Linee guida per l'applicazione dei principi FAIR alla gestione e al riuso dei dati

PARTHENOS

A PROPOSITO DI QUESTA GUIDA

Questo manuale propone una serie di linee guida per allineare gli sforzi di chi produce, gestisce e utilizza i dati nel campo delle scienze sociali e umanistiche e del patrimonio culturale, in modo da renderli il più possibile riutilizzabili.

Le linee guida sono il risultato del lavoro di oltre cinquanta professionisti coinvolti nel progetto PARTHENOS, che hanno esaminato le caratteristiche condivise nell'implementazione di politiche e strategie per la gestione dei dati di ricerca. Sono stati analizzati i risultati derivanti dalla ricerca documentale, da questionari e da interviste a esperti selezionati, e circa un centinaio di documenti di policy relativi alla gestione dei dati (comprese linee guida sui formati, sulle politiche per la revisione dei dati e sulle buone pratiche, adottate sia formalmente che in modo tacito).

Concentrandosi sulla qualità dei (meta)dati e dei repository, il gruppo di PARTHENOS ha estratto 20 linee guida comuni alle diverse discipline.

Per facilitare la consultazione, le linee guida sono state organizzate in base ai principi FAIR Data - Findable, Accessible, Interoperable, Reusable - ovvero rintracciabilità, accessibilità, interoperabilità e riusabilità pubblicati per la prima volta da FORCE11 (2016).
Ciascuna delle linee guida di PARTHENOS comprende delle specifiche raccomandazioni rivolte a due gruppi principali di stakeholder: da una parte i produttori e gli utilizzatori dei dati, dall'altra i gestori degli archivi di dati. Delle icone grafiche indicano a quale stakeholder sono indirizzate le raccomandazioni.

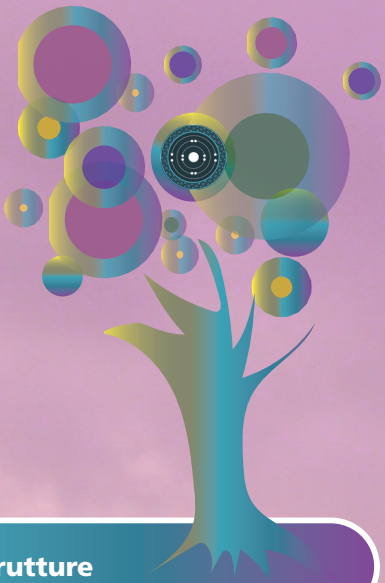
La lampadina mostra le raccomandazioni rivolte ai produttori e agli utilizzatori dei dati, come ad esempio i ricercatori e le comunità di ricerca nel campo della storia, dell'archeologia, degli studi linguistici e delle scienze sociali.

Gli ingranaggi mostrano le raccomandazioni per le infrastrutture di ricerca e gli archivi di dati gestiti dagli istituti di ricerca e dagli istituti culturali.

PARTHENOS è un consorzio formato da 16 istituti e infrastrutture di ricerca europei, il cui obiettivo è incrementare la riusabilità dei dati condividendo standard e procedure per la gestione del ciclo di vita delle risorse digitali migliorando la qualità dei dati e dei repository, e favorendo lo sviluppo di policy per gli Open Data e l'Open Access nell'ambito delle infrastrutture di ricerca e degli istituti dei beni culturali nei campi correlati delle scienze umane e degli studi sociali.

20 LINEE GUIDA

per l'applicazione dei principi FAIR alla gestione e al riuso dei dati



1 Investite in persone e infrastrutture

Un importante presupposto per rendere possibile l'attuazione delle linee guida presenti in questo manuale è investire nelle infrastrutture, assumendo e formando esperti nella gestione dei dati.

Familiarizzate con le buone pratiche riguardanti la gestione dei dati di ricerca. Consultate i moduli di formazione di PARTHENOS oppure il manuale CESSDA Data Management Expert Guide.

Formate esperti nella gestione dei dati e investite nell'infrastruttura tecnica e nel personale.

RINTRACCIABILI (Findable)

I dati di ricerca dovrebbero essere facilmente rintracciabili sia dagli esseri umani sia dai sistemi informatici e dovrebbero basarsi su descrizioni obbligatorie dei metadati che permettano la scoperta di set di dati di interesse per la propria ricerca.

2 Usate identificatori persistenti

L'individuazione dei dati è una condizione necessaria per eseguire qualsiasi operazione, dall'accesso al loro riuso. Per essere rintracciabile (Findable), ogni risorsa e dataset deve essere identificabile in modo univoco e immutabile nel tempo tramite un identificatore persistente (PID). Il PID continua a funzionare anche quando l'indirizzo web di una risorsa cambia. I PID possono essere di vario tipo, ad esempio Handle, DOI, PURL oppure URN.

Citate il PID attribuito al vostro dataset nei risultati della vostra ricerca.

Selezionate lo schema d'identificazione persistente più appropriato e attribuite un PID a ogni risorsa. Utilizzate la Guida PID prodotta da NCDD per decidere quale sia il PID adeguato alla vostra infrastruttura di ricerca.

