Humanidades Digitales: Construcciones locales en contextos globales. Asociación Argentina de Humanidades Digitales, Buenos Aires, 2016.

Recursos digitales para el acceso a los bienes culturales en dominio público.

Gamba Guido, Heidel Evelin, Raia Matías, Acuña Ezequiel, Actis Caporale Carla, De La Hera Diego y Acevedo Melisa.

Cita:

Gamba Guido, Heidel Evelin, Raia Matías, Acuña Ezequiel, Actis Caporale Carla, De La Hera Diego y Acevedo Melisa (2016). *Recursos digitales para el acceso a los bienes culturales en dominio público. Humanidades Digitales: Construcciones locales en contextos globales. Asociación Argentina de Humanidades Digitales, Buenos Aires.*

Dirección estable: https://www.aacademica.org/aahd2016/3

ARK: https://n2t.net/ark:/13683/ey3x/W8t

ASOCIACIÓN ARGENTINA DE HUMANIDADES DIGITALES

Humanidades Digitales: construcciones locales en contextos globales

Actas del I Congreso Internacional de la Asociación Argentina de Humanidades Digitales (AAHD)

FILO:UBA Facultad de Filosofía y Letras

FILO digital Repositorio Institucional

Humanidades Digitales : Construcciones locales en contextos globales : Actas del I Congreso Internacional de la Asociación Argentina de Humanidades Digitales - AAHD / Agustín Berti ... [et al.] ; editado por Gimena del Rio Riande, Gabriel Calarco, Gabriela Striker y Romina De León - 1a ed . - Ciudad Autónoma de Buenos Aires : Editorial de la Facultad de Filosofía y Letras Universidad de Buenos Aires, 2018.

Libro digital, PDF

Archivo Digital: descarga y online

ISBN 978-987-4019-97-4

1. Actas de Congresos. 2. Humanidades. 3. Digitalización. I. Berti, Agustín II. del Rio Riande, Gimena, ed.

CDD 301

Humanidades Digitales. Construcciones locales en contextos globales

Gimena del Rio Riande, Gabriel Calarco, Gabriela Striker y Romina De León (Eds.)
ISBN: 978-987-4019-97-4

› *Índice*

Letras en Argentina

MARTIN, Jonathan y **TORRES**, Diego. *Análisis de patrones en la evolución de wikis*

MARTÍNEZ CANTÓN, Clara Isabel; **DEL RIO RIANDE**, Gimena y **GONZÁLEZ-BLANCO GARCÍA**, Elena. Poetriae. *Una colección de poéticas medievales basada en conceptos métricos únicos y referenciables*

SUED, Gabriela. *Ciudades visibles: estética y temática de tres ciudades iberoamericanas en la red social* Instagram*. Un estudio exploratorio desde las Humanidades Digitales*

III. Educación, políticas públicas, Humanidades Digitales en el aula

DAVICO, María Luz; **LINEARES**, Gabriel y **PEZZUTTI**, Luciana. *Literacidad electrónica en la enseñanza universitaria: cómo, cuándo y dónde*

MUÑOZ, Patricia Alejandra. *Valoración de un proyecto de desarrollo tecnológico y social en la enseñanza de Inglés como lengua extranjera*

PACHECO DE OLIVEIRA, Maria Lívia y **SÁ DE PINHO NETO**, Júlio Afonso. *Brecha digital e o acesso à informação: projetos de inclusão digital*

CASASOLA, Laura. *Experiencia educativa con TIC: Celulares en acción*

DÍAZ, Aída Alejandra y **HUALPA**, Mariela. *Una experiencia de aprendizaje en educación superior mediada por TIC*

FRESCURA TOLOZA, Claudio Daniel. *Computación en la nube en la enseñanza de escritura académica*

LEÁNEZ, Nancy; **LECETA**, Andrea; **MARTÍN**, Marcela y **MORCHIO**, Marcela. *Hacia una reconfiguración del aula de lengua extranjera*

OLAIZOLA, Andrés. *Los escritores vernáculos digitales y el concepto de valor en las escrituras digitales*

CHECHELE, Patricia; **LURO**, Vanesa y **PINTOS ANDRADE**, Esteban. *Afiliarse en la distancia. El ingreso a la educación superior en un entorno virtual de aprendizaje*

ALLÉS TORRENT, Susanna y **DEL RIO RIANDE**, Gimena. *Enseñar edición digital con TEI en español. Aprendizaje situado y transculturación*

IV. Medios, re-mediación, redes sociales

RODRIGUEZ KEDIKIAN, Martín. #100DiasdeMacri. *Analítica cultural en la construcción de los primeros cien días de la presidencia de Mauricio Macri en conversaciones en* Twitter

ALONSO, Julio**; ALAMO**, Sofía; **GONZALEZ OCAMPO**, María Eugenia; **GIAMBARTOLOMEI**, Guido; **MANCHINI**, Lucas y **TOSCANO**, Ayelén. *¿Hacia una algoritmización de los sentimientos?*

DE MIRANDA, Jair Martins. *Samba Global– Do mundo do samba ao samba no mundo*

ORTIZ, María. *Las migraciones en los tiempos del software*

dos SANTOS, Laura. *Arte urbano, de la calle a las redes*

ALAMO, Sofía; **BORDOY**, Giselle; **CHETTO**, Melisa; **IBAÑEZ**, Fernanda, **MIGLIORINI**, Agustina y **GONZALEZ OCAMPO**, María Eugenia. #NiUnaMenos: *Big Data para la comprensión de una problemática de género*

KLIMOVSKY, Pedro. *El documental digital y la representación de lo real*

BERTI, Agustín. *Fotogramas autorizados: La crisis de la noción de obra cinematográfica ante las remasterizaciones*

