

عشر قواعد لبناء خطة فعّالة لإدارة بيانات البحث

ويليام ك. ميتشنر



المخلص:

يقدم المقال عشر قواعد بسيطة لمساعدة الباحثين على إعداد خطة لإدارة البيانات، وهي وثيقة تطلبها معظم الجهات الممولة ضمن مقترح البحث وتصف كيفية التعامل مع البيانات في أثناء المشروع وبعد انتهائه. ويربط المؤلف هذه القواعد بدورة حياة البحث ودورة حياة البيانات. وتشمل القواعد: تحديد متطلبات الجهة الراعية، وتعيين البيانات التي ستُجمع من حيث أنواعها ومصادرها وحجمها وصيغها، وتنظيم البيانات، وتوثيقها بالبيانات الوصفية والمعايير المجتمعية، وضمان جودتها، ووضع استراتيجية سليمة لتخزينها وحفظها، وتحديد سياسات الترخيص والمشاركة والقيود الأخلاقية والقانونية، ووصف طرق نشرها وإتاحتها، وتوزيع الأدوار والمسؤوليات، وإعداد ميزانية واقعية. ويخلص المقال إلى أن الخطة الجيدة «وثيقة حية» تُراجع دوريًا وتُعاد الاستفادة منها في المشاريع اللاحقة، وأن إتاحتها علنًا بترخيص مفتوح وبصيغة مقروءة آليًا يزيد أثرها.

الكلمات المفتاحية: خطة إدارة البيانات، إدارة بيانات البحث، دورة حياة البيانات، البيانات الوصفية، الحفظ الرقمي

مقدمة:

الأوراق البحثية ومنتجات البيانات هي المخرجات الرئيسية للمؤسسة العلمية. وتزداد الجهات الراعية للبحث، الحكومية وغير الحكومية والمؤسسات الخاصة، إدراكًا لقيمة بيانات البحث. ونتيجة لذلك، تشترط معظم الجهات الممولة اليوم تقديم خطط مفصلة لإدارة البيانات ضمن مقترح البحث. وخطة إدارة البيانات (data management plan - DMP) وثيقة تصف كيف ستعامل مع بياناتك في أثناء المشروع، وما الذي سيحدث للبيانات بعد انتهائه. وتغطي هذه الخطط عادةً دورة حياة البيانات كلها أو أجزاء منها - من اكتشاف البيانات (discovery) وجمعها وتنظيمها (مثل جداول البيانات وقواعد البيانات)، مرورًا بضمان الجودة وضبطها، والتوثيق (مثل أنواع البيانات والطرق المخبرية) واستخدام البيانات، وصولًا إلى حفظ البيانات ومشاركتها مع الآخرين (مثل سياسات البيانات

وطرق النشر). وترتبط هذه القواعد بدورتي حياة البحث والبيانات؛ إذ تساعد إدارة البيانات الجيدة على إنجاز مراحل البحث وتوثيق نتائجه. وتخضع خطة إدارة البيانات للتحكيم العلمي، وتستخدم جزئيًا في تقييم جدارة المشروع. كما توثق الخطط أنشطة إدارة البيانات المرتبطة بالمشاريع الممولة، وقد يرجع إليها في أثناء مراجعات الأداء.

في إطار دورة حياة البحث، يختبر كثير من الباحثين (1) أفكارًا وفرضيات عن طريق (2) اقتناء بيانات (3) تُدمج في تحليلات وتمثيلات مرئية متنوعة، فتفضي إلى تفسيرات (4) تُنشر في الأدبيات وتُعمَّم عبر آليات أخرى (مثل العروض في المؤتمرات والمدونات والتغريدات)، وكثيرًا ما تعود بهم إلى (1) أفكار وفرضيات جديدة. وفي أثناء دورة حياة البيانات، يقوم الباحثون عادةً بما يلي: (1) وضع خطة لكيفية إدارة البيانات في أثناء المشروع وبعده؛ و(2) اكتشاف البيانات الموجودة واقتنائها، و(3) جمع بيانات جديدة وتنظيمها؛ و(4) ضمان جودة البيانات؛ و(5) وصف البيانات (أي إسناد البيانات الوصفية إليها)؛ و(6) استخدام البيانات في التحليلات والنماذج والتمثيلات المرئية وغيرها؛ و(7) حفظ البيانات و(8) مشاركتها مع الآخرين (مثل الباحثين والطلاب وصانعي القرار)، بما قد يفضي إلى أفكار وفرضيات جديدة.