3 Citate i dati di ricerca

Se i dati hanno un identificatore persistente e vengono citati in base agli standard utilizzati dalla vostra comunità di ricerca, le risorse o i dataset corrispondenti saranno rintracciabili più facilmente.

Familiarizzate con le linee guide sulla citazione dei dati specifiche del vostro campo di ricerca o disciplina e citate i dati in maniera conforme.

Fornite alle comunità di ricerca informazioni sulle buone pratiche relative alla citazione dei dati e fate in modo che gli utenti possano citare i dati facilmente, per esempio mediante un pulsante che riporta la dicitura "Come citare questo dataset".

4 Usate identificatori persistenti per l'autore

L'identificatore persistente per l'autore (per esempio VIAF, ISNI o ORCID) serve per creare collegamenti fra dataset, pubblicazioni e ricercatori, consentendone il riconoscimento e l'individuazione (Findability).

Distinguetevi da ogni altro ricercatore o gruppo di ricerca. Se non lo avete già fatto, richiedete un numero d'identificazione per l'autore e citatelo nel vostro dataset.

Menzionate l'identificativo dell'autore nei metadati.

5 Scegliete lo schema di metadati più appropriato

I metadati sono essenziali per rendere i dati rintracciabili, soprattutto quelli usati allo scopo di citare e descrivere i dati stessi. Lo schema di metadati consiste in una lista di elementi standard per raccogliere le informazioni di una risorsa, ad esempio il titolo, l'identificativo, il nome dell'autore o una data. Utilizzate schemi di metadati esistenti conformi agli standard internazionali per favorire lo scambio dei dati.

Per permettere la scoperta dei contenuti, descrivete i dati di ricerca nella maniera più uniforme e completa possibile. Includete una quantità sufficiente di informazioni in modo da rendere i dati accessibili e comprensibili nel tempo. Se possibile, utilizzate uno schema di metadati esistente e adatto alle risorse o ai dataset che state descrivendo.

Dichiarate chiaramente quale schema di metadati adoperate e condividetelo con la comunità di ricerca. Per arricchire i dataset al momento del deposito dei dati, prendete in considerazione la possibilità di raccogliere metadati aggiuntivi, ad esempio la provenienza, attraverso un modulo che accompagna la consegna dei dati.

ACCESSIBILI
(Accessible)

I dati di ricerca dovrebbero essere facilmente accessibili e recuperabili con condizioni d'accesso ben definite tramite protocolli di comunicazione standardizzati.

6 Selezionate un repository affidabile

Un repository accreditato offre un'organizzazione affidabile dei dataset. La certificazione garantisce che i dati siano conservati in maniera sicura e che restino disponibili, rintracciabili e accessibili nel lungo termine. Alcuni esempi di certificazione standard sono CoreTrustSeal, Nestor Seal e la certificazione ISO 16363.

Rendete i vostri dati accessibili attraverso un repository affidabile. Inoltre, se utilizzate gli standard raccomandati dagli stessi repository (relativi ai formati dei file, allo schemadi metadati ecc.) sarete certi di aver rispettato tutti i requisiti per rendere i dati FAIR (Findable, Accessible, Interoperable, Reusable).

Dichiarate chiaramente il livello di certificazione sul vostro sito web. Se il vostro repository non è certificato, dichiarate come intendete assicurare la disponibilità, la rintracciabilità, l'accessibilità e il riutilizzo dei dati nel lungo termine.

7 Dichiarate chiaramente l'accessibilità

Le informazioni d'accesso specificano quali sono le condizioni di accesso ai dataset. Quando si depositano i dati in un repository, devono essere chiare le modalità di accesso che si possono selezionare.

Nello scegliere le modalità di accesso, tenete presenti i requisiti legali, le politiche specifiche delle discipline e i protocolli etici. Quando possibile, scegliete l'Open Access. Quando raccogliete dati personali, domandatevi se questi contengano informazioni che possano rivelare l'identità dei partecipanti, a cosa hanno dato il consenso e quali misure sono state attuate per proteggere i dati. Se i vostri dati non possono essere pubblicati in Open Access, i relativi metadati invece dovrebbero esserlo, in modo da permettere la scoperta dei dati.

Incoraggiate la pubblicazione Open Access dei (meta)dati. Dichiarate chiaramente quali sono le modalità di accesso riservate per i (meta)dati sensibili che non sono pubblicamente accessibili. In questo caso, cercate di rendere disponibili i (meta)dati attraverso una procedura di accesso controllata e documentata.

8 Utilizzate l'embargo dei dati solo quando necessario

Durante il periodo di embargo dei dati, solo la descrizione del dataset viene pubblicata. I dati stessi non sono accessibili e i (meta)dati vengono resi disponibili solo dopo un certo periodo di tempo.

Dichiarate chiaramente per quale motivo e per quanto tempo si è reso necessario l'embargo dei dati. Rendete i (meta)dati pubblicamente accessibili appena è possibile.

Specificate se è permesso l'embargo dei dati e quali condizioni si applicano.

