Comunicar, nº 74, v. XXXI, 2023 | Revista Científica de Educomunicación | ISSN: 1134-3478; e-ISSN: 1988-3293
www.revistacomunicar.com | www.comunicarjournal.com

Cómo automatizar la extracción y análisis de información sobre ciencia ciudadana con propósitos educativos

How to automate the extraction and analysis of information for educational purposes

Miriam Calvera-Isabal. Asistente de Investigación, Grupo de Investigación de Tecnologías Interactivas y Distribuidas para la Educación (TIDE), Universidad Pompeu Fabra, Barcelona (España) (miriam.calvera@upf.edu) (https://orcid.org/0000-0003-4117-6953)
Dra. Patricia Santos. Investigadora, Grupo de Investigación de Tecnologías Interactivas y Distribuidas para la Educación (TIDE), Universidad Pompeu Fabra, Barcelona (España) (patricia.santos@upf.edu) (https://orcid.org/0000-0002-7337-2388)
Dr. H.-Ulrich Hoppe. Profesor Emérito, Instituto de Innovación de Sistemas Aplicados Rhine-Ruhr, Duisburg (Alemania) (uh@rias-institute.de) (https://orcid.org/0000-0003-3240-5785)
Cleo Schulten. Asistente de Investigación, Instituto de Innovación de Sistemas Aplicados Rhine-Ruhr, Duisburg (Alemania) (cs@rias-institute.de) (https://orcid.org/0000-0003-3082-6084)

RESUMEN

El interés y la práctica de la ciencia ciudadana (CC) ha aumentado en los últimos años. Esto ha derivado en el uso de páginas web como herramienta de comunicación, recolección o análisis datos o repositorio materiales y recursos. Desde una perspectiva educativa, se espera que al integrar información sobre proyectos de CC en un entorno educativo formal, se inspire a los maestros a crear actividades de aprendizaje. Este, es un caso interesante para usar bots que automaticen el proceso de extracción de datos de webs de CC que ayuden a comprender mejor su uso en contextos educativos. Aunque esta información está disponible públicamente, se deben seguir las reglas de la ley de protección de datos o GDPR. Este artículo tiene como objetivo explicar: 1) cómo la CC se comunica y promueve en los sitios web; 2) cómo se diseñan, desarrollan y aplican los métodos de web scraping y las técnicas de anonimización para recopilar información en línea; y 3) cómo se podrían usar estos datos con fines educativos. Tras el análisis de 72 webs algunos de los resultados son que solo el 24,8% incluye información detallada sobre el proyecto, y el 48,61% incluye información sobre propósitos o materiales educativos.

ABSTRACT

There is an increasing interest and growing practice in Citizen Science (CS) that goes along with the usage of websites for communication as well as for capturing and processing data and materials. From an educational perspective, it is expected that by integrating information about CS in a formal educational setting, it will inspire teachers to create learning activities. This is an interesting case for using bots to automate the process of data extraction from online CS platforms to better understand its use in educational contexts. Although this information is publicly available, it has to follow GDPR rules. This paper aims to explain (1) how CS communicates and is promoted on websites, (2) how web scraping methods and anonymization techniques have been designed, developed and applied to collect information from online sources and (3) how these data could be used for educational purposes. After the analysis of 72 websites, some of the results obtained show that only 24.8% includes detailed information about the CS project and 48.61% includes information about educational purposes or materials.

PALABRAS CLAVE | KEYWORDS

Ciencia ciudadana, aprendizaje informal, algoritmos, automatización, educación, protección de la privacidad.
Citizen science, informal learning, algorithms, automatization, education, privacy protection.

Recibido: 2022-05-30 | Revisado: 2022-06-27 | Aceptado: 2022-07-19 | OnlineFirst: 2022-10-30 | Publicado: 2023-01-01
DOI https://doi.org/10.3916/C74-2023-02 | Páginas: 23-35

1. Introducción y estado del arte

La ciencia ciudadana (CC) es la participación activa de todo tipo de público en tareas de investigación científica (Vohland et al., 2021). Las actividades de CC se organizan típicamente en proyectos y tienen gran presencia en línea a través de las páginas web o plataformas, lo que permite que estas se utilicen para la diseminación de datos, participación o como repositorio de información (Vohland et al., 2021). Hay varias asociaciones de CC: La asociación norteamericana de ciencia ciudadana (CSA-North América, por sus siglas en inglés), la asociación europea de ciencia ciudadana (ECSA, por sus siglas en inglés) y la asociación australiana de ciencia (ACSA, por sus siglas en inglés). Además, hay asociaciones nacionales o regionales como el Observatorio de la ciencia ciudadana (España) o Bürger schaffen Wissen (Alemania) o proyectos individuales como Cities-Health. Se puede encontrar información sobre actividades de CC en páginas webs de institutos de investigación, universidades, museos, etc. La variedad de instituciones de CC demuestra que la comunicación sobre proyectos de CC puede ser a través de diferentes canales (individual, como parte de una web o asociación, así como también a nivel local, regional o a gran escala). A pesar de que el enfoque de comunicación varía a medida que va transcurriendo el proyecto y puede ser diferente para cada tipo, es importante definirlo bien para atraer, retener, motivar e informar a los voluntarios (Vohland et al., 2021; Veeckman et al., 2019). Como concluyó Lin-Hunter et al. (2020) en sus análisis sobre las tareas de los voluntarios definidas en las descripciones de proyectos de CC y sus conexiones con su alfabetización científica, la forma en que los proyectos de CC comunican puede afectar a la participación y podría implicar cambios en la percepción de la ciencia y concienciación del problema sobre el que se investiga.

Históricamente, Internet (a través de páginas web) o la televisión han contribuido al aprendizaje informal y a la comunicación de la ciencia (Stocklmayer et al., 2010). La existencia de varias formaciones en CC demuestra que la comunicación sobre proyectos se puede dar a través de diferentes canales (individuales, como parte de una «red» o asociación local, regional o a gran escala). Los materiales proporcionados por estas plataformas tienen un gran potencial para ser usados con propósitos educativos, especialmente en relación a los Objetivos de Desarrollo Sostenible (ODS) teniendo en cuenta que muchos proyectos de CC abordan cuestiones sobre sostenibilidad (Fraisl et al., 2020; Storksdieck et al. 2016). Sin embargo, aunque múltiples proyectos CC se recogen y se muestran en plataformas nacionales o globales, no existe una base de datos centralizada que contenga esta información (Vohland et al., 2021).