BORDOY, Giselle. *El disco como obra abierta en interacción con las audiencias*

COELHO, Cidarley. *Forma Material Digital: livro e leitura na sociedade contemporânea*

V. Reflexiones sobre/desde/hacia lo digital

VISCARDI, Ricardo. *Actuvirtualidad e inter-rogación: un lugar entre-otros*

ÁLVAREZ GANDOLFI, Federico y **DEL VIGO**, Gerardo Ariel. *Hatsune Miku, una idol digital: entre el* otakismo *y el* waifuismo

SAÁ, Guido. *Reflexiones sobre música y narración: Recursos retóricos y exegéticos musicales en la línea narrativa y el pathos en* BioShock 2 *y* BioShock Infinite

GLUZMAN, Georgina Gabriela. *Algunas reflexiones sobre la* Base de datos de mujeres artistas en Buenos Aires (1924-1939)

DOMINGUEZ HALPERN, Estela**; ALAMO**, Sofía; **ALONSO**, Julio. *Entramados y ciudades. Visibilizando* Baldosas por la Memoria

GÓMEZ, Verónica Paula. *Territorios nacionales, territorialidades ciberespaciales: disputas discursivas sobre la soberanía en la circulación de literatura digital*

RIGAL COLLADO, Pablo Alonso; **MAESTIGUE PRIETO**, Nancy y **GARCÍA VÁZQUEZ**, Maytée. *La narración hipertextual. El reto cubano*

VI. La publicación científica y el Acceso Abierto desde las Humanidades Digitales

TSUJI, Teresa y **CANELLA**, Rubén. *Lenguajes y recursos multimediales para la difusión de la ciencia. Desafíos y oportunidades digitales*

CATALDI, Marcela; **DI CÉSARE**, Victoria; **FERNÁNDEZ**, Néstor; **HERNÁNDEZ**, Alicia; **LIBERATORE**, Gustavo y **VOUTTO**, Andrés. *Sistema taxonómico de organización de los recursos de información autoarchivados en el Repositorio Institucional de la Facultad de Humanidades de la Universidad Nacional de Mar del Plata*

ÁLVAREZ, Leonardo Javier y **CORDA**, María Cecilia. FLACSOAndes *Tesis: comunicación científica de investigaciones realizadas en maestrías y doctorados del sistema* FLACSO

VII. Digitalización, políticas y prácticas, archivo y memoria

AUTHIER, Carlos; **GIORDANINO**, Eduardo y **LUIRETTE**, Carlos. *La preservación de la memoria audiovisual en Argentina*

GAMBA, Guido; **HEIDEL**, Evelin; **RAIA**, Matías; **ACUÑA**, Ezequiel; **ACTIS CAPORALE**, Carla; **DE LA HERA**, Diego y **ACEVEDO**, Melisa. *Recursos digitales para el acceso a los bienes culturales en dominio público*

FLORES MUTIGLIENGO, Jennifer. *Arte y Archivo*

BUGNONE, Ana y **SANTAMARÍA**, Mariana. *La política de democratización del archivo: el caso del Centro de Arte Experimental Vigo*

GAMBA, Guido; **HEIDEL**, Evelin; **RAIA**, Matías; **ACUÑA**, Ezequiel; **ACTIS CAPORALE**, Carla; **DE LA HERA**, Diego y **ACEVEDO**, Melisa. *Digitalización: Una experiencia de campo*

Recursos digitales para el acceso a los bienes culturales en dominio público

GAMBA, Guido / Universidad de Buenos Aires (UBA). Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET) – gambaguido@gmail.com

HEIDEL, Evelin / Museo del Cine Pablo Ducrós Hicken. Universidad de Buenos Aires (UBA) – scannopolis@gmail.com

RAIA, Matías / Universidad de Buenos Aires (UBA) – mhraia@yahoo.com.ar

ACUÑA, Ezequiel / Universidad de Buenos Aires (UBA) – ezequielmac@gmail.com

ACTIS CAPORALE, Carla / Universidad de Buenos Aires (UBA) – carla.actis@gmail.com

DE LA HERA, Diego / Universidad de Buenos Aires (UBA) – delahera@gmail.com

ACEVEDO, Melisa / Universidad de Buenos Aires (UBA) – melisacevedo@gmail.com

» *Palabras clave: dominio público, propiedad intelectual, internet, base de datos, acceso.*

› *Resumen*

El dominio público es el estado que adquieren las obras literarias, científicas y artísticas una vez expirado el derecho de autor. Es, de hecho, el patrimonio cultural de la humanidad y una fuente inagotable para la creación de nuevos bienes intangibles, dado que, finalizado el plazo de monopolio legal sobre la propiedad intelectual, cualquier persona puede utilizar las obras de manera relativamente libre. En otras palabras, no es una caída en desgracia para la obra, sino que, bien al contrario, el dominio público la habilita a tener una circulación mucho más extendida y a un costo mucho más bajo. Así las cosas, el principal problema para tomar real dimensión del valor del dominio público es la ausencia de información. No se sabe, porque no existe ninguna clase de registro, qué obras están en el dominio público: hay que determinar individualmente, obra por obra, autor por autor, cuál es la situación de cada trabajo. De hecho, muchas veces la información sobre la situación patrimonial de la obra puede llegar a perderse y es así como existe un 98% de obras cuya situación patrimonial se desconoce y son calificadas como huérfanas. En este contexto, el presente trabajo busca profundizar en las herramientas digitales disponibles para la creación, publicación y mantenimiento de un registro de obras en dominio público. Más particularmente, se centra en el caso del portal dominiopublico.org.ar, ya que se trata de una plataforma que implementa una serie de recursos interesantes para el caso: desde proponer una estructura informática que facilita la identificación del estado de derechos de autor de los

bienes culturales, hasta la puesta en circulación digital de obras en dominio público, apelando, además, a técnicas semi-automáticas de recolección y procesamiento de información bibliográfica de referencia para hacer más eficiente la tarea.

