de los diferentes algoritmos con cada una de las categorías a identificar.

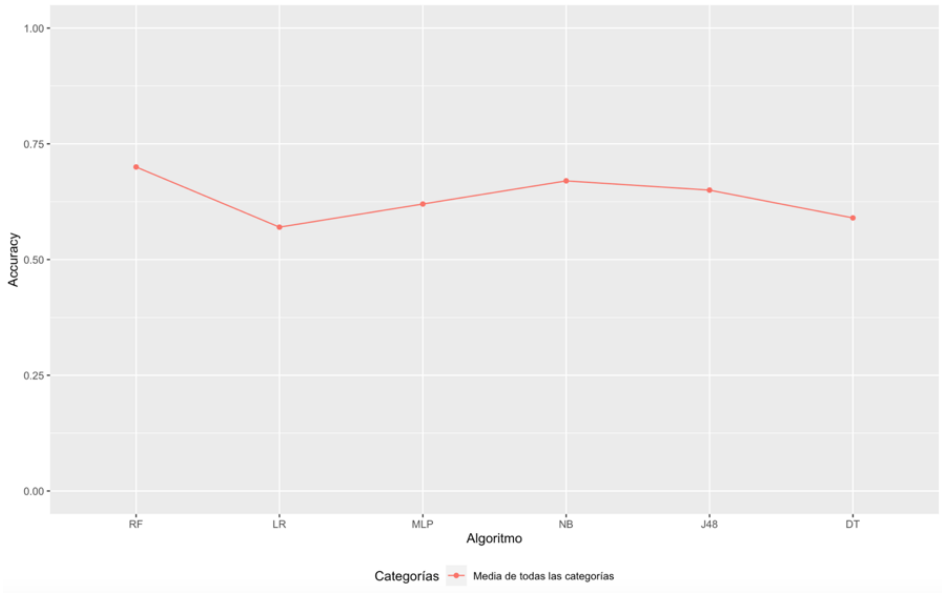
Figura 30. Comparativa de los valores de accuracy en función del algoritmo y la categoría.



Fuente: elaboración propia.

Con el fin de obtener un valor unificado por cada uno de los algoritmos para poder estimar el que mejor rendimiento ofrece a la hora de identificar las categorías, se ha calculado la media de los diferentes valores de *accuracy* en cada una de las categorías. El algoritmo que, de media, ha obtenido mejor resultado a la hora de clasificar todas las categorías ha sido RF (0,7), seguido de NB (0,67), J48 (0,65), MLP (0,62), DT (0,59) y LR (0,57). Estos datos se han plasmado gráficamente en la figura 31.

Figura 31. Comparativa de los valores de accuracy medios de todas las categorías.



Fuente: elaboración propia.

4.7 Conclusiones

La clasificación manual de la producción científica en el campo de HE según las categorías específicas del campo puede convertirse en una tarea ingente, puesto que los investigadores deberían revisar y catalogar uno a uno todos los artículos. La IA permite automatizar este proceso tras un proceso de entrenamiento, que no es otro que la ingesta de un número *n* de artículos ya catalogados, lo que permitiría al sistema inferir las características que debe tener un artículo para ser clasificado de una manera u otra.

El presente capítulo se ha centrado en el diseño e implementación de un sistema de IA que permite automatizar el proceso de clasificación y catalogado de artículos del campo de HE según categorías específicas del campo. Para tal efecto se ha partido de los fundamentos teóricos de la IA, así como de los algoritmos utilizados con más frecuencia en la investigación educativa, y se han realizado una serie de modificaciones a la base de datos Hecumen para poder generar los archivos de entrenamiento en base a títulos, resúmenes y palabras clave de los artículos que, posteriormente, se han suministrado a la herramienta Weka para el entrenamiento de diferentes IAs.

Los resultados muestran que, a pesar de haber realizado el entrenamiento con una muestra reducida (n = 256), el algoritmo RF ha obtenido unos buenos resultados a la hora de clasificar las categorías, con un *accuracy* de 0,7 y una *precision*, *recall* y *f-measure* de 0,69.

Tanto este sistema de clasificación de artículos mediante IA como la misma base de datos Hecumen, descrita en el capítulo 3, serán utilizadas para el estudio bibliométrico del próximo capítulo.

Capítulo 5: La historia de la educación a través de las revistas especializadas. Temáticas, producción científica y bibliometría (1961-2022)

5.1 Introducción

El campo de estudio de la HE ha sufrido un significativo aumento en cuanto a número de investigaciones a lo largo de las últimas 3 décadas. A pesar de que los motivos subyacentes a este aumento, tal como aducen Hernández-Huerta, Payá y Sanchidrián (2019) tienen que ver con la maduración del campo, la irrupción de la edición y distribución electrónica y la exigencia de publicaciones para las acreditaciones de los investigadores, el fenómeno exige de un análisis en profundidad de tal forma que se pueda obtener una perspectiva general del campo de la HE.

Por una parte, el análisis debe aspirar a una visión global, incluyendo un número importante de publicaciones o, al menos, seleccionadas según unos criterios de relevancia en el panorama internacional. Además, sin apartarnos de esta visión global, se deben establecer relaciones y conexiones entre publicaciones, autores, países o contenidos (Hofstetter y Huitric, 2020; Hofstetter et al., 2014; Hernández y Cagnolati, 2015). En este sentido, los estudios bibliométricos, tal como afirman Ellegaard y Wallin (2015), permite identificar el corpus de producciones científicas dentro de un área de conocimiento determinada. En contraste con otros métodos como las revisiones sistemáticas o los meta-análisis, permite resumir la estructura intelectual y bibliométrica de un campo analizando las relaciones sociales y estructurales entre los datos de diferentes investigaciones (Donthu et al., 2021).

Los estudios bibliométricos en el campo de HE no son muy frecuentes. En base a la cadena de consulta *TITLE-ABS-KEY ("history of education" bibliometric)*, Scopus arroja 3 resultados, los cuales ponen el foco en un ámbito muy localizado, tanto geográfica como temáticamente. *About curriculum history: themes, concepts and references of Brazilian researches* (Meira, 2020) se centra en el currículum de Brasil, mientras que *Teaching Drawing through the textbooks of High School Education published in Spain (1915-1990): Bibliometric study of contents* (Cardona, 2010) y *The reception of new education in Spain by means of manuals on the history of education for teacher training colleges (1898-1976)* (Ruiz et al., 2006) se circunscriben a España,

centrando sus investigaciones en unas temáticas concretas (enseñanza de dibujo y manuales respectivamente).

Desde estas coordenadas, especialmente debido al aumento de la producción y a la progresiva internacionalización y construcción de redes transnacionales de investigación en el campo de la HE (Payá et al., 2016; Hernández y Payà, 2022), resulta pertinente plantear un estudio bibliométrico global que permita analizar, tanto su evolución, como el estado de la cuestión de la investigación científica en HE para poder afrontar los retos futuros en el campo con la mayor información posible.

Como novedad y para poder complementar el estudio bibliométrico con los datos específicos del campo de la HE, se pondrán en práctica las herramientas desarrolladas en el marco de la presente tesis, y que han sido descritas en los capítulos 3 y 4.

5.2 Método

El objetivo principal de este estudio se focaliza en el análisis de la producción científica en el campo de la HE en términos cuantitativos desde una perspectiva bibliométrica, así como de la información cualitativa específica del campo. Para tal efecto se han seguido el procedimiento propuesto por Donthu et al. (2021) que consiste en una secuencia de 4 pasos: definición del propósito y alcance de la investigación, selección de las técnicas de análisis bibliométrico, recolección de los datos y realización del análisis bibliométrico en sí mismo y documentación de los hallazgos.

En lo que respecta al primer paso, el propósito de la investigación es analizar la producción científica en el campo de la HE dentro del ámbito de 11 publicaciones especializadas, que han sido seleccionadas según el criterio propuesto por Hernández et al. (2019), como es el hecho de que su gestión editorial se haya profesionalizado en aras de estar acreditadas en el ámbito internacional. El listado propuesto original estaba compuesto por 12 publicaciones, pero la revista *Historical Studies in Education/ Revue d'histoire de l'éducation* no aparece indexada en Scopus, con lo que se eliminó de la muestra objeto de estudio, cuyo listado puede observarse en la tabla 11. En esta tabla, además, figura el país de origen en el que la revista es editada. En el curso de la investigación se constató que los artículos de algunos años no estaban indexados en Scopus, cuestión que también queda reflejada en la misma tabla.

Tabla 11. Revistas incluidas en la muestra.

Revista	**País**	**Años no indexados**
Childhood in the Past: An International Journal	Reino Unido	-
Espacio, Tiempo y Educación	España	-

Histoire de l'éducation	Francia	-
História da Educação	Brasil	-
Historia Social y de la Educación/ Social and Education History	España	-
Historia y Memoria de la Educación	España	-
History of Education & Children's Literature	Italia	-
History of Educacion. Journal of the History of Education Society	Reino Unido	1998; 2002-2004
History of Education Quarterly	Reino Unido	1968-1970; 1973; 1976-1977; 1979; 1982; 1987-1998; 2001; 2004-2007
History of Education Review	Reino Unido	1981; 1983-2011
Paedagogica Historica: International Journal of the History of Education	Reino Unido	1986-1989; 2000; 2005

Fuente: elaboración propia.

Las técnicas de análisis bibliométrico seleccionadas han sido el análisis de la producción a lo largo de los años por cada una de las fuentes, el estudio de citaciones globales, los autores y las colaboraciones entre los mismos, los países de afiliación, instituciones y colaboraciones, junto con los idiomas. También se ha llevado a cabo un análisis temático a través del estudio de las palabras clave más frecuentes y su evolución. Específicamente, gracias al desarrollo de Hecumen y de su funcionalidad para catalogar temáticas y épocas estudiadas, se ha realizado un análisis temático y otro de las épocas estudiadas.

Para los dos últimos pasos del procedimiento, se han exportado la totalidad de los artículos indexados en Scopus el 21 de marzo de 2022 y, posteriormente, se han incorporado en la base de datos Hecumen, cuyo desarrollo ha sido descrito en el capítulo 3, y que permite tanto la grabación de datos básicos bibliográficos, como otros dispersos y contingentes que no comparten todos los artículos. Además, permite el etiquetado de información cualitativa específica de la HE, como son las categorías de los artículos, épocas o periodos históricos estudiados. Así pues, Hecumen ha sido la herramienta utilizada para la realización del análisis bibliométrico.

5.3 Resultados

5.3.1 Análisis de la producción y fuentes

En primer lugar, se han extraído las métricas relacionadas con las publicaciones en términos generales. La muestra está compuesta por un total de 5217 publicaciones, desglosadas según la clasificación de Scopus en *article* (n = 3879 y 74,35%), *review* (n = 1084 y 20,78%), *editorial* (n = 86 y 1,65%), *conference paper* (n = 76 y 1,46%), *note* (n = 69 y 1,32%), *erratum* (n = 15 y 0,29%), *letter* (n = 7 y 0,13%) y *article in press* (n = 1 y 0,02%).

Un total de 4053 autores han contribuido a la producción de esta muestra en 6642 apariciones. Un total de 4022 artículos fueron escritos por un único autor (77,10%), mientras que 1195 trabajos se firmaron por varios autores (22,90%). En base a estos datos, se calcularon los índices de autores por artículo (0,78), los artículos por autor (1,29) y el índice de co-autores (cociente resultante de dividir las apariciones de los autores entre el total que artículos), que se situó en 1,27. Además, se calculó el índice de colaboración (Elango y Rajendran, 2012), situado en 0,44, que es el cociente entre los autores de documentos firmados por varios autores entre los artículos elaborados por varios autores. Los datos de este análisis descriptivo pueden observarse en la tabla 12.

Tabla 12. Análisis descriptivo de publicaciones y autores.

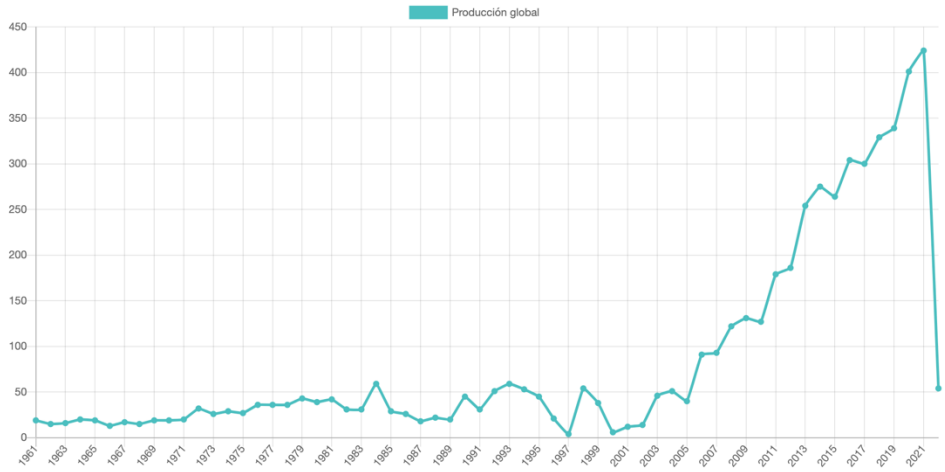
Dato	**Valor**
Publicaciones	5217
Autores	4053

Aparición de autores	6642
Artículos escritos por un único autor	4022
Artículos escritos por varios autores	1195
Autores de artículos escritos por un único autor	3526
Autores de artículos escritos por varios autores	527
Índice de autores por artículo	0,78
Índice de artículos por autor	1,29
Índice de co-autores	1,27
Índice de colaboración	0,44

Fuente: elaboración propia.

A continuación se ha analizado la evolución de la producción global a lo largo de los años. La horquilla de años analizada es la que ofrecen los datos indexados por Scopus, cuyo primer artículo data de 1961 y el resultado se ha representado gráficamente en la figura 32.

Figura 32. Evolución de la producción global. Nota: datos recogidos hasta el 21 de marzo de 2022.



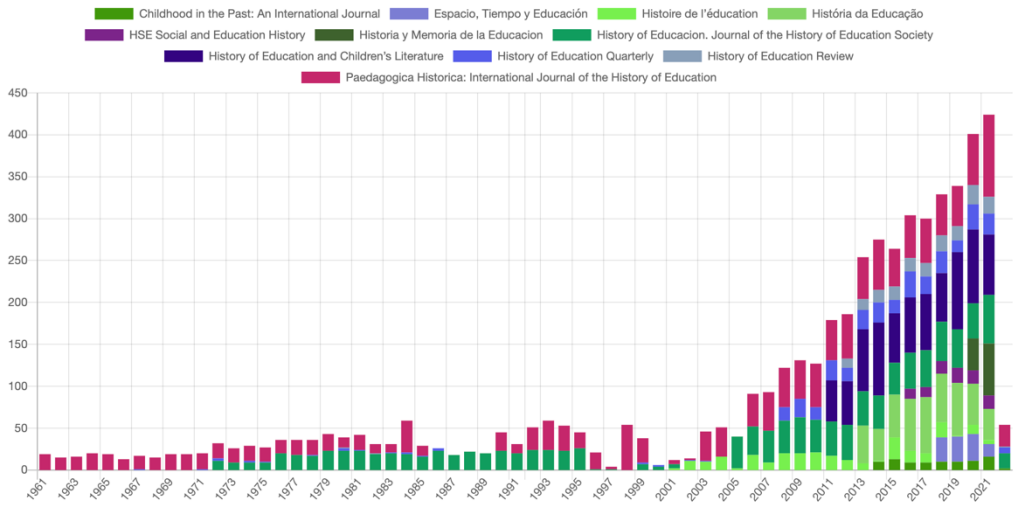
Fuente: elaboración propia.

Tal como se puede observar en la figura 32, tras un periodo de crecimiento discreto desde 1961 hasta 2006, la producción científica se ajusta a la ley de Price de crecimiento exponencial (Price, 1963) desde 2006 hasta la actualidad, de tal manera que la producción en el campo se duplicó cada 7 años aproximadamente.

Sin embargo, se debe tener en cuenta que cada revista comenzó su labor editorial en diferentes momentos, con lo que el número de revistas que contribuían a la producción científica en HE cada año variaba.

Para ilustrar esta variación de número de fuentes a lo largo de los años, se ha confeccionado la figura 35 en la que se puede analizar tanto el año de comienzo de la labor editorial de cada revista, como el número de publicaciones con las que contribuyeron en cada momento.

