سیف الدین صالح الریالات
شاهر مومانی

راهنمای مبتدیان برای استفاده از
داده های دسترسی باز



مروّت گیوی

انتشارات ترجمک

راهنمای مبتدیان برای استفاده از داده های دسترسی باز

دسترسی باز و داده بزرگ

توجه :

کتاب حاضر حاصل زحمات **مروت گیوی** می‌باشد. فایل کتاب حاوی اطلاعات **DRM** (مدیریت حقوق دیجیتال) است. وقتی برای اولین بار فایل را باز می کنید، کد شناسایی کتاب به همراه آدرس **IP** سیستم شما ذخیره شده و زمانیکه آنلاین شوید، به سرور انتشارات ترجمک انتقال می یابد.

خواهشمند است به حقوق نگارنده و انتشارات ترجمک احترام گذاشته و از توزیع بدون مجوز فایل کتاب اجتناب نمایید. شما با خرید و دانلود این کتاب موافقت نموده اید که اطلاعات فایل **DRM** به سرور انتشارات ترجمک انتقال یابد و در صورت محرز شدن نقض حقوق صاحب اثر، کلیه خسارات حاصله در طی فرآیند حقوقی و مطابق قانون حمایت حقوق مؤلفان و مصنفان و هنرمندان و ناشران جمهوری اسلامی (مصوب دوازده اسفند ۱۳۶۵ یا بعد از آن) از شما دریافت شود.

از اینکه با عرضه مقرون به صرفه کتاب های الکترونیک و شکوفایی انتشارات ترجمک همیاری می کنید، سپاسگزاریم.

انتشارات ترجمک

https://tarjomac.ir/shop/

سیف الدین صالح الریالات و شهر مومانی

راهنمای مبتدیان برای استفاده از داده های دسترسی باز

ترجمه

مروّت گیوی

ویراست اول

بهار ۱۴۰۴

انتشارات ترجمک

شماره کتابشناسی ملی	:	۱۰۰۶۹۳۰۹
شابک	:	978-622-5126-01-5 ، ۲۲۵۰۰۰۰ ریال
عنوان و نام پدیدآور	:	راهنمای مبتدیان برای استفاده از داده‌های دسترسی باز/سیف‌الدین صالح‌الریالات، شاهر مومانی؛ [مترجم] مروت گیوی.
مشخصات نشر	:	همدان، ۱۴۰۴، ۱۳۰ص
یادداشت	:	عنوان اصلی: A beginner's guide to using open access data,2019
موضوع	:	دسترسی به داده‌های باز، Open scholarship، تحقیق -- روش‌شناسی، Research -- Methodology
سرشناسنامه		گیوی، مروت، ۱۳۴۷–، مترجم
رده بندی کنگره	:	۱۸۱Q [AZ۱۰۱]
رده بندی دیویی	:	۵۰۷/۱۲

شناسنامه کتاب

نام کتاب: راهنمای مبتدیان برای استفاده از داده های دسترسی باز

ترجمه: مروّت گیوی

ناشر: انتشارات ترجمک

صفحه آرایی: انتشارات ترجمک

طراحی جلد: محمدحسین گیوی

نوبت چاپ: اول، ۱٤۰٤

قیمت (چاپی): ۲۲۵۰۰۰ تومان

قیمت (ایبوک): ۱۲۵۰۰۰ تومان

چاپ: گروه نشر الکترونیک ترجمک

شابک (پرینت): ۹۷۸–٦۲۲–۵۱۲٦–۰۱–۵

شابک (ایبوک): ۹۷۸–٦۲۲–۵۱۲٦–۰۲–۲

تلفن تماس: ۰۹۱۸۱۵۰٦۱۰۰

تارنمای اینترنتی: https://tarjomac.com

ISBN : 978-622-5126-01-5

ISBN : 978-622-5126-02-2



نویسنده:

سیف الدین صالح الریالات و شهر مومانی

مترجم:

مروّت گیوی

شابک ترجمه فارسی:

شابک (پرینت): ۹۷۸-۶۲۲-۵۱۲۶-۰۱-۵

شابک (ایبوک): ۹۷۸-۶۲۲-۵۱۲۶-۰۲-۲

سال انتشار: ۲۰۲۰ – ورکس (چاپ دوم)

کلیه حقوق این اثر برای انتشارات ترجمک محفوظ است. این اثر شامل کلیه قوانین حمایت حقوق مؤلفان و مصنفان و هنرمندان و ناشران جمهوری اسلامی می باشند (مصوب دوازده اسفند ۱۳۶۵ یا بعد از آن)، و بهره برداری بدون مجوز از آن به هر شکلی ممنوع است.