Entre los beneficios educativos potenciales que las actividades de CC podrían tener, vemos mejoras en el conocimiento y entendimiento científico, el desarrollo de habilidades técnicas/científicas, motivación hacia carreras STEM y valores como sostenibilidad o respeto por el medio ambiente (Hiller & Kitsantas, 2014; Bonney et al., 2016; Kobori et al., 2016; Vohland et al., 2021). Aunque los proyectos de CC no suelen tener como objetivo principal fomentar la alfabetización y el conocimiento científico de los/las ciudadanos/as, a menudo desarrollan materiales educativos o llevan a cabo actividades de formación para preparar a los participantes en las tareas científicas que llevarán a cabo cómo puede ser el recoger o clasificar datos (Bonney et al., 2009). Cada vez más, desde las instituciones, se promueve la participación de las escuelas en los proyectos de CC (ej., la Oficina de Ciencia Ciudadana de Barcelona ha hecho un llamamiento para que las escuelas de Barcelona participen en proyectos de CC: https://bit.ly/3cB1lMH), y va en aumento. No obstante, todavía queda mucho por conocer sobre cómo la CC puede integrarse en escuelas como guía o fuente de inspiración para que los docentes creen actividades alineadas con la investigación que se lleva a cabo y los problemas sociales que abordan los proyectos de CC. Todos los materiales y datos generados por proyectos de CC se pueden utilizar para que los estudiantes aprendan sobre temas específicos o apoyen la actividad docente de los profesores. Este es un trabajo en el cual educadores y científicos deben trabajar juntos, ya que la comunicación de la ciencia (a través de talleres, actividades de aprendizaje o conversaciones informales) puede tener un impacto en el entendimiento y conocimiento público de los hechos científicos (Bickford et al., 2012; Stocklmayer et al., 2010).

Dada la presencia masiva y disponible en línea de información sobre proyectos y actividades de CC, parece prometedor usar técnicas de análisis computacional para generar conocimientos específicos sobre el funcionamiento y la evolución de las actividades de CC. Estas herramientas se han utilizado en diferentes ámbitos, especialmente la extracción de datos de forma masiva de sitios web y su almacenamiento en bases

de datos (Diouf et al., 2019). Son pocos los ejemplos en los que estas técnicas se hayan utilizado en el campo de la CC (Ponti et al., 2018). Desde una perspectiva europea, hay interés en conocer mejor el rol de la CC en ciencia y en la sociedad, ej., la distribución y contribución por regiones, distribución por disciplina, así como la importancia de la comunicación científica en el campo de la CC y el impacto en la educación. Todavía falta por conocer cómo los proyectos de CC se distribuyen para seguir desarrollando y apoyando tipos específicos de CC (Warin & Delaney, 2020). El trabajo explicado aquí es parte del proyecto europeo CS Track (https://cstrack.eu/) que opera en esta línea de investigación. Para este propósito, CS Track se basa en una combinación de técnicas de análisis web y métodos clásicos de estudios sociales. CS Track ha construido una base de datos que contiene información sobre 4.949 proyectos de CC que se han recogido de diferentes páginas web. Esta es la base de datos que contiene la información de la extracción que se lleva a cabo y del futuro enriquecimiento de la información descriptiva relacionada con los proyectos de CC. Todos los datos centralizados nos permitirán conocer más sobre cómo la CC se comunica en línea y ampliar nuestro conocimiento en cuáles son sus conexiones con la educación.

En este artículo, explicamos cómo se ha construido un punto de información centralizado sobre CC utilizando como base el contenido sobre CC distribuida en diferentes páginas web. Esto nos permite ver cuáles son las diferencias y similitudes entre las estructuras de datos que las diferentes páginas web tienen para compartir los datos. Como parte de esta extracción y análisis de datos, hemos intentado identificar el potencial que estos tienen con propósitos educativos.

En este trabajo, hemos sido conscientes de las limitaciones impuestas legítimamente por los principios de privacidad y protección de datos, en especial el Reglamento General Europeo de Protección de Datos (GDPR, por sus siglas en inglés). El objetivo del GDPR es dar a los/las ciudadanos/as control sobre sus datos personales y hacer cumplir la anonimización de los datos a menos que no haya un consentimiento individual específico. Un grupo de datos se considera anónimo si una persona solo puede re identificarse (Gruschka et al., 2018). A pesar de que se han extraído datos que describen las características de los proyectos de CC, a veces, en estos, se incluye información personal entre los textos ya sea de manera directa o indirecta. El trabajo presentado en este artículo ha sido guiado por los siguientes objetivos de investigación:

- (OI1) Diseñar e implementar un algoritmo que automáticamente extraiga datos sobre plataformas de CC y los almacene en un único punto central (base de datos). Los datos extraídos deberían estar alineados con el estándar de datos PPSR, y extenderlo si fuera necesario.
- (OI2) Encontrar la solución técnica para cumplir con los requerimientos de GDPR en este contexto.
- (OI3) Identificar potenciales usos educativos de los datos recogidos.

2. Metodología y selección de datos

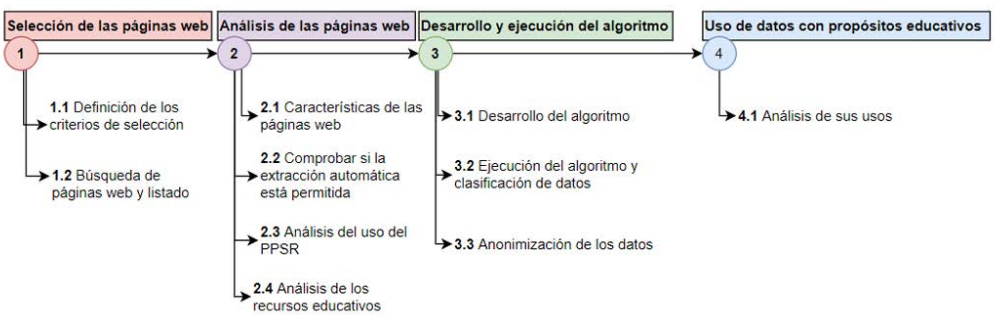
Las fuentes de información de este estudio fueron páginas web que contenían información sobre proyectos de CC. Se aplicó el siguiente criterio de inclusión para identificar datos en línea de proyectos de CC (nuestra unidad de análisis):

- Las páginas web contienen una lista con información de proyectos de CC o son páginas web de un único proyecto de CC.
- De Europa, países asociados o se llevan a cabo online.
- Se permite la extracción de datos tanto automáticamente como manualmente.

En la primera fase, todos los miembros del consorcio realizaron una búsqueda manual de las páginas web que podían contener información sobre CC en las regiones europeas. Después, se exploraron manualmente cada referencial para identificar cuáles contenían información específica sobre proyectos de CC y seguían los criterios definidos anteriormente. La identificación, selección y análisis de las páginas webs se hizo manualmente y como resultado se obtuvo una lista de 72 sitios web. Esta lista puede ser extendida en las siguientes iteraciones. Es posible que no se hayan identificado todas las páginas webs existentes que sigan el criterio definido, pero las más relevantes sí se han seleccionado. El análisis manual de las páginas webs tuvo dos objetivos: (1) identificar cómo se comunica la información sobre proyectos de CC en línea, qué elementos constituyen la información que se comparte, la distribución geográfica

de las páginas web o los idiomas utilizados y, (2) entender la/s estructura/s técnica/s que se utilizan para compartir los datos y cómo estas se alinean con el estándar de metadatos PPSR. La Figura 1 muestra el proceso que se ha seguido durante la investigación.