قدّمت مقالات سابقة في سلسلة «عشر قواعد بسيطة» (Ten Simple Rules) في مجلة (PLOS Computational Biology) إرشادات بشأن الحصول على المنح [1]، وكتابة الأوراق البحثية [2]، وعرض نتائج البحث [3]، والعناية بالبيانات العلمية [4]. وأعرض هنا عشر قواعد بسيطة يمكن أن تساعد في توجيه عملية إعداد خطة فعّالة لإدارة بيانات البحث - وهي الأساس الذي تقوم عليه نتائج المشروع وأوراقه البحثية ومنتجات بياناته. وأركز على المبادئ والممارسات التي ستُفضي إلى خطة لإدارة البيانات يسهل على الآخرين فهمها، ويستطيع فريقك البحثي وضعها موضع التطبيق. وفضلًا عن ذلك، سيساعد اتباع القواعد العشر البسيطة على ضمان أن تكون بياناتك آمنة وقابلة للمشاركة، وأن يحقق مشروعك أقصى عائد على استثمار الجهة الممولة.

القاعدة 1: حدّد متطلبات الجهة الراعية للبحث:

تطوّر المجتمعات البحثية عادةً أساليبها ومقارباتها المعيارية الخاصة لإدارة البيانات ونشرها. وبالمثل، كثيرًا ما تكون لدى الجهات الراعية للبحث توقعات محددة جدًا بشأن خطة إدارة البيانات. فعلى سبيل المثال، تموّل كلّ من مؤسسة ويلكم ترست (Wellcome Trust)، ومؤسسة غوردون وبيتي مور (Gordon and Betty Moore Foundation - GBMF)، والمعاهد الوطنية للصحة في الولايات المتحدة (NIH)، والمؤسسة الوطنية للعلوم في الولايات المتحدة (NSF) أبحاث علم الأحياء الحاسوبي، لكنها تختلف اختلافًا ملحوظًا في متطلباتها لخطة إدارة البيانات. فمؤسسة (GBMF) مثلًا تشترط أن يعدّ المرشحون للمنح، بالتعاون مع مسؤول البرنامج لديها، خطة شاملة لإدارة البيانات تجيب عن عشرات الأسئلة المحددة. وفي المقابل، متطلبات المعاهد الوطنية للصحة أقل تفصيلًا بكثير، وتطلب أساسًا من المرشحين للمنح أن يشرحوا كيف ستشارك البيانات، أو أن يقدموا أسبابًا لتعذر مشاركتها. وفوق ذلك، قد تكون لجهة راعية واحدة (مثل المؤسسة الوطنية للعلوم) متطلبات مختلفة تُحدّد لكل قسم وبرنامج داخل المؤسسة. ولاحظ أن متطلبات الخطة قد لا تُسمّى بهذا الاسم؛ فمثلًا، تركز إرشادات المعاهد الوطنية للصحة أساسًا على مشاركة البيانات، وترد في وثيقة عنوانها «NIH Data Sharing Policy and Implementation» (Guidance) «سياسة المعاهد الوطنية للصحة لمشاركة البيانات وإرشادات تنفيذها».

ويمكن توفير وقت وجهد كبيرين بالبداية بفهم المتطلبات التي تضعها الجهة التي تقدّم إليها المقترح. وعادةً ما تقدّم الجهات الراعية للبحث متطلبات خطة إدارة البيانات إما في الدعوة العامة لتقديم المقترحات (request for proposals - RFP)، وإما في دليل لمقترحات المنح عبر الإنترنت. كما أن موقعي (DMPTool) و(DMPonline)

موردان بالغا القيمة، يقدّمان متطلبات الخطط المحدّثة لدى الجهات الممولة (للولايات المتحدة والمملكة المتحدة على التوالي) في صورة قوالب تصحبها عادةً نصائح مشروحة لملء القالب. ويتضمن موقع (DMPTool) أيضًا خططًا نموذجية عديدة نشرها مستخدموه. وتعطي هذه الأمثلة فكرة عن عمق التفاصيل واتساعها مما يُدرج عادةً في الخطة، وكثيرًا ما توحى بأفكار جديدة يمكن إدراجها في خطتك.

وبصرف النظر عما إذا كنت قد قدّمت مقترحات سابقًا إلى برنامج تمويل بعينه، من المهم دائمًا الاطلاع على أحدث دعوة لتقديم المقترحات، وعلى موقع الجهة الراعية، للتحقق مما إذا كانت المتطلبات قد تغيّرت مؤخرًا وكيف تغيّرت. وفوق ذلك، لا تتردد في التواصل مع مسؤول البرنامج (أو مسؤوليه) المذكورين في دعوة بعينها لمناقشة متطلبات الجهة الراعية، أو لمعالجة أسئلة محددة تنشأ وأنت تعدّ خطة إدارة البيانات لمشروعك المقترح. وتذكر أن الهدف الأساسي ينبغي أن يكون إعداد خطة مفيدة لمشروعك. ولذلك يمكن لخطط إدارة البيانات الجيدة أن تتضمن، وكثيرًا ما تتضمن بالفعل، معلومات تفوق الحد الأدنى الذي تشترطه الجهة الراعية. ولاحظ مع ذلك أن بعض الجهات الراعية تقيد طول خطط إدارة البيانات (مثل حد أقصى قدره صفحتان)؛ وفي هذه الحالات يمكن تقديم خلاصة لخطة الأشمل، وقد يُسمح بإدراج ملحق أو ملف تكميلي أو رابط.