انتشارات ترجمک: سال ۱۴۰۳

همدان، صندوق پستی ۱۱۷-۶۵۹۱۵ tarjomac@gmail.com



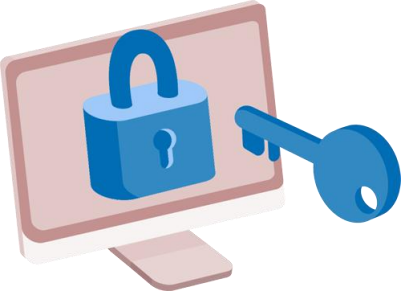
درباره نویسندگان

دکتر سیف الدین صالح الریالات یک پزشک در گروه چشم پزشکی در دانشگاه اردن در امان است. او یکی از محققان جوان مشهور در اردن است که در چندین رشته تحقیقاتی زیست پزشکی و عمومی مقالاتی منتشر کرده است. دکتر الریالات سخنران اصلی چندین کنفرانس بین المللی زیست پزشکی بود. او چندین دوره تحقیقاتی در زمینه تحقیقات زیست پزشکی برای دانشجویان پزشکی، دستیاران و متخصصان عالی ارائه کرد که در پایان دوره انتظار می رود دانشجویان مطالعات خود را با استفاده از داده های دسترسی آزاد منتشر کنند. او همراه با انتشارات فراوانش، اولین دارنده اختراع اردنی در معاهده بین المللی شرکت ثبت اختراع (PCT) است. (https://orcid. org/0000-0001-5588-9458)

پروفسور شاهر مومانی یک محقق برجسته علمی در دانشگاه اردن در امان است و بیش از ۲۵۰ مقاله با داوری همتا در مجلات بین المللی موسسه اطلاعات علمی (ISI) با کیفیت بالا تالیف یا همکاری کرده است. مومانی در سال های ۲۰۱٤، ۲۰۱۵ و ۲۰۱٦ توسط تامسون رویترز به فهرست معتبر محققان با استناد بالا منصوب شد. او تنها دانشمندی در جهان عرب است که برای سه سال متوالی برای این افتخار معتبر انتخاب شده است. بر اساس پایگاه داده اسکوپوس و گوگل اسکالر، مومانی بالاترین شاخص هیرش (h-index) و تعداد استنادات را در اردن به دست آورده است.

مومانی علاوه بر نشان عالی رتبه دوم ملک عبدالله دوم ابن الحسین، افتخارات و جوایز زیادی از جمله جایزه محقق ممتاز علی مانگو در اردن (۲۰۱٦)، جایزه پژوهشگر ممتاز در اردن (۲۰۱۲)، جایزه پژوهشگر ممتاز علمی در دانشگاه اردن، جایزه علمی آموزش و پرورش اسلامی اردن (۲۰۱۲) جایزه (۲۰۰۸)، جایزه اسکوپوس برای دانشمندان اردن (۲۰۰۹)، جایزه پژوهشگر ممتاز در دانشگاه موتا (۲۰۰۹)، جوایز TWAS برای دانشمندان جوان در کشورهای در حال توسعه، علوم دانشگاهی جهان سوم (۲۰۰۰)، و جایزه کمیسیون ملی علوم اردن برای آموزش، فرهنگ، و ۲۰۰۸. مومانی در سال ۲۰۱٦ توسط بسیاری از محققان و مؤسسات در سراسر جهان عرب نامزد دریافت جایزه نوبل فیزیک شد. (https://orcid.org/0000-0002-6326-8456)

لنا دبلیو ملکاوی یک پزشک در گروه رادیولوژی و پزشکی هسته ای در دانشگاه اردن در امان است. دکتر ملکاوی یک پیشگام در زمینه تحقیقات اخلاقی در اردن است که در آنجا مشاوره هایی در زمینه اخلاق زیست پزشکی در تحقیق و عمل ارائه می دهد. او دارای چندین مطبوعات منتشره در زمینه تحقیقات زیست پزشکی و داده های باز است.



مقدمه

پیشرفت های چشمگیر و اکتشافات بی وقفه، زیست پزشکی را به طور فزاینده ای به زندگی روزمره مرتبط کرده است. همگام شدن با همه این پیشرفت ها، حتی برای محققان فعال، یک کار دلهره آور است. تقاضای آشکاری برای بررسی‌های مختصر و خلاصه‌های ترکیبی از موضوعات زیست‌پزشکی برای دانشجویان تحصیلات تکمیلی، دانشجویان کارشناسی، اساتید، محققان زیست‌پزشکی، متخصصان پزشکی، سیاست‌گذاران علم و عموم مردم وجود دارد. با درک این نیاز مبرم، شرکت سی آر سی پرس مجموعه‌های راهنمای جیبی علوم بیومدیکال را با هدف اصلی ارائه بررسی‌های پیشرفته و معتبر از موضوعات مختلف با استفاده از قالب‌های کوتاه و خوانا در نظر گرفته شده برای مخاطبان گسترده ایجاد کرد. مجلدات این مجموعه به اصول و مفاهیم علوم طبیعی و هنرهای آزاد، به ویژه آنهایی که به پزشکی زیستی و رفاه انسان مربوط می شوند، می پردازد و ادغام می کند. مجلدات آینده از بیوشیمی، اخلاق زیستی، زیست سلولی، ژنتیک، ایمونولوژی، میکروبیولوژی، زیست شناسی مولکولی، علوم اعصاب، انکولوژی، انگل شناسی، آسیب شناسی و ویروس شناسی و سایر رشته های مرتبط خواهد آمد.

