

Perfiles de investigadores en ORCID: confiabilidad de los datos sobre resultados de investigación

Borja González-Albo^{1*}, María Bordons²

¹Centro de Ciencias Sociales y Humanas (CCHS) – Consejo Superior de Investigaciones Científicas (CSIC) *borja.gonzalezalbo@cchs.csic.es.

²Instituto de Filosofía (IFS) – Consejo Superior de Investigaciones Científicas (CSIC)

Introducción

ORCID se ha consolidado como el sistema de identificación personal más empleado por la comunidad investigadora (Bello & Galindo-Rueda, 2020), convirtiéndose en un estándar de facto, cuyo su uso se ha planteado para la evaluación de las trayectorias científicas (Arroyo-Machado et al., 2025). Permite a los investigadores mantener un CV digital, estableciendo vínculos con fuentes confiables para la importación de datos y fomentando las conexiones entre autores, editores, financiadores y otros actores implicados en el sistema de comunicación científica (ORCID, 2023).

El análisis de la actividad científica precisa de datos de calidad, por lo que la limpieza y curación de los mismos es una fase esencial para obtener indicadores consistentes (Nguyen et al., 2022, Repiso & Cabezas-Clavijo, 2025) y evitar problemas como las duplicaciones de registros o los errores de atribución (Valderrama-Zurián et al., 2015, Franceschini et al., 2016 o Sauvayre, 2022). El análisis formal de la incertidumbre en los resultados debido a errores en los datos es importante y rara vez se tiene en cuenta en los estudios cuantitativos (Doner, 2024).

Objetivo Identificar posibles publicaciones duplicadas en los perfiles ORCID de los investigadores, y analizar el tiempo de demora en la inclusión de las publicaciones. Ambos aspectos influyen sobre la confiabilidad de los perfiles.

Método

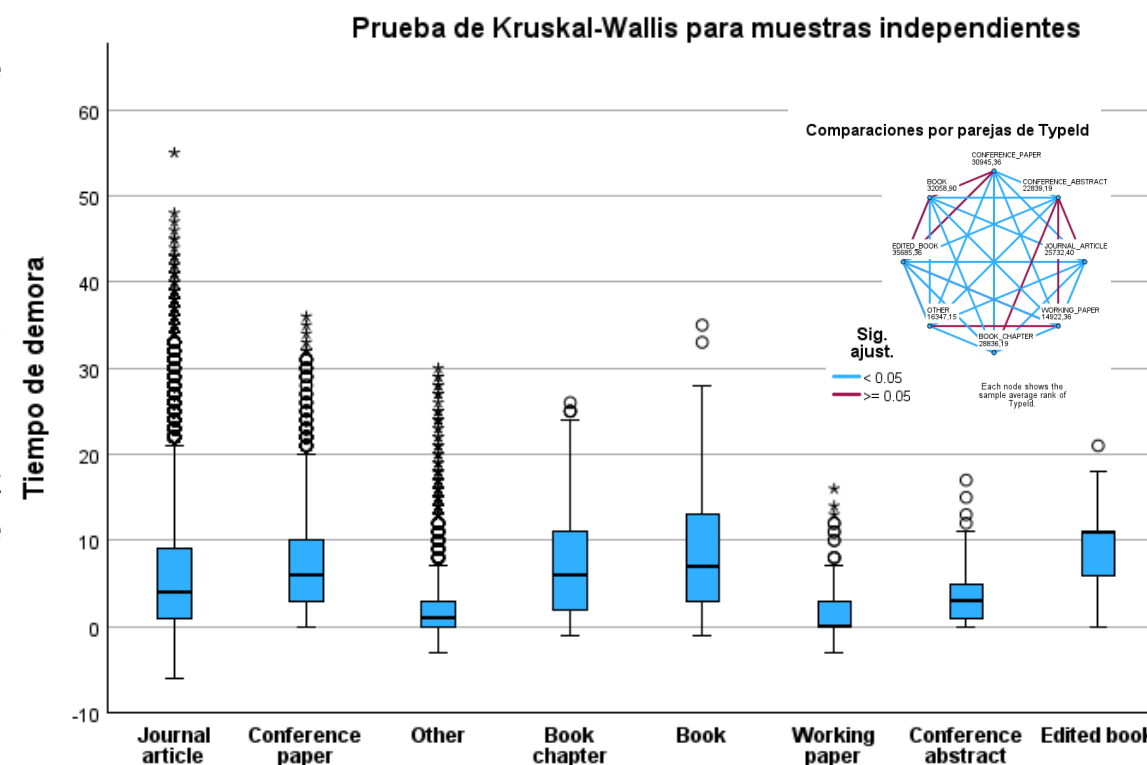
- Objeto de estudio: Perfiles ORCID de investigadores receptores de ayudas del *European Research Council* (ERC) adscritos a una institución española entre 2014 y 2020.
- Etapas: (1) Descarga de la sección “Works” de sus perfiles mediante la API pública de ORCID. (2) Construcción de una base de datos relacional con sus publicaciones. (3) Análisis de las duplicaciones por comparación de las cadenas de texto de los títulos de las publicaciones mediante un script en Python que utiliza distintos algoritmos de proximidad (comparación de palabras de *sklearn* y modificaciones del algoritmo *Gestalt pattern matching*). (4) Análisis de la demora de inclusión de las publicaciones en ORCID basado en el año de publicación del documento y en el año de incorporación en el perfil.