القاعدة 2: عيّن البيانات التي ستُجمع:

يعتمد كل مكّون من مكّونات خطة إدارة البيانات على معرفة كمية البيانات التي ستُجمع وأنواعها. وحجم البيانات مهم بوضوح، إذ إن إدارة 10 تيرابايت من البيانات تكلف عادةً أكثر من إدارة 10 ميغابايت من حيث البنية التحتية ووقت العاملين. لكن خصائص أخرى للبيانات تؤثر أيضًا في التكاليف، وكذلك في البيانات الوصفية، واستراتيجيات ضمان جودة البيانات وحفظها، بل وفي سياسات البيانات. وستتضمن الخطة الجيدة معلومات كافية لفهم طبيعة البيانات التي تُجمع، ومنها:

- الأنواع: من الخطوات الأولى الجيدة إعداد قائمة بأنواع البيانات المختلفة التي تتوقع جمعها أو إنشاءها. وقد يشمل ذلك النصوص، وجداول البيانات، والبرمجيات والخوارزميات، والنماذج، والصور والأفلام، والملفات الصوتية، وسجلات المرضى. ولاحظ أن كثيرًا من الجهات الراعية للبحث تعرّف البيانات تعريّفًا واسعًا يشمل المجموعات المادية، والبرمجيات والشفرات، والمواد المنهجية.
- المصادر: قد تأتي البيانات من الملاحظة البشرية المباشرة، والأجهزة المخبرية والميدانية، والتجارب، والمحاكاة، وتجميع بيانات من دراسات أخرى. وقد يهتم المحكّمون والجهات الراعية اهتمامًا خاصًا بمعرفة ما إذا كانت البيانات مملوكة ملكية خاصة، أو مجمعة من دراسات أخرى، أو تتعلق بمشاركين بشر، أو خاضعة بأي وجه آخر لقيود على استخدامها أو إعادة توزيعها.
- الحجم: يمكن أن يؤثر الحجم الإجمالي للبيانات والعدد الإجمالي للملفات المتوقعة جمعها في جميع أنشطة إدارة البيانات الأخرى.
- صيغ البيانات والملفات: تتغير التقنية، وقد تتقدم قريبًا صيغ المقبولة اليوم. ومن الخيارات الجيدة الصيغ غير الاحتكارية، القائمة على معايير مفتوحة، والمعتمدة على نطاق واسع والمفضّلة لدى المجتمع العلمي (مثل تفضيل صيغة القيم المفصولة بفواصل (CSV) على صيغة (Excel) (.xls, .xlsx)). ويُرَجَّح أن تظل البيانات متاحة على المدى البعيد إذا كانت غير مضغوطة وغير مشفرة، ومخزّنة باستخدام ترميزات محارف معيارية مثل (UTF-16).

وقد لا تكون الأنواع والمصادر والأحجام والصيغ الدقيقة للبيانات معروفة مسبقًا، بحسب طبيعة البحث وتفردته. وفي هذه الحالة يكون الحل تحديث الخطة على مراحل مع تطور المشروع (انظر القاعدة 9).

القاعدة 3: حدّد كيف ستُنظّم البيانات:

متى اتضح حجم البيانات التي ستُجمع وأنواعها، تكون الخطوة البديهية التالية تحديد كيفية تنظيم البيانات وإدارتها. ففي كثير من المشاريع سيُنتج عدد قليل من جداول البيانات يمكن إدارته بفاعلية باستخدام برامج جداول البيانات التجارية أو مفتوحة المصدر مثل (Excel) و (OpenOffice Calc). وقد تتطلب أحجام البيانات الأكبر وقيود الاستخدام استعمال أنظمة إدارة قواعد البيانات العلائقية (RDBMS) لجداول البيانات المترابطة، مثل (ORACLE) أو (MySQL)، أو نظام معلومات جغرافية (GIS) لطبقات البيانات الجغرافية المكانية، مثل (ArcGIS) أو (GRASS) أو (QGIS).

ويمكن أن تملأ التفاصيل المتعلقة بكيفية تنظيم البيانات وإدارتها صفحات كثيرة من النص، بل ينبغي في الواقع تدوينها مع تطوّر المشروع. غير أن من المفيد جدًّا عند صياغة خطة إدارة البيانات التركيز في البداية على أنواع المنتجات التي ستُستخدم، وربما أسمائها. وينبغي أن تكون الأدوات البرمجية المستخدمة في المشروع ملائمة للمهام المتوقعة. فبرنامج جداول البيانات مثلًا لن يكفي لمشروع يتوقع أن يُنتج تيرا بايتات من البيانات، وقد يكون نظام متطور لإدارة قواعد البيانات العلائقية مبالغة لا داعي لها في مشروع لن يُنشأ فيه سوى عدد قليل من جداول البيانات الصغيرة. وفوق ذلك، قد تنطوي المشاريع المعتمدة على نظام معلومات جغرافية أو نظام لإدارة قواعد البيانات العلائقية على تكاليف برمجية كبيرة وجهد في التصميم والبرمجة ينبغي التخطيط لهما ورصد الميزانية لهما مسبقًا (انظر القاعدتين 9 و 10). وبحسب متطلبات الجهة الراعية وقيود المساحة، قد يكون من المفيد أيضًا تحديد قواعد تسمية الملفات، والمعرفات الفريدة الدائمة (مثل معرفات الكائنات الرقمية (DOIs))، وضبط الإصدارات (لكلٍّ من البرمجيات ومنتجات البيانات).