جلد فعلی به استفاده از داده های دسترسی باز به عنوان منبعی برای بورسیه های پیشرفته، به ویژه برای تحقیقات زیست پزشکی می پردازد. در طول سه دهه گذشته، از زمان شکل‌گیری اینترنت در سال ۱۹۹۱، تعداد زیادی از سازمان‌های دولتی، مؤسسات تحقیقاتی و مؤسسات دانشگاهی در سراسر جهان پایگاه‌های اطلاعاتی در زمینه‌های مختلف علم و فناوری ایجاد کرده‌اند. امروزه بسیاری از این پایگاه‌های اطلاعاتی از طریق پورتال‌ها و وب‌سایت‌های منفرد این سازمان‌ها به‌طور رایگان در شبکه جهانی وب در دسترس هستند. داده‌های دسترسی آزاد، مانند فرادادهای کارآزمایی بالینی، ژنوم، پروتئوم، میکروبیوم و متابولوم از بسیاری از مخازن، فرصت‌های روزافزونی را برای استفاده از این داده‌ها برای تحقیق فراهم می‌کنند. این

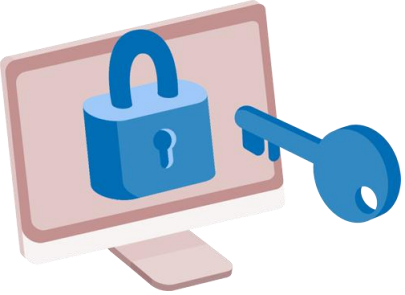
مجلد مختصر راهنمایی عملی برای نحوه استفاده از داده‌های دسترسی آزاد در تحقیقات، از تولید ایده‌های پژوهشی تا انتشار در مجلات ارجاع‌شده، ارائه می‌کند. هم محققان جوان و هم محققین با سابقه می توانند از این کتاب برای ارتقای مهارت های خود در زمینه منابع داده، دسترسی به داده های باز، ملاحظات اخلاقی، تجزیه و تحلیل داده ها برای رسیدن به نتیجه، نوشتن دست نوشته ها، انتشار و حتی تبلیغات پس از انتشار استفاده کنند.

کتاب حاضر نیز از هدف مجموعه تبعیت می کند: ساده کردن، خلاصه کردن و ترکیب یک موضوع پیچیده به طوری که خوانندگان بتوانند بدون نیاز به انجام جستجوهای وقت گیر پیشینه موضوع خود به اصل موضوع برسند. ما از پیشنهادات و توصیه‌های خوانندگان و اعضای جامعه زیست‌پزشکی برای موضوعات آینده در این مجموعه استقبال می‌کنیم و از دیگر کارشناسان می‌خواهیم که به‌عنوان نویسندگان/ویراستاران جلد آینده خدمت کنند.

لیجوان یوان[1]

بلکسبورگ، ویرجینیا، ایالات متحده آمریکا

[1] Lijuan Yuan



پیشگفتار

تحقیقات پزشکی چهره مراقبت های بهداشتی را تغییر داده است، امید به زندگی را حدود ۲۰ سال افزایش داده، منجر به ریشه کنی آبله شده و سرطان را به یک بیماری قابل درمان تبدیل کرده است. از سوی دیگر، افرادی که چنین تحقیقات تغییر دهنده زندگی را انجام دادند، قبل از آن در چندین مطالعه شرکت کرده بودند. نکته این است که قبل از درخواست بودجه برای مطالعه احتمالاً پیشگامانه خود، باید مراحل انجام تحقیق را بدانید. اما چگونه می توان بدون پشتوانه مالی، یا گاهی بدون راهنمایی مناسب، تحقیق کرد، که در اکثر کشورهای در حال توسعه چنین است؟ کتاب حاضر با هدف پاسخ به این سوال از طریق راهنمایی محققان تازه کار در مورد چگونگی استفاده از داده های دسترسی باز در انجام تحقیقات خود، نوشته شده است. **داده‌های باز**[1] داده‌هایی هستند که دیگران می‌توانند آزادانه از آنها استفاده نمایند و در حوزه کاری ما، استفاده از آن برای اهداف پژوهشی است.

این کتاب با فصل ۱ آغاز شده و به بازنگری داده های باز، اهمیت آنها برای محققان و نمونه هایی از نحوه استفاده از آنها در تحقیق می پردازد. پس از آن، محققان را در مورد منابع داده باز موجود آگاه می‌کند و دو نوع اصلی را تعریف می‌کند: داده‌های باز در دسترس صریح، که می‌توانند مستقیماً توسط محققانی که مایل به تجزیه و تحلیل آن هستند دانلود شوند و داده‌های باز در دسترس تلویحی، که محققین را ملزم به انجام یک یا چند مرحله پردازش برای دریافت اطلاعات می‌کند. پس از به دست آوردن داده ها، اکنون باید به آن رسیدگی کرده و ایده های تحقیقاتی ایجاد کنید. یک گام مهم در تولید ایده و انجام تحقیقات، مرور پیشینه است و این بحث نکاتی را در مورد کارآمدترین روش انجام این کار ارائه می‌کند. اکنون که تصمیم

[1] Open Data

گرفتید به یک مجموعه داده خاص علاقمند هستید، زمان دسترسی به آن فرا رسیده است و این فصل کل فرآیند دسترسی به داده های باز را توضیح می دهد. در نهایت، ملاحظات اخلاقی وجود دارد که باید در طول تحقیق خود در مورد داده های باز به خاطر داشته باشید. در پایان این فصل، آموزش نحوه ایجاد یک پروژه تحقیقاتی و دسترسی به داده های باز که برای اجرای آن نیاز دارید ارائه شده است، که در آن یا از مثال ارائه شده پیروی می کنید یا در طول این کتاب نمونه خود را ایجاد می‌کنید.