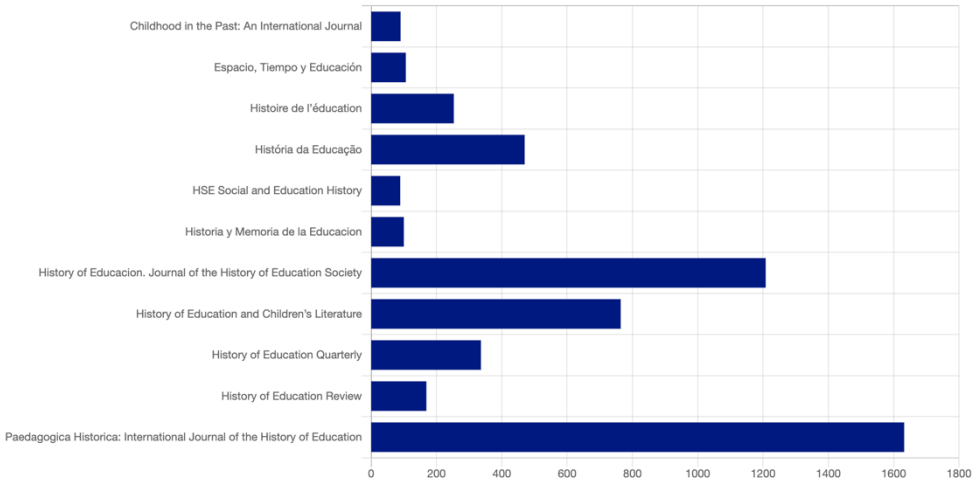
Figura 33. Producción por revista y año. Nota: datos recogidos hasta el 21 de marzo de 2022.



Fuente: elaboración propia.

Teniendo en cuenta esto, la aportación de cada revista en términos globales a la producción científica analizada en la muestra está encabezada por las revistas *Paedagogica Historica: International Journal of the History of Education* e *History of Educacion. Journal of the History of Education Society*, con 1632 documentos (31,28%) y 1208 (23,16%) respectivamente. En tercer lugar encontramos a *History of Education and Children's Literature* (n = 764 y 14,64%), seguida de *História da Educação*, con 470 documentos y un 9,01% de la muestra. El resto de revistas están por debajo del 10% de la producción total de la muestra y se pueden consultar en la figura 34.

Figura 34. Producción por revista. Fuente: elaboración propia



Fuente: elaboración propia.

5.3.2 Análisis de las citaciones

Una vez analizada la muestra, tanto en términos globales como en aportaciones por cada una de las publicaciones, se ha procedido a descender el nivel de análisis para centrarse en los artículos, que han sido cribados en función del dato de las citaciones globales que han recibido. Desde esta óptica, los tres artículos que más citaciones globales han recibido han sido *Social darwinistic ideas and the development of women's education in England, 1880–1920* (Dyhouse, 1976), *The examination, disciplinary power and rational schooling* (Hoskin, 1979) y *A new education for a new era: The contribution of the conferences of the New Education Fellowship to the disciplinary field of education 1921–1938* (Brehony, 2004).

El listado de los 10 artículos más citados, junto con el número de citas globales que han recibido, los autores y la revista donde fue publicado se puede encontrar en la tabla 13.

Tabla 13. Artículos más citados globalmente.

Artículo	**Citaciones**	**Autores**	**Revista**
Social darwinistic ideas and the development of women's education in England, 1880–1920	80	Dyhouse (1976)	History of Educacion. Journal of the History of Education Society
The examination, disciplinary power and rational schooling	72	Hoskin (1979)	History of Educacion. Journal of the

			History of Education Society
A new education for a new era: The contribution of the conferences of the New Education Fellowship to the disciplinary field of education 1921–1938	64	Brehony (2004)	Paedagogica Historica
The mental hygiene movement, the development of personality and the school: the medicalization of American education	55	Cohen (1983)	History of Education Quarterly
The feminization of teaching in the nineteenth century: A comparative perspective	47	Albisetti (1993)	History of Education. Journal of the History of Education Society
Educational sciences, morality and politics: International educational congresses in the early twentieth century	45	Fuchs (2004)	Paedagogica Historica

The historiography of British Imperial education policy, Part II: Africa and the rest of the colonial empire	44	Whitehead (2005)	History of Education. Journal of the History of Education Society
Child Bioarchaeology: Perspectives on the Past 10 Years	41	Mays et al. (2017)	Childhood in the Past
Men and women and the rise of professional society: the intriguing history of teacher educators	40	Heward (1993)	History of Education. Journal of the History of Education Society
The historiography of British Imperial education policy, Part I: India	40	Whitehead (2005)	History of Education. Journal of the History of Education Society

Fuente: elaboración propia.

5.3.3 Autores y representación por sexo

En lo que respecta a los autores de los artículos analizados, en primer lugar se han identificado aquellos cuya producción ha sido mayor, distinguiendo la misma entre investigaciones realizadas en solitario y aquellas realizadas en colaboración con otros investigadores. Dentro de este último grupo, se han contabilizado aquellos artículos en los que el investigador firmaba el artículo en primer lugar y, en base a estos datos, se ha calculado el factor de dominancia (Kumar y Kumar, 2008), que es el cociente entre las veces en las que ha aparecido como primer firmante entre el número de artículos que ha escrito en colaboración con otros autores. Los datos de los 10 autores más productivos han sido plasmados en la tabla 14.

Tabla 14. Autores más productivos.

Autor	**Artículos**	**Artículos individuales**	**Artículos en colaboración**	**Primer autor**	**Factor de dominancia**
Sani, Roberto	23	9	14	2	0,14
Ascenzi, Anna	18	4	14	14	1
Brickman, William W.	18	18	0	0	-
Caroli, Dorena	12	7	5	4	0,8
Bakker, Nelleke	11	8	3	1	0,33

Grosvenor, Ian	10	2	8	4	0,5
Pomante, Luigiaurelio	10	8	2	0	0
Chiosso, Giorgio	7	4	3	1	0,33
Montecchiani, Sofia	7	6	1	0	0
Bastos, Maria Helena Camara	6	5	1	1	1

Fuente: elaboración propia.

En cuanto a la representación de cada sexo en este listado de los 10 investigadores más productivos en la muestra de datos analizada, encontramos a 5 investigadoras (Anna Ascenzi, Nelleke Bakker, Maria Helena Camara Bastos, Dorena Caroli y Sofia Montecchiani), frente a 5 investigadores (William W. Brickman, Giorgio Chiosso, Ian Grosvenor, Luigiaurelio Pomante y Roberto Sani), lo que sitúa un porcentaje paritario del 50% entre las 10 investigadoras e investigadores más productivos.

También se han analizado aquellas parejas de autores que suelen publicar juntos, así como el número de artículos que han publicado. Para ello se han permutado todas las posibilidades de parejas en todos los artículos firmados por varios autores en un proceso combinatorio sin repetición. El total de las 10 parejas que más artículos han publicado juntos se refleja en la tabla 15.

Tabla 15. Parejas de autores que más han publicado juntos.

Autores	**Total artículos**
Almeida, D.B.; De Azevedo Ramil, C.	3
Beadie, N.; Gottesman, I.	3
Gottesman, I.; Williamson-Lott, J.	3
Roderick, G.W.; Stephens, M.D.	3
Almeida, D.B.; Luchese, T.Â.	2
Anderson, J.D.; Span, C.M.	2
Bakker, N.; Smit, M.	2
Bartholomew, D.J.; Lawn, M.	2
Beadie, N.; Bowman, M.	2
Beadie, N.; Frizell, T.	2

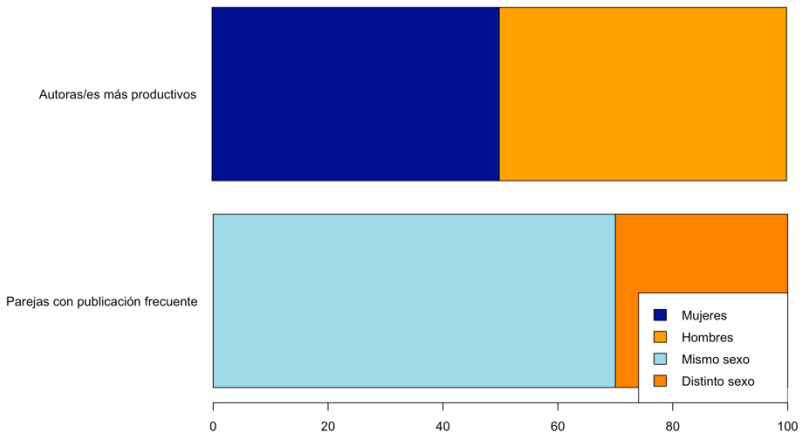
Fuente: elaboración propia.

Siguiendo con el análisis de representación por sexo, pero ahora centrado en las parejas de autores que publican con más frecuencia, destaca que hay el mismo número de investigadoras que de investigadores (con 16 autores distintos en el listado, encontramos a 8 autoras y 8 autores, lo que sitúa porcentajes idénticos de 50%). Las autoras son Almeida, Bakker, Beadie, De Azevedo Ramil, Frizell, Luchese, Smit y Williamson-Lott. Respecto a los autores encontramos a Anderson, Bartholomew, Bowman, Gottesman, Lawn, Roderick, Span y Stephens.

En base a estos datos, se ha analizado la frecuencia de las parejas formadas por hombres-hombres o mujeres-mujeres, frente a las veces en las que se trataba de parejas de mujeres-hombres. De entre las 10 parejas de investigadores que han publicado con mayor frecuencia juntos, en 7 ocasiones la pareja estaba formada por 2 hombres o 2 mujeres, mientras que en los 3 casos restantes las parejas estaban constituidas por mujer-hombre. Esto sitúa las frecuencias en 70% para parejas del mismo sexo, frente a un 30% de distinto sexo.

Este análisis de la representación por sexo se ha representado en la figura 35.

Figura 35. Representación por sexo.



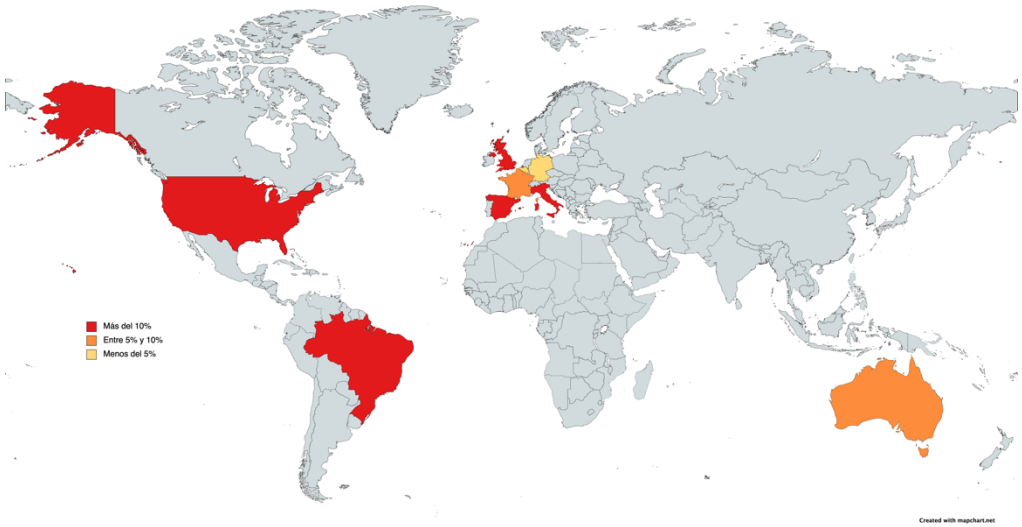
Fuente: elaboración propia.

5.3.4 Países, colaboración entre países y afiliaciones

El análisis de los países se ha realizado siguiendo dos líneas complementarias: en la primera se han contabilizado en términos globales los países de afiliación de todas las apariciones de autores que han producido la muestra. En la segunda, se ha tenido en cuenta solo los autores que figuraban como autores de contacto.

En lo referente a la primera parte del análisis, los 10 países con más aparición de autores, según sus afiliaciones, han sido Reino Unido (833 apariciones de autores), Estados Unidos (n = 713), Brasil (n = 633), España (n = 584), no especificado[1] (n = 538), Italia (n = 524), Francia (n = 346), Australia (n = 327), Alemania (n = 238) y Bélgica (n = 188). El resto de países suman, en total, 1718 apariciones de autores. Su representación porcentual puede observarse en la figura 36.

Figura 36. Aparición de autores por países.



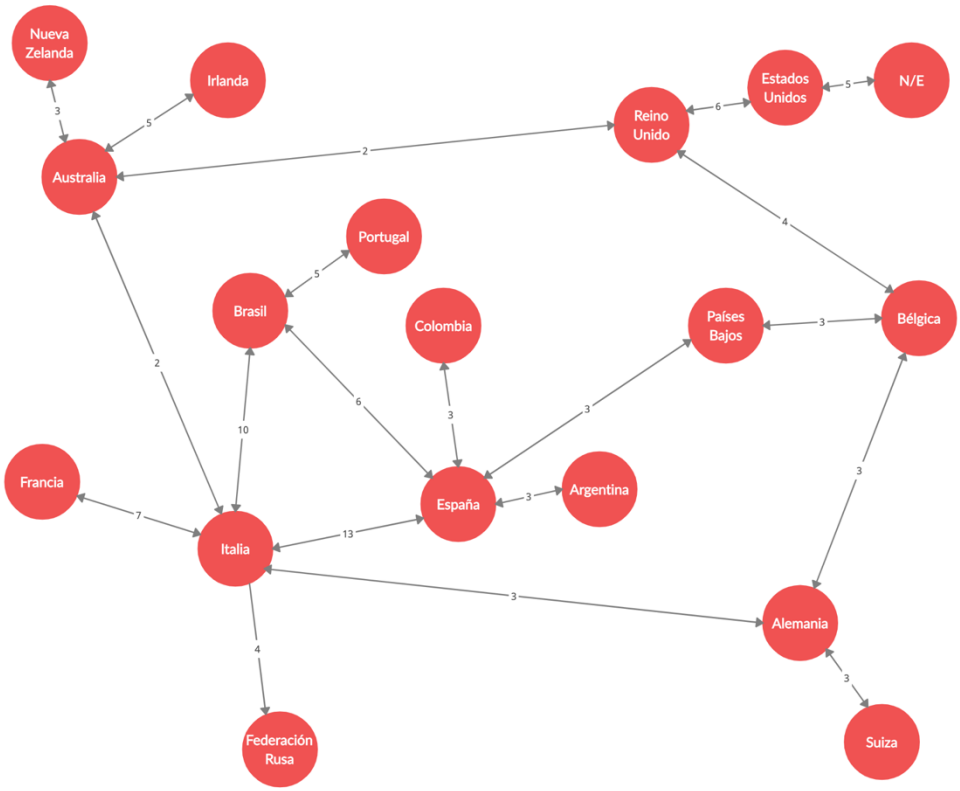
Fuente: elaboración propia.

A continuación se han identificado las parejas de países que han colaborado con mayor frecuencia en la producción de investigaciones. Para ello, tal como se realizó en la colaboración por parejas de autores, se ha realizado un proceso combinatorio sin

[1] La categoría "no especificado" incluye aquellos registros en los que Scopus no aporta ningún tipo de información en el campo de afiliación.

repetición en el que se han permutado todas las parejas de países posibles, lo que ha permitido identificar las 20 parejas de países que han trabajado juntos con más frecuencia. Los resultados quedan plasmados en la figura 37.

Figura 37. Parejas de países que más han colaborado. NOTA: El número en cada flecha indica la cantidad de colaboraciones que han tenido.



Fuente: elaboración propia.

Tal como se ha indicado anteriormente, en la segunda parte de este análisis referido a los países de afiliación y colaboración entre países, se han identificado a los autores que figuraban como investigadores de contacto de cada uno de los artículos. A través de los campos de afiliaciones de estos autores de contacto, se ha confeccionado una tabla con aquellos países que contaban con más artículos con autores de contacto y, adicionalmente, se han contabilizado cuántos de sus artículos se habían escrito con la colaboración de varios países (MCP) o en un solo país (SCP). Los tres primeros países con mayor número de artículos con autor de contacto fueron Reino Unido (n = 641 y 12,27%), Estados Unidos (n = 545 y 10,45%) e Italia (n = 363 y 6,96%). Los valores

absolutos de los 10 primeros países con más artículos con autor de contacto han sido reflejados en la tabla 16 y su representación porcentual gráfica en la figura 38.

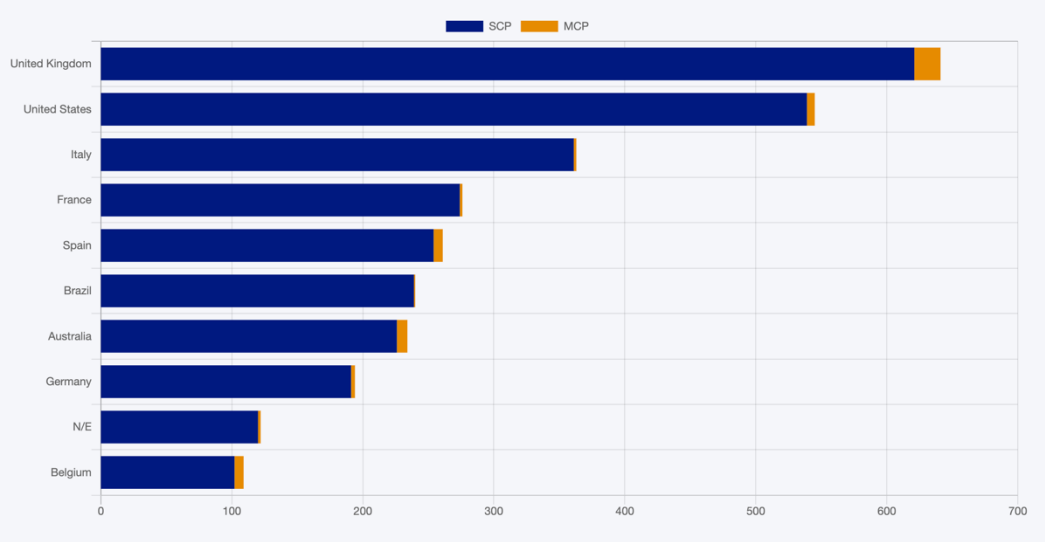
Tabla 16. Países con más artículos con autor de contacto.