› *El dominio público en Argentina y la necesidad de hacer las cuentas*

Antes de pasar a los detalles técnicos respecto del proyecto *Dominio Público de Argentina*, es necesario preguntarse para qué armar una base de datos de estas características. La tarea de construir una base de datos con todos los autores argentinos que alguna vez publicaron o editaron algún tipo de obra, ¿no es acaso una tarea hercúlea y de proporciones desmesuradas? ¿Vale la pena? ¿Cuál es el sentido de emprender una tarea de estas características en un mundo donde cada vez más la información se produce directamente en digital?

La razón de ser de esta base de datos es primordialmente política. En efecto, el dominio público no es simplemente el estado en el que ingresan las obras una vez que expiran los plazos de monopolio del derecho de autor. El dominio público es nuestra antigua plaza pública, es el estado que existía de manera previa a que existieran las formas privadas sobre el conocimiento, el arte y la cultura. Todo era dominio público hasta que en 1710 se forjó la primera ley de copyright, el Estatuto de la Reina Ana. El dominio público era enorme porque abarcaba todas las obras y porque era la norma para la circulación de la cultura y no la excepción. Y era, además, el espacio donde todos éramos libres para jugar, crear y recrear con las obras. Reivindicar el dominio público es una forma de reivindicar la existencia de un espacio que nos fue expropiado.

Desde 1710 hasta hoy la situación ha cambiado radicalmente. La extensión de los privilegios monopólicos de la propiedad intelectual tanto en plazos como en áreas que abarca, su fijación en normas internacionales (cuya violación incluye la posibilidad de un Estado de ser demandado en tribunales extranjeros con escaso escrutinio democrático) y el avance voraz de un puñado de corporaciones multinacionales (para privatizar el conocimiento y la cultura), han debilitado el concepto del dominio público y su importancia estratégica para el ejercicio de derechos humanos fundamentales.

Normalmente se entiende al dominio público como el estado al que ingresan las obras literarias, artísticas y científicas una vez que han expirado los plazos de monopolio del derecho de autor (70 años post-mortem en Argentina). Que ingresen al dominio público implica que cualquier persona puede utilizar esas obras para crear nuevas obras, trasladarlas a diferentes formatos, traducirlas, intervenirlas y también digitalizarlas y ponerlas a disposición del público, entre otras actividades. Así, el dominio público se convierte en un interesante territorio para la experimentación artística, para la divulgación en general y para

la puesta a disponibilidad de obras culturales.

Sin embargo, conocer el momento en que una obra ingresa al dominio público no es tan simple como parece. En efecto, el Convenio de Berna y el Acuerdo sobre Aspectos de la Propiedad Intelectual relacionados con el Comercio (ADPIC), reconocen el principio jurídico de la *protección automática*. Este principio implica que la obra recibe protección de manera automática en el momento de su fijación en un soporte, exista o no un trámite formal en una oficina de derecho de autor. Esto quiere decir que la opción por defecto para todas las obras es *todos los derechos reservados*, independientemente de la real voluntad del autor, e implica en la práctica que uno debe solicitarle al titular de los derechos de autor permiso para cualquier clase de uso que quiera darle a la obra.

Los perjuicios de semejante protección son evidentes, especialmente para las instituciones culturales, que no tienen la capacidad ni los recursos para solicitar permiso por cada uso que quieran hacer de una obra. La mayoría de los países resolvieron esta situación tempranamente, fijando excepciones en sus leyes de derecho de autor para permitir a las instituciones culturales y educativas que hagan uso de las obras sin necesidad de pedir permiso al titular de los derechos, y en la enorme mayoría de los casos, sin exigir una remuneración como contrapartida por el uso. Argentina es, sin embargo, uno de los 21 países, de los 186 adheridos a la OMPI (la Organización Mundial de la Propiedad Intelectual) que no tiene excepciones educativas ni de uso justo.

Pero, aún más, con el tiempo el principio de protección automática o *copyright por default* generó una situación que no fue prevista por los redactores del Convenio de Berna: la aparición de las obras huérfanas, es decir, obras que no se sabe su estado legal simplemente porque se desconoce quién es su titular o bien cuál es su situación. ¿Vive aún el titular? ¿Falleció? Y si falleció, ¿hace cuánto? ¿O acaso su obra se encuentra en manos de un albacea o heredero que gestiona los derechos? ¿Y dónde se lo podrá localizar? Estas preguntas, que pueden parecer las elucubraciones de una mente peregrina, se corresponden en realidad con alrededor del 98% de las obras, según estimaciones de la oficina de copyright de los Estados Unidos.

La situación empeoró aún más en los últimos años, con la extensión desmedida de los plazos de protección de las obras. De los iniciales 14 años post-publicación del Estatuto de la Reina Ana en la Gran Bretaña de 1710, a los 30 años post-mortem que dieron la mayoría de los países en sus primeras leyes modernas de derecho de autor, se pasó a 50 post-mortem como piso mínimo en virtud del Convenio de Berna. Hoy la mayoría de los países se encuentran en 70 años post-mortem.