فصل دوم با ارائه مروری بر تحقیقات و فواید انجام آن، علاوه بر اصلاح برخی باورهای غلط مربوطه، آغاز می شود. انجام صحیح تحقیق قبل از اینکه بتوانید به ایجاد یک مقاله یا دستنوشته فکر کنید نیاز به برنامه ریزی خوبی دارد، بنابراین ما اجزای مراحل برنامه ریزی و اجرا را در طول یک مطالعه تحقیقاتی توضیح می دهیم. دانستن نوع مطالعه‌ای که انجام می دهید، خروجی با کیفیت بالا را تضمین می کند، و این فصل انواع اصلی تحقیق و یک نمای کلی برای هر نوع را پوشش می دهد. بعلاوه، این بحث یک نمای کلی آینده نگر از چرخه زندگی یک مطالعه ارائه می دهد، به طوری که شما برای مراحل بعدی تحقیق خود آماده شوید. در نهایت، تحقیق باید برای رعایت استانداردهای اخلاقی بالا، موضوعی که در این فصل به تفصیل به آن پرداخته شده است، انجام شود. انجام تحقیق با استفاده از داده های باز کمی متفاوت از تولید داده های نو است، بنابراین، در این فصل، بر این تفاوت ها تاکید می کنیم. در پایان این فصل، آموزشی وجود دارد که در آن پروتکل ایده حاصله از فصل ۱ را می نویسیم. به شما توصیه می کنیم یک پروتکل برای داده هایی ایجاد کنید که باید در آن فصل به آنها دسترسی داشته باشید.

در فصل ۳، بینشی در مورد یک جنبه مهم تحقیق به دست خواهید آورد: تجزیه و تحلیل آماری. ما با مروری بر اصول اولیه تجزیه و تحلیل آماری شروع می کنیم، وظیفه ای که برای محققان ضروری است، هم برای تفسیر مقالاتی که می خوانید و هم برای تجزیه و تحلیل داده های خود. پس از آن، ما یک نمودار ساده برای کمک به تجزیه و تحلیل داده های خود ارائه می دهیم که شما را در انتخاب آزمون های آماری که باید بر روی فرضیه خود به ساده ترین شکل ممکن انجام دهید، راهنمایی می کند. اگرچه فرمول های آزمون و تجزیه و تحلیل های آماری پیچیده خارج از حوصله این کتاب است، اگر مایلید در مورد این موضوع بیشتر بدانید،

فهرستی از منابع را در انتهای فصل ارائه می دهیم. برای انجام اولین مطالعه تحقیقاتی خود نیازی به آمارشناس زیستی ندارید، اما توصیه می کنیم پس از اتمام تجزیه و تحلیل خود با یک آمارشناس مجرب مشورت کنید.

در نهایت، در فصل ٤، مراحل نگارش دست‌نوشته خود را طبق جدیدترین دستورالعمل‌های گزارش‌دهی به شما آموزش می‌دهیم، جایی که ما بر روی نکات نوشتن دست‌نوشته مخصوص تحقیقات داده‌های باز تمرکز می‌کنیم. هنگام نوشتن دست‌نوشته خود، توصیه می‌کنیم بطور گام به گام به همان ترتیبی که در این کتاب ارائه شده است پیش بروید. پس از تکمیل دستنوشته خود، احتمالاً می خواهید آن را منتشر کنید. بخش پایانی این فصل اطلاعاتی را ارائه می دهد که باید در یک مجله قوی منتشر کنید که به بهترین وجه با محدوده مطالعه شما مطابقت دارد. همانطور که استاد مومانی فرمودند: «یک مطالعه عالی که در یک مجله ضعیف منتشر شود به یک مطالعه ضعیف تبدیل می شود و یک مطالعه ضعیف منتشر شده در یک مجله قوی به یک مطالعه قوی تبدیل می شود.»

دکتر کریستین داربی[1] کتاب را به طور کامل بررسی کرد و نظرات بسیار ارزشمندی را برای بهبود محتوای آن در اختیار ما قرار داد. طراحی فیگورهای این کتاب توسط حسین الریالات[2] انجام شده است.

سیف الدین صالح الریالات[3]

شهر مومانی[4]

[1] Dr. Christine Darby
[2] Hussein Al-Ryalat
[3] Saif Aldeen S. AlRyalat
[4] Shaher M. Momani