País	**Artículos con autor de contacto**	**SCP**	**MCP**
Reino Unido	641	621	20
Estados Unidos	545	539	6
Italia	363	361	2
Francia	276	274	2
España	261	254	7
Brasil	240	239	1
Australia	234	226	8
Alemania	194	191	3
N/E	122	120	2

Bélgica	109	102	7

Fuente: elaboración propia.

Figura 38. Países con más artículos con autor de contacto.



Fuente: elaboración propia.

Para finalizar este apartado, se han estudiado las instituciones según el campo de afiliación de todos los autores firmantes de los artículos. A este respecto cabe mencionar que el primer lugar de la lista se ha descartado, ya que era el que aunaba las apariciones de todos aquellos autores que no habían especificado una institución en su campo de afiliación (n = 500). Además, los puestos 2º (n = 101), 4º (n = 24) y 11º (n = 10) se han unificado, ya que a pesar de estar tipificados de manera diferente se refieren a la misma institución. El resultado de dicho tratamiento de datos para las 10 instituciones con más artículos puede observarse en la tabla 17.

Tabla 17. Instituciones con más artículos.

Institución	**Artículos**
Department of Education, Cultural Heritage and Tourism, University of Macerata, Italy	135
Department of Education, University of Groningen, Groningen, Netherlands	25
School of Education, University of Birmingham, Birmingham, United Kingdom	23
Melbourne Graduate School of Education, The University of Melbourne, Melbourne, Australia	17
School of Education, University College Dublin, Dublin, Ireland	14
Institute of Education, University of London, United Kingdom	13
Department of Education Science, University of Roma Tre, Italy	11

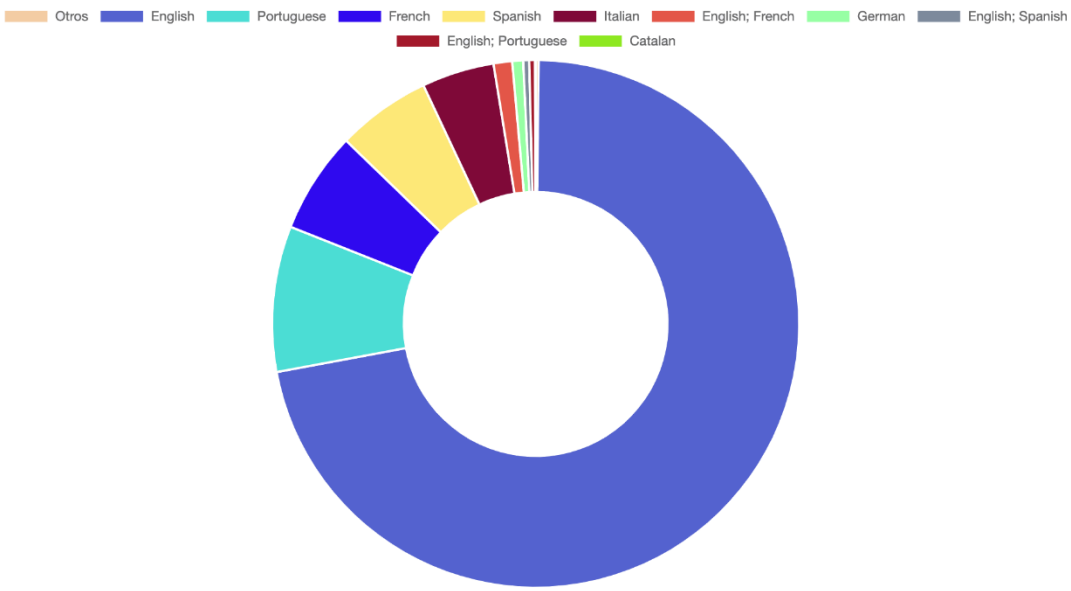
Sydney School of Education and Social Work, University of Sydney, Sydney, Australia	10
Service d'Histoire de l'Éducation (INRP), France	9
Universidade de São Paulo (USP), São Paulo/SP, Brazil	9

Fuente: elaboración propia.

5.3.5 Idiomas

En cuanto a los idiomas que se han utilizado en los artículos de la muestra, el Inglés es el predominante (n = 3795 y 73,76%), seguido del Portugués (n = 390 y 7,58%) y el Francés (n = 304 y 5,91%). La representación porcentual completa de todos los idiomas de la muestra puede analizarse en la figura 39.

Figura 39. Idiomas de la muestra.



Fuente: elaboración propia.

5.3.6 Análisis de las palabras clave

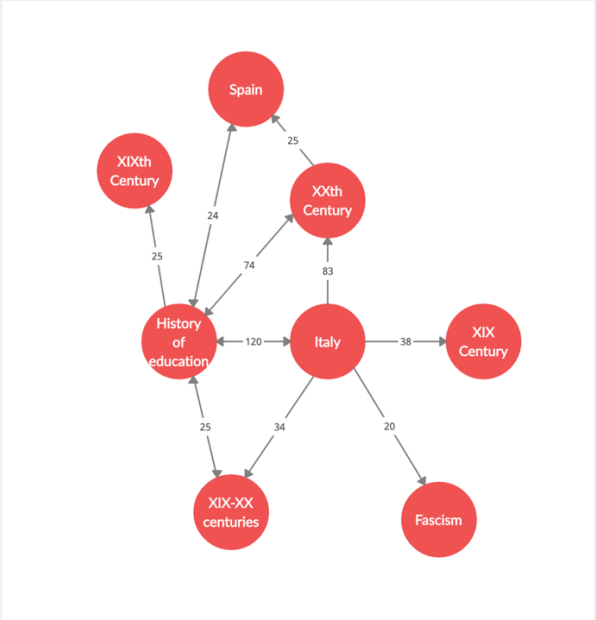
En este apartado se ha procedido al análisis de las palabras clave con las que se había etiquetado cada uno de los artículos.

Para la primera parte del estudio de las palabras clave, se han analizado las parejas de palabras clave de toda la muestra que aparecían con mayor frecuencia siguiendo el procedimiento combinatorio sin repetición descrito anteriormente. Los resultados se han procesado de la siguiente manera:

- Se han unificado las siguientes parejas por similitud: *Italy; XXth Century* (n = 54) e *Italy; XX century* (n = 29). *History of education; XXth Century* (n = 53) e *History of education; XX century* (n = 21).
- Se ha eliminado la pareja *education; History* (n = 47) por su obviedad en la muestra.

El resultado se ha representado gráficamente en la figura 40.

Figura 40. Parejas de palabras clave que aparecen con mayor frecuencia. NOTA: El número en cada flecha indica la cantidad de veces que ha aparecido la pareja.



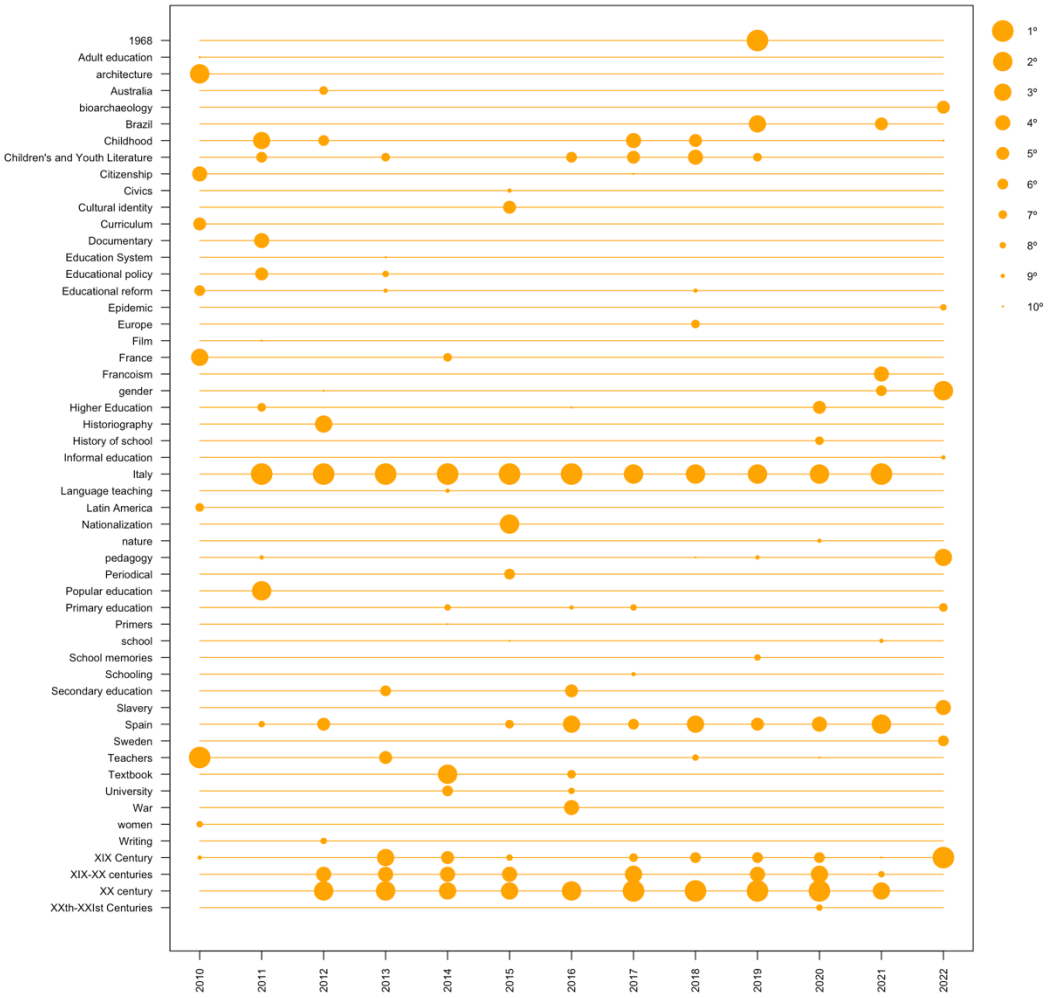
Fuente: elaboración propia.

En la segunda parte del estudio de las palabras clave se ha confeccionado una figura que muestra la evolución de las 10 palabras clave más utilizadas desde el año 2010. Para su realización se han obtenido, en primer lugar, las 10 palabras clave más utilizadas por cada uno de los años, teniendo en cuenta el siguiente procesamiento de datos:

- Se han eliminado las palabras clave *history*, *education* e *history of education* por su obviedad en el campo de estudio.
- Se han unificado las siguientes grupos de palabras clave por similitud: *nineteenth century* y *XIX Century*; *Primary education*, *Primary school* y *Primary schooling*; *teacher education*, *Teacher training* y *Teachers*; *Universities* y *University*; *XIX-XX centuries*, *XIX-XXth Centuries* y *XIXth-XXth Centuries*; *XIX Century* y *XIXth Century*; *XX century* y *XXth Century*; *Children's literature* y *Children's and Youth Literature*; *child* y *Childhood*.

  El resultado de dicha evolución se ha representado gráficamente en la figura 41.

Figura 41. Evolución de las 10 palabras clave más utilizadas desde 2010.



Fuente: elaboración propia.

5.3.7 Análisis temático

Tal como se ha ido exponiendo a lo largo de la presente tesis, los estudios bibliométricos realizados hasta el momento no permiten un análisis de temáticas específicas de HE de una manera automatizada. Esto es así porque dicha información no está contenida en los metadatos bibliográficos de los registros. Sin embargo, el desarrollo de Hecumen, descrito en el capítulo 3, ha permitido el clasificar una serie de artículos, lo que abre las puertas a la identificación de las temáticas específicas más estudiadas en HE.

El número de artículos catalogados para este análisis no fue la totalidad de la muestra, sino la submuestra de aquellos que ya habían sido clasificados. Así pues, a fecha 26 de octubre de 2022, la muestra de artículos catalogados era de n = 1863, lo que supone un 42,64% de la muestra total del estudio bibliométrico. Su desglose por revista puede encontrarse en la tabla 18.

Tabla 18. Desglose de artículos catalogados por revista.

Revista	**Artículos catalogados**
History of Educacion. Journal of the History of Education Society	262
History of Education and Children's Literature	763
History of Education Quarterly	169
História da Educação	467
Paedagogica Historica: International Journal of the History of Education	202

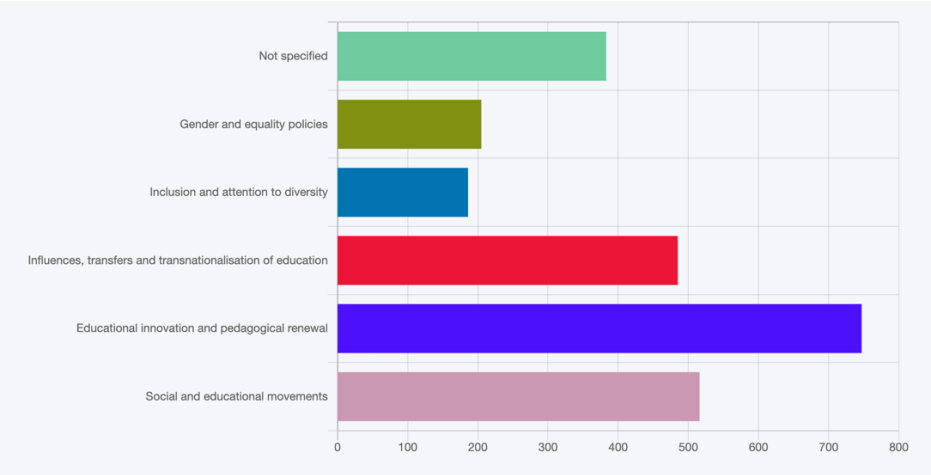
Fuente: elaboración propia.

Las temáticas entre las cuales se clasificaron los artículos son las especificadas previamente en la tabla 3. Si se comparan los datos de los artículos catalogados en el momento de la elaboración de esa tabla (junio de 2022) y los correspondientes al momento de elaboración de este análisis temático (octubre de 2022) se puede observar un aumento significativo (pasando de n = 256 a n = 1863). Esto es debido al proceso de catalogación realizado durante los 5 meses de diferencia entre ambas fechas, lo que hace que el presente análisis temático sea más preciso al contar con una muestra más elevada.

En base a la submuestra actual de artículos catalogados, de n = 1863, el análisis temático arrojó que la categoría más investigada en las revistas especificadas en la tabla 18 fue "Innovación educativa y renovación pedagógica" (29,62%), seguida de

"Movimientos sociales y educativos" (20,46%), "Influencias, transferencias y transnacionalización de la educación" (19,23%), un 15,19% de los artículos se identificaron como "No especificado". Las temáticas menos estudiadas fueron "Género y políticas de igualdad" (8,13%) e "Inclusión y atención a la diversidad" (7,37%). En la figura 42 se pueden identificar gráficamente estos porcentajes.

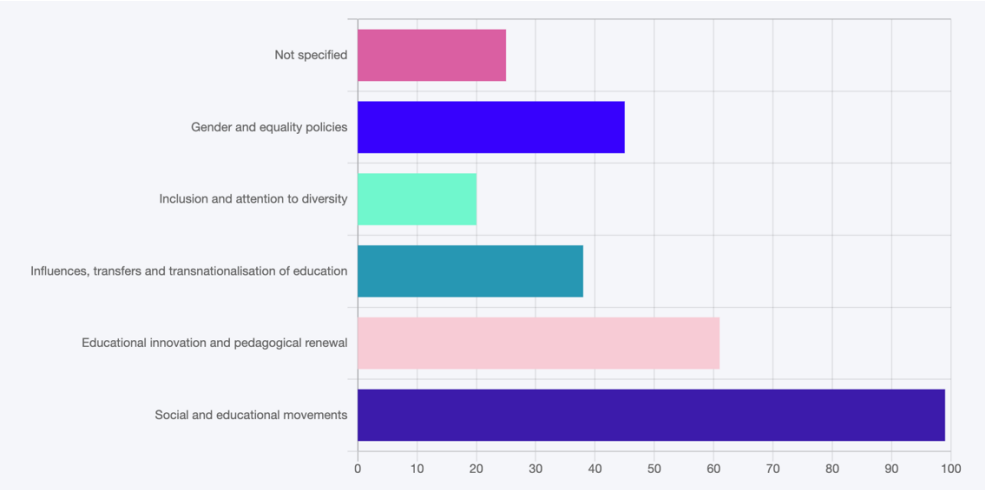
Figura 42. Análisis temático global.



Fuente: elaboración propia.