Esta cifra no le sugiere absolutamente nada a la mayor parte de la gente. Sin embargo, esto quiere decir que la obra de un escritor argentino muy prolífico y de importancia fundamental en la vida cultural argentina, como Juan Filloy, estará protegida durante más o menos unos 150 años –suponiendo que hubiera empezado a producir su obra alrededor de los 30. La mayoría de los que están leyendo este artículo –e incluso quienes lo escriben– probablemente estén muertos para el año 2071, momento en que la obra de Juan Filloy

entrará finalmente al dominio público.

El problema es que nadie pone las cifras de la protección del derecho de autor en estos términos, porque implicaría dar un privilegio monopólico por el equivalente a una eternidad, salvo que se crea en la reencarnación o que nuestra única fuente para hacer cultura en el siglo XXI sean las fantásticas obras de Giordano Bruno. Más aún, estamos considerando únicamente a nuestra exquisita cultura letrada, porque si se consideran los pocos medios posmodernos agraciados, (como el cine, la radio, la televisión y más recientemente el código, donde hay más de un autor y titular de derechos involucrados) es fácil darse cuenta de que esta protección es literalmente la eternidad.

Es probable, además, que en 140 años nadie recuerde quiénes eran, cuándo fallecieron, dónde se puede encontrar a sus herederos, y otras cuestiones semejantes respecto a la mayoría de los autores que recibieron semejante protección. Incluso que en ese momento la obra tampoco tenga un contexto de lectura que la haga inteligible (en algunos casos ni siquiera encontrará un soporte donde pueda leerse). Pero incluso sin hacer futurología, lo cierto es que actualmente no sabemos cuál es el estado de más del 98% de la cultura del siglo XX porque la protección automática nos impide hacer uso de esas obras y porque no existe ninguna clase de registro ni información consistente sobre qué obras efectivamente están en dominio público.

La paradoja en Argentina es que además tenemos una figura muy particular, llamada *dominio público pagante*, según la cual estamos obligados a pagar por el uso de las obras que se encuentran en dominio público, a excepción de los usos sin ánimo de lucro. Esta figura fue instituida en Argentina en el año 1958, a través del decreto ley 1224[1], mediante el cual se creaba el Fondo Nacional de las Artes y el dominio público pagante como forma de financiarlo. El dinero de dicha figura efectivamente va a parar al Fondo Nacional de las Artes, luego de que las sociedades de gestión colectiva respectivas hayan hecho los descuentos correspondientes por la generosa tarea de recolectar el dinero, en concepto de *gastos de administración*.

Sin embargo, ni el Fondo Nacional de las Artes, ni las sociedades de gestión colectiva, ni mucho menos la Dirección Nacional de Derecho de Autor, pueden decir de manera fehaciente qué autores se encuentran en el dominio público y qué autores no. Aún más, ese registro, si existe, no es público.

Es razonable preguntarse en este punto: ¿qué tan complicado puede ser hacer la cuenta sabiendo cuándo falleció un autor? No es que se necesite saber la fecha de muerte de un autor específico y a partir de ahí hacer un cálculo. El problema es de carácter ontológico; lo que se necesita es saber realmente qué es de dominio público, las obras que abarca, los autores que contiene, todas las obras que una persona puede editar, publicar o digitalizar el año que viene. Y lo más importante: tomar dimensión real de que extender estúpidamente los plazos de ingreso de las obras al dominio público implica encerrar miles de obras en el oscurantismo de la propiedad privada por dos o tres herederos (o peor aún, por dos o tres

[1] Véase http://servicios.infoleg.gob.ar/infolegInternet/verNorma.do?id=37242.

discográficas) que se benefician de ello.

Un buen mecanismo para entender este problema es darle una vuelta de tuerca a una analogía que con frecuencia y con poca justicia se hace en el campo del derecho de autor. Algunos autoralistas malintencionados suelen decir que el derecho de autor es como la propiedad de una casa: si podemos heredarles a nuestros hijos la propiedad de una casa, ¿por qué no heredarles el derecho de autor? Pero, por supuesto, para que el Estado les reconozca la titularidad de la casa, o de un automóvil, les pide que se registren en una oficina especial donde consta que ustedes son los propietarios de ese bien y donde también se registran las transacciones de los mismos (compra y venta) y registra también si el titular del bien falleció. En el caso del derecho de autor, el Estado provee de un privilegio que no se registra en ningún lado y del que no se deja constancia de su estado a medida que se efectúan operaciones sobre esa obra, como compra, venta, traspaso hereditario o cambio de titularidad. Esto es como tener un automóvil sin papeles abandonado en la calle sin que el Estado ni nadie lo pueda mover de ese lugar. Nadie sabe a quién le pertenece, pero nadie tampoco lo puede mover, así que en defensa del sacrosanto privilegio del derecho de autor lo mejor es dejarlo ahí hasta que termine de oxidarse por completo. Esta es la verdadera catástrofe cultural de la que tenemos que seguir conversando.

A esta falta de información consistente y unificada es a la que intentamos darle respuesta mediante la construcción de una base de datos de autores de Argentina en el dominio público, inspirados en la experiencia del grupo de *Autores.Uy*[2]. El sitio Dominio Público[3] busca convertirse en una base de datos de fácil consulta para saber si las obras de un autor se encuentran en dominio público o no, y en una segunda instancia en una biblioteca de obras en el dominio público.

› *De un problema teórico a un conflicto metodológico*

La pregunta que sigue a continuación se refiere al cómo construir la base de datos del dominio público. Para ello, en una primera instancia necesitábamos un CMS (Content Management System, sistema de gestión de contenidos) que permitiera una interfaz amigable para la carga de datos y para la presentación de la información y fuentes bibliográficas secundarias digitalizadas (enciclopedias, diccionarios biobibliográficos, diccionarios biográficos, referidos a autores de todas las disciplinas artísticas).