محتوا در یک نگاه

عنوان **صفحه**

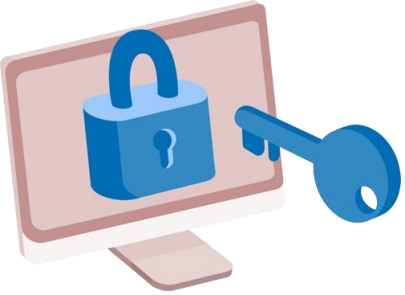
فصل ۱- داده باز ۲

فصل ۲- مبانی زمینه تحقیق ۳۳

فصل ۳- آماری که نیاز دارید ۶۶

فصل ۴- از پروژه پژوهشی تا مقاله ۱۰۳

واژه نامه ۱۳۲



فهرست مندرجات کتاب

مقدمه ..... ج

پیشگفتار ..... ه

فصل ۱- داده باز ۲

بازنگری داده ۲

داده های دردسترس ۷

ایده هایی از داده باز ۱۴

بازنگری (خواندن عمیق) ۱۸

دسترسی به داده های باز ۲۲

ملاحظات اخلاقی در مورد داده های باز ۲۳

آموزش کار با داده ۲۸

منابع فصل ۳۰

فصل ۲- مبانی زمینه تحقیق ۳۳

بازنگری تحقیق ۳۳

اجزای یک مطالعه ۳۷

انواع مطالعه ۴۰

چرخه اجرای یک مطالعه ۵۰

مبانی اخلاقی ۵۳

آموزش پروتکل ۶۰

منابع فصل ۶۳

فصل ۳- آماری که نیاز دارید ۶۶

دانش آماری ضروری ۶۶

تحلیل های آماری ۷۵

آموزش فصل ۹۴

منابع فصل ۱۰۱

فصل ۴- از پروژه پژوهشی تا مقاله ۱۰۳

نوشتن دست نوشته ۱۰۳

انتشار دست نوشته ۱۱۵

آموزش ۱۲۶

منابع فصل ۱۳۰

واژه نامه ۱۳۲



داده باز

سیف الدین صالح الریالات[1]

بازنگری داده

در سال ۲۰۱۳، کل جهان ۱.٤۸ تریلیون دلار برای تحقیق هزینه کرد و ایالات متحده و چین به تنهایی حدود ۱ تریلیون دلار هزینه کرده اند (ISSC, IDS, and UNESCO, 2016). از سوی دیگر، هزینه‌های تحقیقاتی کشورهای کم درآمد و متوسط تا کم درآمد جهان بسیار کم است، ظاهراً به این دلیل که اولویت‌های دیگری برای صرف پول وجود دارد (World Bank, 2018). این فقدان بودجه تحقیقاتی کافی منجر به این باور نادرست در بین محققان و دانشگاهیان جهان در حال توسعه شده که انجام تحقیق در این مناطق دشوار است (Horton, 2000). بعلاوه، نهادهای تأمین مالی معمولاً از محققانی که درخواست پول می کنند تجربه قبلی طلب می کنند و حتی اگر محققان بتوانند بودجه دریافت کنند، برای یافتن کارکنان تحصیل کرده برای حمایت از کار خود با مشکل مواجه خواهند شد. بنابراین، یک دور باطل آغاز می شود که محققان نمی توانند مطالعاتی انجام دهند زیرا نمی توانند بودجه دریافت کنند، که نیاز به تجربه تحقیقاتی دارد. نمونه‌ای از این وضعیت در آفریقا اتفاق می‌افتد، جایی که اکثر دانشگاه‌ها حتی برای اهداف آکادمیک، تاکید یا بودجه کافی برای تحقیق ندارند (Horton, 2000)، که منجر به کمبود تجربه پژوهشی کارکنان آنها می‌شود (Vernon et al., 2018).

1 Saif Aldeen Saleh AlRyalat

جمع‌آوری داده‌های با کیفیت بالا برای هر مطالعه مرحله ای است که به بیشترین پول و تجربه و همچنین زمان نیاز دارد (Dicks et al., 2014). این اصول جمع آوری داده ها به آسانی در کشورهای در حال توسعه در دسترس نیست، بنابراین مانع پیشرفت تحقیقات در این کشورها (دور باطل فوق الذکر) می شود. برای کاهش هزینه‌های هنگفت تحقیق در کشورهای در حال توسعه، محققان و دانشگاه‌ها باید به داده‌های جمع‌آوری‌شده توسط سایر محققان با تجربه‌تر که بودجه کافی برای حمایت از کار خود دارند، متکی باشند. این مدل از تحقیق که مبتنی بر استفاده از داده های دیگران برای انجام تحقیق است، تحت عنوان تحقیق داده های دسترسی باز (پژوهش داده باز)[1] شناخته می‌شود.

تحقیقات داده باز اخیراً به‌عنوان مدل جدیدی از تحقیق ظهور کرده است، مدلی که در گذشته به دلیل کمبود منابع، فناوری سیستم‌ها برای دسترسی به داده‌ها، تخصص بیوانفورماتیک و زیرساخت‌های قانونی برای تسهیل اشتراک‌گذاری؛ امکان‌پذیر نبود (Bertagnolli et al., 2017). از سوی دیگر، انتظار می‌رود که رویکرد داده‌های باز برای دستیابی به پتانسیل خود به پشتیبانی خوبی نیاز داشته باشد. برای نمایش هرچه بیشتر اهمیت ترویج و پشتیبانی از داده‌های باز، در سال ۲۰۰۳ دانشگاه روچستر، یکی از دانشگاه‌های با رتبه بالا در نیویورک، یک آرشیو دیجیتالی (یعنی مخزن) راه‌اندازی کرد که برای به اشتراک گذاشتن پایان‌نامه‌ها، پیش‌چاپ، مقاله‌های کاری، عکس‌ها، موسیقی‌ها و هر نوع دیگری از داده‌های دیجیتالی که سرمایه‌گذاران دانشگاه می‌توانند تولید کنند، طراحی کرد.