Resultados

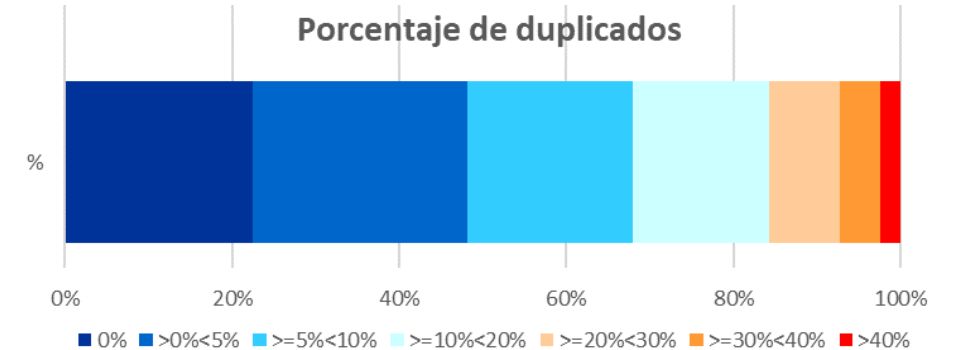
346 investigadores ERC (**330** con algún documento en su perfil). **56.353** items descargados de la sección Works (**171** items/investigador).

Tiempo de demora en la inclusión de las publicaciones

- El 30% de los documentos se incluyen en ORCID en el mismo año o al año siguiente de su publicación, y el 50% lo hacen en los primeros 4 años.
- La inclusión más rápida se observa en *working papers* y artículos de revista, mientras que los libros y capítulos de libro presentan la mayor demora.
- Las organizaciones sin ánimo de lucro (ej. Cross-ref) muestran la mayor rapidez incorporando documentos, por encima de otras fuentes como las bases de datos comerciales (WoS, Scopus), las instituciones científicas o los investigadores.
- La demora tiende a reducirse con el tiempo: en los últimos 10 años $\frac{2}{3}$ partes de los documentos se han introducido el mismo año de su publicación.



Publicaciones duplicadas en los perfiles



- En el 78% de los perfiles de autores en ORCID se identifica alguna publicación duplicada. Hay baja presencia de duplicados (<5%) en un 50% de los autores, pero alta (>20%) en un 15% de los mismos, por lo que es importante:
 - ORCID: Mejorar los procedimientos de identificación de duplicados
 - Investigadores: Asignar identificadores persistentes a los resultados de investigación siempre que sea posible, para asegurar su visibilidad, trazabilidad y adecuado reconocimiento en bases de datos, evitando errores como la presencia de duplicados
- La ausencia de datos sobre las fuentes de las publicaciones en las descargas mediante la API pública dificulta la identificación de duplicaciones en los perfiles.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación español a través del Proyecto RESPONSIBLE (PID2021-128429NB-I00), financiado por MCIN/AEI/10.13039/501100011033/El Fondo Europeo de Desarrollo Regional (FEDER), Unión Europea (UE). Agradecemos la contribución de José Ortega Gómez mediante el desarrollo de los algoritmos de proximidad.