Los compañeros de *Autores.Uy* ya habían configurado un sitio *Drupal*, por lo que simplemente personalizamos la instalación con la que ellos estaban trabajando y lo subimos a un servidor provisto por la organización USLA (Usuarios de Software Libre de Argentina).

Lo siguiente era digitalizar las fuentes. En una primera instancia, reunimos más de 25 fuentes bibliográficas secundarias y las digitalizamos con escáneres de libros *Do It Yourself*,

[2] Véase http://autores.uy.
3 Véase http://dominiopublico.org.ar.

escáneres hechos con madera, cámaras de fotos y controlados a través de una *RaspberryPi*[4].

¿Qué es un escáner de libros Do It Yourself*?*

Es un escáner que, como lo indica su nombre, cualquiera puede hacer. En su versión más simple, es una estructura diseñada para sostener un libro, dos cámaras de fotos y una lámpara: el material de la estructura puede ser algo tan pedestre como cajas de cartón hasta extrusiones de aluminio en sus versiones más complejas.

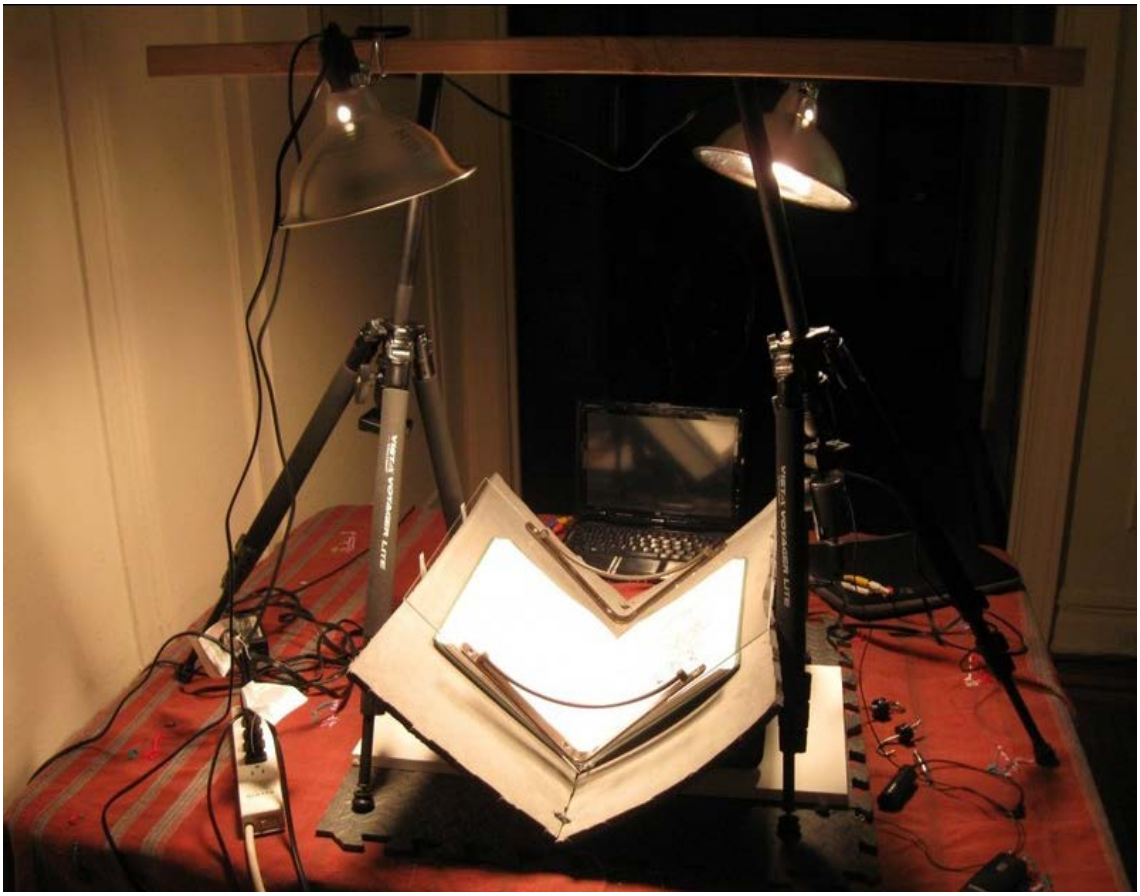

Figura 1. Ejemplo de un escáner construido con cartón. Fuente: foro DIY Book Scanner.

