Automatización de la integración de la producción científica en los sistemas institucionales de gestión de la investigación

Inmaculada Bravo García

DIRECTOR

José Antonio Merlo Vega

PLAN DE INVESTIGACIÓN

PROGRAMA DE DOCTORADO FORMACIÓN EN LA SOCIEDAD DEL CONOCIMIENTO

UNIVERSIDAD DE SALAMANCA

FECHA

31 de mayo de 2021

INTRODUCCIÓN

Un Sistema de Gestión de la Investigación (Current Research Information System, CRIS) es una herramienta que centraliza la gestión de toda la producción científica de una institución, principalmente los proyectos de investigación y las publicaciones científicas (Simons, 2017). Permite la visibilidad y la difusión de la actividad investigadora, facilita entre otros, los procesos de evaluación para obtener financiación, la realización del CV del investigador, la generación de informes de gestión.

El mantenimiento de esta información, en algunas instituciones, se realiza de modo manual recayendo sobre el investigador la responsabilidad de su actualización y curación (CRUE, 2018, Terán, 2015).

En nuestro caso de estudio, la USAL, la base de datos que centraliza la información es Universitas XXI, los proyectos de investigación son gestionados y alimentados por la Agencia de Gestión de la Investigación (AGI), pero la gestión de la producción científica recae en el investigador, que la actualiza manualmente, sin validación posterior por el servicio de Bibliotecas (Iribarren-Maestro, 2018) y sin apoyo por parte de los Servicios Informáticos para facilitar la alimentación automatizada procedente de fuentes de datos bibliográficas.

La investigación se basará en el análisis de varias bases de datos de producción científica existentes (como Dialnet, Web of Sciencie, Research Gate, Google Scholar, etc.), para sistematizar la sincronización entre las mismas y las bases de datos internas de la universidad de Salamanca tratando de dar visibilidad, fiabilidad y validez a estas últimas (Azeroual 2018, 218, 2019, 2020). Uno de los principales indicadores del prestigio de las universidades, es precisamente su producción científica (Codina, 2016; Azeroual, 2018). Disponer de esta información depurada, centralizada y favorecer su visibilidad es un factor positivo y prioritario para mejorar en los distintos rankings de producción científica de universidades tanto nacionales como internacionales (González-Pérez et al., 2017).

La tesis persigue también simplificar los procedimientos de alimentación de datos, (Andrade, 2007) ya que una de las principales quejas de los investigadores docentes es el esfuerzo burocrático al que son sometidos constantemente, ya que se les solicita en numerosas ocasiones (y de manera no unificada) información detallada sobre cada una de sus aportaciones científicas para distintos tipos de solicitudes, donde se necesita recabar todo o parte de su historial científico-investigador.

Se parte de la hipótesis de que, si se mantuviesen dichas aportaciones en las bases de datos institucionales de la propia universidad de modo semiautomático, disminuiría considerablemente este esfuerzo burocrático y repetitivo, pero también se mantendría actualizado el curriculum vitae de cada investigador con más facilidad y eficacia organizativa (García-Peñalvo, 2018).

La información acerca de las publicaciones científicas también es requerida habitualmente en varios procesos administrativos de la universidad, como el análisis de carga docente e investigadora del PDI, o las distintas solicitudes para la concesión de ayudas de programas propios de la USAL. Estos procesos en la actualidad suponen un importante esfuerzo por parte de múltiples actores, lo que obliga a su vez al investigador a repetir esa misma información en diferentes formatos. Además, requiere la posterior validación de toda esa información recabada en la pertinente base de datos.

Por todo ello, si dispusiéramos previamente de la información sobre la producción científica de los investigadores de la Universidad de Salamanca, de tal modo que está fuese rápidamente recopilada y estuviese actualizada y validada constantemente en nuestras bases de datos, estos procesos se

simplificarían significativamente, lo que redundaría en una mayor eficacia y fiabilidad lo que, a su vez, aliviaría considerablemente el trabajo burocrático del propio investigador a la hora de tener actualizado su currículum normalizado, para poder exportarlo a diferentes formatos y manejarlo en función de sus intereses y demandas.