Figura 1. Proceso de selección y verificación (para sitios web que contienen información de proyectos de CC)

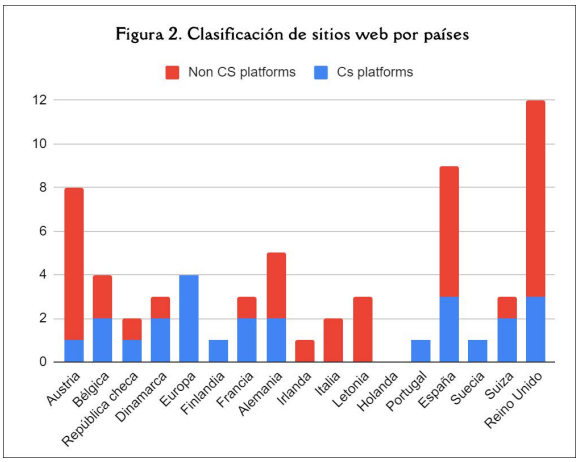


3. La presencia de la ciencia ciudadana en las plataformas en línea

En esta sección se explica la segunda fase del proceso (Figura 1), los resultados y conclusiones obtenidos del análisis. Clasificamos las páginas web en dos categorías: Plataformas de CC (29 páginas web) y Plataformas No-CC (43 páginas web). Las plataformas de CC son plataformas digitales que comparten información sobre los proyectos de CC, actividades, eventos, materiales o recursos, noticias sobre el campo, se utilizan como herramienta de comunicación (ej., comentarios o foros) o en ocasiones también se utilizan como herramientas de participación (Sanz et al., 2019). El primer objetivo de las plataformas No-CC es que no informan sobre CC, pero han sido creadas como una herramienta de comunicación, como repositorio o incluso permiten a los usuarios que interactúen entre ellas. Entre aquellas que clasificamos como No-CC podemos encontrar diversidad de asociaciones (Asociación Helmholtz), museos (El museo de historia natural del Reino Unido) o institutos de investigación (ICM Divulga). En la descripción de estas páginas web se usa terminología como «centro nacional de redes de comunicación científica», «Museo», «organización independiente sin ánimo de lucro» u «Oficina de Transferencia de Investigación» para definir asociaciones u organizaciones.

La clasificación de plataformas de CC se ha llevado a cabo siguiendo el criterio propuesto por Vohland et al., (2021) el cual, diferencia entre cinco tipos de plataformas. Debido a los requisitos definidos para la selección de plataformas, se ha añadido la categoría «Plataforma de CC mundial» para aquellas que contienen información de proyectos de todo el mundo. Después de analizar la descripción de las plataformas, las hemos categorizado en: Plataformas comerciales para iniciativas de CC (2 páginas web), plataformas de CC para proyectos específicos (8 páginas web), plataformas de CC para temas científicos específicos (2 páginas web), plataformas de CC nacionales (15 páginas web), plataforma europea de CC (1 página web) y plataforma de CC mundial (2 páginas web). En estos textos hemos leído términos como «portal de ciencia ciudadana» o «centro en línea de ciencia ciudadana», los cuales se han utilizado para identificar la página web como plataforma de CC y otros como «centro de ciencia ciudadana» o «red de ciencia ciudadana» que hacen referencia a las asociaciones de CC que coordinan las páginas web. Es común que los proyectos de CC utilicen las páginas web como herramientas de participación, por esta razón podemos leer que las plataformas de CC utilizan términos como «simulador» o «herramienta web».

Europa es un continente en el cuál las culturas e idiomas coexisten. Para entender la distribución de las páginas web en Europa, estas se han analizado desde dos puntos de vista: la localización geográfica y los idiomas disponibles. 17 de los 44 países han sido identificados en la lista de páginas web. La Figura 2 muestra la distribución de los países en relación a los dos tipos de plataforma. Todas las plataformas en línea consideradas mundiales como SciStarter (https://scistarter.org/), red iNaturalist (https://www.inaturalist.org/), Zooniverse (https://www.zooniverse.org/) e Instant wild (https://instantwild.zsl.org/intro) se han excluido del análisis porque, aunque se podría asignar a cada una un único país, comparten información sobre proyectos o iniciativas alrededor del mundo.



Para entender mejor la distribución geográfica y la extensión ciudadana a la que podrían llegar, es importante también entender la diversidad lingüística de Europa. Varias plataformas en línea facilitan el uso de más de un idioma. Por ejemplo, el 29,7% de las plataformas facilitan el uso de dos idiomas (Iteritalia), el 8,1% de las plataformas facilitan el uso de tres idiomas (OpenSystems UB) y el 4,1% de las plataformas facilitan el uso de más de tres idiomas (EU-Citizen science). El 58,1% de las plataformas solo disponen del contenido en un idioma (Desqbre). Tal y como recoge la Carta de los derechos fundamentales de la Unión Europea (European Union, 2010), «la unión debe respetar (...) la diversidad lingüística» así como la «discriminación basada en el (...) idioma (...) estará prohibida». La UE tiene 24 idiomas oficiales y otros regionales o minoritarios. 17 de los 24 idiomas se han cubierto por las páginas web seleccionadas. Además, se han identificado tres idiomas regionales (euskera o catalán) y uno extra comunitario (árabe). Las plataformas de CC cubren el 84,3% de los idiomas utilizados en Europa.

Tabla 1. Clasificación de las plataformas en base al «Mercado», «Afiliación» y «Centralización»

	Cero lados		**Un lado**		**Múltiples lados**			
Mercado	**Todas las webs**	**Plataformas de CC**	**Todas las webs**	**Plataformas de CC**	**Todas las webs**	**Plataformas de CC**		
	20,83%	17,24%	52,78%	41,38%	26,39%	41,38%		
Afiliación			**Registro**					
			Todas las webs		**Plataformas de CC**			
			43,06%		37,93%			
			Subscripción					
			Todas las webs		**Plataformas de CC**			
			51,39%		55,17%			
					Sin transacción		**Con transacción**	
					Todas las webs	**Plataformas de CC**	**Todas las webs**	**Plataformas de CC**
					79,17%	72,41%	20,83%	27,95%
							Inversión	
							0%	
Centralización					**Centralizado**			
					Todas las webs	**Plataformas de CC**		
					33,33%	51,72%		
					Descentralizado			
					Todas las webs	**Plataformas de CC**		
					66,67%	48,28%		

Aunque indicamos el país de cada página web, cuando leemos las descripciones, nos damos cuenta que el 35,6% muestra información sobre las regiones que cubre (ej. en «en Flandes» (Citizen Science Vlaanderen) o «Globalmente»). Las páginas web cubren áreas regionales o nacionales (como por ejemplo la plataforma de CC de Barcelona (Barcelona)), europea (EU-Citizen science platform) y todas las áreas del mundo (iNaturalist). La región geográfica cubierta por las plataformas en línea está alineada con el idioma disponible.