شش ماه تحقیق دانشگاه و محققین آن را متقاعد کرد که بایگانی آنلاین در دسترس عموم به خوبی مورد استقبال قرار گرفته است. در زمان راه‌اندازی مخزن، کتابداران دانشگاه نگران بودند که سیل داده‌هایی که آپلود می‌شد ممکن است فضای ذخیره‌سازی موجود را تحت تأثیر قرار دهد. شش سال بعد، مخزن ۲۰۰۰۰۰ دلاری عمدتاً خالی بود (Nelson, 2009).

برای ترویج تحقیقات داده باز، یک جایزه سالانه برای به رسمیت شناختن اشتراک گذاران داده‌ای که داده‌هایشان به روش‌های تاثیرگذار استفاده مجدد شود، ایجاد شد: جایزه Research Symbiont (www.researchsymbionts.com). همچنین فراخوانی برای مجلات برای ترویج تحقیقات

[1] Open-Access Data Research (Open Data Research)

داده‌های باز با انتشار تحقیقات انجام شده با استفاده از داده‌های باز، به‌ویژه مجلاتی که قبلاً مطبوعات اولیه را منتشر کرده‌اند، وجود داشت (Byrd, 2017). علاوه بر این، مطالعه ای نشان داد که نرخ استناد در میان مقالاتی که داده های آنها به طور آشکار در دسترس است، بالاتر است و آنهایی که دریافتند در دسترس قرار دادن داده ها به صورت آشکار، نرخ استناد به مقاله را افزایش می دهد، باعث انگیزه بیشتر و افزایش تحقیقات داده های باز گردید (Piwowar et al., 2007).

داده های باز ممکن است آزادانه توسط هر کسی استفاده، استفاده مجدد و توزیع مجدد قرار گیرد (Dietrich et al., 2009). داده های باز دارای سه ویژگی اصلی هستند:

- ☑ **قابلیت دسترسی**: داده ها باید به طور کلی در دسترس باشند و نه بیشتر از هزینه تولید مجدد معقول، ترجیحاً با دانلود از طریق اینترنت. داده ها همچنین باید به روشی راحت و قابل تغییر، ترجیحاً به شکل یادگیری ماشینی (به عنوان مثال، صفحه گسترده مایکروسافت اکسل) در دسترس باشند.
- ☑ **استفاده مجدد و توزیع مجدد**: داده ها باید تحت شرایطی ارائه شوند که اجازه استفاده مجدد و توزیع مجدد آنها را بدهد، از جمله تلفیق با سایر مجموعه داده ها (به عنوان مثال، ترکیب دو مجموعه داده در یک تم برای ایجاد یک مجموعه بزرگتر).
- ☑ **مشارکت همگانی**: همه باید بتوانند از داده ها استفاده، استفاده مجدد و توزیع مجدد کنند.

مفهوم داده های باز بیشتر مربوط به حوزه اقتصاد است، جایی که دولت ها و شرکت ها می توانند از داده های باز برای بهبود عملکرد و محصولات خود استفاده کنند. یک مثال می تواند تولیدکنندگانی باشد که از داده های به دست آمده از محصولات خود در حال حاضر در بازار برای بهبود توسعه محصولات بعدی خود و ایجاد پیشنهادات پس از فروش نوآورانه استفاده می کنند (Manyika et al., 2011). دانشمندانی که از داده‌های مصرف‌کنندگان مراقبت‌های بهداشتی جمع‌آوری‌شده توسط شرکت‌های داروسازی، پرونده های مؤسسات بهداشتی، شرکت‌های بیمه و سایر مقامات محلی و ملی برای انجام تحقیقات استفاده می‌کنند، همین مفهوم را می‌توانند بکار بگیرند. مثال زیر بیشتر روشن می کند که چگونه محققان می توانند از داده های باز بهره ببرند.

در سال ۲۰۱۰، پروتکل یک مطالعه بزرگ به نام کارآزمایی مداخله فشار خون سیستولیک (SPRINT) نهایی شد و بیان داشت که شامل ۹۳۶۱ بیمار از ۱۰۲ مرکز در ایالات متحده بوده و آنها به مدت هشت سال پیگیری شده اند (Ambrosius et al., 2014). هزینه کل این مطالعه منفرد تقریباً ۱۵۷ میلیون دلار (NHLBI, n.d.) برآورد شد (بیش از بودجه تحقیقاتی موجود برای چندین کشور در مجموع). هدف SPRINT مقایسه نتایج کاهش فشار خون سیستولیک استاندارد در مقابل مفرط (SBP) بود. ٤٦٧٨ بیمار را برای کاهش شدید SBP (SBP < 120 میلی‌متر جیوه) و ٤٦٨٣ بیمار دیگر را برای کاهش استاندارد SBP (SBP <140 میلی‌متر جیوه) انتخاب کرد. در پایان مطالعه، چهار ماه بعد، بیش از ۱٤۰ مقاله جدید با استفاده از این مجموعه داده توسط محققان سراسر جهان، از جمله کشورهای در حال توسعه، بدون هزینه اضافی نوشته شد.