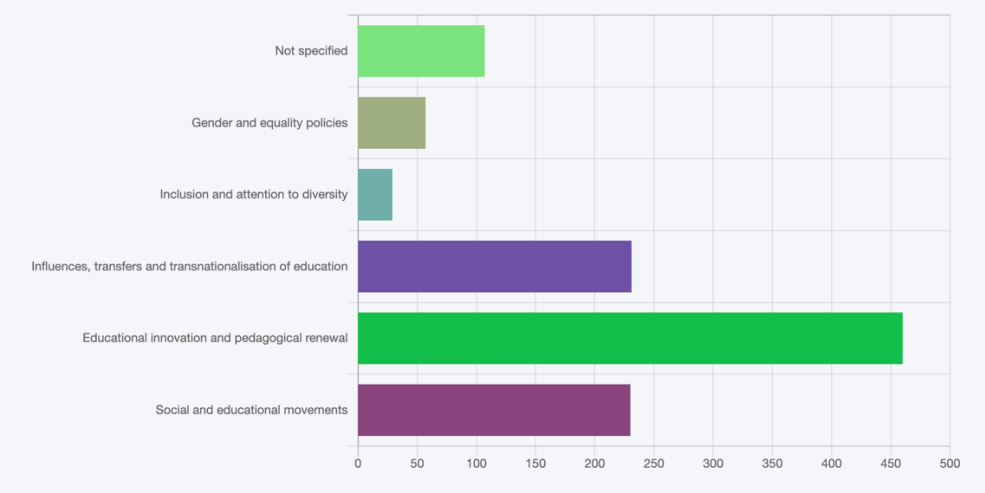
Sin embargo, un análisis pormenorizado en función de la revista refleja unos resultados diferentes. En *History of Educacion. Journal of the History of Education Society*, la mayoría de los estudios publicados se clasificaron dentro de la categoría "Movimientos sociales y educativos", al igual que los artículos publicados en *History of Education Quarterly*. El caso de *History of Education and Children's Literature* refleja los datos medios obtenidos en cuanto a la categoría mayoritaria, así como *História da Educação*. En el caso de *Paedagogica Historica: International Journal of the History of Education*, la mayoría de artículos fueron clasificados dentro de la categoría "No especificado".

Figura 43. Análisis temático de History of Educacion. Journal of the History of Education Society.



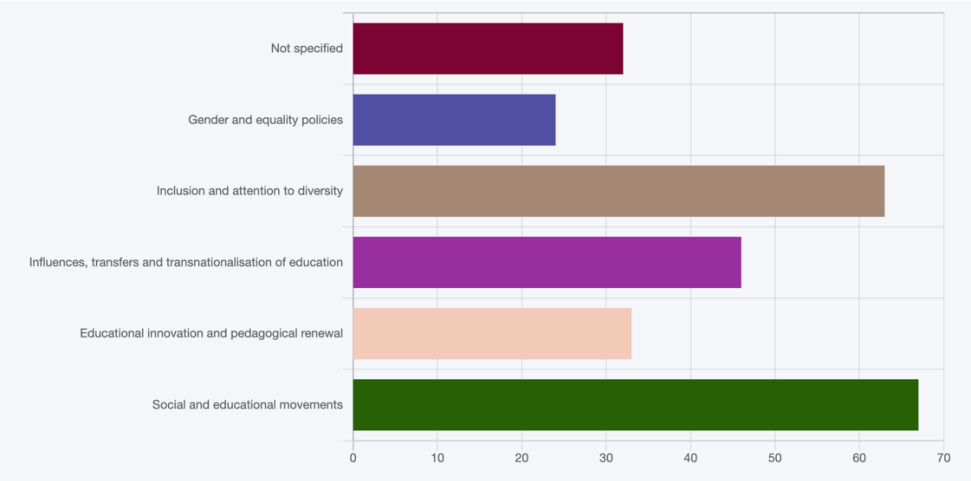
Fuente: elaboración propia.

Figura 44. Análisis temático de History of Education and Children's Literature.



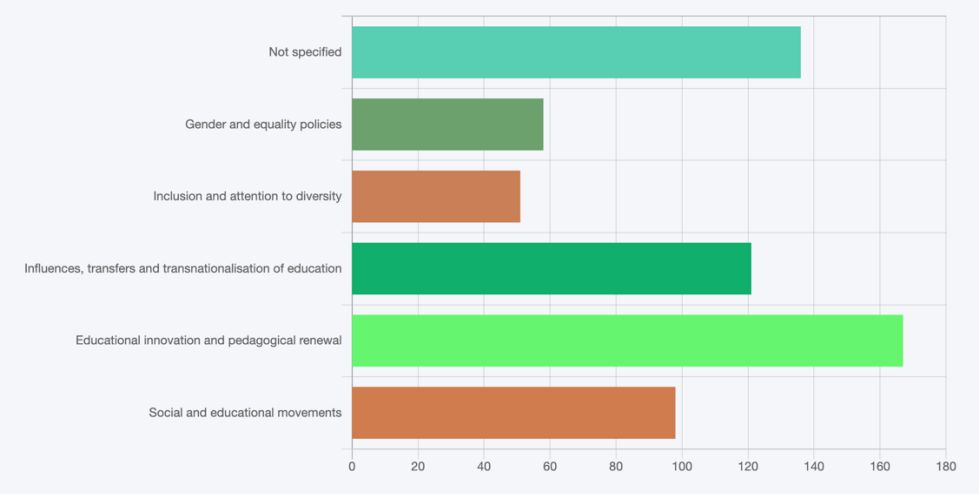
Fuente: elaboración propia.

Figura 45. Análisis temático de History of Education Quarterly.



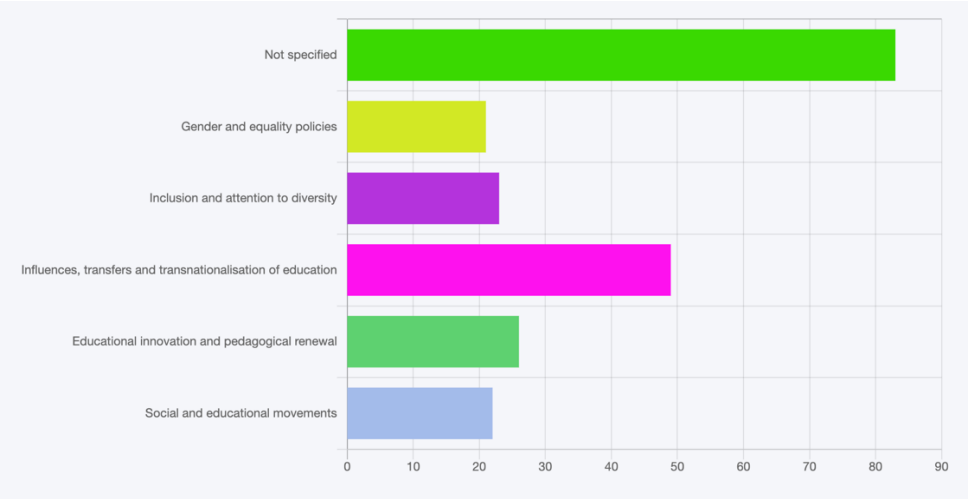
Fuente: elaboración propia.

Figura 46. Análisis temático de História da Educação.



Fuente: elaboración propia.

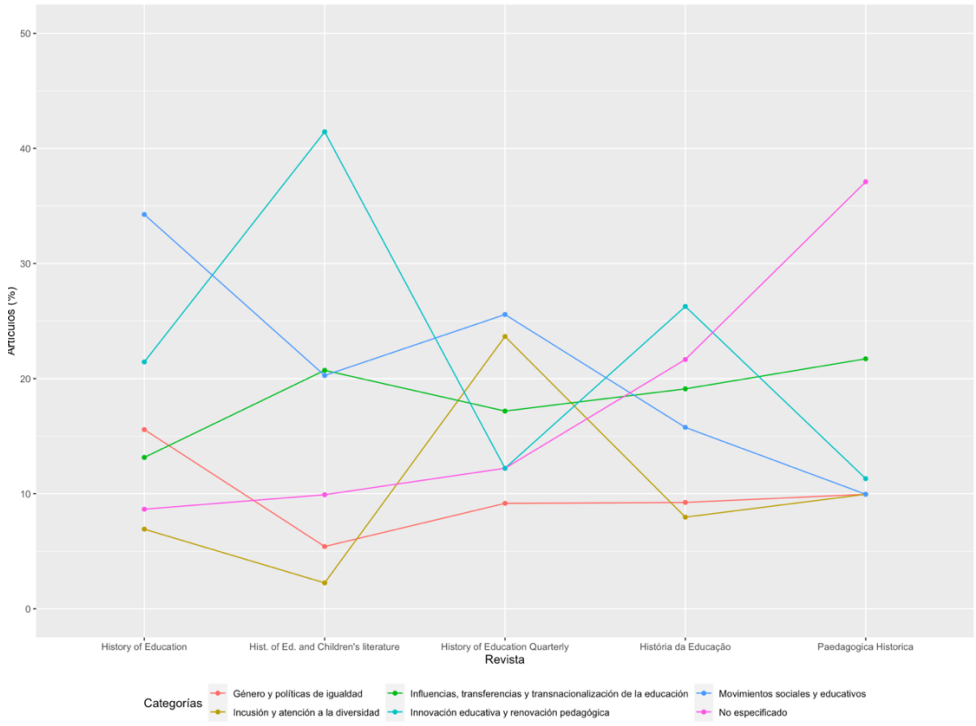
Figura 47. Análisis temático de Paedagogica Historica: International Journal of the History of Education.



Fuente: elaboración propia.

Una comparativa gráfica de los porcentajes de los artículos a los que cada revista a dedicado a cada una de las categorías analizadas en el estudio se puede observar en la figura 48.

Figura 48. Comparativa porcentual de las categorías por revista.



Fuente: elaboración propia.

5.3.8 Análisis de las épocas estudiadas

Para el análisis de las épocas estudiadas se ha partido de la misma submuestra del apartado anterior, la que ha sido reflejada en la tabla 18, pero en este caso el estudio se ha focalizado en los siglos más estudiados en las investigaciones. En términos globales, las épocas más estudiadas han sido el siglo XX (30,16%), seguida del siglo XIX (14,69%) y, en tercer lugar, el siglo XXI (4,05%). Los porcentajes globales de las 10 épocas más estudiadas puede observarse en la tabla 19. En este caso, no se desglosará por revistas, ya que la distribución de las primeras épocas más estudiadas se mantiene a penas sin cambios entre publicaciones.

Tabla 19. Análisis de épocas global.

Siglo	**Porcentaje artículos**
XX d.C.	30,16%
XIX d.C.	14,69%
XXI d.C.	4,05%
No especificado	2,71%
XVIII d.C.	2,21%
XVI d.C.	1,15%

XVII d.C.	1,13%
XV d.C.	0,32%
XIV d.C.	0,16%
XIII d.C.	0,14%

Fuente: elaboración propia.

5.5 Conclusiones

Durante las tres últimas décadas se ha producido un importante aumento en cuanto al número de investigaciones en el campo de la HE. Esto, a pesar de favorecer la internacionalización y la creación de redes transnacionales de investigación, también aumenta la complejidad de la comprensión de la disciplina en términos globales.

Debido a estas circunstancias, junto con el reducido número de estudios al respecto, en el presente capítulo se realiza un análisis bibliométrico de 11 revistas relevantes internacionalmente de HE desde 1961 hasta la actualidad. Se ha analizado la evolución de la producción a lo largo de los años por fuente, las citaciones, los autores y colaboración entre ellos, los países de afiliación, instituciones y colaboraciones, así como los idiomas.

También se ha estudiado la representación por sexo de autoras y autores más productivos. Además, se ha realizado un análisis de las combinaciones de palabras clave más frecuentes, junto con la evolución de las temáticas más investigadas en la última década. Los resultados muestran un punto de inflexión en la producción a partir de 2006, con un crecimiento exponencial desde entonces, así como unos bajos índices de colaboración entre autores y países. También muestran la relevancia en la última década de la investigación histórico-educativa por países y épocas, destacando la HE contemporánea especialmente.

En cuanto al análisis temático y de épocas estudiadas, posibilitado gracias el desarrollo de la herramienta Hecumen, que permite el catalogado de los artículos según su temática y las épocas estudiadas, ha arrojado que la categoría más investigada fue "Innovación educativa y renovación pedagógica" (17,17%), seguida de "Movimientos sociales y educativos" (11,86%), "Influencias, transferencias y transnacionalización de la educación" (11,15%), un 8,8% de los artículos se identificaron como "No especificado". Las temáticas menos estudiadas fueron "Género y políticas de igualdad" (4,71%) e "Inclusión, transferencias y transnacionalización de la educación" (4,28%). Sin embargo, este orden variaba en función de la época estudiada. Las tres épocas más estudiadas

fueron, sin apenas variaciones entre revista, el siglo XX (30,16%), seguida del siglo XIX (14,69%) y, en tercer lugar, el siglo XXI (4,05%).

Para este estudio se han utilizado las herramientas específicas para HE descritas en los capítulos 3 y 4.

Capítulo 6: Discusión y conclusiones

6.1 Introducción

La presente tesis ha sido desarrollada dentro del marco de la ayuda predoctoral PRE2020-093276 (financiada por MCIN/AEI/10.13039/501100011033 y por FSE invierte en tu futuro), y forma parte del proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global" (ayuda PID2019-105328GB-I00) que, entre otros, persigue el objetivo de cartografiar la producción científica internacional en HE, estudiar la producción historiográfica durante los últimos veinticinco años y desarrollar herramientas para la investigación en red a través de las posibilidades ofrecidas por las tecnologías de la información y la comunicación.

La investigación realizada ha explorado alguno de estos retos, y ha ofrecido unas soluciones con un acercamiento a las ciencias de la computación, no sin antes establecer, en el capítulo 2, un marco teórico de la cienciometría, los estudios bibliométricos y las limitaciones a las que se enfrentan los investigadores cuando se tratan de aplicar al campo de la HE. En concreto, se ha constatado que la ausencia de información específica de HE en los registros bibliográficos, como son las temáticas y las épocas estudiadas, invisibiliza parte de la riqueza de las investigaciones en HE.

Se ha mostrado el tipo de información bibliométrica que se puede obtener mediante los registros bibliográficos tradicionales en las bases de datos generalistas, y se ha explorado el camino a seguir a la hora de extraer esta información específica del contenido de los artículos con un breve acercamiento a la biblioteconomía y a la hermenéutica. Sin embargo, se ha constatado la imposibilidad de registrar esta información en las bases de datos generalistas, lo que ha mostrado el camino a seguir: el diseño e implementación de una base de datos específica de HE.

La herramienta, llamada Hecumen, ha sido desarrollada desde cero. El proceso completo mediante la metodología DBR, desde sus requisitos iniciales hasta la elaboración final de los principios de diseño, pasando por un proceso iterativo donde se han ido fundamentando cada una de las decisiones de diseño, ha sido descrito exhaustivamente en el capítulo 3. El resultado es una base de datos *online*, accesible por

el equipo de investigadores del proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global", que permite, además de la importación de registros bibliométricos desde otras bases de datos generalistas, el etiquetado y catalogado de cada uno de los artículos según sus temáticas y las épocas estudiadas.

Sin embargo, este último punto planteó una problemática logística: el proceso de catalogación debía realizarse manualmente, por un número elevado de investigadores, que debían acceder a los contenidos de cada artículo, uno a uno, para extraer e interpretar la información respecto a la temática y a las épocas estudiadas. Dada la eclosión de producción científica en HE de los últimos años, el número de artículos a revisar, incluso de un número reducido de revistas, resultaba enorme. Así pues, se planteó buscar soluciones, de nuevo en un acercamiento a las ciencias de la computación.

Este punto de partida es el núcleo del capítulo 4, donde se explora la posibilidad de utilizar la IA para tal tarea. Los problemas de clasificación son uno de los retos habituales a los que se enfrenta la IA, con lo que se planteó el investigar la posibilidad de utilizar una IA para la clasificación automática de artículos. Para ello se contextualizó la tecnología de la IA, además de describir los procesos de entrenamiento de una IA para la clasificación de artículos y se realizaron diferentes propuestas de diseño de IAs, implementando una de ellas con una primera muestra de entrenamiento reducida (n = 256), que obtuvo una efectividad del 70% a la hora de clasificar los artículos de media.

En base a todo este proceso de investigación se elaboró el capítulo 5, en el que se realizaba un estudio bibliométrico (basándose en la contextualización del capítulo 2), con la herramienta Hecumen (descrita en el capítulo 3) y que, como novedad, realizaba un análisis temático y de épocas estudiadas, según la clasificación automática mediante IA del capítulo 4, además del resto de análisis habituales en este tipo de estudios.

En el siguiente apartado se procederá a discutir cada uno de estos capítulos en detalle para, finalmente, explicitar las limitaciones de la tesis, proponer una serie de prospectivas y establecer unas conclusiones finales.