القاعدة 4: اشرح كيف ستوثّق البيانات:

لا معنى يُذكر، أو لا معنى البتة، لصفوف وأعمدة من الأرقام والمخارف ما لم تُوثّق بطريقة ما. فالبيانات الوصفية - أي التفاصيل المتعلقة بماهية البيانات ومكان جمعها وزمانه وسببه وكيفية جمعها ومعالجتها وتفسيرها - توفر المعلومات التي تمكّن من اكتشاف البيانات والملفات واستخدامها والاستشهاد بها على النحو الصحيح. وتشمل البيانات الوصفية أوصافًا لكيفية تسمية البيانات والملفات وبنيتها المادية وتخزينها، إضافة إلى تفاصيل عن التجارب والطرق التحليلية وسباق البحث، والقاعدة العامة أن فائدة البيانات وطول عمرها يرتبطان ارتباطًا مباشرًا بمدى اكتمال البيانات الوصفية وشمولها. وقد يتفاوت مقدار الجهد المبذول في إنشاء بيانات وصفية شاملة متفاوتًا كبيرًا بحسب تعقّد البيانات وأنواعها وحجمها.

ويمكن أن تقوم استراتيجية توثيق سليمة على ثلاث خطوات. أولًا، حدّد أنواع المعلومات التي ينبغي تسجيلها لتمكين باحث مثلك من اكتشاف بياناتك والوصول إليها وتفسيرها واستخدامها والاستشهاد بها. ثانيًا، تبين ما إذا كان ثمة مخطط أو معيار للبيانات الوصفية معتمد لدى المجتمع البحثي (أي مجموعات مفضلة من عناصر البيانات الوصفية) يمكن اعتماده. ومن الأمثلة على ذلك أن صيغًا مختلفة من النموذج المجرد لمبادرة دبلن كور للبيانات الوصفية (Dublin Core Metadata Initiative Abstract Model) تُستخدم لأنواع كثيرة من البيانات وغيرها من الموارد، وأن المعيار (ISO) (المنظمة الدولية للتقييس) 19115 يُستخدم للبيانات الجغرافية المكانية، وأن صيغة الملفات (ISA-Tab) تُستخدم للبيانات الوصفية التجريبية، وأن لغة البيانات الوصفية البيئية (Ecological Metadata

(Language - EML) تُستخدم لأنواع كثيرة من البيانات البيئية. وفي حالات كثيرة يوصي مستودع بيانات أو أرشيف مستهدف، أو منظمة مهنية في المجال، بمعيار محدد لمحتوى البيانات الوصفية. ثالثاً، حدّد الأدوات البرمجية التي يمكن استخدامها لإنشاء محتوى البيانات الوصفية وإدارته (مثل (Metavist) و(Morpho)). وفي غياب أدوات قائمة، يمكن إدراج ملفات نصية (مثل (readme.txt)) تتضمن البيانات الوصفية ذات الصلة أو إدراج هذه البيانات في ترويسات ملفات البيانات.

ومن أفضل الممارسات تكليف شخص مسؤول بتعهّد دفتر مخبري إلكتروني تُحفظ فيه جميع تفاصيل المشروع. ويُستحسن أن يراجع عضو آخر في الفريق هذا الدفتر وينقّحه بانتظام، وأن تُنسخ منه نسخ احتياطية (انظر القاعدتين 6 و9). وتشكّل البيانات الوصفية المسجلة في الدفتر الأساس للبيانات الوصفية التي سَقرن بمنتجات البيانات المزمع تخزينها وإعادة استخدامها ومشاركتها.

القاعدة 5: صف كيف ستضمن جودة البيانات:

يشير ضمان الجودة وضبطها (QA/QC) إلى العمليات المستخدمة لقياس جودة المنتجات (مثل البيانات والبرمجيات وغيرها) وتقييمها وتحسينها. وقد يلزم اتباع إرشادات محددة لضمان الجودة وضبطها بحسب طبيعة الدراسة والجهة الراعية للبحث؛ وهذه المتطلبات، إن وُجدت، تُذكر عادةً في الدعوة لتقديم المقترحات. وعلى أي حال، من الممارسات الجيدة أن تصف تدابير ضمان الجودة وضبطها التي تعتمزم استخدامها في مشروعك. وقد تشمل هذه التدابير أنشطة التدريب، ومعايرة الأجهزة واختبارات التحقق، وإدخال البيانات المزدوج المعقّى، والمقاربات الإحصائية والمربئية لاكتشاف الأخطاء. ويمكن أن تكون للمقاربات البسيطة للاستكشاف البياني للبيانات (مثل مخططات الانتشار ورسم الخرائط) قيمة لا تُقدّر بثمن في اكتشاف الحالات الشاذة والأخطاء.