مدل داده باز برای دستیابی به حداکثر سود حاصل از تحقیقات پیشنهاد شد. به جای صرف میلیون‌ها دلار برای انتشار تنها یک مطالعه، داده‌های باز اجازه می‌دهند تا یافته‌های جدید و انتشارات جدید از آن مجموعه داده‌ها استخراج شود. نوآوری های محققان ممکن است در سایر نقاط جهان، از جمله کشورهای در حال توسعه، اطلاع رسانی و استفاده شود. به همین دلیل، در مارس ۲۰۱۷، مجله پزشکی نیوانگلند (NEJM)[1] چالشی را برای محققان برای انجام یک مطالعه جدید از داده‌های جمع‌آوری شده از مطالعه SPRINT آغاز کرد (Burns and Miller, 2017). برنده مسابقه NEJM یک ماشین حساب ریسک–منفعت وزنی برای بررسی مزایا و معایب درمان مفرط بیماران مبتلا به فشار خون ایجاد کرد. نفر دوم، متشکل از دانشجویان پزشکی، تنها برای بخشی از جمعیت درگیر در کارآزمایی SPRINT تجزیه و تحلیل انجام داد و به این نتیجه رسید که درمان فشرده در بیمارانی که بیماری مزمن کلیوی داشتند، آسیب بیشتری ایجاد می کند و دشوارتر است. برنده رتبه سوم، یک ماشین حساب ریسک برای کمک به تصمیم گیری پزشکانی که بیماران مبتلا به فشار خون بالا را درمان می کنند، ایجاد کرد (Burns and Miller, 2017). اگر بیش از ۱٤۰ مطالعه انجام شده با استفاده از داده‌های SPRINT نیاز به جمع‌آوری داده‌های تازه مختص خود داشتند، با در نظر گرفتن این که هزینه متوسط دستیابی به داده‌های با کیفیت بالا در مورد فشار خون بالا حدود ۱ میلیون دلار خواهد بود، هزینه جمع آوری داده های جدید بیش از ۱٤۰ میلیون دلار می شد (درگاه تحقیقات و نوآوری انگلستان (https://gtr.ukri.org)).

[1] New England Journal of Medicine

نمونه دیگری از مدل داده باز، مطالعه قلب فرامینگهام[1] (Dawber et al., 1951) بود، یک مطالعه کوهورت که در سال ۱۹٤۸ تحت مدیریت موسسه ملی قلب، ریه و خون (NHLBI)[2] با هدف شناسایی عوامل یا ویژگی های مشترکی که در بیماری قلبی عروقی نقش دارند، آغاز شد. قبل از این مطالعه، تقریباً هیچ چیز در مورد اپیدمیولوژی بیماری قلبی عروقی، فشار خون بالا یا تصلب شرائین شناخته نشده بود. بسیاری از دانش رایج در مورد بیماری قلبی، مانند اثرات بالقوه رژیم غذایی، ورزش و داروهای رایج مانند آسپرین بر خطرات قلبی عروقی، بر اساس این مطالعه طولی است. داده های این مطالعه در حال انجام برای محققان آینده در دسترس است. اکنون، بیش از ۱۰۰۰ مطالعه پزشکی مرتبط با مطالعه قلب فرامینگهام، بدون اینکه نویسندگان آنها هیچ گونه هزینه ای برای داده کرده باشند، منتشر شده است. این مطالعات میلیون ها دلار هزینه داشت و برخی از آنها توسط نویسندگانی از کشورهای در حال توسعه و محققان اولیه شغلی انجام شد که اگر مجبور به تولید داده های خود بودند، بودجه کافی برای مطالعات به دست نمی آوردند.

این دو مثال نشان می‌دهند که چگونه استفاده از داده‌های باز گاهی بهترین راه برای انجام تحقیق است. مدل استفاده از داده های باز برای انجام تحقیق، گزینه های عالی است اگر خود را در یکی از سه صحنه زیر بیابید:

۱. ایده ای در مورد آنچه می خواهید مطالعه کنید دارید، اما منابع لازم را برای انجام مطالعه در موسسه خود ندارید. علاوه بر این، یک مطالعه داده باز بزرگ قبلاً در مورد همین موضوع در گذشته انجام شده است. به عنوان مثال، شما دوست دارید بیماری آلزایمر را مطالعه کنید اما منابع لازم برای انجام آن را ندارید. هرچند شما به یک مطالعه داده باز درباره آلزایمر به نام ابتکار تصویربرداری عصبی بیماری آلزایمر (ADNI)[3] دسترسی دارید.

۲. ایده خاصی ندارید، اما به طور کلی علاقمند به انجام تحقیقات در یک تخصص خاص هستید (مثلاً مغز و اعصاب). در این مورد، می‌توانید به دنبال منابع داده باز در زمینه مغز و اعصاب بگردید و به منابعی دسترسی پیدا کنید که فکر می‌کنید بیشتر با علایق شما مطابقت دارد. شما ایده خود را از آن منبع تولید خواهید کرد.