6.2 Discusión

6.2.1 Discusión sobre principios cienciométricos como soporte a la investigación en historia de la educación

Los orígenes de la cienciometría (acuñada por Vasily Vasilyevich Nalimov por primera vez con el término *naukometria,* y con Price y Garfield como precursores en las décadas de los 60 y 70) responden, entre otras cuestiones, a la necesidad de valorar cuantitativamente la influencia de la producción científica (Garfield, 2009). Para lograr este objetivo se sentaron las bases de los actuales índices de citaciones que estimulan, condicionan y encorsetan, a partes iguales, la producción científica a nivel mundial. Nalimov no solo acuñó el término por primera vez, sino que también propuso usar las citaciones como indicadores (Wouters, 1999).

Tal como se ha mostrado, este planteamiento no estuvo exento de críticas, puesto que se estaba invisibilizando a determinada producción científica que, a pesar de su influencia, no era citada con frecuencia (Ardanuy, 2012). En un sentido similar, las críticas ponían a la cienciometría en un lugar al servicio de las agencias que fiscalizan la labor científica, como una herramienta de sometimiento y control, lo que podía cuestionar la falta de determinación epistemológica de la disciplina (Millán et al., 2017). En la misma línea crítica Masic y Jankovic (2021) exponen las prácticas de las autocitas que manipulan artificialmente el impacto de un artículo al ser citado por sus mismas autoras y autores en otros artículos.

Al margen de este debate, los estudios bibliométricos emergen dentro del marco de la cienciometría como la disciplina que estudia la producción escrita en base a unos procesos matemáticos para obtener una serie de indicadores. Para ello recurre a los registros bibliográficos (Garfield, 1970), que categorizan la producción científica según una serie reducida de campos, y que resumen una muestra de artículos en unos pocos indicadores, lo que permite realizar estudios comparativos. Esto es, los indicadores sintetizan un número elevado de datos extraídos de los registros bibliográficos de las

muestras de artículos. Así pues, no se pueden obtener indicadores de aspectos que no estén cubiertos por los datos de los registros bibliográficos.

Esto es especialmente problemático cuando tratamos de acercar la bibliometría a la HE. En este caso nos encontramos con una serie de características propias en los estudios de HE que no quedan reflejados en los registros bibliográficos. Los estudios de HE, en tanto que investigan unos procesos educativos insertos en unas coordenadas espacio-temporales (Guichot, 2006), estudian indirectamente determinadas épocas cuya información no está incluida en los registros bibliográficos, sino que forma parte del contenido del artículo (se ha establecido la diferencia entre fechas internas, contenidas en el mismo artículo, y fechas externas, que son las indicadas en los registros bibliográficos, que reflejan el momento de la publicación, no de la época estudiada). Lo mismo sucede con las temáticas, puesto que en un estudio pueden coexistir temáticas diametralmente opuestas a lo comúnmente esperado según la época estudiada. La primera información, la de las fechas internas, es sencilla de obtener simplemente accediendo a la mayoría de los títulos de los artículos (tal como se ha mostrado descriptivamente en la tabla 1), pero la segunda, referida a las temáticas, requiere un análisis en profundidad de los textos ya que hay que realizar un proceso de comprensión e interpretación del mismo.

El acceso a los contenidos de los textos, su comprensión e interpretación para su posterior clasificación, nos ha obligado acercarnos a la biblioteconomía, que desde su génesis se ha enfrentado a esta problemática, tal como coinciden Orera (1995) y Garrido (1990). También nos hemos aproximado a la hermenéutica en su acepción más popular, entendida como la interpretación y comprensión de los textos (Grondin, 2014), pero sin acercarse a las complejidades planteadas por Dilthey respecto al análisis de la psique del autor durante la producción del artículo (Sánchez, 2019). Sin embargo, debido a la prolija producción científica de las últimas décadas añade una dificultad enorme a la hora de clasificar la producción científica en HE para ponerla a disposición de la comunidad. La otra de las dificultades identificadas ha sido que las bases de datos generalistas no permiten añadir este tipo de información cualitativa, lo que ha abierto las puertas al desarrollo específico de Hecumen para resolver dicha eventualidad.

6.2.2 Discusión sobre el desarrollo de una base de datos específica para historia de la educación

El desarrollo de este capítulo ha tenido como objetivo el acercamiento de los procedimientos de las TIC al campo de la HE, con el fin de dar respuesta a las necesidades derivadas de la eclosión de producción científica en el campo desde mediados de los 90, y que se articulan alrededor de tres ejes: la dispersión de la investigación en HE en diversas bases de datos generalistas, la imposibilidad de catalogar con información cualitativa específica del campo a los artículos y la necesidad de establecer un punto de encuentro virtual entre los investigadores de HE.

Para tal efecto, en la primera parte del capítulo se ha descrito el desarrollo de la herramienta Hecumen, dentro de la metodología del DBR de cuatro fases, que ha implicado dos iteraciones, cada una de ellas con un diseño anterior y un análisis posterior (Design-based Research Collective, 2003). Esta metodología conecta al investigador con las problemáticas del mundo real, ya que se produce una colaboración estrecha entre profesional e investigador (Amiel y Reeves, 2008). El inicio del desarrollo se ha fundamentado en la información obtenida en una entrevista semiestructurada con los investigadores principales del proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global", lo que permitió establecer cuatro requisitos iniciales que debía cumplir la aplicación. Estos eran que el sistema debía permitir el almacenaje de datos bibliográficos de artículos extraídos de revistas de HE, así como la posibilidad de aparición o no de determinados metadatos (R1); la herramienta debía ser accesible por diferentes investigadores desde distintos lugares geográficos que, además, contaran con un sistema de permisos para acceder a determinadas partes de la aplicación (R2 y R3). Por último, la aplicación debía permitir a los investigadores catalogar y etiquetar los artículos en función de las categorías, los periodos históricos o las épocas estudiadas (R4).

A continuación, se procedió a diseñar la base de datos con la dificultad de la contingencia de algunos metadatos (una de los aspectos de R1), que añadía un nivel de complejidad al diseño de la base de datos. Esto planteaba varios caminos a seguir a la

hora de diseñar la base de datos. Uno de ellos consistía en incluir tantos campos como metadatos pudieran haber, pero dicha opción se rechazó ya que implicaba prever campos desconocidos lo que es imposible y, en el caso de tener que añadirlos en sucesivas versiones de Hecumen, implicaría cambiar el esquema de la base de datos y todas las partes de la aplicación que se comunican con ella (Xie et al., 2013). Cerrada et al. (2000) proponen una segunda opción que consiste en recurrir a la propiedad de la herencia, de tal forma que se deben escribir nuevas entidades que heredan de una entidad base genérica. Esta misma línea es la seguida en proyectos como el de Rashid y Chitchyan (2003), pero adolece de la misma problemática de la primera opción, es decir, que hay que reescribir partes de la aplicación cuando se crean nuevas entidades derivadas en el modelo de datos (Xie et al., 2013). La tercera opción se fundamenta en el uso de una base de datos no-SQL, orientadas a documentos y que no cuentan con una estructura fija de datos. Aunque este enfoque resuelve la problemática de la contingencia de los datos, ya que el esquema de datos se puede ir modificando según las necesidades cambiantes de los datos a almacenar, Feng (2006) alerta de que la combinación de datos entre diferentes entidades puede ser mucho más compleja y con serias implicaciones en el rendimiento. Puesto que la combinación de datos entre diferentes entidades resulta fundamental en este tipo de bases de datos, se descartó esta opción. La cuarta alternativa para la contingencia de datos consistía en recurrir al modelo EAV, en el que las tablas tienen un aspecto "vertical", ya que todas las entidades, atributos y valores están relacionadas entre sí mediante tablas intermedias. Este enfoque permite cambiar la forma en la que la base de datos almacena la información con simples cambios en registros, y sin la necesidad de cambiar su estructura (Ganslandt et al., 1999). Batra et al. (2018) continúan en la misma línea al afirmar que es el modelo más utilizado en bases de datos que almacenan datos de salud, y que estructuralmente es muy parecida a cómo se pueden almacenar datos bibliográficos contingentes. Por su parte, otros autores como Batra et al. (2018) alertan de que este modelo no es el más eficiente a la hora de realizar búsquedas.

Con todo ello, para el diseño de la base de datos de Hecumen se optó por una combinación entre el modelo EAV con un esquema relacional clásico para aquellos datos

que no iban a ser contingentes. El diseño final de la base de datos, tras el proceso de refinamiento sucesivo DBR, queda plasmado en el diagrama ER de la figura 14.

Una vez definido el diagrama ER, se construyó la aplicación *web* que permitiría interactuar con la base de datos, basándose en el *framework web* Laravel. A pesar de sus similitudes con otros *frameworks* como Symfony, Laravel implementa exclusivamente el modelo *middleware* de desarrollo en capas (Laaziri et al., 2019), mediante el cual se pueden añadir funcionalidades a la aplicación sin realizar cambios estructurales en la misma, lo que le da una gran flexibilidad al desarrollo. Al igual que se optó por el modelo EAV en combinación con el relacional clásico para el modelo de datos por su flexibilidad, este mismo criterio es el que se aplicó a la hora de escoger Laravel.

Finalmente, el desarrollo concluyó con la enumeración de cuatro principios de diseño en base a la información obtenida en cada una de las fases del DBR.

El primero ha sido que el diseño de bases de datos orientadas a almacenar información bibliográfica deben tener prevista la contingencia y dispersión de algunos metadatos, así como la posibilidad de que aparezcan nuevos metadatos en futuros artículos que obliguen al rediseño, tanto del modelo de datos, como de la aplicación misma. Se han valorado diferentes opciones para resolver esta problemática, como el uso de campos vacíos, las bases de datos no-SQL o las relaciones de herencia en las entidades, optando finalmente por la combinación de un diseño relacional clásico junto con un modelo EAV.

El segundo principio de diseño corrobora que el optar por un diseño de aplicación *web* resuelve la problemática del acceso por parte de diferentes investigadores simultáneamente a una misma base de datos compartida, además de que es un flujo de trabajo con el que ya están habituados la mayoría de investigadores por el uso de otras plataformas *online*.

En lo que respecta al desarrollo de la aplicación en sí mismo, el tercer principio de diseño enuncia que el uso de *web frameworks* da rigor garantizando una arquitectura coherente, lo que permite optimizar el tiempo de desarrollo y centrarse en

funcionalidades específicas, dando por sentadas aquellas que deben estar implementadas en cualquier aplicación *web*, ya que las aporta el *framework*.

El último principio de diseño gira alrededor de focalizar el esfuerzo de desarrollo, en un primer estadio, en un PMV, alejándose de los objetivos de la aplicación completa, lo que permite obtener información de los usuarios que posteriormente revertirá al enriquecimiento de la aplicación en sucesivas fases. Esto es especialmente relevante, como se ha podido constatar durante el desarrollo de la aplicación, con equipos de desarrollo pequeños (incluso exclusivamente individuales), donde hay determinadas decisiones de diseño para las que se debe ser flexible, ya que las necesidades del proyecto pueden cambiar. En este sentido, centrarse en un PMV permite reducir las implicaciones de desarrollo, tener un prototipo sobre el que trabajar y que mejorar con la información obtenida por su uso diario (Ries, 2011).

Una vez desarrollada la base de datos Hecumen se ha constatado que el proceso de catalogación y etiquetado de artículos según su temática y épocas estudiadas es costoso y complejo, ya que se tiene que revisar el contenido de los artículos individualmente. A pesar de haber diseñado Hecumen para resultar una herramienta ágil y dinámica, que permite al investigador centrarse solo en la información que le interesa en cada momento, el catalogado de miles de artículos manualmente es una actividad que consume una gran cantidad de recursos humanos.

Es por esto que se ha planteado el acercamiento, de nuevo, a la IA dentro de las ciencias de la computación, para explorar la posibilidad de la clasificación automática de textos con un sistema de IA.

6.2.3 Discusión sobre la inteligencia artificial para automatizar la clasificación de artículos de historia de la educación

La exploración de las posibilidades del uso de la IA para la clasificación temática y de épocas estudiadas en la producción científica de IA, desarrollada a lo largo del capítulo 4, ha tenido que enfrentarse a una serie de dificultades, para los cuales se han analizado en cada momento los pros y los contras, optando siempre por aquella que respondiera de la manera más eficaz a los objetivos del proyecto.

Respecto a las dificultades detectadas, en primer lugar hay que tener en cuenta que cada una de las categorías no tenía el mismo grado de balanceamiento entre las muestras positivas y las negativas. Una muestra desbalanceada es problemática para los sistemas de clasificación, tal como han expuesto diversos autores (Su et al., 2006; Jiang et al., 2012), ya que el algoritmo sesga hacia la categoría más representada. Así pues, la horquilla de la muestra de la investigación oscilaba entre los extremos de la categoría "No especificado", la más balanceada con una proporción de 33,20% entre la clasificación positiva y la negativa y, en el polo opuesto, la categoría "Género y políticas de igualdad", la más desbalanceada con una proporción del 8,20%. Puesto que la situación ideal para un algoritmo de clasificación es que la muestra esté balanceada en un 50% entre muestras positivas y negativas, se ha optado por la técnica del *undersampling* (Lemaître et al., 2017), en el que se han eliminado aleatoriamente el número necesario de muestras de la categoría mayoritaria con el fin de equipararla a la minoritaria, con lo que se ha conseguido alcanzar la citada cifra.

El poder modificar el código fuente de la herramienta Hecumen ha sido fundamental para poder implementar las modificaciones necesarias para generar los archivos ARFF requeridos por Weka. En este sentido se ha seguido el criterio de que la IA imitara el procedimiento de los investigadores a la hora de clasificar un artículo: analizar el título, resumen y palabras clave para dirimir la categoría a la que pertenecía, pero desde una perspectiva cuantitativa en la que se han contabilizado las *n* palabras más repetidas en los 3 campos, de tal forma que la IA fuera entrenada para identificar patrones subyacentes a la aparición de determinada combinación de palabras. Esto se

ha basado en la ley de Zipf, que plantea que en un artículo se tiende a utilizar un número mínimo de diferentes palabras (Araujo, 2006) y que, al ordenarlas por su frecuencia de aparición, resultan ser las palabras clave que permiten clasificar un artículo en una categoría determinada. Un ejemplo de este tipo de archivos generados por la modificación implementada en Hecumen se ha presentado en la figura 23, mientras que su equivalente con palabras se ha plasmado en la figura 24. El proceso propuesto para la transformación de palabras en valores únicos queda plasmado en el esquema ER de la figura 20, de tal forma que cada palabra tiene un equivalente numérico único y una tabla intermedia relaciona estos valores numéricos con los artículos en los que aparecen.

Para el proceso de entrenamiento se ha optado por construir una IA independiente para cada una de las categorías a identificar. De esta forma, a cada una de las IA se le suministraba una muestra de artículos en los que los correspondientes a la categoría a identificar estaban etiquetados como clasificación positiva, mientras que el resto (reducidos mediante el citado *undersampling*) se consideraban clasificación negativa. Por cada una de estas IA se han realizado 6 experimentos, uno por cada uno de los algoritmos que, tras la revisión de la literatura de IA aplicada a la educación, resultaban ser más frecuentes en diversas investigaciones. Estos han sido RF, NB, J48, MLP, DT y LR, ampliamente utilizados (Landa et al., 2021; Castrillón et al., 2020; Gil et al., 2018; Jokhan et al., 2022; Mourdi et al., 2020).

Tras el entrenamiento de cada uno de estos algoritmos individualmente con cada una de las categorías se han extraído los variables de TP, TN, FP y FN (representados en diversas matrices de confusión), lo que ha permitido calcular las métricas de *accuracy*, *precision*, *recall* y *f1-score*. Si nos centramos en el valor de *accuracy* que, al tratarse de muestras balanceadas, resulta significativo para medir el rendimiento global de cada una de las IA, podemos identificar que el valor más alto (0,95) ha sido el obtenido por el algoritmo RF dentro de la categoría “No especificado”. Por el contrario, el valor más bajo (0,35) se ha dado también con el algoritmo RF, pero en la categoría “Movimientos sociales y educativos”.

Con los datos de rendimiento de las IAs a la hora de clasificar cada una de las categorías, tal como puede observase en la figura 30, la categoría que ha sido más fácilmente reconocida ha sido "No especificado", mientras que la que menos corresponde a "Movimientos sociales y educativos" con unos datos sensiblemente inferiores. El resto de categorías a identificar, "Género y políticas de igualdad", "Inclusión y atención a la diversidad", "Influencias, transferencias y transnacionalización de la educación" e "Innovación educativa y renovación pedagógica" se han situado con unos valores intermedios entre los dos extremos, siempre por encima de 0,5 de *accuracy*, a excepción de los picos inferiores que marcan "Género y políticas de igualdad" con el algoritmo LR e "Inclusión y atención a la diversidad" con el DT.

Puesto que el objetivo de los experimentos era analizar cada uno de los algoritmos en términos globales, independientemente de la categoría que fueran a clasificar, se han calculado las medias de *accuracy* de todas las categorías. Así, tal como refleja la figura 31, el algoritmo RF, basado en árboles de decisión, ha sido el que mejor rendimiento ha obtenido, con un valor de 0,7. De hecho, su rendimiento ha sido superior que el de los otros dos algoritmos basados en árboles de decisión, como DT y J48, con *accuracies* de 0,65 y 0,59 respectivamente. El algoritmo LR ha sido el que ha obtenido los resultados más bajos, con una *accuracy* de 0,57.