A partir de la misma información que se ha utilizado para clasificar las páginas web, se han extraído términos clave para asignar las áreas de investigación correspondientes. Para identificar las categorías, se

ha seleccionado terminología como «protegiendo nuestro planeta», «ciencia utilizada en la investigación de la ciencia criminal, análisis de laboratorio y la presentación de evidencia científica dentro de los tribunales» o «servicios meteorológicos y geográficos». Las plataformas se han clasificado en seis grandes áreas de investigación definidas en Web of Science Core Collection (Clarivate analytics, 2022): Artes y humanidades (1,37%), Ciencias de la vida y biomedicina (50,68%), Ciencias físicas (1,37%), Ciencias sociales (2,74%), Tecnología (0%) y Todas (43,84%).

3.1. Funcionalidades y aplicaciones de las páginas web

Para esta investigación, aplicamos manualmente la taxonomía de plataformas definidas por Derave et al. (2020). Aunque define siete categorías, hemos analizado solo las tres primeras debido a que las plataformas seleccionadas están orientadas a la participación y comunicación y centraremos nuestra atención en ello.

- Mercado: es la primera categoría que define el número de grupos de usuarios. Hemos incluido el término cero lados. De las páginas web seleccionadas, hemos identificado: 15 páginas web cero lados (sin interacciones entre usuario, solo entre ellos y los gestores de la web), 38 un lado (los usuarios interacciones con la plataforma e indirectamente con otros usuarios vía comentarios o publicaciones) y 19 de múltiples lados (interacciones entre usuarios y la plataforma). Las plataformas de CC son, comúnmente, un lado o múltiples lados, aunque a veces se utilizan como herramientas de participación, por lo que es común que acepten que los usuarios interactúen entre ellos (vía foros o comentarios) (Tabla 1).
- Afiliación: hace referencia a cómo los usuarios interactúan con otros y la web. Hay muchas funcionalidades diseñadas para dar a los usuarios la oportunidad de estar conectados (ej. foros, boletines informativos, etc.) y pueden combinarse en una página web. Hemos identificado 31 páginas web que permiten el registro, 37 que permiten la subscripción, 19 que permiten la creación de contenido o comentarios, 15 que permiten transacciones y 0 que permiten inversión. Para las plataformas de múltiples lados es común involucrar a usuarios en añadir comentarios o crear contenido, pidiendo registro para su uso y dándoles la opción de estar conectados e informados vía suscripción. Sin embargo, para los otros tipos, el registro no es un requisito imprescindible pero la subscripción es muy recomendable para estar conectados. La herramienta más utilizada para la interacción entre usuarios es el foro, pero solo 8 páginas tienen uno (5 de las cuales son plataformas de CC). La segunda opción para que los usuarios interactúen es añadir comentarios (no mensajes directos). Solo 6 de las plataformas permiten este tipo de funcionalidades. Como son páginas web utilizadas para investigaciones científicas o su comunicación, la diseminación de resultados y la captación de participantes es importante. Hay páginas específicas creadas para compartir noticias o blogs con información: 47 disponen de ellas, siendo 20 plataformas de CC.
- Centralización: la tercera categoría está alineada con la segunda porque indica la forma en la que los usuarios se conectan con otros. Hay 48 páginas web descentralizadas mientras que 13 son centralizadas.

3.2. Información en línea de la ciencia ciudadana

En esta sección se explica cómo la información sobre CC se comparte en las páginas web desde dos puntos de vista: cómo se estructuran los datos y qué tipo de datos se comparten. El análisis es necesario para el desarrollo del algoritmo y el diseño de la base de datos.

A pesar de que cada página web sigue su propio diseño y estructura de datos, es común que todas tengan una página principal, una página con la lista de proyectos y enlaces indexados a otras páginas con la información asociada a cada proyecto de CC (Figura 3). No obstante, hay una excepción con las plataformas sobre proyectos específicos de CC ya que contienen información sobre un único proyecto distribuida en páginas, no como la lista de proyectos individuales que podemos encontrar en las otras páginas web.

Figura 3. Navegación de las webs y estructura de datos

Como primer paso, el set de datos original se clasificó de acuerdo a tres tipos de páginas web (Figura 3):

- Estructurada: la información se organiza en categorías (como por ejemplo en https://eu-citizen.science/).
- Semiestructurada: alguna información se organiza en categorías y otra en párrafos (como la página web de https://www.zooniverse.org/).
- No-estructurada: no se organiza la información en categorías (como por ejemplo en https://www.scivil.be/en).

De las 72 páginas web, podemos identificar 13 con datos estructurados, 27 semiestructurados y 34 no-estructurados. Varios grupos de trabajo de asociaciones de CC (Grupo de trabajo de datos y metadatos (CSA) o Grupo de Trabajo sobre «Datos, Herramientas y Tecnología» (ECSA)) se centran en promover la estandarización de los datos de CC. Este es el caso, por ejemplo, del modelo de datos Participación Pública en la Investigación Científica - Núcleo (PPSR-Core, por sus siglas en inglés) que propone la estandarización de los datos y trabaja en promover que sea aceptado y utilizado por las páginas web (Bowser et al., 2017). En relación a la información de proyectos de CC, para analizar cómo estos datos se comparten y definir una estructura común para almacenarlos en la base de datos, ha sido necesario identificar cómo la información asociada a los proyectos de CC se puede clasificar de acuerdo a los atributos del estándar de metadatos PPRS-Core y otros creados.

Hemos analizado cómo este estándar se ha seguido en las páginas web y hemos identificado que está bien definido el Título (comúnmente es lo primero que se ve y es más grande que otros textos) y la Descripción (debajo del título). Para el resto, hemos creado un diccionario de término que contienen 19 categorías (15 incluidas en el estándar de metadatos y 4 añadidas a este) basándonos en términos o secciones similares identificadas en las páginas web analizadas:

- Redes sociales: el nombre de la red social («twitter» o «facebook») o términos generales («blog», «REDES SOCIALES» o «PERFILES EN REDES SOCIALES»).
- Recursos en línea: extensiones de formatos de fichero («.pdf») o términos generales («OTROS RECURSOS DEL PROYECTO» o «Desktop»).
- Herramientas y materiales: solo una expresión seleccionada «Tipo de medios».
- Aplicaciones utilizadas: repositorio de nombre de aplicaciones utilizadas, aunque podría integrarse también en la categoría de herramientas y materiales («play.google» o «apple.com») o términos generales («Mobile»).
- Imágenes: extensión de los ficheros («.jpg», «.png», «.JPG» o «.jpeg»).