La universidad de Salamanca dispone del Repositorio Institucional GREDOS (Ferreras Fernández, 2016) que podría ser alimentado de manera semiautomática desde estas bases de datos institucionales (González-Pérez et al., 2018; Díaz del Río, 2014), cuando la información fuera fiable y completa, redundando en la visibilidad e impacto de dichas publicaciones e incrementando la contribución a la ciencia abierta (Ramírez-Montoya et al., 2018; Hernández-Pérez, 2019,2020; De-Castro, 2019).

HIPÓTESIS DE TRABAJO Y PRINCIPALES OBJETIVOS A ALCANZAR

La tesis doctoral parte de la hipótesis de que es posible automatizar los procesos de integración de la información bibliográfica mediante la ingesta desde bases de datos estructuradas hacia los sistemas de gestión de la investigación.

Esta hipótesis se estructura en preguntas de investigación, objetivos generales y objetivos específicos.

Preguntas de investigación:

¿Es posible centralizar la información de forma semiautomatizada y unificada, recabando datos masivos a partir de los resultados de investigación de los miembros PDI de una universidad y recogiendo dicha información proveniente de distintas fuentes y bases de datos para que se encuentre constantemente actualizada en un portal de investigación único?

¿Cuál sería la carga de trabajo que recaería en cada uno de los actores: Investigador, Servicios Informáticos, Bibliotecas y AGI?

¿Se puede simplificar el actual sistema de gestión de la Universidad de Salamanca?

¿La información almacenada en las bases de datos de Universitas XXI investigación, es completa, está actualizada, puede reutilizarse?

¿Es posible integrar en UXXI-Investigación la información relevante que ofrecen las grandes bases de datos bibliográficas como Scopus, Web of Science o Dialnet?

¿Cómo son los portales institucionales de difusión y divulgación de la producción científica?

¿Qué parte de la información interna de Universitas XXI-Investigación, es aconsejable mostrar y difundir en el Portal del Investigador de la Universidad de Salamanca?

Objetivos generales:

Analizar los modelos de introducción de la información bibliográfica en los sistemas institucionales de gestión de la investigación.

Establecer una metodología para la comunicación entre los sistemas de gestión de la investigación y las bases de datos bibliográficas.

Sistematizar los procesos de integración de recursos externos de información bibliográfica para su ingesta en un sistema de gestión de la investigación.

Analizar las posibilidades de creación de un portal que visibilice la producción científica de una Universidad a partir de un sistema de gestión de la investigación.

Proponer un modelo de integración aplicable a la Universidad de Salamanca.

Objetivos específicos:

Establecer la metodología de exportación de datos bibliográficos desde bases de datos científicas empleadas en las convocatorias de evaluación de la investigación.

Identificar las posibilidades de integración de la información bibliográfica externa en el sistema Universitas XXI Investigación empleado por la Universidad de Salamanca.

Establecer procedimientos para la ingesta y depuración de la producción científica en el sistema Universitas XXI.

Analizar los procedimientos para facilitar al PDI la actualización continua de su currículum de forma semiautomatizada.

Diseñar un sistema para la Universidad de Salamanca que simplifique las gestiones administrativas centrales y que permita encontrar los medios para revertir la información actualizada de ida y vuelta en todos procesos de gestión interna de la Universidad.
Proponer un modelo automático de difusión de los datos en el portal institucional de la investigación a partir de la información contenida en el sistema de gestión de la investigación Universitas XXI.

METODOLOGÍA A UTILIZAR

La investigación parte del establecimiento del análisis del estado de la cuestión, de cada uno de los temas implicados en la presente investigación, para lo que se utilizará la metodología bibliográfica.