Tal como se ha mencionado previamente, el dato más bajo de *accuracy* se ha dado en la categoría "Movimientos sociales y educativos", pero si observamos la tabla 10 podemos identificar que esta capacidad tan baja de clasificación no se ha dado exclusivamente con el algoritmo RF. De hecho, el rendimiento general de cualquiera de los algoritmos a la hora de identificar esta categoría en ningún momento ha pasado de 0,5, cuestión que contrasta con el rendimiento de los mismos algoritmos en el resto de categorías. Si bien es cierto que la categoría que ha sido mejor identificada ha sido aquella que tenía el mayor número de artículos ("No especificado", con n = 85), la categoría "Movimientos sociales y educativos" no era la que menos artículos tenía. De hecho, tanto "Inclusión y atención a la diversidad" como "Género y políticas de igualdad" tenían valores más bajos (n = 29 y n = 21 respectivamente) y sus valores han sido sensiblemente mejores: la media de las *accuracy* de todos los algoritmos para la

categoría “Inclusión y atención a la diversidad” era de 0,59, la de “Género y políticas de igualdad” se situaba en 0,60, mientras que la de “Movimientos sociales y educativos” descendía hasta la cifra de 0,43. Una razón de esto la podemos encontrar en la lógica que subyace al proceso de clasificación en categorías. Si tenemos en cuenta que se basa en los patrones de aparición de determinadas palabras utilizadas con más frecuencia en los artículos, se puede inferir que una baja capacidad de clasificación de esta categoría se deba a que no se repiten demasiadas palabras entre diferentes artículos, lo que hace imposible la tarea de que la IA identifique patrones. Esto se fundamentaría en un alto grado de especificidad en cada uno de los artículos que, centrándose en movimientos sociales y educativos concretos, haría que las palabras de mayor frecuencia de aparición estuvieran circunscritas al movimiento estudiado concreto, diluyendo la posibilidad de aparición de palabras comunes a la categoría y, por tanto, de que emergieran patrones identificables para la clasificación de los artículos de esta temática. A tenor de los resultados, esta cuestión se muestra mucho más determinante para que la IA obtenga un buen rendimiento, que el número de artículos suministrados en el proceso del entrenamiento.

6.2.4 Discusión sobre la historia de la educación a través de revistas especializadas

Para la realización del estudio bibliométrico del capítulo 5 se ha seguido el procedimiento planteado por Donthu et al. (2021) en 4 pasos. La muestra del estudio está compuesta por 5217 publicaciones, cuyas dos tipologías más frecuentes son los artículos (74,35%) seguidos de las *reviews* (20,78%). A pesar de que el documento más antiguo de la muestra data de 1961, la producción no se distribuye de una manera homogénea a lo largo de los 61 años estudiados. De hecho, tal como se puede observar en la figura 32, la producción se mantiene por debajo de la cota de 60 investigaciones por año hasta 2006 (a excepción de 2004, que contó con 51 documentos). Precisamente en 2006 se multiplicó la producción por más del doble respecto al año anterior (91 documentos frente a 40). Desde ese momento el crecimiento se ha mantenido con una tendencia creciente hasta hoy.

De acuerdo al crecimiento analizado y según las etapas propuestas por Price (1963), entre el año 1961 y 2006 el campo de HE se encontraba en el momento de los "precursores", caracterizado por un escaso crecimiento. A partir de 2006 podemos observar una duplicación de la producción cada 7 años aproximadamente, lo que permitiría ubicar el campo en la etapa de "crecimiento exponencial".

Una explicación para este punto de inflexión acaecido en 2006 lo encontramos en la figura 33, en la que se muestra una eclosión de publicaciones que aumentaron significativamente la producción en el campo de HE. De hecho, entre 2006 y 2022 se escribieron 3873 documentos, frente a los 1344 documentos producidos entre 1961 y 2005, lo que supone que un 74,24% de la producción total de la muestra fue realizada en tal solo 17 de los 61 años que cubre la investigación (es decir, en menos de un tercio del tiempo que cubre la investigación).

Resulta pertinente realizar un análisis del comportamiento de las revistas más veteranas ante la eclosión de las nuevas publicaciones (aquellas con artículos indexados anteriores a 1998, como son *Paedagogica Historica*, *History of Education. Journal of the History of Education Society* e *History of Education Quaterly*, dejando al margen *History*

of Education Review debido a la falta de indexación de sus artículos en Scopus anteriormente a 2012).

En lo que respecta a la primera revista en cuanto a producción, *Paedagogica Historica*, se identifican tres periodos en su producción anual. En el primero, que cubre desde el año 1961 hasta 1983, su producción se mantiene con 20 artículos o menos al año. A partir de 1984 se inicia un periodo con muchas oscilaciones en su producción (con cotas superiores, como la de 54 documentos en 1998). En la última etapa identificada, a partir de 2006, se consolida su tendencia creciente con el pico máximo de 2021 con 87 documentos (un incremento en la producción de 67,30% respecto a los 52 documentos de 2020).

En el caso de *History of Education. Journal of the History of Education Society*, si anteriormente al punto de inflexión se mantenía con una producción constante por debajo de los 30 documentos por año, a partir de 2005 aumenta su producción a una media de 42 artículos por año (calculado en el periodo entre 2005 y 2021).

El caso de *History of Education Quaterly* es incluso más llamativo. Si, según los datos de Scopus, anteriormente a 2008 nunca había publicado más de 3 artículos por año, a partir de este año su producción pasa a una media de 21 artículos por año (entre 2008 y 2021).

Así pues, la aparición de nuevas publicaciones a partir de 2006 no supuso una merma en la producción de las referidas revistas más veteranas sino que, al contrario, tuvo como consecuencia un aumento notable en su producción anual. Dicho de otro modo, no se repartió la misma producción que se venía realizando entre más fuentes, sino que la producción global de HE se vio enriquecida por el aumento de revistas.

Las dos fuentes que más investigaciones han aportado a la muestra han sido *Paedagogica Historica: International Journal of the History of Education* e *History of Educacion. Journal of the History of Education Society*. Entre las dos suponen un 54,44% de la muestra. Sin embargo, su distribución a lo largo de los años no ha sido la misma,

ya que los artículos más antiguos de la primera datan de 1961, mientras que los de la segunda comienzan en 1972.

El protagonismo de estas dos revistas en cuanto a la producción también se traslada a las citaciones globales de los artículos. De los 10 artículos más citados, un 60% fueron publicados en *History of Educacion. Journal of the History of Education Society* mientras que un 20% lo hicieron en *Paedagogica Historica: International Journal of the History of Education*.

Los países de afiliación mayoritarios de los autores han sido Reino Unido (16,19%), Estados Unidos (13,86%), Brasil (12,30%), España (11,35%) e Italia (10,18%). En total, estos 5 países suman el 63,88% de los autores de la muestra. Hay ciertos cambios en lo que se refiere al análisis de las afiliaciones de los autores de contacto. Los dos primeros lugares los siguen manteniendo Reino Unido y Estados Unidos, pero en esta ocasión seguidos de Italia, Francia y España. En lo referente a la colaboración entre países, los datos constatan una baja colaboración ya que, de entre los países mencionados previamente, Reino Unido fue el que más artículos escribió en colaboración con otros países, y lo hizo en tan solo un 3,12% de las ocasiones. España fue el siguiente país que más ha colaborado con otros países (2,68%), seguido de Estados Unidos (1,10%). Los porcentajes de colaboración entre países (MCP) de Italia, Brasil y Francia están por debajo del 1%.

A pesar de estas tasas de colaboración entre países tan bajos, se ha analizado cuáles eran las parejas de países que colaboraban con más frecuencia. En la figura 6 se pueden observar tres clústers: el primero de ellos está formado por Brasil-España-Federación Rusa-Francia-Italia-Portugal-Argentina-Colombia, en el que las mayores colaboraciones se dieron entre Italia-Brasil y España-Italia. El segundo está compuesto por Australia-Irlanda-Nueva Zelanda, con 5 colaboraciones entre Australia e Irlanda, y 3 entre Australia y Nueva Zelanda. El tercer clúster lo componen Estados Unidos-Bélgica-Reino Unido-Países Bajos-Alemana-Suiza y las afiliaciones no especificadas. Sin embargo, estos tres clústers no han elaborado investigaciones de una manera totalmente aislada, sino que hay algunas parejas de países que los enlazan. Así pues,

tanto Italia-Alemania (3 colaboraciones) como España-Países Bajos (3 colaboraciones) unen el primer y el tercer clúster, mientras que Australia-Reino Unido (2 colaboraciones) hacen lo respectivo entre el segundo y el tercer clúster.

A nivel de autores, se han identificado 4053 autores únicos, que han aparecido un total de 6642 ocasiones. Un elevado porcentaje de artículos (77,10%) ha ha sido escrito por un solo autor, mientras que tan solo un 22,90% ha sido producido por la colaboración de varios investigadores. Esta baja tendencia a la colaboración en la muestra analizada cristaliza en el índice de autores por artículo (0,78), el índice de co-autores (1,27) y el índice de colaboración (0,44). Estas cifras contrastan con las aportadas por otros estudios bibliométricos centrados en una vertiente más técnica de la educación, como el de Roda-Segarra et al. (2022), en el que se pueden apreciar índices de autores por artículo de 2,61, así como índices de co-autores de 3,04. Lo mismo sucede con el índice de colaboración si se compara con el 2,24 obtenido por Mengual-Andrés et al. (2020).

El autor que cuenta con más publicaciones dentro de la muestra objeto de estudio es Roberto Sani. Ha escrito un total de 23 artículos, siendo un 39,13% individuales frente al 60,87% producidos en colaboración con otros autores. Sin embargo, respecto a estos últimos, solo en 2 ocasiones firmó como primer autor, con lo que su índice de dominancia se sitúa en 0,14. El segundo puesto en cuanto a producción lo ostenta Anna Ascenzi, con 18 artículos, de los que un 22,22% fueron individuales y un 77,78% colectivos, aunque en este caso en todas las ocasiones firmó como primera autora, lo que le da un factor de dominancia de 1. El tercer autor es William W. Brickman, con 18 artículos, todos ellos individuales.

En cuanto a las parejas de autores que suelen escribir artículos en colaboración, teniendo en cuenta la amplia horquilla temporal del estudio de 61 años, las cifras son bajas. Solo 4 parejas de autores comparten autoría con 3 artículos cada una. Estos son Almeida-De Azevedo Ramil, Beadie-Gottesman, Gottesman-Williamson-Lott y, por último, Roderick-Stephens. El resto de parejas de la muestra descienden a, como mucho, 2 artículos por cada pareja.

La representación por sexo entre los autores más productivos y las parejas de autores que publican con más frecuencia revelan una paridad ideal: entre los 10 autores más productivos hay el mismo número de hombres que de mujeres; en lo que se refiere a las 10 parejas de autores que publican juntos con más frecuencia, un un 70% está compuesto por parejas del mismo sexo, mientras que el 30% restante son de sexo distinto.

Si combinamos los datos de los autores más productivos con los del número de investigaciones por país, podemos desgranar la densidad de producción por autor respecto a su país de afiliación, es decir, el porcentaje de producción que realizó respecto a la producción total del país. Atendiendo a esta densidad de producción, el autor que más contribuyó a la producción científica de su país fue Bakker, que produjo el 7,10% de la producción total de Países Bajos. Poniendo el foco en Italia, 6 de los 10 autores más productivos son italianos. Estos autores escribieron, en total, un 14,71% de la producción italiana. William W. Brickman produjo el 2,52% de Estados Unidos. Por su parte, Ian Grosvenor aportó el 1,20% de la producción científica de Reino Unido, mientras que María Helena Cámara Bastos lo hizo con un 0,95% respecto a la producción total brasileña.

El idioma más utilizado en los artículos de la muestra, tal como se puede observar en la figura 8, es el Inglés (supone el 73,76% de la muestra), con una caída abrupta del Portugués al 7,58%, seguido del Francés con el 5,91%. Teniendo en cuenta que solo un 30,05% de las afiliaciones de los autores corresponden a Reino Unido y Estados Unidos, resulta evidente la preponderancia del Inglés en el campo de la HE a pesar de que sus autores no estén afiliados a países angloparlantes.

También se han analizado las palabras clave de los artículos de la muestra. Esto se ha hecho de dos maneras diferentes: por una, se han cuantificado en la totalidad de la muestra las parejas de palabras clave más utilizadas, tras haber unificado algunas y eliminado otras por su obviedad, con el fin de identificar bloques temáticos más frecuentes. El resultado, que se puede observar en la figura 40, muestra a *History of education-Italy* (n = 120) como la pareja más frecuente, seguida de *Italy-XXth Century* (n

= 83) e *History of education-XXth Century* (n = 74). La figura muestra la imbricación de estas palabras clave entre sí, de tal forma que *XXth Century* también aparece muy relacionada con *Spain* (n = 25). *Italy* también se conecta con estudios de *XIX Century* (n = 38).

La segunda parte del estudio de palabras clave se ha focalizado en una submuestra, compuesta por las palabras clave de los artículos publicados desde 2010, con la finalidad de mostrar la evolución en el interés de determinados temas en la última década en el campo de HE. Así pues, se han detectado las 10 palabras clave más utilizadas en cada uno de los años (con el procesamiento previo descrito en el apartado 3.6) y se ha representado la figura 41. Estas palabras clave más representativas de los últimos años se pueden agrupar en tres grandes bloques: cronológicas, geoespaciales y conceptuales.

El grupo de las palabras clave que hace alusión a aspectos cronológicos incluiría las siguientes: *1968, nineteenth century, XIX Century, XIX-XX centuries, XX century* y *XXth-XXIst Centuries*. En lo que respecta al grupo de las palabras clave geoespaciales encontraríamos *Australia, Brazil, Europe, France, Italy, Latin America, Spain* y *Sweden*. Por último, el grupo más numeroso es el de las palabras clave referidas a conceptos, con *adult education, architecture, bioarchaeology, Childhood, Children's and Youth Literature, Citizenship, Civics, Cultural identity, Curriculum, Documentary, Education System, Educational policy, Educational reform, Epidemic, Film, Francoism, gender, Higher education, Historiography, History of school, Informal education, Language teaching, Nationalization, nature, pedagogy, Periodical, Popular education, Primary education, Primers, school, School memories, Schooling, Secondary education, Slavery, Teachers, Textbook, University, War, women* y *Writing*. Tal como se puede observar en cuanto al número de elementos de cada grupo, destaca la tendencia de los investigadores a utilizar términos conceptuales (72,73% del total), frente a aspectos geoespaciales (16,36%) o cronológicos (10,9%) a la hora de seleccionar palabras clave.

Dentro de la clasificación de las palabras clave cronológicas, destaca que 1968 solo aparezca en 2019 y de una manera muy protagonista, entre los primeros puestos

de las más utilizadas ese año. La razón de esto la podemos encontrar en la publicación en *History of Education and Children's Literature*, en el número 2 del volumen 14, del monográfico sobre los movimientos estudiantiles que tuvieron lugar en la década de 1960 (Rico y Huerta, 2019), compuesto por 10 trabajos.

También llama la atención la aparición del término *epidemic* en 2022. A pesar de aparecer en los últimos puestos dentro de la lista de las palabras clave más utilizadas durante ese año, su mera presencia es llamativa dentro del campo de investigación de la HE, por lo actual de las implicaciones en educación de la situación derivada de la pandemia de COVID-19.

En cuanto a los términos que han estado presentes de una manera más constante a lo largo de los años, hay que mencionar el mantenimiento del interés en los temas relacionados con Italia, el siglo XX, la combinación de siglos XIX y XX y, en menos medida, exclusivamente el siglo XIX. También destaca el aumento progresivo de artículos relacionados con España (Hernández et al., 2020) desde 2015, así como la entrada en la lista en el último año de temas relacionados con Brasil, esclavitud o género, este último entendido como construcción cultural y diferenciado del término *women*, también dentro de la lista pero cuyo interés se circunscribe al año 2010.

Por último, en el análisis temático y de épocas realizado, específico por la posibilidad que se ha implementado en Hecumen de añadir esta información cualitativa a los registros bibliográficos, ha revelado que, en la muestra de 1863 artículos clasificados pertenecientes a las revistas *History of Educacion. Journal of the History of Education Society, History of Education and Children's Literature, History of Education Quarterly, História da Educação* y *Paedagogica Historica: International Journal of the History of Education*, la temática más estudiada ha sido la de "Innovación educativa y renovación pedagógica". Si consideramos la mitad de los artículos de la muestra, esta ha sido dedicada a las temáticas (en orden) "Innovación educativa y renovación pedagógica" y "Movimientos sociales y educativos". En total, estas 2 categorías han aparecido en el 50,08% de los artículos. Destaca el salto tan abrupto en cuanto al interés

porcentual dedicado a las temáticas "Género y políticas de igualdad" e "Inclusión y atención a la diversidad", que suman en conjunto un 15,50% de la muestra.