4 Véase http://diybookscanner.org/es/index.html.

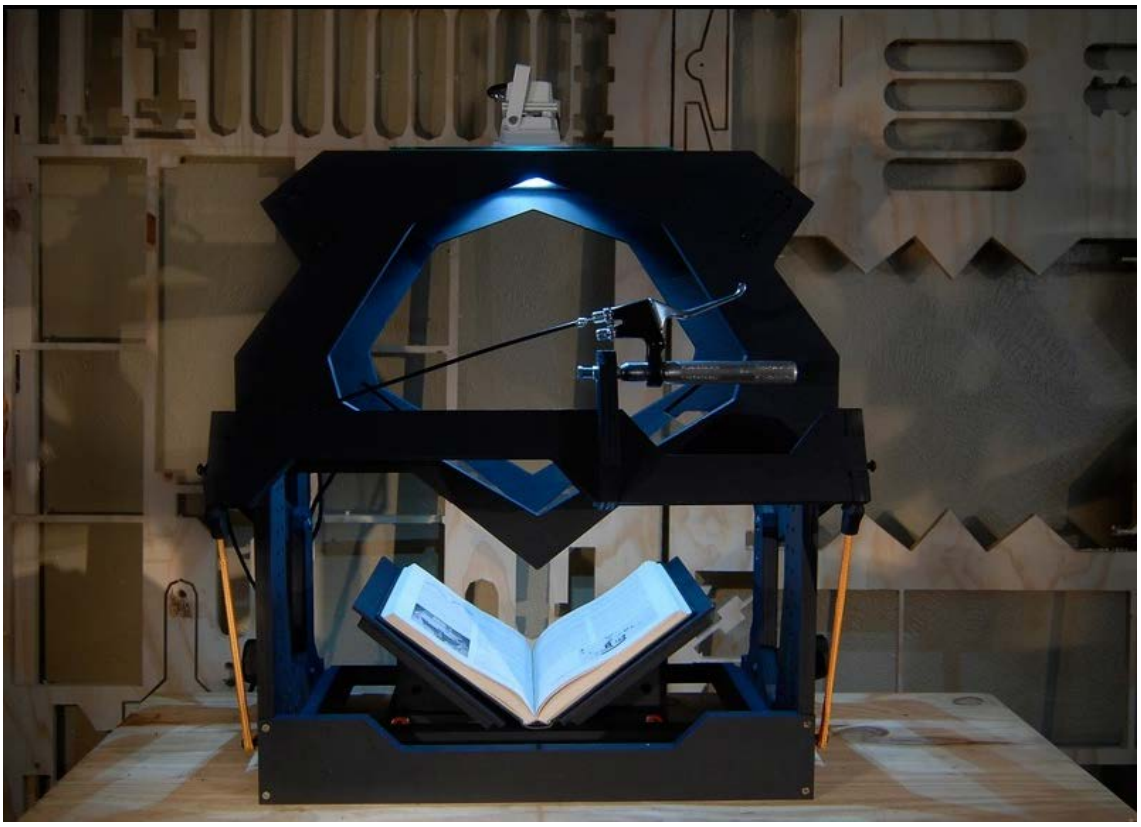

Figura 2. Modelo Hackerspace*. Fuente: foro* DIY Book Scanner*.*

Las instrucciones para construir este escáner de libros se encuentran en Internet, donde existe también una comunidad de colaboradores muy activa que se dedica a pensar y probar nuevos diseños, en el foro de los escáneres *Do It Yourself*[5]. En particular, el modelo que usamos para digitalizar es *El Archivista*, del que se pueden encontrar instrucciones muy detalladas para construirlo[6].

Las ventajas de este tipo de escáneres por sobre los escáneres comerciales y/o de cama plana es evidente. La primera característica que sobresale es la rapidez con la que se puede digitalizar. Un libro de 400 páginas toma entre 20 y 30 minutos, dependiendo del estado en el que se encuentre el ejemplar. Por otro lado, no deforma el libro ya que la cama superior del libro está diseñada para que el ejemplar pueda abrirse de manera gentil, sin dañar el lomo ni la encuadernación.

Frente a los escáneres comerciales de libros, se destaca su precio (con unas cámaras Canon ELPH160, de 20 megapíxeles, cuesta entre 7.000 y 8.000 pesos argentinos). Otra ventaja es la facilidad para reemplazar partes: todos los materiales con los que está hecho el escáner se encuentran fácilmente en una ferretería de barrio y reemplazar las partes no

[5] Véase http://diybookscanner.org/forum.
[6] Véase http://www.diybookscanner.org/archivist/.

requiere conocimiento técnico especializado. Las cámaras también pueden cambiarse fácilmente si sale un modelo mejor en el mercado o si hay ventajas comparativas entre dos modelos distintos.

Pero quizás, la principal y más importante característica sea el espíritu con el que está hecho el escáner. En efecto, tanto los diseños como el software de control y de post-procesamiento están liberados con una licencia libre, tanto de hardware libre como de software libre. Esto implica que no sólo se puede estudiar cómo está hecho el software (y el modelo), sino que también se puede modificarlo, copiarlo y distribuirlo, respetando la licencia original.

El escáner se maneja a través de una *RaspberryPi*, una pequeña computadora del tamaño de un celular promedio. El sistema operativo se instala sobre una tarjeta microSD. En este caso, el controlador del escáner se llama *PiScan* (y puede descargarse desde el sitio del proyecto en *GitHub*[7] que incluye un manual en castellano). Las cámaras tienen que ser cámaras Canon, ya que funcionan con un complemento especial llamado *Canon Hack Development Kit*[8], un desarrollo no oficial de Canon que se instala sobre la tarjeta SD de la cámara y permite eliminar las limitaciones de software impuestas por el fabricante del dispositivo. En concreto, esto significa que uno puede ampliar las opciones de control del dispositivo (tomar fotos en el formato nativo de la cámara, raw, aplicar filtros de colorimetría, ejecutar comandos sobre la cámara, entre otras opciones).

¿Por qué es necesaria una computadora especial, la *RaspberryPi*, para manejar el escáner? Esta es una pregunta válida para un dispositivo que busca ser económico. Aunque la *Raspberry* no tenga un precio excesivo (en Argentina están alrededor de 1000 pesos). Lo cierto es que en los primeros tiempos de la digitalización con escáneres *Do It Yourself*, uno de los principales obstáculos era el hecho de que cada biblioteca y cada usuario contaban siempre con una computadora diferente, con arquitecturas diferentes y con sistemas operativos diferentes. La comunidad técnica reunida alrededor del foro de los escáneres DIY trató de desarrollar diferentes soluciones para este problema. Finalmente, la opción más simple fue diseñar este software de control que permitiera que el escáner pudiera funcionar como un *stand-alone device*, es decir, un dispositivo autónomo que no necesita de una configuración manual individual. Las fotos se toman y se digitaliza el libro, la información se baja a un pendrive y el pendrive se conecta a la computadora con las imágenes en JPEG.