القاعدة 6: قدّم استراتيجية سليمة لتخزين البيانات وحفظها:

من الأخطاء الشائعة لدى الباحثين قليلي الخبرة (بل ولدى كثير من ذوي الخبرة) افتراض أن حواسيبهم الشخصية ومواقعهم الإلكترونية ستبقى إلى الأبد. فهم يفلون عن نسخ بياناتهم احتياطياً بانتظام في أثناء المشروع، ولا يرون فائدة أرشفة البيانات في موقع آمن على المدى البعيد. لكن لا مفر من أن تضع الأوراق، وتتعلّل الأقراص الصلبة، وتنكسر الروابط (URLs)، وتتلف الأشرطة وغيرها من الوسائط، فتصبح البيانات غير متاحة للاستخدام، لا لمنشئها ولا لغيرهم. ولذلك فإن تخزين البيانات وحفظها في صميم أي خطة جيدة لإدارة البيانات. فأمعن النظر في ثلاثة أسئلة:

• إلى متى ستظل البيانات متاحة؟

• كيف ستُخزّن البيانات وتُحمى طوال مدة المشروع؟

• كيف ستُحفظ البيانات وتُتاح للاستخدام في المستقبل؟

تعتمد الإجابة عن السؤال الأول على عدة عوامل. أولاً، تبيّن ما إذا كانت للجهة الراعية للبحث أو لمؤسستك الأم أي متطلبات محددة. فعادةً لا يلزم الاحتفاظ بجميع البيانات، وما يلزم الاحتفاظ به لا يلزم الاحتفاظ به إلى الأبد. ثانياً، انظر في القيمة الذاتية للبيانات. فأرصاد الظواهر التي لا يمكن تكرارها (مثل الأحداث الفلكية والبيئية) قد تحتاج إلى التخزين إلى أجل غير مسمى. أما بيانات التجارب التي يسهل تكرارها فقد لا تحتاج إلى التخزين إلا لمدة قصيرة.

وقد لا يلزم في عمليات المحاكاة سوى تخزين الشفرة المصدرية والشروط الابتدائية وبيانات التحقق. وإلى جانب شرح كيفية انتقاء البيانات للتخزين القصير المدى والحفظ الطويل المدى، تذكر أن تبرز أيضًا خطتك للبيانات الوصفية المصاحبة والشفرة والخوارزميات ذات الصلة التي ستيح للآخرين تفسير البيانات واستخدامها (انظر القاعدة 4).

ضع خطة سليمة لتخزين البيانات وحمايتها طوال عمر المشروع. ومن المقاربات الجيدة تخزين ثلاث نسخ على الأقل في موقعين على الأقل متباعدين جغرافيًا (مثل الموقع الأصلي كحاسوب مكتبي، وقرص صلب خارجي، وموقع بعيد واحد أو أكثر)، واعتماد جدول منتظم لنسخ البيانات (أي النسخ الاحتياطي). وقد تشمل المواقع البعيدة مختبر متعاون خارج الموقع، أو مستودعًا مؤسسيًا (institutional repository) (مثل مستودع قسمك أو جامعتك أو مؤسستك إن كان في مبنى آخر)، أو خدمة تجارية مثل تلك التي تقدمها (Amazon) و (Dropbox) و (Google) و (Microsoft). وينبغي أن يتضمن جدول النسخ الاحتياطي أيضًا اختبارات للتأكد من إمكان استرجاع ملفات البيانات المخزنة.

ومن المرجح أن يتطلب ضمان أفضل فرصة للوصول إلى البيانات بعد 20 عامًا من انتهاء عمر المشروع حلًا أكثر متانة (أي السؤال 3 أعلاه). فاطلب المشورة من الزملاء وأمناء المكتبات لتحديد مستودع بيانات مناسب لمجالك البحثي. فكل من التخصصات تتعهد مستودعات خاصة، مثل (GenBank) لبيانات تسلسل النوكليوتيدات، وبنك بيانات البروتين (Protein Data Bank) لتسلسلات البروتين. وبالمثل، تستضيف جامعات ومؤسسات كثيرة مستودعات مؤسسية، وثمة مستودعات عديدة للبيانات العلمية العامة مثل (Dryad) و (figshare) و (Zenodo). وبدلاً من ذلك، يمكن البحث بسهولة عن المستودعات التخصصية والعامة الاستخدام عبر فهارس على الإنترنت مثل (re3data.org) (أي سجل مستودعات بيانات البحث (Registry of REsearch data REpositories) و (BioSharing). وكثيرًا ما يُعدّ من الممارسات الجيدة إيداع الشفرات في مستودع مضيف مثل (GitHub) المتخصص في إدارة الشفرات المصدرية، وكذلك بعض أنواع البيانات مثل الملفات الكبيرة والبيانات الجدولية. ودون أي سياسات خاصة بالمستودع (مثل خصوصية البيانات وأمنها، واشتراط تقديم الشفرة المرتبطة) وتكاليف إيداع البيانات وتنسيقها (curation) ونسخها الاحتياطي، مما ينبغي إدراجه في خطة إدارة البيانات وفي ميزانية المقترح.