Sin embargo, este orden temático, que se ha calculado de una manera global, difiere si entramos en detalle de cada una de las revistas. Así pues, la pareja de temáticas "Género y políticas de igualdad" e "Inclusión y atención a la diversidad" ha sido estudiada en el 32,83% de los artículos de la revista *History of Education Quarterly.* En el resto de revistas, la pareja de temáticas tiene unas cifras más bajas, siendo la menor de todas en la revista *History of Education and Children's Literature*, que apareció en un 7,72% de los artículos.

Por último, en lo que respecta a las épocas estudiadas en esta submuestra de artículos y revistas, un 48,90% de las investigaciones se centraron en los siglos XX, XIX y XXI, en este orden, aunque hay un salto abrupto entre el estudio del siglo XIX (14,69% de los artículos) y el XXI, que desciende hasta un 4,05%.

6.3 Limitaciones y prospectiva

En cuanto a las limitaciones de la presente tesis, en lo que respecta al desarrollo de Hecumen cabe destacar que solo se ha implementado la importación de los registros bibliográficos de Scopus. Cada una de las bases de datos generalistas que permiten exportar datos bibliográficos tiene su propio formato, lo que obliga a implementar sistemas de importación específicos para cada una de ellas. Además, si se desea mantener la base de datos de Hecumen actualizada, hay que verificar que estos formatos no hayan cambiado, puesto que en ese caso la importación sería incorrecta. Esto hace que el sistema sea dependiente, tanto a priori como a posteriori, de los formatos de archivo de exportado de las bases de datos generalistas.

Otra de las limitaciones la podemos encontrar en que la muestra de artículos clasificados para el entrenamiento de las IAs era reducida (n = 256), hecho que debe ser tenido en cuenta a la hora de analizar los resultados, ya que los sistemas de entrenamiento de IAs se basan en el suministro de un número de muestras varios órdenes de magnitud superiores. A pesar de esta limitación, tal como se ha descrito, los resultados han sido positivos porque el rendimiento en las clasificaciones ha sido elevado y se debe tener en cuenta que con una muestra superior de artículos clasificados, dichos rendimientos incluso podrían mejorar.

Centrándonos en el estudio bibliométrico, la limitación fundamental del estudio radica en que hay que tener presente la asuencia de datos de determinados años debido a su inexistencia en Scopus, cuestión que se ha reflejado en la tabla 11, aunque esta limitación queda circunscrita solo a una parte de las revistas y a unos años concretos. Otra de las limitaciones radica en que se ha utilizado el *software* de desarrollo propio Hecumen, para el cual no se han implementado todas las funcionalidades con las que cuentan otras herramientas específicas como Bibliometrix. Es por esto que algunos análisis, como por ejemplo la comparativa entre las citas globales y las citas locales (dentro del *dataset*), no han podido ser realizadas.

Las posibilidades prospectivas a partir de este punto son amplias y variadas. En primer lugar, el desarrollo que se ha mostrado de Hecumen corresponde a la parte privada de importado y catalogación de artículos. Para poder plantear la plataforma para la comunidad global de investigadores de HE se debe desarrollar la parte pública, con las funcionalidades de búsqueda por temáticas y épocas estudiadas, así como otras opciones para mantenerse al día de los últimos artículos importados en la base de datos. Esta parte de la aplicación es vital de cara a visibilizar el trabajo que se está realizando en el proyecto "Connecting History of Education. Redes internacionales, producción científica y difusión global".

Por otra parte, las IAs especializadas en la clasificación temática debe ser conectada con Hecumen para facilitar la tarea del etiquetado y clasficación. Además, se debe implementar una IA especializada en la identificación de épocas estudiadas, ya que esta parte ha quedado al margen de la presente investigación.

Al respecto del entrenamiento de las IAs, cabe analizar comparativamente si el hecho de suministrar solo algunos de los metadatos estudiados (sólo título, resumen o palabras clave), o ciertas permutaciones de ellos (título+resumen, título+palabras clave…) mejora la capacidad de clasificación de la IA. También resulta de interés el variar el valor de las *n* palabras con más frecuencia de aparición en el artículo, para comprobar cuáles son los valores que mejores resultados ofrecen.

Con esta identificación de la mejor configuración de la IA para la identificación temática y de épocas, una línea prospectiva obvia es volver a realizar el entrenamiento de la IA para identificar si mejora el porcentaje de eficacia del sistema, con el objetivo de poder acercarse al 100% de éxito en la identificación de los aspectos específicos en la producción científica en HE.

6.4 Conclusiones

La presente tesis comenzaba formulando una pregunta: ¿se puede analizar cuantitativamente la complejidad contenida en las investigaciones en HE? Para responderla, la investigación ha realizado un recorrido que ha partido, en primer lugar, de la cienciometría, sus orígenes, evolución y problemáticas. Independientemente de las críticas que ha recibido por este aspecto, es innegable la importancia de esta ciencia y sus índices en la comunidad científica internacional hoy en día.

Posteriormente se han identificado las particularidades que tienen las investigaciones en HE, como son las temáticas y las épocas estudiadas, y que quedan al margen de la información bibliográfica contenida en los registros de las bases de datos generalistas. Se ha planteado un acercamiento a la biblioteconomía y a la hermenéutica para la interpretación y la extracción de estos datos.

Así pues, la primera conclusión de esta tesis es que, teniendo en cuenta que tanto las temáticas como las épocas estudiadas son dos objetos clave de los estudios de la HE, cualquier estudio que se limite a utilizar los datos bibliográficos tradicionales no va a reflejar completamente la riqueza de estas investigaciones. Esto es así porque los datos bibliográficos tradicionales no registran este tipo de información, con lo que los datos sobre épocas y temáticas estudiadas quedan invisibilizados en los estudios bibliométricos sobre HE.

Esta conclusión ha llevado a la siguiente problemática analizada en la presente tesis: ¿cómo se pueden almacenar estos datos específicos junto con la información bibliográfica tradicional? La respuesta ha sido el desarrollo de una base de datos específica para HE con la metodología DBR, llamada Hecumen, y que permite aunar tanto la información bibliográfica tradicional, como las valoraciones cualitativas sobre estos datos específicos de HE. Se ha descrito su desarrollo, sopesando en cada punto las diferentes opciones que se podían seguir, incidiendo especialmente en la problemática referida a la contingencia y dispersión de algunos metadatos a almacenar en la base de datos, lo que ha obligado a analizar diferentes aproximaciones teóricas para el

modelado de dicha base de datos. Tal como exige la metodología DBR, el desarrollo debe finalizar con unos principios de diseño que constituyen las conclusiones de dicha etapa de la tesis.

Así pues, la segunda conclusión de la tesis consiste en que las bases de datos diseñadas para almacenar datos bibliográficos deben tener en cuenta la contingencia y dispersión de algunos metadatos, y deben estar previstas a la ampliación de los mismos en el futuro sin que tenga demasiadas implicaciones en el código de la herramienta. Tras el análisis de diferentes alternativas, la conclusión ha sido que la combinación del modelo EAV junto con un modelo relacional permite integrar la rapidez y facilidad del uso del último, junto con la flexibilidad y adaptabilidad del primero.

La tercera conclusión es que, cuando los requisitos del desarrollo de un *software* es que debe ser accesible por diversos investigadores e investigadoras simultáneamente, trabajando sobre una misma base de datos, el implementarlo como una aplicación *web* es una ventaja respecto al desarrollo como aplicación de escritorio. Y no solo respecto a su uso en remoto y simultáneo, sino también por la facilidad de publicar las correcciones de errores o nuevas versiones y mejoras.

La cuarta conclusión se basa en que, una vez decidido que la aplicación iba a ser desarrollada como aplicación *web*, el uso de *web frameworks* facilita el desarrollo de la herramienta y optimiza el tiempo necesario a invertir, ya que el desarrollador puede centrarse en las funcionalidades específicas, dando por sentadas las cuestiones básicas sobre seguridad, comunicación, rutas o sobre los patrones de diseño subyacentes.

También relacionado con el tiempo necesario para desarrollar una herramienta de este tipo, la quinta conclusión afirma que centrarse en un PMV en lugar de abordar un desarrollo completo desde el inicio, permite obtener un prototipo funcional que, además, permite ser testado por los investigadores e investigadoras, lo que facilita las sucesivas iteraciones dentro del DBR.

Una vez desarrollado el PMV de Hecumen, ha surgido una problemática relacionada con la complejidad de catalogar manualmente un volumen grande de

artículos de HE. La pregunta, en este caso, ha sido: ¿se puede automatizar el proceso de clasificación de artículos según los parámetros específicos de HE?

De nuevo, nos hemos acercado a las ciencias de la computación, en concreto a la IA aplicada a la resolución de problemas de clasificación. Se ha conceptualizado la tecnología de la IA, así como los diferentes algoritmos, realizando diversos experimentos para poder identificar el más adecuado para la clasificación de textos. Además, se ha propuesto la metodología a seguir basada en la identificación de las *n* palabras más repetidas en cada artículo, lo que ha sido la información necesaria para nutrir el aprendizaje de diversas IAs conexionistas. Los resultados han revelado que, con una muestra relativamente reducida de n = 256 artículos para el entrenamiento, el algoritmo RF ha permitido clasificar los artículos por categorías con unos valores de *accuracy* (0,7), *precision*, *recall* y *f-measure* (0,69) aceptables.

En este contexto, la sexta conclusión está en la línea de diversos autores y autoras expuestos a lo largo de la tesis, y es que las muestras desbalanceadas provocan un sesgo de detección hacia la submuestra más representada y, al mismo tiempo, reducen la capacidad de detección de la submuestra menos representada. Es por esto que hay que recurrir a diversas técnicas para lograr una muestra balanceada previamente al entrenamiento de la IA.

Atendiendo a los resultados y teniendo en cuenta la muestra reducida con la que se trabajó, se puede afirmar como séptima conclusión que la identificación de las palabras con mayor frecuencia de aparición, basado en la ley de Zipf, permite obtener unos resultados buenos a la hora de clasificar textos.

En relación con esto, la octava conclusión extraída de la investigación expone que el algoritmo RF, que es un tipo de algoritmo de árboles de decisión, ha obtenido los mejores resultados para clasificar los textos en base a las palabras de mayor frecuencia de aparición.

La novena y última conclusión dentro del contexto de la IA afirma que la categoría "Movimientos sociales y educativos" ha sido la que peor rendimiento ha

obtenido a la hora de ser clasificada, probablemente porque cada artículo contenga un conjunto de palabras de mayor frecuencia de aparición específicas del movimiento social concreto estudiado. Esto implica que dichas palabras no se repetirán en otros artículos de la misma categoría, pero que estudian diferentes movimientos sociales. Al no emerger patrones entre los artículos, la IA no es capaz de identificar los artículos dentro de esta categoría.

Con toda esta base metodológica y de herramientas, para finalizar, se ha elaborado un estudio bibliométrico de 11 revistas especializadas de HE de relevancia internacional, cubriendo la ventana temporal desde 1961 hasta la actualidad. Para ello se ha recurrido a las herramientas desarrolladas previamente. La muestra ha estado formada por 5217 artículos, mediante los cuales se ha realizado un análisis de la evolución de la producción, así como el estudio de las citaciones globales, los autores, su colaboración y la representación por sexo entre los más productivos y productivas. También se han analizado los países más frecuentes en las afiliaciones, las instituciones y los idiomas. Se ha profundizado en el análisis de los temas a través del estudio de las palabras clave, tanto por sus apariciones en parejas como por la evolución en su interés a lo largo de la última década.

La décima conclusión es que se ha detectado un punto de inflexión en 2006, año en el que se produce una eclosión en la producción científica debido al aumento significativo de publicaciones del campo de HE. El mayor peso en cuanto a artículos analizados en la muestra recae en las revistas *Paedagogica Historica: International Journal of the History of Education* e *History of Educacion. Journal of the History of Education Society*, que aportan un 54,44%, siendo la segunda de ellas la que publicó el 60% de los 10 artículos más citados a nivel global. Reino Unido, Estados Unidos, Brasil, España e Italia, en este orden, son los primeros 5 países en cuanto a afiliaciones de los autores.

En cuanto a la undécima conclusión, destacan los bajos índices de colaboración en el campo de HE entre autores, puesto que un 77,10% de los artículos fueron escritos

en solitario. También se ha constatado que la representación de cada sexo en cuanto a las autoras y autores más productivos es paritaria.

La duodécima conclusión se centra en el idioma de las investigaciones. A este respecto, resulta muy llamativo que, a pesar de que solo el 30,05% de las afiliaciones corresponden a autores de Reino Unido y Estados Unidos, el 73,76% de los artículos de la muestra está redactado en inglés, lo que refleja la absoluta preponderancia del inglés en el campo de la HE a pesar de las bajas afiliaciones a países de habla inglesa.

El análisis de las palabras clave ha mostrado temáticas que se pueden agrupar en 3 grandes núcleos: cronológicos, geoespaciales y conceptuales, siendo este último el más recurrido por parte de los investigadores a la hora de catalogar sus estudios.

En base a este análisis de palabras clave se ha establecido la decimotercera conclusión, que afirma que los siglos más estudiados han sido el XIX y el XX, y se ha detectado una constancia a lo largo de los años en temas relacionados con Italia, así como un aumento progresivo de artículos relacionados con España.

De manera específica para este estudio bibliométrico y basándose en las posibilidades que ofrece la herramienta Hecumen, se ha realizado un análisis temático y de épocas en una submuestra de las revistas estudiadas en el análisis bibliométrico global.

Así pues, la decimocuarta y última conclusión muestra que la mitad de los artículos de la submuestra analizada trataban las temáticas de "Innovación educativa y renovación pedagógica" y "Movimientos sociales y educativos", siendo las temáticas "Género y políticas de igualdad" e "Inclusión y atención a la diversidad" las menos estudiadas, aunque con ciertas diferencias en función de la revista analizada. En cuanto a las épocas, de una manera constante se han estudiado los siglos XX, XIX y XXI.

Para terminar, la finalidad de esta tesis ha sido aumentar la comprensión del campo de la HE desde una perspectiva global, así como el dotar a la comunidad de investigadoras e investigadores de HE de una serie de herramientas que permitan

analizar el pasado y presente de las investigaciones con la finalidad de poder abordar los retos futuros de las investigaciones en HE.

Referencias

Aborisade, O., & Anwar, M. (2018). Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 269-276). IEEE. DOI: 10.1109/IRI.2018.00049

Albisetti, J. C. (1993). The feminization of teaching in the nineteenth century: a comparative perspective. *History of Education, 22*(3), 253-263. DOI: 10.1080/0046760930220305

Aldabas-Rubira, E. (2002). Introducción al reconocimiento de patrones mediante redes neuronales. In *IX Jornades de Conferències d'Enginyeria Electrònica del Campus de Terrassa, Terrassa, España, del 9 al 16 de Diciembre del 2002*.

Alshaikh, K., Bahurmuz, N., Torabah, O., Alzahrani, S., Alshingiti, Z., & Meccawy, M. (2021). Using Recommender Systems for Matching Students with Suitable Specialization: An Exploratory Study at King Abdulaziz University. *International Journal of Emerging Technologies in Learning (iJET), 16*(3), 316-324. DOI: 10.3991/ijet.v16i03.17829

Amiel, T., & Reeves, T. C. (2008). Design-Based Research and Educational Technology: Rethinking Technology and the Research Agenda. *Educational Technology & Society, 11*(4), 29–40

Apté, C., Damerau, F., & Weiss, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS), 12*(3), 233-251.

Araújo Ruiz, J.A., & Arencibia Jorge, R. (2002). Informetría, bibliometría y cienciometría: aspectos teórico-prácticos. *Acimed, 10*(4), 5-6.

Araujo, C.A. (2006). Bibliometría: evolução histórica e questões atuais. *Questão, 12*(1), 11-32.

Ardanuy, J. (2012). *Breve introducción a la bibliometría*. Universitat de Barcelona.

Arini, F. Y., Arifudin, R., & Aris, M. (2019). Applied structured database in a small project. Trabajo presentado en Journal of Physics: Conference Series, 1321(3). DOI: 10.1088/1742-6596/1321/3/032130

Asimov, I. (2006). *Fundación*. Bruguera.

Ávila-Toscano, J. H. (2018). El estudio de la ciencia y sus productos. Aproximación a la sociología de la ciencia y el conocimiento científico. In *Cienciometría y bibliometría. El estudio de la producción científica: Métodos, enfoques y aplicaciones en el estudio de las Ciencias Sociales* (pp. 27-48). Corporación Universitaria Reformada.

Batra, S., Sachdeva, S., & Bhalla, S. (2018). Entity attribute value style modeling approach for archetype based data. *Information, 9*(1), 2. DOI: 10.3390/info9010002

Beck, M.T. (1978). Editorial statements. *Scientometrics, 1*(1), 3-4.

Blanco, N., & Pirela, J. (2016). La complementariedad metodológica: Estrategia de integración de enfoques en la investigación social. *Espacios Públicos, 19*(45),97-111.