Una vez que el libro está digitalizado, se utilizan varios softwares de post-procesamiento que nos devuelven la información en un formato que podamos utilizar para la tarea de *scraping*. En particular, el más importante de estos software es el *ScanTailor*[9], un simpático programa diseñado especialmente para trabajar con un gran número de imágenes digitalizadas. El programa nos devuelve los JPEG originales que toman las cámaras en un TIFF blanco y negro. Este programa está licenciado con una licencia GPL y se puede descargar de

[7] Véase https://github.com/Tenrec-Builders/pi-scan.
[8] Véase http://chdk.wikia.com/wiki/CHDK.
[9] Véase http://scantailor.org.

manera libre y gratuita en el sitio del *ScanTailor*.

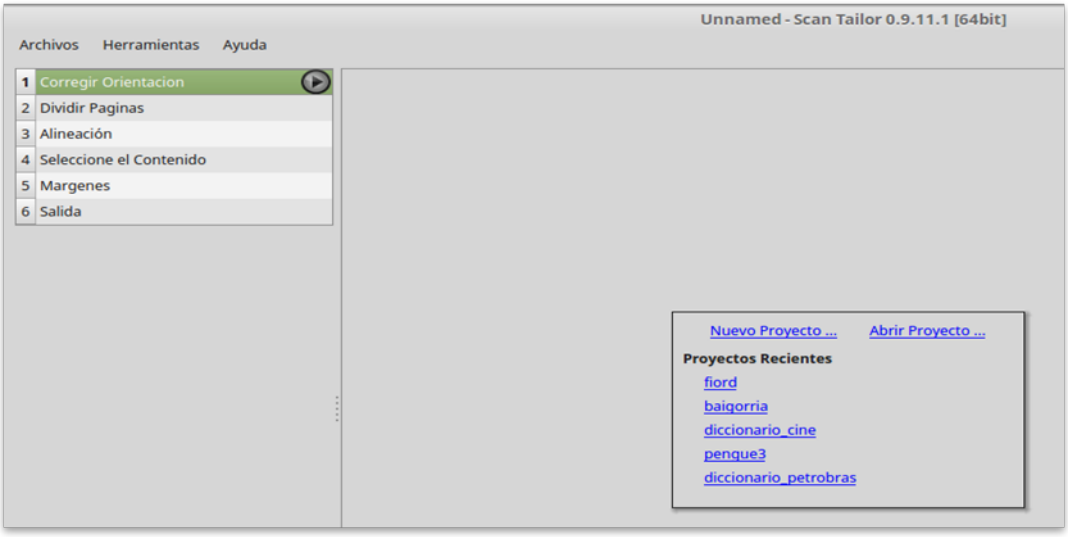

Figura 3. Pantalla del ScanTailor. Fuente: Propia de los autores.

La idea original de la base de datos era subir las fuentes para que los voluntarios del proyecto manualmente fueran ingresando los datos. Sin embargo, a medida que íbamos trabajando con los diccionarios, encontramos que los más nuevos y contemporáneos mostraban una cierta estructura en la presentación de la información de modo tal que ofrecían la posibilidad de hacer *scraping* automático de textos.

› *El* scraping *de textos*

La carga automática de autores a la base de datos comprende tres etapas: 1. reconocimiento óptico de caracteres, 2. identificación y extracción de campos, y 3. integración con la base.

1. Reconocimiento óptico de caracteres:
   Las fuentes digitalizadas mejor estructuradas fueron analizadas con el motor Tesseract para el reconocimiento óptico de caracteres (OCR, por sus siglas en inglés). Tesseract[10] es un software de código abierto. Como resultado de aplicar el OCR, obtuvimos un documento hOCR, un estándar abierto que utiliza XML en forma de HTML para almacenar no sólo el texto digitalizado, sino también información sobre su distribución en la página e índices de confianza para cada palabra reconocida, entre otros atributos. A partir de este documento hOCR se identifican y extraen los campos de interés en la etapa siguiente.

[10] Véase: https://github.com/tesseract-ocr/

2. Identificación y extracción de campos:

   Como se indicó más arriba, notamos que algunas fuentes muestran una gran regularidaden la presentación de la información biográfica de los autores. En algunos casos esta regularidad ocurre a nivel de la distribución del texto en la página (por ejemplo, si las líneas que comienzan con el nombre de un autor sistemáticamente tienen sangría mayor que las otras), y en otros a nivel del texto mismo (por ejemplo, si los años de nacimiento y muerte invariablemente figuran entre paréntesis, separados con un guión). Por lo tanto en esta etapa empleamos el documento hOCR obtenido en la etapa previa, aprovechando que contiene no sólo el texto sino también información sobre su disposición en la hoja.

   Para llevar a cabo esta tarea decidimos escribir un programa que identificara y extrajera loscampos de interés conociendo sus patrones regulares de presentación particulares para cada fuente. Si bien hay lenguajes quizá más especializados para el trabajo con cadenas de texto, como Perl, elegimos Python por su popularidad y versatilidad.

   En una primera versión del programa, estos campos fueron identificados principalmente haciendo uso de expresiones regulares. Una expresión regular es una secuencia de caracteres que forma un patrón de búsqueda y permite hallar partes de un texto que cumplen con dicho patrón. Estos patrones de búsqueda son a la vez estrictos pero suficientemente flexibles para tolerar posibles errores del motor OCR. En versiones recientes, aún en desarrollo, aprovechamos las capacidades del lenguaje para explorar árbol XML y el lenguaje de consulta XPath para simplificar el código y hacerlo más fácilmente adaptable a distintas fuentes.