القاعدة 7: حدّد سياسات البيانات في المشروع:

رغم ما قد يكون نزوعًا طبيعيًا إلى تجنب مسائل السياسات والقانون، لا يملك الباحثون ترف ذلك حين يتعلق الأمر بالبيانات. فالجهات الراعية للبحث، والمؤسسات التي تحتضن البحث، والعلماء، لكلّ منهم دور في تعزيز السلوك المسؤول والأخلاقي والتزام به. ولذلك تشترط جهات راعية كثيرة أن تتضمن خطط إدارة البيانات نصوصًا صريحة للسياسات المتعلقة بكيفية إدارة البيانات ومشاركتها. ومن هذه السياسات:

- ترتيبات الترخيص أو المشاركة المتعلقة باستخدام مواد موجودة مسبقًا؛
- خطط الاحتفاظ بالبيانات والشفرة وغيرها من المواد وترخيصها ومشاركتها وفرض فترات حظر عليها (embargo)، أي تقييد استخدام الآخرين لها مدة من الزمن؛
- القيود القانونية والأخلاقية على الوصول إلى بيانات المشاركين البشر وغيرها من البيانات الحساسة واستخدامها.

ولسوء الحظ، كثيرًا ما تبدو السياسات والقوانين مربكة أو متناقضة، أو تكون كذلك بالفعل. وفوق ذلك، قد لا تنطبق السياسات السارية داخل مؤسسة واحدة أو في بلد ما على غيرها. وعند الشك، استشر مكتب البحوث الممولة في مؤسستك، أو مجلس المراجعة المؤسسي (Institutional Review Board) المختص، أو مسؤول البرنامج (أو مسؤوليه) المعيّنين للبرنامج الذي تتقدم إليه طلبًا للدعم.

ورغم هذه التحفظات، يمكن عادةً وضع سياسة سليمة باتباع بضع خطوات بسيطة. أولًا، إذا كانت مواد موجودة مسبقًا، مثل البيانات والشفرات، سَتستخدم، فحدّد ترتيبات الترخيص والمشاركة ذات الصلة وأدرج وصفًا لها في خطة إدارة البيانات. واطرح كيف تُستخدم برمجيات أو مكتبات برمجية من أطراف ثالثة في إنشاء البرمجيات الجديدة وإصدارها. ولاحظ أن قوانين الملكية وحقوق الملكية الفكرية (IPR) ولوائح ضبط التصدير قد تحدّ من مدى إمكان مشاركة الشفرات والبرمجيات.

ثانيًا، اشرح كيف ومتى ستُتاح البيانات وغيرها من منتجات البحث. واحرص على شرح أي فترات حظر أو تأخير، مثل التأخير لأسباب تتعلق بالنشر أو ببراءات الاختراع. ومن الممارسات الشائعة إتاحة البيانات على نطاق واسع وقت النشر، أو، في حالة طلاب الدراسات العليا، وقت منح الدرجة العلمية. وكلما أمكن، طبّق تنازلات أو تراخيص معيارية عن الحقوق، مثل تلك التي وضعتها (Creative Commons - الإبداعي) (ODC) (Open Data Commons) (CC)، لتوجيه الاستخدام اللاحق للبيانات وغيرها من المنتجات الفكرية. فخصة (CC0) ورخصة (ODC) للإهداء إلى الملك العام (ODC Public Domain Dedication and License) مثلًا تعززان المشاركة والاستخدام غير المقيد للبيانات. أما التراخيص والتنازلات غير المعيارية فقد تكون عائقًا كبيرًا أمام إعادة الاستخدام.

ثالثًا، اشرح كيف ستُعامل بيانات المشاركين البشر وغيرها من البيانات الحساسة (مثل لوائح أبحاث صحة الإنسان الواردة في قانون نقل التأمين الصحي والمساءلة الأمريكي (US Health Insurance Portability and Accountability Act)). وتشترط جهات راعية كثيرة أن يطلب الباحثون العاملون في أبحاث المشاركين البشر موافقة مسبقة من مجلس المراجعة المؤسسي المختص، أو أن يحصلوا عليها، قبل تقديم مقترح المنحة، وأن يحصلوا عليها بالتأكيد قبل الشروع في البحث الفعلي. وقد تشترط الموافقات الحصول على موافقة مستنيرة، أو إخفاء هوية أصحاب البيانات، أو تقييد الاستخدام بطريقة ما.