- Localización geográfica: términos generales utilizados («Geographical», «Geographic Scope», «WHERE», «Ubicación», «places», «Project Location» o «Location») o términos específicos para las regiones o áreas («Country», «PROVINCIA» o «País»).
- Estado: términos generales («Project Status», «Status» o «ESTADO DEL PROYECTO»).
- Metodología – Tareas de los participantes: términos generales («Participation Tasks» o «Tasks») y preguntas abiertas sobre la participación («HOW TO GET STARTED», «RELACIÓN CON LA CIENCIA CIUDADANA» o «¿Cómo participan los voluntarios/as?»).
- Fecha de inicio: términos generales («Start Date», «FECHA DE INICIO DEL PROYECTO» o «Projektstart»).
- Financiación o apoyo económico: términos generales («Sponsor», «TOTAL EXPENSE» o «Project Funding»).
- Campo de investigación: términos generales («Fields of Science», «TOPICS», «ÁREA DE CONOCIMIENTO»).
- Tiempo de desarrollo: términos generales («Intended Outcomes», «IDEAL FREQUENCY», «When?» o «Période»).
- Objetivos principales: términos generales («Goal», «Waarom doe je mee?» o «Objet»).
- Edad de los participantes: solo un término «IDEAL AGE GROUP».
- Perfil de los participantes: términos generales («Wie kan meedoen?», «Usuarios», «/people/», «PÚBLICO AL QUE SE DIRIGE EL PROYECTO», «INTEGRANTES DEL PROYECTO», «INTEGRANTES», «Who can take part?», «Public», «Project Partners» o «Users»).
- Lugar de desarrollo: términos generales para explicar el espacio o área de desarrollo de las actividades de investigación («SPEND THE TIME», «Region», «Ubicación», «ÁMBITO DE ACTUACIÓN» o «Type of activity»).
- Tiempo de dedicación: términos generales para explicar cuánto tiempo tienen que dedicar los participantes («AVERAGE TIME» o «How long will it take?»).
- Información de contacto: «@» se utiliza en las direcciones de correo electrónico.
- Fecha de actualización de los datos: solo el término «PROJECT UPDATED».
- Programa principal o persona al cargo: en esta categoría se combina información sobre el creador, coordinador o gestor junto con las asociaciones que apoyan o colaboran («PRESENTED BY», «Wie organiseert het?», «/researcher/», «Creado por», «Administradores de proyecto», «Administrador de proyecto», «MIEMBROS DEL EQUIPO», «OTROS GRUPOS O INSTITUCIONES COLABORADORES», «OTRAS PERSONAS O ENTIDADES COLABORADORAS», «Project Manager», «Project Co-ordinator» o «Kontakt»).

4. Desarrollo y ejecución del algoritmo

Nuestro objetivo es tener toda la información almacenada en la base de datos, así que es esencial escoger la más adecuada basándonos en el tipo de datos extraídos de las plataformas seleccionadas. En esta sección se explica el proceso seguido en la tercera fase (sección 2).

4.1. Selección de la base de datos

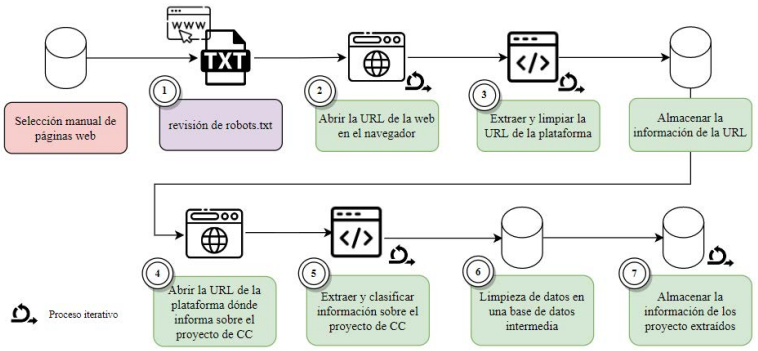
Para seleccionar la base de datos, se han comparado bases de datos relacionales (aquellas que aceptan lenguaje de consulta estructurado, (SQL por sus siglas en inglés)) y bases de datos no relacionales (aquellas que no aceptan SQL) (Li & Manoharan, 2013). Se seleccionó la base de datos MongoDB en su versión 4.2 ya que: puede almacenar datos estructurados y no estructurados; puede crecer fácilmente; la estructura de la base de datos puede cambiar independientemente de otros datos, otras colecciones y documentos (la estructura de los datos depende de cómo se ha descrito el proyecto) y admite consultas y consumo de datos.

4.2. Desarrollo y ejecución

El algoritmo se adapta a los tres tipos de estructura web. Para acceder a la información de proyectos de CC, aplicamos un proceso en dos pasos: primero para acceder a la página principal donde se listan los proyectos y, segundo, para seleccionar proyectos de CC individualmente para consultar su información.

La Figura 4 muestra el proceso que sigue el algoritmo. Se desarrolló en el lenguaje de programación Python (usando las librerías selenium, b4soup, requests y PyMongo). En este proceso, fue necesario tener en cuenta si las páginas web tenían una interfaz de programación de aplicaciones (API, por sus siglas en inglés) (ej. la EU-Citizen science e iNaturalist), lo que permitía la extracción automática de los datos desde la base de datos origen.

Figura 4. Proceso seguido por el algoritmo



El protocolo de exclusión de robots regula, para los bots, el acceso al Código Fuente de las páginas web (Kolay et al., 2008). Se define por cada página web y se informa en el fichero robots.txt que es accesible vía la URL de la web. Las restricciones combinan el bloqueo total o parcial del contenido para todos o ciertos robots. Esta información se comprueba antes de la extracción de datos. Hemos identificado: 4 páginas web que han definido un protocolo de exclusión de robots; 20 páginas web que contienen restricciones, pero no para el algoritmo creado o el contenido específico que se extrae y 1 página web cuyo protocolo no permite la extracción de ficheros «.pdf». Para cada web, el robot busca el código fuente, que ha sido identificado manualmente con anterioridad, y extrae elementos específicos de la web (Parvez et al., 2018). Además, las acciones de los usuarios como pulsar un botón se tienen que replicar automáticamente porque algunas páginas que contienen información sobre proyectos de CC tienen más de una pestaña.

4.3. Clasificación, limpieza y almacenamiento de los datos

La clasificación de los datos de proyectos de CC se hace automáticamente buscando palabras clave, símbolos o frases definidas previamente (sección 3.3). Se realiza un proceso de limpieza para borrar los datos que no son correctos o los elementos sin clasificar y la información final se almacena en la base de datos. Para evitar duplicados, se comprueba si el título del proyecto existe en la base de datos. Se añade información adicional como la fecha de almacenaje, fecha de actualización (en caso de que ya exista), origen o ID antes de almacenarlo. La información obtenida a partir de la API se tiene que mapear a la estructura de la base de datos (no se necesita limpieza). Este proceso puede reproducirse cuando sea necesario, por lo que el número de proyectos puede aumentar. La Figura 5 muestra un ejemplo de la información almacenada.

Hemos utilizado el reconocimiento de entidad nombrada (NER por sus siglas en inglés) junto con las reglas de entidad para identificar números de teléfono, direcciones de correo electrónico y cuentas personales basándonos en patrones de expresiones regulares dadas (RegEx). El algoritmo comprueba si los nombres de personas aparecen junto con datos personales encontrados por la regla de entidad. El nombre puede, en algunos casos mantenerse sin ser anonimizados como, por ejemplo, si existe un artículo de Wikipedia para el nombre identificado consideramos que es una persona conocida y su nombre tiene un significado más allá de nombrar a una persona (Albert Einstein) o, si es un nombre común que se identifica con una única persona.