[1] Framingham Heart Study
[2] National Heart, Lung, and Blood Institute
[3] Alzheimer's Disease Neuroimaging Initiative

۳. می خواهید محققان جدیدی را تربیت کنید که قبلاً مقاله ای منتشر نکرده اند. به جای صرف منابع برای جمع آوری داده ها برای اهداف آموزشی، می توانید به سادگی از مخازن داده های باز برای آموزش آنها با حداقل هزینه استفاده کنید.

داده های دردسترس

به طور کلی دو نوع اصلی داده باز وجود دارد:

- به صراحت در دسترس است (به عنوان مثال، آماده در قالب صفحات گسترده قابل دانلود اکسل).
- به طور ضمنی در دسترس است (به عنوان مثال، لزوم یک مرحله اضافی برای قرار دادن داده‌ها در قالب صفحات گسترده اکسل).

داده های باز با دسترسی صریح

داده‌های باز صریحاً در دسترس قابل دسترسی و استفاده هستند، جایی که می‌توانید مستقیماً آن‌ها را در قالب صفحات گسترده مایکروسافت اکسل (یا هر نرم‌افزار موجود دیگری) دانلود کنید، یا باید از طریق مدل دروازه‌بان به آن‌ها دسترسی پیدا کنید. این نوع برخلاف داده‌های باز موجود به طور ضمنی است، جایی که شما باید قبل از رسیدن به فرم نهایی قابل تجزیه و تحلیل (به عنوان مثال، صفحه گسترده اکسل) مقداری پردازش انجام دهید.

این داده‌ها معمولاً از کشورها، مؤسسات یا مخازن در دسترس عموم برای استفاده فردی ارائه می‌شوند. نمونه ای از پایگاه داده در سطح کشور، مخزن داده باز ایالات متحده (www.data.gov) است، که در آن دولت به بیش از ۲۵۰۰۰۰ مجموعه داده با ارزش بالا در مورد تمام اطلاعات دولتی دسترسی دارد. در سطح موسسات، دانشگاه ییل پروژه YODA (http://yoda.yale.edu) را راه اندازی کرد تا داده‌های تولید شده از مطالعات بالینی را برای سایر محققان ارائه دهد. در سطح فردی، Figshare (www.figshare.com) به کاربران اجازه می‌دهد فایل‌ها را با هر فرمتی آپلود کنند تا در مرورگر مخزن پیش‌نمایش شوند.

اگرچه داده ها به طور گسترده در دسترس هستند، اما با توجه به کیفیت یا روش جمع آوری آنها متفاوت هستند. همیشه باید قبل از استفاده از داده ها، کیفیت آنها را بررسی کنید، که می توانید با استفاده از نکات زیر ارزیابی کنید:

۱. **کیفیت خود مخزن**: به عنوان مثال، داده های سپرده شده در مخازن مؤسسه ملی بهداشت (NIH) قبلاً از نظر کیفیت بررسی شده اند.
۲. **تحقیق اولیه ای که با استفاده از داده ها منتشر شده است**: اگر پژوهش اولیه در یک مجله با کیفیت بالا منتشر شود، داده های آن در مقایسه با تحقیقات منتشر شده در مجله های با کیفیت پایین‌تر از کیفیت بالاتری برخوردار است (توجه: برای راهنمایی در مورد بررسی کیفیت یک مجله، به بخش ٤-٢ در فصل ٤ مراجعه کنید.)
۳. **پروتکلی که جزئیات نحوه جمع‌آوری داده‌ها را نشان می‌دهد**: معمولاً با داده‌های اصلی ذخیره می‌شود.

از آنجایی که اولین گام برای انتخاب داده ها بررسی کیفیت مخزن است، ما در مورد سه مخزن داده با کیفیت بالا صحبت خواهیم کرد که معتقدیم نقطه شروع شما برای انتخاب داده ها برای استناد به تحقیق شما خواهد بود. اول؛ NIH مستقر در ایالات متحده با بودجه سالانه بیش از ۳۰ میلیارد دلار، بزرگترین سرمایه گذار تحقیقات در زمینه بهداشت و درمان و علوم زیستی در جهان است (Li et al., 2017). تقریباً تمام داده‌های پروژه‌های تحقیقاتی تأمین‌شده توسط NIH با کیفیت بالا هستند و باید داده‌ها را در یک مخزن پشتیبانی شده توسط NIH (سیاست‌های اشتراک‌گذاری داده NIH) سپرده گذاری کنند. کمیته هماهنگی انفورماتیک بیومدیکال Trans-NIH (BMIC) مسئول هماهنگی بین مراکز NIH و ارائه داده ها از هر مرکز در یک صفحه است.

مخزن دوم FAIRsharing (www.FAIRsharing.org) است که یک پورتال داده قابل جستجو و مبتنی بر وب وابسته به دانشگاه آکسفورد است. هر مخزن لیست شده در این سایت برای اطمینان از کیفیت بررسی می شود.

مخزن سوم، re3data (https://www.re3data.org)، توسط یک سازمان غیرانتفاعی به نام Science Europe پشتیبانی می شود. این تضمین می‌کند که مخازن فهرست‌شده در آن دارای شناسه‌های دائمی، ابرداده، دسترسی و استفاده از داده‌ها، خوانایی ماشینی و نگهداری طولانی‌مدت هستند. مخازن را می‌توان بر اساس موضوع، کشور یا نوع محتوا در یک رابط کاربر پسند مرور کرد.