Para conseguir recabar información de publicaciones en las distintas bases de datos, que arrojan resultados de los méritos de investigación de los PDI, e integrarla en las bases de datos internas de la institución, sería necesario acudir a la metodología comparativa, analizando y comparando sistemáticamente los datos y formatos que encontramos en las distintas fuentes (Mesias, 2010, Biesenbender 2019), buscando paralelismos a utilizar, desviaciones a enmendar y lagunas a completar (Azeroual, 2019).

Será necesario proponer la aproximación de la solución informática, se utilizará la metodología DSR (Design Science Research) “es un enfoque de investigación riguroso que propone la construcción de artefactos para brindar una solución útil y efectiva a un problema de un dominio dado” (Rivera 2016).

En el análisis de cargas de trabajo de los diferentes actores y la contabilización y comparación del número de artículos científicos actuales y potenciales se acudirá a la metodología cuantitativa, basada en las técnicas bibliométricas.

La investigación doctoral empleará diferentes metodologías de la investigación, al tratarse de un estudio que por una parte requiere técnicas más tradicionales (investigación bibliográfica e investigación bibliométrica), pero también técnicas de análisis de casos y, sobre todo, investigación tecnológica, ya que la demostración de la hipótesis requiere de desarrollo informático.

MEDIOS Y RECURSOS MATERIALES DISPONIBLES

Este trabajo se desarrolla en el programa de Doctorado Formación en la Sociedad del Conocimiento (García-Peñalvo, 2013, 2014, 2017; García-Peñalvo et al., 2020), siendo su portal la principal herramienta de comunicación y visibilidad de los avances (García-Holgado et al., 2015; García-Peñalvo et al., 2019).

Debido al tema de la investigación que se realizará, se empleará tanto los recursos bibliográficos de la Universidad de Salamanca, como las plataformas objeto de la investigación. En síntesis, los medios y recursos son los siguientes:

Bases de datos bibliográficas especializadas y recursos de investigación disponibles en la Universidad de Salamanca.

Bases de datos de Universitas XXI Investigación, Universitas XXI Recursos Humanos y Universitas XXI Académico.

Bases de datos de Dialnet, Dialnet Métricas y Portal de investigación Dialnet (convenio Universidad de La Rioja-Fundación Dialnet y Universidad de Salamanca) (Mateo, 2015).

APIS de Web of Science y Scopus (mediante convenios firmados por el servicio de bibliotecas de la USAL).

Información abierta sobre investigación de los PDI de la USAL en Google Scholar.

Equipo informático y autorización para conexión remota a esas distintas bases de datos.

Máquinas virtuales con el software de desarrollo necesario para llevar a cabo todos estos desarrollos informáticos.

Software:

- Servidores de bases de datos:
  - Oracle
  - Vertica
  - MySql
- Clientes gráficos de bases de datos:
  - TOAD
  - DBaver
- Herramienta de integración de datos ETL:
  - Talend Open Studio
- Lenguajes de programación:

Phyton
Java
SQL

PLANIFICACIÓN TEMPORAL

La planificación del proyecto de investigación doctoral estará dividida en las diferentes fases que definimos a continuación:

PRIMERA ANUALIDAD

- Acotación del objeto de Estudio
- Marco teórico y SLR (Systematic Literature Review)
- Formulación de las hipótesis de partida y las preguntas de investigación asociadas a dichas hipótesis
- Desarrollo de la herramienta metodológica.

SEGUNDA ANUALIDAD

- En esta segunda anualidad, utilizando la metodología comparativa se realizará un análisis exhaustivo de los datos de las diferentes fuentes objeto de estudio. Se propondrá un marco que permita sincronizar información entre las fuentes externas analizadas y las internas.

Se dividirá el tratamiento de las publicaciones en tres fases temporales, abordando la primera y la segunda fase en esta anualidad y dejando la FASE 3 para la tercera anualidad.

FASE 1: Tratar publicaciones científicas de la última anualidad hasta ese momento.