   Este programa también incluye rutinas de corrección automática de sitios de nacimiento y muerte ante errores menores del motor OCR, y rutinas de validación en todos los campos según patrones esperados.

   Como resultado, se obtiene un documento CSV en el que cada línea contiene información sobre un autor: nombre completo, seudónimos, género, años y lugares de nacimiento y muerte, y disciplina. Se incluye también el índice de confianza del motor OCR para algunos campos, errores de validación y número de página en la fuente original para facilitar la revisión manual por los voluntarios del proyecto.

   La primera versión del programa se encuentra especialmente adaptada para una fuente particular, pero versiones posteriores serán más fácilmente adaptables e incluso genéricas.

3. Integración con la base:

   Para evitar la generación de autores duplicados, desarrollamos un segundo programa que

busca integrar armónicamente la nueva información obtenida en la etapa previa con las fichas previamente cargadas. Éste toma un volcado actualizado de la base de datos y agrega la nueva información a los autores evidentemente preexistentes, entendiendo por autores evidentemente preexistentes a aquellos con exactamente el mismo nombre completo y la misma fecha de nacimiento. El programa también señala campos en conflicto entre la nueva información y la anterior para revisión manual posterior.

Además, el programa cuenta con un conjunto de algoritmos que buscan reconocer autores preexistentes aún cuando nombres o fechas de nacimiento no coinciden exactamente. Por ejemplo, nombres idénticos con distintas fechas de nacimiento, variantes de nombre (Roberto Arlt y Roberto Emilio Arlt), y nombres similares (habitualmente producto de errores del motor OCR), entre otros. Estos posibles autores preexistentes son señalados para confirmación posterior durante la revisión manual.

Ambos programas se encuentran en activo desarrollo, son de código abierto bajo una licencia GPLv3 y están disponibles online[11]. En la documentación del proyecto se encuentra también un manual del tipo de tareas que deben realizar quienes revisan manualmente la información.

› *Próximos pasos y futuro del proyecto*

Hay fuentes que, por cierto, no pueden ser tratadas de manera automática, en especial las fuentes más antiguas, que en general no suelen tener los datos de manera estructurada. En concreto, esto implica que el proyecto necesita aún voluntarios para poder ingresar la información de las fuentes que no pueden ser cargadas de manera automática.

De manera paralela, el objetivo del proyecto es convertirse en una biblioteca del dominio público de Argentina. La primera instancia es la búsqueda de las obras: hay material de autores argentinos en dominio público digitalizados por la Biblioteca Nacional de España, por Internet Archive[12], y por un sinnúmero de instituciones extranjeras, que se pueden *linkear* dentro de la base.

[11] Véase https://github.com/autoresar/autores/.
[12] Véase http://archive.org.

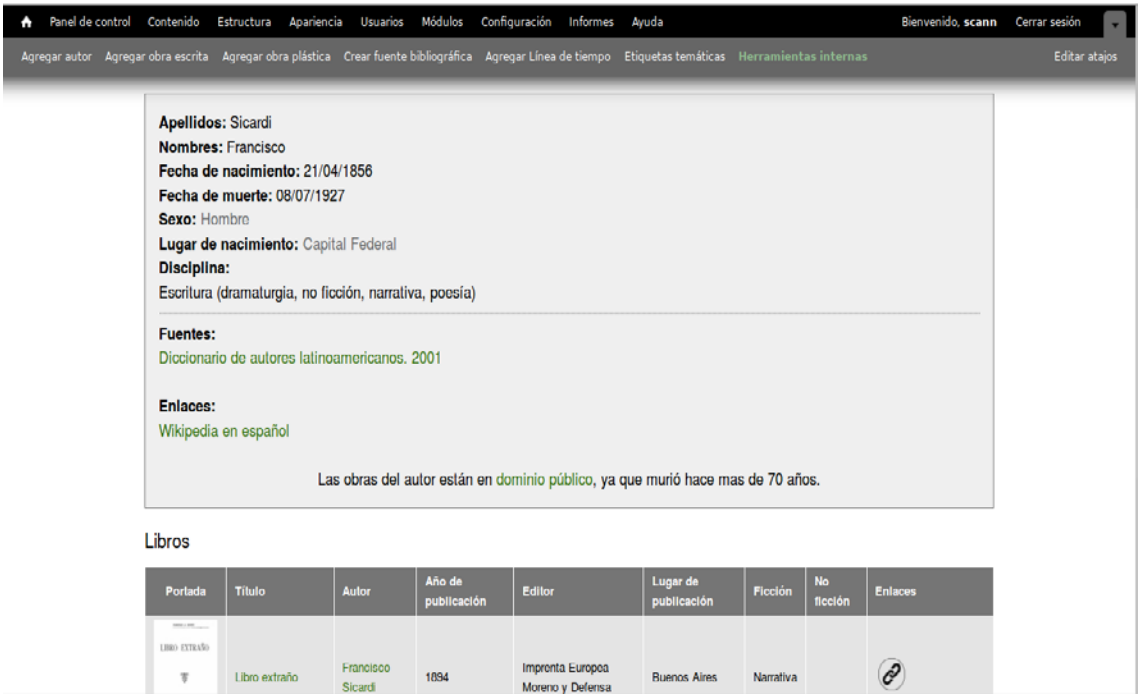

Figura 4. Las obras de Francisco Sicardi linkeadas*. Fuente: Propia de los autores.*

En una segunda instancia, el objetivo será digitalizar obras que se encuentren en dominio público y que sean difíciles de encontrar en Internet (hay muchas obras que cumplen esta condición). Nuestro principal interés es contribuir a destacar el enorme patrimonio cultural argentino.