Brehony, K. J. (2004). A new education for a new era: the contribution of the conferences of the New Education Fellowship to the disciplinary field of education 1921–1938. *Paedagogica historica, 40*(5-6), 733-755. DOI: 10.1080/0030923042000293742

Brindha, T., & Murugesapandian, N. (2016). Scientometrics Tools and Techniques: An Overview. *Shanlax International Journal of Arts, Science & Humanities, 4*(2), 91-92.

Brookshear, J. G. (1993). *Teoría de la computación: lenguajes formales, autómatas y complejidad*. Adison Wesley.

Bruno, V., Tam, A., & Thom, J. (2005). Characteristics of web applications that affect usability: a review. In P*roceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (pp. 1-4).

Budiman, E., Jamil, M., Hairah, U., & Jati, H. (2017). Eloquent object relational mapping models for biodiversity information system. In *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)* (pp. 1-5). IEEE.

Cardona, E. C. (2010). Teaching Drawing through the textbooks of High School Education published in Spain (1915-1990): Bibliometric study of contents. *Revista de Educacion, 352*, 517-544.

Castrillón, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Formación universitaria, 13*(1), 93-102. DOI: 10.4067/S0718-50062020000100093

Cerrada Somolinos, J.A., Collado Machuca, M.E., Gómez Palomo, S.R., & Estivariz López, J.F. (2000). In J.A. Cerrada Somolinos (Coord.), *Introducción a la ingeniería del software* (pp. 35-83). Centro de Estudios Ramón Areces.

Chernyi, A. I. (2009). The ISI Web of Knowledge, a modern system for the information support of scientific research: a review. *Scientific and Technical Information Processing, 36*(6), 351-358. DOI: 10.3103/S0147688209060069

Cohen, S. (1983). The mental hygiene movement, the development of personality and the school: The medicalization of American education. *History of Education Quarterly, 23*(2), 123-149.

Colorado, Y. S., & Anaya, O. P. (2018). La evaluación de la actividad científica: Indicadores bibliométricos. In *Cienciometría y bibliometría. El estudio de la producción científica: Métodos, enfoques y aplicaciones en el estudio de las Ciencias Sociales* (pp. 96-118). Corporación Universitaria Reformada.

Cucuzza, H. R. (1996). Hacia una redefinición del objeto de estudio de la Historia Social de la Educación. In *Historia de la educación en debate* (pp. 125-148). Editorial Miño y Dávila.

Curry, E., & Grace, P. (2008). Flexible self-management using the model-view-controller pattern. *IEEE software, 25*(3), 84-90. DOI: 10.1109/MS.2008.60

Design-based Research Collective (2003). Design-based research: an emerging paradigm for educational inquiry. *EducRes, 32* (1), 5–8. DOI: 10.3102/0013189X032001005

Dobrev, D. (2012). A definition of artificial intelligence. arXiv preprint arXiv:1210.1568.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research, 133*, 285-296. DOI: 10.1016/j.jbusres.2021.04.070

Dyhouse, C. (1976). Social Darwinistic ideas and the Development of women's education in England, 1880–1920. *History of Education*, *5*(1), 41-58. DOI: 10.1080/0046760760050105

Elango, B., & Rajendran, P. (2012). Authorship Trends and Collaboration Pattern in the Marine Sciences Literature: A Scientometric Study. *Int. J. Inf. Dissem. Technol*, 166-169.

Ellegaard, O., & Wallin, J.A. (2015). The bibliometric analysis of scholarly production: How great is the impact?. *Scientometrics, 105*, 1809–1831. DOI: 10.1007/s11192-015-1645-z

Feng, Y. (2006). PROC SQL: When and How to Use It?. Trabajo presentado en *Proceedings of 2006 NESUG Conference.*

Frank, E., Hall, M.A., & Witten, I.H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Fuchs, E. (2004). Educational sciences, morality and politics: international educational congresses in the early twentieth century. *Paedagogica historica, 40*(5-6), 757-784. DOI: 10.1080/0030923042000293751

Ganslandt, T., Mueller, M., Krieglstein, C. F., Senninger, N., & Prokosch, H. U. (1999). A flexible repository for clinical trial data based on an entity-attribute-value model. In *Proceedings of the AMIA Symposium* (p. 1064). American Medical Informatics Association.

Garfield, E. (1970). Citation indexing for studying science. *Nature, 227*(5259), 669-671.

Garfield, E. (1979). Mapping the structure of science. *Citation Indexing: Its Theory and Applications in Science, Technology, and Humanities, 2*, 98-147.

Garfield, E. (2009). From the science of science to Scientometrics visualizing the history of science with HistCite software. *Journal of Informetrics, 3*(3), 173-179. DOI: 10.1016/j.joi.2009.03.009

Garrido Arilla, M. R. (1990). Contienda por el control documentario: etapas pretécnica y técnica en catalogación. *Boletín de la Asociación Andaluza de Bibliotecarios, 6*(19), 6.

Gil, D., Fernández-Alemán, J. L., Trujillo, J., García-Mateos, G., Luján-Mora, S., & Toval, A. (2018). The effect of green software: a study of impact factors on the correctness of software. *Sustainability, 10*(10), 3471. DOI: 10.3390/su10103471

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315-323). JMLR Workshop and Conference Proceedings.

Green, R. (1996). The design of a relational database for large-scale bibliographic retrieval. *Information Technology and Libraries, 15*(4), 207-221.

Grondin, J. (2014). *¿Qué es la hermenéutica?*. Herder Editorial.

Guichot Reina, V. (2006). Historia de la educación: reflexiones sobre su objeto, ubicación epistemológica, devenir histórico y tendencias actuales. *Revista Latinoamericana de Estudios Educativos, 2*(1), 11-51.

Gutiérrez Fernández, A., Guerrero Higueras, G. R., Guerrero Higueras, Á. M., Conde González, M. Á., & Fernández Llamas, C. (2020). Evaluación del resultado académico de los estudiantes a partir del análisis del uso de los Sistemas de Control de Versiones. *RIED. Revista iberoamericana de educación a distancia*.

Hernández Huerta, J. L. & Payà, A. (2022). International Standing Conference for the History of Education (ISCHE): Networks, internationalisation, and scientific communication. *El Futuro Del Pasado*. DOI: 10.14201/fdp.28213

Hernández Huerta, J. L., Cagnolati, A. (2015). En la Historia de la Educación. La gestión editorial, las revistas de Historia de la Educación y Espacio, Tiempo y Educación. *History of Education & Children's Literature, 10*(1), pp. 39-55.

Hernández Huerta, J. L., González Gómez, S., Pérez Miranda, I. (2020). History of Education in the Iberian Peninsula (2014-2019). Societies, journals and conferences in Spain and Portugal. *Histoire de l'éducation*, (154), pp. 177-206. DOI: 10.4000/histoire-education.5730

Hernández Huerta, J. L., Payà, A., & Sanchidrián, C. (2019). El mapa internacional de las revistas de historia de la educación. *Bordón. Revista de pedagogía, 71*(4), 45-64. DOI: 10.13042/Bordon.2019.69624

Hesse, L.A.C. (2010). *Bibliothéconomie: Ou Nouveau Manuel Complet Pour L'arrangement, La Conservation Et L'administration Des Bibliothèques*. Nabu Press.

Heward, C. (1993). Men and women and the rise of professional society: the intriguing history of teacher educators. *History of Education, 22*(1), 11-32. DOI: 10.1080/0046760930220102

Hofstetter, R., Fontaine, A., Huitric, S., Picard, E. (2014). Mapping the discipline history of education. Paedagogica Historica, 50(6), pp. 871-880. DOI: 10.1080/00309230.2014.948017

Hofstetter, R., Huitric, S. (2020). La Carte et le Miroir. Ancrages, Enjeux et Horizons de l'Histoire de l'Education. Histoire de l'éducation, (154), pp. 9-48. DOI: 10.4000/histoire-education.5485.

Hoskin, K. (1979). The examination, disciplinary power and rational schooling. *History of Education*, *8*(2), 21-135. DOI: 10.1080/0046760790080205

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications, 39*(1), 1503-1509. DOI: 10.1016/j.eswa.2011.08.040

Jokhan, A., Chand, A. A., Singh, V., & Mamun, K. A. (2022). Increased digital resource consumption in higher educational institutions and the artificial intelligence role in informing decisions related to student performance. *Sustainability, 14*(4), 2377. DOI: 10.3390/su14042377

Kammerer, K., Reichert, M., & Pryss, R. (2021). Ambalytics: A scalable and distributed system architecture concept for bibliometric network analyses. *Future Internet, 13*(8), 203. DOI: 10.3390/fi13080203

Kamruzzaman, S. M. (2010). Text classification using artificial intelligence. arXiv preprint arXiv:1009.4964.

Kappel, G., Pröll, B., Reich, S., & Retschitzegger, W. (2006). *Web engineering.* Wiley.

Karamcheti, K. (2007). Design and implementation of bibliographic database.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. *American documentation, 14*(1), 10-25. DOI: 10.1002/asi.5090140103

Knuth, D.E. (1997). *The art of computer programming, volume 1 (3rd Ed.): fundamental algorithms*. Addison Wesley Longman Publishing Co. Inc.

Koenig, M. E. D., & Bookstein, A. (1995). Fifth Biennial Conference of the International Society for Scientometrics and Informatrics. NJ: Learned Information.

Kuhn, T. S. (2019). *La estructura de las revoluciones científicas*. Fondo de cultura económica.

Kumar, S., & Kumar, S. (2008). Collaboration in Research Productivity in Oil Seed Research Institutes of India. In *Proceedings of the International Conference on Webometrics, Informetrics and Scientometrics, Berlin, Germany, 28 July 2008* (pp. 1–18).

Laaziri, M., Benmoussa, K., Khoulji, S., Larbi, K. M., & El Yamami, A. (2019). A comparative study of laravel and symfony PHP frameworks. *International Journal of Electrical and Computer Engineering, 9*(1), 704.

Landa, B. D., Romero, R. M., & Rodriguez, W. J. M. (2021). Rendimiento académico de estudiantes en Educación Superior: predicciones de factores influyentes

a partir de árboles de decisión. *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales, 23*(3), 616-639. DOI: 10.36390/telos233.08

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research, 18*(1), 559-563.

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. *Expert systems with applications, 39*(12), 11303-11311. DOI: 10.1016/j.eswa.2012.02.063

Lott, J., & Patterson, D. (2007). *Action Script 3. Patrones de diseño*. Ediciones Anaya Multimedia.

Martínez, M. (2005). *El paradigma emergente: Hacia una nueva teoría de la racionalidad científica*. México: Trillas.

Masic, I., & Jankovic, S. M. (2021). Inflated co-authorship introduces bias to current scientometric indices. *Medical Archives, 75*(4), 248. DOI: 10.5455/medarh.2021.75.248-255

Mays, S., Gowland, R., Halcrow, S., & Murphy, E. (2017). Child bioarchaeology: Perspectives on the past 10 years. *Childhood in the Past, 10*(1), 38-56. DOI: 10.1080/17585716.2017.1301066

McCulloch, G. (2011). *The struggle for the history of education*. Routledge.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115-133.

Meira, L. M. D. (2020). About curriculum history: themes, concepts and references of Brazilian researches. *Revista Brasileira de Educação, 25*. DOI: 10.1590/S1413-24782020250051

Meneses, J., & Rodríguez-Gómez, D. (2011). El cuestionario y la entrevista.

Mengual-Andrés, S., Chiner, E., & Gómez-Puerta, M. (2020). Internet and people with intellectual disability: A bibliometric analysis. *Sustainability, 12*(23), 10051. DOI: 10.3390/su122310051

Millán, J. D., Polanco, F., Ossa, J. C., Béria, J. S., & Cudina, J. N. (2017). La cienciometría, su método y su filosofía: Reflexiones epistémicas de sus alcances en el siglo XXI. *Revista Guillermo De Ockham, 15*(2), 17-27. DOI: 10.21500/22563202.3492

Mira, J., Delgado, A.E., Boticario, J.G., & Díez, F.J. (2003). *Aspectos básicos de la inteligencia artificial*. Sanz y Torres.

Moreno, C. M. (1997). Técnicas bibliométricas aplicadas a las estudios de usuarios. *Revista general de información y documentación, 7*(2), 41-41.

Moschovakis, Y. N. (2001). What is an algorithm?. In *Mathematics unlimited—2001 and beyond* (pp. 919-936). Springer, Berlin, Heidelberg.

Mourdi, Y., Sadgal, M., Fathi, W. B., & El Kabtane, H. (2020). A machine learning based approach to enhance MOOC users' classification. *Turkish Online Journal of Distance Education, 21*(2), 47-68. DOI: 10.17718/tojde.727976

Orera, L. O. (1995). Evolución histórica del concepto de biblioteconomía. *Revista General de información y Documentación, 5*(2), 73.

Payà, A., & Hernández Huerta, J. L. (2019). Student movements of the «long 1960s». Steps towards the cultural revolution, social change and political transformation. *History of Education and Children's Literature, 14*(2), 13-20.

Payà, A., Duart, J. M., & Mengual-Andrés, S. (2016). Histoedu, redes sociales e historia de la educación: el pasado pedagógico desde el presente educativo. *Education in the Knowledge Society*, *17*(2), 55-72. DOI: 10.14201/eks20161725572

Piroska, H., Gyula, F., & Ioan-Cosmin, S. (2012). Data storage for smart environment using non-SQL databases. In *2012 IEEE 8th International Conference on Intelligent Computer Communication and Processing,* 305-308.

Prabhat, A., & Khullar, V. (2017). Sentiment classification on big data using Naïve bayes and logistic regression. In *International conference on computer communication and informatics (ICCCI)* (pp. 1–5).

Price, D.J.S. (1963). *Little science, big science*. Columbia University Press.

Price, D.J.S. (1965) The Science of Science. *Bulletin of the Atomic Scientists*, 21:8, 2-8. DOI: 10.1080/00963402.1965.11454842

Rashid, A., & Chitchyan, R. (2003). Persistence as an Aspect. In *2nd International Conference on Aspect-Oriented Software Development*, Boston, MA. 120-129. DOI: 10.1145/643603.643616

Rich, E., & Knight, K. (1991). *Learning in neural network*. McGraw-Hill.

Rico, A.P., & Motilla Salas, X. (2016). Web 2.0, social networks and the history of education in Spain: creating a scientific collaborative space (Histoedu. net). *Web 2.0, social networks and the history of education in Spain: creating a scientific collaborative space (HistoEdu. net)*, 249-263.

Ries, E. (2011). *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Currency.

Roda-Segarra, J., Mengual-Andrés, S., & Martínez-Roig, R. (2022). Using Virtual Reality in Education: a bibliometric analysis. *Campus Virtuales, 11*(1), 153-165. DOI: 10.54988/cv.2022.1.1006

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review, 65*(6), 386.

Ruiz, J., Rabazas, T., & Ramos, S. (2006). The reception of new Education in Spain by means of Manuals on the history of Education for teacher training colleges (1898–1976). *Paedagogica historica, 42*(1-2), 127-141. DOI: 10.1080/00309230600552070

Ruiz, M. E., & Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Information retrieval, 5*(1), 87-118.

Salas, J. A. (2019). Historia general de la educación.

Sánchez Meca, D. (2019). *Iniciación a la Teoría del Conocimiento*. Dykinson.

Schleiermacher, F. (2002). *Schleiermacher: Lectures on Philosophical Ethics*. Cambridge University Press.

Su, J.S., Zhang, B.F., & Xu, X. (2006). Advances in machine learning based text categorization. *Ruan Jian Xue Bao (Journal of Software), 17*(9), 1848-1859.

Supaartagorn, C. (2011). PHP Framework for database management based on MVC pattern. *AIRCC's International Journal of Computer Science and Information Technology, 3*(2), 251-258.

Valle, S. (2018). Retos de las Ciencias Sociales en la producción científica. In *Cienciometría y bibliometría. El estudio de la producción científica: Métodos, enfoques y aplicaciones en el estudio de las Ciencias Sociales* (pp. 49-76). Corporación Universitaria Reformada.

Whitehead, C. (2005). The historiography of British imperial education policy, Part I: India. *History of Education, 34*(3), 315-329. DOI: 10.1080/00467600500065340

Whitehead, C. (2005). The historiography of British Imperial education policy, Part II: Africa and the rest of the colonial empire. *History of Education, 34*(4), 441-454. DOI: 10.1080/00467600500138147

Wittgenstein, L. (1989). Conferencia sobre Ética. In C. Gómez Sánchez (Ed.), *Ética. Doce textos fundamentales del siglo XX* (pp. 137-150). Alianza editorial.

Wouters, P. F. (1999). *The citation culture* (Doctoral dissertation, Universiteit van Amsterdam).

Xie, Z., Frimpong, E. A., & Lee, S. (2013). FishTraits version 2: Integrating ecological, biogeographic and bibliographic information. Trabajo presentado en *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries; 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2013*, Indianapolis, IN. 447-448. DOI: 10.1145/2467696.2467791

Yu, H. R. (2015). Design and implementation of web based on Laravel framework. *ACSR-Advances in Comptuer Science Research volume, 6*, 302. DOI: 10.2991/iccset-14.2015.66