Consistirá fundamentalmente en tres acciones divididas en tres grupos de información sobre publicaciones de investigación:

GRUPO I. Intersección entre las bases de datos en estudio.
- Generación de un archivo con diferencias entre ambas bases de datos
- Corregir las diferencias
- Validar publicaciones

GRUPO II. Publicaciones sólo existentes en las bases de datos internas de la USAL: Universitas-XXI
- Generar archivo con campos RIS de estas publicaciones
- Validarlas en UXXI
- Crearlas en DIALNET

GRUPO III. Publicaciones nuevas, solo existentes en las bases de datos de DIALNET
- Generar archivo con estas publicaciones
- Importar a UXXI-INV con el nuevo procedimiento de importación desarrollado, quedando realizada la validación de la publicación de modo automático.

FASE 2: Definición del procedimiento de funcionamiento para tratar las publicaciones mediante su actualización, validación y visibilización en el futuro portal del investigador de la USAL.

TERCERA ANUALIDAD

Continuando con el ciclo iniciado en la segunda anualidad se prosigue con la última fase:
FASE 3: Definición y ejecución del procedimiento para tratar las publicaciones anteriores a la última anualidad, para hacer una revisión histórica, desde la fecha de partida, que se decida, de todas las publicaciones anteriores a las tratadas en la FASE 1.

Participar en congresos de referencia internacional sobre gestión de la investigación, como euroCRIS Conference.

Participar en el foro técnico TECNIRIS haciendo partícipe a las universidades españolas y centros de investigación de los avances realizados y proponer en los grupos de RedIRIS la creación de un grupo de trabajo centrado en la automatización en la gestión de la producción científica.

Publicación de un artículo sobre la interoperabilidad entre el sistema de gestión de la investigación y el Repositorio institucional en revistas interesadas en temas afines, como la revista European Journal of Higher Education, United Kingdom, Taylor and Francis Ltd. la cual en la actualidad se sitúa en el cuartil Q2 de Scimago Journal Rank (SJR). Según el enfoque final del primer artículo se seleccionarán revistas de los primeros cuartiles de las categorías Library & Information Science de Scimago Journal & Country Rank (SJR) o Information Science & Library Science de Journal Citation Reports.

CUARTA ANUALIDAD

Discusión de resultados.
Participación en foros internacionales como euroCRIS.

Publicación en revistas de impacto, y difusión de artículos sobre el proceso diseñado.
Una de estas publicaciones se centrará en el análisis cuantitativo de las publicaciones corregidas, validadas e importadas automáticamente con el sistema propuesto en la tesis, así como las publicaciones insertadas en el Repositorio Institucional y su influencia en la evolución en los rankings de producción científica, se propondrá publicarlo en la revista Scientometrics, Springer Netherlands, la cual en la actualidad se sitúa en el cuartil Q1 de Scimago Journal Rank (SJR). En todo caso, se seleccionarán revistas similares una vez finalizado el artículo.

QUINTA ANUALIDAD
Elaboración de las conclusiones y la bibliografía de referencia.
Elaboración de los Anexos de la investigación.
Publicación en revistas de impacto, y difusión de artículos sobre los resultados obtenidos.
Uno de estos artículos se publicará en la revista International Journal of Information Management, United Kindong, Elsevier Ltd., la cual en la actualidad se sitúa en el cuartil Q1 de Scimago Journal Rank (SJR). Igualmente, se tendrán en cuenta revistas Q1 y Q2 de SJR y JCR en las que tenga cabida por su temática.

REFERENCIAS

Andrade, J. A. (2007). Structuration to research in information systems. 54, 15.

Azeroual et al. - (2018)—Analyzing data quality issues in research informat.pdf.