Otra razón puede ser que el nombre de una persona sea el mismo que el de un proyecto ya que, hemos encontrado que algún proyecto consiste en analizar datos sobre una única persona o familia. Ambas

razones se ignoran si se menciona información del número de teléfono, el correo electrónico o la cuenta personal en la misma frase.

Hemos optado por la pseudo anonimización de las partes de texto seleccionadas mediante hash repetido (Kasgar et al., 2012) agregando el tipo de datos previamente atribuido en el texto. De esta forma, las personas, así como los datos personales no se pueden identificar, pero el texto se puede leer y se pueden seguir estableciendo conexiones entre múltiples proyectos a través de las personas.

Figura 5. Información del Proyecto Neureka almacenada en la base de datos


```
{
    "_id": {
        "$oid": "5efc3ae868516f3f3e568278"
    },
    "TITLE": "The Neureka Project",
    "WEB": ["https://www.neureka.ie/", "WEBSITE The Neureka Project"],
    "SOCIAL MEDIA": ["https://www.youtube.com/watch?v=TkWtYafAAT4", "https://twitter.com/@neurekaApp"],
    "DESCRIPTION": ["Neureka is a collection of fun brain games and challenges that allows you to solve cogr
    "Plat Id": ["65", "17"],
    "Plat country": ["USA", "EU"],
    "Insert date": "2020-07-01",
    "APPS": ["https://play.google.com/store/apps/details?id=com.gillanlab.neureka.beta", "https://apps.apple
    "GEOGRAPHICAL LOCATION": ["WHERE Online"],
    "METHODOLOGY": ["HOW TO GET STARTED Download neureka from either the Google Play Store or Apple App Stor
    "INVESTMENT": ["TOTAL EXPENSE 0.00"],
    "TOPICS": ["Biology", "Psychology", "Health & Medicine", "brain", "dementia", "depression", "neuroscienc
    "DEVELOPMENT TIME": ["IDEAL FREQUENCY Twice per day"],
    "DEDICATION TIME": ["AVERAGE TIME 1 hour"],
    "DEVELOPMENT SPACE": ["SPEND THE TIME Indoors"],
    "PLATFORM UPDATE DATE": ["PROJECT UPDATED 06/22/2020, 08:25 pm GMT+2", "2021-03-03T10:35:12.479221Z"],
    "Url platform": ["https://scistarter.org/the-neureka-project", "https://eu-citizen.science/projects"],
    "MAIN PROGRAM OR PERSON IN CHARGE": ["PRESENTED BY Gillan Lab at Trinity College Dublin", "Trinity Colle
    "MAIN OBJECTIVES": ["GOAL Improve mental health and dementia research"],
    "TOOLS AND MATERIALS": ["MATERIALS A smartphone (either Android or Apple)"],
    "PARTICIPANTS PROFILE": ["SPECIAL SKILLS None"],
    "Wp2 Id": ["3", "17"],
    "Language": ["English"],
    "STATUS": ["Active"],
    "START DATE": ["2020-07-01"],
    "END DATE": ["2025-12-31"],
    "LATITUDE": ["53.343667"],
    "LONGITUDE": ["-6.254445"],
    "COUNTRY": ["IE"],
    "IMAGE": ["https://eu-citizen.science/media/media/images/2021-02-08_020202020404_546_NEUREKA%20LOGO.png"
    "Date update": ["2021-03-19", "2021-04-08", "2021-05-14"],
    "MAIL": null
}
```


5. Resultados

En nuestro análisis, hemos incluido 4 nuevas categorías a las 43 incluidas en el estándar PPSR-Core: 11 campos obligatorios, 21 opcionales y 15 nuevos atributos identificados en las páginas web. Esta clasificación tiene como objetivo facilitar el análisis automático de los datos en los siguientes pasos. El problema principal identificado es la inconsistencia cuando se reporta información sobre proyectos de CC en línea, ya que muchas de las páginas web o plataformas no siguen los estándares de metadatos. Los atributos más comunes (obligatorios) son el título, la descripción, el enlace a la web (si existe), las cuentas en redes sociales, información de contacto o tema del proyecto. El 91,56% de ellos están informados. Toda la información relacionada con el origen de datos, idioma, fecha de almacenamiento y otra información necesaria para la gestión de los datos siempre se informa, pero no se considera para el análisis. Los tributos menos comunes (definidos como opcionales o nuevos creados por nosotros) como la geolocalización o la edad de los participantes se informan en un 24,8%. Las técnicas de minería de datos se han utilizado para la extracción de información de los textos en otros campos. Una fuente de información importante es la descripción del proyecto, importante para que se puedan aplicar técnicas para que automáticamente extraigan información y completen otros atributos de la base de datos. Por ejemplo, como se explica en secciones anteriores, NER se ha utilizado para identificar el nombre de las personas, pero también se puede utilizar para obtener información sobre la organización, números cardinales, países, ciudades o estados, etc.

Otros métodos computacionales orientados a la clasificación de datos se pueden aplicar a los datos para añadir categorías o crear grupos que fácilmente organicen los datos para que sea más sencillo filtrarlos o darles contexto (Roldán-Álvarez et al., 2021). Al extraer palabras clave o significado a los textos, se pueden crear nuevas categorías como: ODS, objetivos de aprendizaje o áreas de investigación. El

resultado de la aplicación de métodos, se puede utilizar para dar apoyo a profesores para seleccionar temas que tratar en el aula y recoger inspiración para crear actividades de aprendizaje.

6. Discusión y conclusiones

La variedad de páginas web que muestran información sobre proyectos de CC es también una reflexión de cómo de variada es la CC. Involucrar al ciudadano/a también implica adaptarse a las diferentes formas de comunicación, tanto por el lenguaje o por la región en los que ocurre. Las páginas web son buenas herramientas para la comunicación y para compartir con otros la información, pero también para la participación. Aun así, todavía hay trabajo que hacer para aumentar el acceso público a los datos para que la CC sea bien conocida y aumentar el interés de los/las ciudadanos/as a participar en la investigación. Las plataformas de CC deben tener en cuenta el estar alineadas con el estándar de datos PPST_Core y otros. Normalizar todas las estructuras de datos y la información compartida mejora la experiencia de usuario en las páginas web, así como facilita en ellas la búsqueda.