القاعدة 8: صف كيف ستُنشر البيانات:

إن أحكم خطط الحفظ وسياسات مشاركة البيانات لا تضمن وحدها إتاحة بيانات المشروع. وسيطمئن المحكمون والجهات الراعية للبحث إلى أن الأمر لن يكون كذلك إذا فصلت كيف ومتى ستُنشر منتجات البيانات للآخرين، ولا سيما الأشخاص من خارج مجموعتك البحثية. وهناك أساليب تكتفي بالاستجابة للطلب وأخرى تبادر إلى نشر البيانات. فالأساليب التي تستجيب للطلب تشمل نشر البيانات على موقع المشروع أو الموقع الشخصي، أو إرسالها بالبريد أو البريد الإلكتروني عند الطلب، وإن كان الأخير قد يكون إشكاليًا عند التعامل مع بيانات كبيرة وقيود عرض النطاق. أما الأساليب الأكثر مبادرة ومتانة، وهي المفضلة فتشمل: (1) نشر البيانات في مستودع أو أرشيف مفتوح (انظر القاعدة 6)؛ و(2) تقديم البيانات (أو مجموعات فرعية منها) ملاحقًا أو ملفات تكميلية لمقالات الدوريات، كما هو شائع في مجموعة دوريات (PLOS)؛ و(3) نشر البيانات والبيانات الوصفية والشفرات ذات الصلة في صورة «ورقة بيانات» (data paper) [5]. ويمكن نشر أوراق البيانات في دوريات مختلفة، منها (Scientific Data) (من مجموعة

(Nature) للنشر، و(GeoScience Data Journal) (من منشورات Wiley) نيابة عن الجمعية الملكية للأرصاد الجوية)، و(GigaScience) (منشور مشترك بين BioMed Central) و(Springer) يدعم البيانات الضخمة من تخصصات كثيرة في علم الأحياء وعلوم الحياة).

وتتضمن خطة النشر الجيدة بضع عبارات موجزة. اذكر متى ستُتاح منتجات البيانات، وكيف، وما هي. وعمومًا، يُشجّع على إتاحة البيانات إلى أقصى حد وبأقل قيود ممكنة وقت النشر أو عند إتمام المشروع. والمقاربات الأكثر استباقية الموصوفة أعلاه مفضلة كثيرًا على إرسال البيانات بالبريد أو البريد الإلكتروني، ومن المرجح أن توفر وقتًا ومالًا كبيرين على المدى البعيد، لأن تنسيق البيانات ومشاركتها سيحظيان بدعم الدوريات والمستودعات أو الأرشفات المناسبة. وفوق ذلك، تقدّم دوريات ومستودعات كثيرة إرشادات وآليات لكيفية استشهاد الآخرين ببياناتك على النحو الصحيح، بما في ذلك معرفات الكائنات الرقمية وصيغ الاستشهاد الموصى بها؛ وهذا يساعد على ضمان حصولك على الفضل في منتجات البيانات التي تنشئها. وتذكر أن البيانات ستكون أيسر استخدامًا وتفسيرًا لك وللآخرين إذا نُشرت بأساليب معيارية غير احتكارية، وإذا صحبتها البيانات الوصفية والشفرة المرتبطة المستخدمة في معالجة البيانات.

القاعدة 9: وزّع الأدوار والمسؤوليات:

تحدد خطة إدارة البيانات الشاملة بوضوح أدوار كل فرد ومؤسسة مشاركة في المشروع ومسؤولياتهم. وقد تشمل الأدوار جمع البيانات، وإدخالها، وضمان الجودة وضبطها، وإنشاء البيانات الوصفية وإدارتها، والنسخ الاحتياطي، وإعداد البيانات وتقديمها إلى أرشيف، وإدارة الأنظمة. وانظر في الوقت المخصص ومستويات الخبرة اللازمة لدى العاملين. ففي المشاريع الصغيرة والمتوسطة، يستطيع طالب واحد أو باحث مشارك في مرحلة ما بعد الدكتوراه يجمع البيانات ويعالجها أن يتولى بسهولة معظم مهام إدارة البيانات أو جميعها. وفي المقابل، قد تستفيد المشاريع الكبيرة المتعددة الباحثين من تخصيص موظف متفرغ أو موظفين متفرغين لإدارة البيانات.

وتعامل مع خطة إدارة البيانات بوصفها وثيقة حية، وعُد إليها كثيرًا (مثلًا كل ثلاثة أشهر). وكلف عضوًا في فريق المشروع بتنقيح الخطة بما يعكس أي تغييرات جديدة في البروتوكولات والسياسات. ومن الممارسات الجيدة تتبع أي تغييرات في سجل للمراجعات يذكر تواريخ إجراء أي تغييرات على الخطة مع تفاصيل تلك التغييرات، بما في ذلك من أجراها.