Azeroual, O., Saake, G., & Abuosba, M. (2019). ETL Best Practices for Data Quality Checks in RIS Databases. Informatics, 6(1), 10. https://doi.org/10.3390/informatics6010010

Azeroual, O., Saake, G., Abuosba, M., & Schöpfel, J. (2020). Data Quality as a Critical Success Factor for User Acceptance of Research Information Systems. Data, 5(2), 35. https://doi.org/10.3390/data5020035

Azeroual, O., Saake, G., & Schallehn, E. (2018). Analyzing data quality issues in research information systems via data profiling. International Journal of Information Management, 41, 50-56. https://doi.org/10.1016/j.ijinfomgt.2018.02.007

Azeroual, O., Saake, G., & Wastl, J. (2018). Data measurement in research information systems: Metrics for the evaluation of data quality. Scientometrics, 115(3), 1271-1290. https://doi.org/10.1007/s11192-018-2735-5

Azeroual, O., & Schöpfel, J. (2019). Quality Issues of CRIS Data: An Exploratory Investigation with Universities from Twelve Countries. Publications, 7(1), 14. https://doi.org/10.3390/publications7010014

Biesenbender, S., & Herwig, S. (2019). Support structures to facilitate the dissemination and implementation of a national standard for research information – the German case of the Research Core Dataset. Procedia Computer Science, 146, 131-141. https://doi.org/10.1016/j.procs.2019.01.088

Codina, L. (2016). Evaluación de la ciencia: Tan necesaria como problemática. El Profesional de la Información, 25(5), 715. https://doi.org/10.3145/epi.2016.sep.01

CRUE Universidades españolas. (2018). Estado de la cuestión de los CRIS en las universidades españolas. Subgrupo de Acceso Abierto. Línea II del Plan Estratégico de REBIUN. 36.

De-Castro, P. (2019). Progresos recientes en sistemas de gestión de la información científica. Anuario ThinkEPI, 13. https://doi.org/10.3145/thinkepi.2019.e13e04

Díaz del Río, Luis. (2014). *Integración del Sistema de Gestión de la Investigación (CRIS) con un Repositorio Institucional. El modelo de la Universidad Carlos III de Madrid*. 84.

Ferreras Fernández, T. (2016). Visibilidad e impacto de la literatura gris científica en repositorios institucionales de acceso abierto. Estudio de caso bibliométrico del repositorio Gredos de la Universidad de Salamanca [Universidad de Salamanca]. https://doi.org/10.14201/gredos.132444

García-Holgado, A., García-Peñalvo, F. J., & Rodríguez-Conde, M. J. (2015). Definition of a technological ecosystem for scientific knowledge management in a PhD Programme. In G. R. Alves & M. C. Felgueiras (Eds.), Proceedings of the Third International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'15) (Porto, Portugal, October 7-9, 2015) (pp. 695-700). ACM. https://doi.org/10.1145/2808580.2808686

García-Peñalvo, F. J. (2013). Education in knowledge society: A new PhD programme approach. In F. J. García-Peñalvo (Ed.), Proceedings of the First International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'13) (Salamanca, Spain, November 14-15, 2013) (pp. 575-577). ACM. https://doi.org/10.1145/2536536.2536624

García-Peñalvo, F. J. (2014). Formación en la sociedad del conocimiento, un programa de doctorado con una perspectiva interdisciplinar. Education in the Knowledge Society, 15(1), 4-9.

García-Peñalvo, F. J., Sarasa Cabezuelo, A., & Sierra Rodríguez, J. L. (2014). Innovando en los Procesos de Ingeniería. Ingeniería como Medio de Innovación. *VAEP-RITA, 2*(1), 26-28.