Teniendo la información clave sobre proyectos de CC en un único repositorio y bien documentada, puede mejorar la participación de los/las ciudadanos/as en proyectos de investigación y su interés, así como el campo de la CC. En esta línea, el proyecto CS Track,además de desarrollar conocimientos en el campo de la CC, ha abierto una nueva perspectiva en cómo los métodos computacionales se pueden aplicar para centralizar todos los datos en una única base de datos con propósitos de investigación. Hay todavía trabajo por hacer para analizar y aplicar métodos de minería de datos con el objetivo de obtener más información para las categorías que quedan por completar. Sin embargo, los métodos de minería de datos son poco efectivos sin una buena y detallada descripción del proyecto de CC. Para ello, es necesario involucrar a científicos y expertos en comunicación (Roche et al., 2020) y seguir guías ya definidas por expertos (como la propuesta por Veeckman et al., 2019) para buenas acciones de comunicación. Para tener impacto en el ámbito de la educación, es esencial estar alineado con el currículo oficial del nivel educativo al que haga referencia.

6.1. ¿Cómo identificar contenido con potencial beneficio educativo?

Las plataformas y páginas web de CC pueden proporcionar contenido que puede ser un recurso muy poderoso para el aprendizaje y la enseñanza. Un primer estudio exploratorio desarrollado por Calvera-Isabal et al. (2021) ha explorado tres plataformas de CC y encontró que los materiales y datos relacionados con proyectos de CC de páginas web existentes tienen el potencial de apoyar a profesores en su práctica (Asensio-Pérez et al., 2014). Publicaciones anteriores han establecido que la exploración de datos tiene impacto en la concienciación e interés de los estudiantes y promueve la discusión abriendo nuevas perspectivas sobre cómo trabajar las matemáticas en la educación formal (Saddiqa et al., 2019). De los datos clasificados en categorías, los profesores encontrarán una poderosa fuente de conocimiento científico para filtrar (por ejemplo, basado en las áreas de investigación, los ODS o el aprendizaje de habilidades). Teniendo los datos clasificados por áreas de investigación permitirá a los profesores entender el campo que el proyecto está investigando. La información sobre los ODS que abordan problemas reales a nivel mundial, se puede integrar en las actividades de aprendizaje para motivar a los estudiantes a aprender más y también concienciar sobre, por ejemplo, sostenibilidad o ecología (Massa et al., 2011; Djonko-Moore et al., 2018).

Aunque la CC se está introduciendo en educación y tiene potencial de hacerlo de otras formas, solo el 48,61% de las páginas web contiene materiales educativos o información relacionada con el aprendizaje. En el caso de las plataformas de CC, este dato es muy positivo ya que el 55,17% dispone de recursos. Aquellos que permiten la participación de ciudadanos/as, tienen páginas específicas con materiales educativos y preguntas frecuentes (ej. Zooniverse). Plataformas específicas sobre CC se utilizan como herramientas de comunicación, pero también como repositorio para toda la información y documentación que desarrollan (Luonto-Liiton Kevätseuranta). Las otras páginas que no son plataformas de CC, si están orientadas a educación (ej. universidades o asociaciones sobre educación o aprendizaje), entonces comparten materiales específicos para profesores o educadores. Si no lo están, no es frecuente que compartan este tipo de recursos. La aplicación de técnicas computacionales avanzadas y tener toda la

información centralizada, puede utilizarse para completar la información extraída: problemas reales, áreas de investigación, disciplinas científicas, habilidades, etc.

Es común que los profesores integren tecnología para apoyar el aprendizaje o mejorarlo. Por esta razón, las herramientas y el contenido desarrollado por los proyectos de CC podría integrarse en la clase como herramienta para desarrollar actividades, para participar en CC o incluso para dar apoyo a los profesores durante la preparación de las clases. Respecto al potencial uso de los datos en contextos educativos cubierto en este artículo, es necesario trabajar más en identificar cómo comunicar (al nivel de los datos/información que se deben compartir en las páginas web de CC) para estrechar la conexión que podría haber entre CC y los entornos educativos formales. Algunas oportunidades que surgen de este análisis son el uso de la información de proyectos de CC en contextos educativos (como inspirar a profesores en la creación de actividades de aprendizaje) o la participación de las escuelas en el proyecto (como las iniciativas que ya se están llevando a cabo). Todavía es necesario analizar los materiales que necesitan los profesores y en qué medida toda la información centralizada y disponible para que se explore puede tener un impacto en el conocimiento científico que tienen los profesores o en sus habilidades pedagógicas, que podrían derivar en una mejor relación de los estudiantes hacia la ciencia (Chan & Yung, 2018).

Finalmente, la aplicación de algoritmos y la colección de información masiva permite la unificación de los datos en un único origen que puede ser usado potencialmente con propósitos educativos. Por esta razón, como trabajo futuro, se podría desarrollar una plataforma digital que comunique información sobre CC para dar apoyo a la creación de actividades en la clase.

Contribución de Autores

Idea, M.C, P.S, U.H; Revisión de la literatura (estado del arte), M.C, P.S, U.H, C.S; Metodología, M.C, P.S, U.H; Análisis de datos, M.C, C.S; Resultados, M.C, P.S; Discusión y conclusiones, M.C, P.S, U.H, C.S; Redacción (borrador final), M.C, P.S, U.H, C.S; Revisión final, P.S., U.H; Diseño de proyecto y apoyos, M.C, P.S, U.H.

Apoyos

Este trabajo ha sido subvencionado en parte por PID2020-112584RB-C33 financiado por MCIN/AEI/10.13039/501100011033, por 'CS Track: Expanding our knowledge on Citizen Science through analytics and analysis' H2020-SwafS-2019-1 proj. ref: 872522 y por el programa Ramón y Cajal (P. Santos).

Referencias

Asensio-Pérez, J.I., Dimitriadis, Y., Prieto, L.P., Hernández-Leo, D., & Mor, Y. (2014). From idea to VLE in half a day: METIS approach and tools for learning co-design. In *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 741-745). ACM. https://doi.org/10.1145/2669711.2669983

Bickford, D., Posa, M.R.C., Qie, L., Campos-Arceiz, A., & Kudavidanage, E.P. (2012). Science communication for biodiversity conservation. *Biological Conservation*, *151*(1), 74-76. https://doi.org/10.1016/j.biocon.2011.12.016

Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., & Shirk, J. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, *59*(11), 977-984. https://doi.org/10.1525/bio.2009.59.11.9

Bonney, R., Phillips, T.B., Ballard, H.L., & Enck, J.W. (2016). Can citizen science enhance public understanding of science? *Public Understanding of Science*, *25*(1), 2-16. https://doi.org/10.1177/0963662515607406

Bowser, A., Brenton, P., Stevenson, R., Newman, G., Schade, S., Bastin, L., Parker, A., & Oliver, J. (2017). *Citizen Science Association Data & Metadata Working Group: Report from CSA 2017 and Future Outlook*. European Commision. https://bit.ly/3IS85Sl

Calvera-Isabal, M., Varas, N., & Santos, P. (2021). Computational techniques for data science applied to broaden the knowledge between citizen science and education. In *Proceedings of the 18th International Conference on Cognition and Exploratory Learning in the Digital Age (CELDA 2021)* (pp. 219-226). IADIS Press. https://doi.org/10.1007/978-3-030-65657-7