وقد يهتم المحكمون والجهات الراعية اهتمامًا خاصًا بمعرفة كيف سيُقيّم الالتزام بخطة إدارة البيانات ويثبت، وكذلك كيف ستُدار البيانات وتُتاح بعد انتهاء المشروع، ومن سيتولى ذلك. وفيما يتعلق بالأمر الأخير، كثيرًا ما يكفي إدراج إحالة إلى السياسات والإجراءات التي يتبعها المستودع الذي تعتزم إيداع بياناتك فيه. واحرص على ذكر أي إسهامات من عاملين من خارج المشروع، مثل أي دعم تقدمه مؤسستك في المستودعات أو إدارة الأنظمة أو النسخ الاحتياطي أو التدريب أو الحوسبة العالية الأداء.

القاعدة 10: أعد ميزانية واقعية:

إن إنشاء بيانات عالية الجودة وإدارتها ونشرها ومشاركتها جزء من مؤسسة البحث في القرن الحادي والعشرين بقدر ما هو نشر النتائج. وإدارة البيانات ليست أمرًا جديدًا، بل هي أمر يقوم به جميع الباحثين بالفعل. ومع ذلك، من الأخطاء الشائعة في إعداد خطة إدارة البيانات نسيان رصد ميزانية للأنشطة. فإدارة البيانات تستغرق وقتًا وتكلف مالًا من حيث البرمجيات والأجهزة والعاملين. فراجع خطتك وتأكد من وجود بنود في الميزانية لدعم الأشخاص الذين

يديرّون البيانات (انظر القاعدة 9)، وكذلك لتغطية الأجهزة والبرمجيات والخدمات اللازمة. وتواصل مع مستودع البيانات المفضّل (انظر القاعدة 6) لكي تُرصد الرسوم والخدمات اللازمة في الميزانية على النحو المناسب. وبقدر ما تسمح المساحة، ييسر الأمر على المحكّمين بالإحالة إلى بنود أو أقسام محددة في الميزانية وفي صفحات تسويق الميزانية. وسيلحظ المحكّمون ذوو الخبرة المكوّنات غير الممولة، لكنهم سيذكرون أيضًا أن حجم الاستثمار في إدارة البيانات، زيادةً أو نقصانًا، يعتمد على طبيعة البحث وأنواع البيانات.

الخاتمة:

ينبغي أن توفر لك خطة إدارة البيانات، وللآخرين، خريطة طريق سهلة الاتباع ترشد إلى كيفية التعامل مع البيانات طوال عمر المشروع وبعد إتمامه وتشرح ذلك. والقواعد العشر البسيطة المعروضة هنا مصمّمة لمساعدتك على كتابة خطة جيدة منطقية وشاملة، تنال رضا المحكّمين والجهات الراعية للبحث، ويمكنك وضعها موضع التطبيق إذا مَوّل مشروعك. وتوفر خطة إدارة البيانات وسيلة لنقل المعلومات إلى فريق مشروعك وتحديد التوقعات منه، في مرحلتها المقترحة والتخطيط للمشروع، وكذلك في اجتماعات فريق المشروع لاحقًا حين يكون المشروع جاريًا. ومع ذلك، لا توجد خطة كاملة. فالخطط تتحسن بالاستخدام. وأفضل الخطط «وثائق حية» تُراجَع دوريًا وتُنقَح عند الضرورة وفق الاحتياجات وأي تغييرات في البروتوكولات (مثل البيانات الوصفية، وضمان الجودة وضبطها، والتخزين)، والسياسات، والتقنية، والعاملين، كما يُعاد استخدامها، بحيث تُدمج أنجح أجزاء الخطة في المشاريع اللاحقة. والخطة العلنية المقروءة آليًا والمرخصة ترخيصًا مفتوحًا أرجح كثيرًا أن تُدمج في المشاريع المستقبلية وأن يكون لها أثر أكبر؛ وهذه الشفافية المتزايدة في عملية تمويل البحث (مثل نشر المقترحات وخطط إدارة البيانات) يمكن أن تساعد الباحثين والجهات الراعية على اكتشاف البيانات والمتعاونين المحتملين، والتثقيف بشأن إدارة البيانات، ورصد الامتثال للسياسات [6].

المراجع:

1. Bourne PE, Chalupa LM (2006) Ten simple rules for getting grants. PLoS Comput Biol 2(2): e12. pmid:16501664
2. Zhang W. (2014) Ten simple rules for writing research papers. PLoS Comput Biol 10(1):e1003453. pmid:24499936
3. Bourne PE. (2007) Ten simple rules for making good oral presentations. PLoS Comput Biol 3(4): e77. pmid:17500596
4. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Comput Biol 10(4): e1003542. pmid: 24763340
5. Chavan V, Penev L (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinf 12 (Suppl 15):S2.
6. Mietchen D (2014) The transformative nature of transparency in research funding. PLoS Biol 12(12): e1002027. pmid:25549343