García-Peñalvo, F. J. (2018). Identidad digital como investigadores. La evidencia y la transparencia de la producción científica. *Education in the Knowledge Society, 19*(2), 7-28. https://doi.org/10.14201/eks2018192728

García-Peñalvo, F. J., García-Holgado, A., & Ramírez-Montoya, M. S. (2020). Introduction for the TEEM 2020 Doctoral Consortium track. In F. J. García-Peñalvo (Ed.), Proceedings TEEM'20. Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (Salamanca, Spain, October 21st - 23rd, 2020). ACM. https://doi.org/10.1145/3434780.3436704

García-Peñalvo, F. J., Rodríguez-Conde, M. J., Verdugo-Castro, S., & García-Holgado, A. (2019). Portal del Programa de Doctorado Formación en la Sociedad del Conocimiento. Reconocida con el I Premio de Buena Práctica en Calidad en la modalidad de Gestión. In A. Durán Ayago, N. Franco Pardo, & C. Frade Martínez (Eds.), Buenas Prácticas en Calidad de la Universidad de Salamanca: Recopilación de las I Jornadas. REPOSITORIO DE BUENAS PRÁCTICAS (Recibidas desde marzo a septiembre de 2019) (pp. 39-40). Salamanca, España: Ediciones Universidad de Salamanca.

González-Pérez, L. I., Glasserman Morales, L. D., Ramírez-Montoya, M. S., & García-Peñalvo, F. J. (2017). Repositorios como soportes para diseminar experiencias de innovación educativa. In M. S. Ramírez-Montoya & J. R. Valenzuela González (Eds.), *Innovación Educativa. Investigación, formación, vinculación y visibilidad* (pp. 259-272). Síntesis.

González-Pérez, L. I., Ramírez-Montoya, M. S., & García-Peñalvo, F. J. (2018). Identidad digital 2.0: Posibilidades de la gestión y visibilidad científica a través de repositorios institucionales de acceso abierto. In J. A. Merlo Vega (Ed.), *Ecosistemas del Conocimiento Abierto* (pp. 197-206). Ediciones Universidad de Salamanca.

González-Pérez, L. I., Ramírez-Montoya, M. S., García-Peñalvo, F. J., Gibrán Ceballos, H., & Juárez Ibarra, E. A. (2018). RITEC & CRIS: Interoperabilidad para visibilidad y medición del impacto de la producción científica energética. In M. S. Ramírez-Montoya & A. Mendoza-Domínguez (Eds.), *Innovación y sustentabilidad energética: Implementaciones con cursos masivos abiertos e investigación educativa* (pp. 55-73). Narcea.

Hernández-Pérez, T. (2019). El Plan S: Hacia el acceso abierto sin revistas híbridas. Anuario ThinkEPI, 13. https://doi.org/10.3145/thinkepi.2019.e13e06

Hernández-Pérez, T. (2020). Acceso abierto a las publicaciones científicas como objetivo y. 3.

Iribarren-Maestro, I. (2018). Bibliometría y bibliotecas universitarias: ¿matizando el perfil profesional? Anuario ThinkEPI, 12, 142. https://doi.org/10.3145/thinkepi.2018.15

Mateo, F. (2015). Producción científica en español en humanidades y ciencias sociales. Algunas propuestas desde Dialnet. El Profesional de la Información, 24(5), 509. https://doi.org/10.3145/epi.2015.sep.01

Mesias, Norma. (2010). *Generalidades Sobre el Marc, el Dublin Core y la Normalización de la Información Bibliográfica*.

Ramírez-Montoya, M. S., García-Peñalvo, F. J., & McGreal, R. (2018). Shared Science and Knowledge. Open Access, Technology and Education. *Comunicar, 26*(54), 1-5.

Simons, E., Jetten, M., Messelink, M., van Berchum, M., Schoonbrood, H., & Wittenberg, M. (2017). The Important Role of CRIS's for Registering and Archiving Research Data. The RDS-project at Radboud University (the Netherlands) in Cooperation with Data-archive DANS. Procedia Computer Science, 106, 321-328. https://doi.org/10.1016/j.procs.2017.03.031

Rivera, B., Becker, P., Papa, F., & Olsina, L. (2016). Hacia la Evaluación y Mejora de Software dirigidas por Metas Multinivel y Estrategias. 15.

Terán, Á. R. (2015). Sistemas de Gestión de la Investigación: Aproximación a los CRIS Institucionales. 71.