Chan, K.K.H., & Yung, B.H.W. (2018). Developing pedagogical content knowledge for teaching a new topic: More than teaching experience and subject matter knowledge. *Research in Science Education*, *48*, 233-265. https://doi.org/10.1007/s11165-016-9567-1

Clarivate analytics (Ed.) (2022). *Web of Science Core Collection Help*. https://bit.ly/3ts2IZl

Derave, T., Sales, T.P., Gailly, F., & Poels, G. (2020). Towards a reference ontology for digital platforms. In G. Dobbie, U. Frank, G. Kappel, S. W. Liddle, & H. C. May (Eds.), *Conceptual modeling. ER 2020. Lecture notes in computer science* (pp. 289-302). Springer. https://doi.org/10.1007/978-3-030-62522-1_21

Diouf, R., Sarr, E.N., Sall, O., Birregah, B., Bousso, M., & Mbaye, S.N. (2019). Web Scraping: State-of-the-Art and Areas of Application. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 6040-6042). https://doi.org/10.1109/BigData47090.2019.9005594

Djonko-Moore, C.M., Leonard, J., Holifield, Q., Bailey, E.B., & Almughyirah, S.M. (2018). Using culturally relevant experiential education to enhance urban children's knowledge and engagement in science. *Journal of Experiential Education*, *41*(2), 137-153. https://doi.org/10.1177/1053825917742164

European Union (Ed.) (2010). *Charter of fundamental rights of the European Union. Official Journal of the European Union C83*, *53*, *380*, volume 83. https://bit.ly/3PFo605

Fraisl, D., Campbell, J., See, L., Wehn, U., Wardlaw, J., Gold, M., Moorthy, I., Arias, R., Piera, J., Oliver, J.L., Masó, J., Penker, M., & Fritz, S. (2020). Mapping citizen science contributions to the UN sustainable development goals. *Sustainability Science*, *15*(6), 1735-1751. https://doi.org/10.1007/s11625-020-00833-7

Gruschka, N., Mavroeidis, V., Vishi, K., & Jensen, M. (2018). Privacy issues and data protection in big data: A case study analysis under GDPR. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 5027-5033). IEEE. https://doi.org/10.1109/BigData.2018.8622621

Hiller, S.E., & Kitsantas, A. (2014). The effect of a horseshoe crab citizen science program on middle school student science performance and STEM career motivation. *School Science and Mathematics*, *114*(6), 302-311. https://doi.org/10.1111/ssm.12081

Kasgar, A.K., Agrawal, J., & Sahu, S. (2012). New modified 256-bi MD5 algorithm with SHA. Compression Function. *International Journal of Computer Applications*, *42*(12), 47-51. https://doi.org/10.5120/5748-7956

Kobori, H., Dickinson, J.L., Washitani, I., Sakurai, R., Amano, T., Komatsu, N., Kitamura, W., Takagawa, S., Koyama, K., Ogawara, T., & Miller-Rushing, A.J. (2016). Citizen science: a new approach to advance ecology, education, and conservation. *Ecological Research*, *31*(1), 1-19. https://doi.org/10.1007/s11284-015-1314-y

Kolay, S., D'alberto, P., Dasdan, A., & Bhattacharjee, A. (2008). A larger scale study of robots. txt. In *Proceedings of the 17th international conference on World Wide Web* (pp. 1171-1172). https://doi.org/10.1145/1367497.1367711

Li, Y., & Manoharan, S. (2013). A performance comparison of SQL and NoSQL databases. In *2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)* (pp. 15-19). IEEE. https://doi.org/10.1109/PACRIM.2013.6625441

Lin-Hunter, D.E., Newman, G.J., & Balgopal, M.M. (2020). Citizen scientist or citizen technician: A case study of communication on one citizen science platform. *Citizen Science: Theory and Practice*, *5*(1). https://doi.org/10.5334/cstp.261

Massa, N., Dischino, M., Donnelly, J.F., & Hanes, F.D. (2011). Creating real-world problem-based learning challenges in sustainable technologies to increase the STEM Pipeline. In *2011 ASEE Annual Conference & Exposition*. https://doi.org/10.18260/1-2–17678

Parvez, M.S., Tasneem, K.S.A., Rajendra, S.S., & Bodke, K.R. (2018). Analysis of different web data extraction techniques. In *2018 International Conference on Smart City and Emerging Technology (ICSCET)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICSCET.2018.8537333

Ponti, M., Hillman, T., Kullenberg, C., & Kasperowski, D. (2018). Getting it right or being top rank: Games in citizen science. *Citizen Science: Theory and Practice*, *3*(1). https://doi.org/10.5334/cstp.101

Roche, J., Bell, L., Galvão, C., Golumbic, Y.N., Kloetzer, L., Knoben, N., Laakso, M., Lorke, J., Mannion, G., Massetti, L., Mauchline, A., Pata, K., Ruck, A., Taraba, P., & Winter, S. (2020). Citizen science, education, and learning: challenges and opportunities. *Frontiers in Sociology*, *5*, 613814-613814. https://doi.org/10.3389/fsoc.2020.613814

Roldán-Álvarez, D., Martínez-Martínez, F., & Martín, E. (2021). Citizen science and open learning: A Twitter perspective. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 6-8). IEEE. https://doi.org/10.1109/ICALT52272.2021.00009

Saddiqa, M., Larsen, B., Magnussen, R., Rasmussen, L.L., & Pedersen, J.M. (2019). Open data visualization in Danish schools: A case study. *Complex Systems Informatics and Modeling Quarterly*, *21*, 1-20. https://doi.org/10.24132/CSRN.2019.2902.2.3

Sanz, F., Gold, M., & Mazzonetto, M. (2019). *D2.3: Platform functionality requirements & specification report*. Zenodo. https://doi.org/10.5281/zenodo.3612808

Stocklmayer, S.M., Rennie, L.J., & Gilbert, J.K. (2010). The roles of the formal and informal sectors in the provision of effective science education. *Studies in Science Education*, *46*(1), 1-44. https://doi.org/10.1080/03057260903562284

Storksdieck, M., Shirk, J.L., Cappadonna, J.L., Domroese, M., Göbel, C., Haklay, M., Miller-Rushing, A.J., Roetman, P., Sbrocchi, C., & Vohland, K. (2016). Associations for citizen science: Regional knowledge, global collaboration. *Citizen Science: Theory and Practice*, (2), 1-1. https://doi.org/10.5334/cstp.55

Veeckman, C.M., Talboom, S., Gijsel, L., Devoghel, H., & Duerinckx, A. (2019). *Communicatie bij burgerwetenschap: Een praktische handleiding voor communicatie en betrokkenheid bij citizen science*. SCIVIL. https://bit.ly/3PKQz50

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., & Samson, R. (2021). *The science of citizen science*. Springer Nature. https://doi.org/10.1007/978-3-030-58278-4

Warin, C., & Delaney, N. (2020). *Citizen science and citizen engagement. Achievements in Horizon 2020 and recommendations on the way forward*. European Commision. https://doi.org/10.2777/05286