9 Utilizzate protocolli di scambio standardizzati

Utilizzando protocolli di scambio standardizzati, le infrastrutture di ricerca possono rendere i (meta)dati pubblicamente accessibili e disponibili per l'harvesting, per esempio dai motori di ricerca, migliorandone così l'accessibilità.

Utilizzate protocolli standard quali SWORD, OAI-PMH, ResourceSync e SPARQL. Convertite gli schemi di (meta)dati in file XML o RDF. Mantenete e pubblicate un registro dei protocolli degli end-point indicando il percorso per accedere e pubblicare i dati.

Per velocizzare la scoperta e rivelare nuove conoscenze, sia gli esseri umani sia i sistemi informatici devono poter facilmente combinare i dati di ricerca con altri dataset.

INTEROPERABILI (Interoperable)

10 Stabilite API ben documentate ed eseguibili da computer

API (Application Programming Interface) ben documentate ed eseguibili da computer – attraverso una raccolta di procedure, protocolli e strumenti per sviluppare software applicativi – permettono di indicizzare, recuperare e combinare automaticamente i (meta)dati da diversi repository.

Documentate bene le API in modo che siano disponibili gli schemi dei modelli dei (meta)dati adottati. Valutate se fornire esempi su come estrarre i dati da diversi end-point e combinarli in nuovi dataset, utilizzabili per nuove ricerche.

11 Utilizzate vocabolari ben definiti

La descrizione degli elementi dei metadati dovrebbe essere conforme alle linee guida adottate dalle varie comunità di ricerca che adoperano vocabolari open, ben definiti e conosciuti. Tali vocabolari riportano l'esatto significato dei concetti e delle qualità che i dati rappresentano.

Sin dall'inizio del vostro progetto di ricerca utilizzate i vocabolari pertinenti al vostro settore per arricchire e strutturare i risultati della vostra ricerca in maniera conforme.

Fornite alla comunità di ricerca esempi di vocabolari basati sulle specifiche dell'ambito di ricerca.

12 Documentate i modelli di metadati

Documentare chiaramente i modelli di metadati permette agli sviluppatori di confrontare ed eseguire mappature tra i diversi schemi di metadati.

Pubblicate i modelli di metadati usati dalla vostra infrastruttura di ricerca. Documentate le specifiche tecniche e definite le classi (gruppi di risorse che hanno proprietà in comune) e le proprietà (elementi che esprimono gli attributi di una sezione dei metadati oltre alla relazione fra le diverse parti dei metadati). Elencate le proprietà obbligatorie e quelle raccomandate per la mappatura dei metadati.

13 Stabilite e adoperate standard interoperabili

Usare standard adottati da una comunità ampia incrementa la possibilità di condividere, riusare e combinare tra loro le raccolte di dati.

Verificate quali standard sono utilizzati nei repository in cui intendete depositare i vostri dati. Strutturate la vostra raccolta dati secondo questo formato sin dall'inizio del vostro progetto di ricerca.

Dichiarate chiaramente quali standard sono utilizzati dal vostro istituto, condividete con la vostra comunità e garantite il mantenimento soprattutto dal punto di vista dell'interoperabilità. METS e CIDOC-CRM sono alcuni validi esempi.

14 Stabilite procedure per migliorare la qualità dei dati

Per migliorare la qualità dei (meta)dati e di conseguenza la loro l'interoperabilità, stabilite procedure automatiche per ripulire, caricare, gestire e arricchire i (meta)-dati.

Stabilite procedure per minimizzare il rischio di errori nella raccolta dati, per esempio selezionando la data da un calendario invece di inserirla a mano.

Investite in strumenti per ripulire i (meta)dati e per convertire i dati in formati standardizzati e interoperabili. Sviluppate flussi di lavoro e soluzioni software per eseguire procedure automatiche, ad esempio impiegando strumenti di apprendimento automatico.

15 Stabilite e adoperate formati sostenibili nel lungo periodo

Tutti i file depositati in un repository dovrebbero essere in un formato internazionale standardizzato in modo da assicurare l'interoperabilità a lungo termine per quanto concerne la riusabilità, accessibilità e sostenibilità dei dati.

Valutate formati sostenibili sin dall'inizio del vostro progetto di ricerca. Seguite le raccomandazioni del gestore del repository adoperandone i formati preferiti, indipendenti da software, sviluppatori e venditori specifici.

Incoraggiate l'utilizzo di formati adatti all'archiviazione a lungo termine, ad esempio i file PDF-A, CSV e MID/MIF. Fornite una panoramica dettagliata e facile da reperire dei formati accettati.

I dati di ricerca dovrebbero essere già predisposti per ricerche e trattamenti futuri, specificando che i risultati potranno essere replicati e che la nuova ricerca sarà effettivamente fondata sui risultati precedenti.

RIUTILIZZABILI (Reusable)

16 Documentate sistematicamente i dati

I dati dovrebbero essere documentati in maniera sistematica per rendere evidente ciò che ci si aspetta da un dataset o da un repository. La trasparenza su ciò che si trova o non si trova nei dati conferisce affidabilità e di conseguenza favorisce anche il riuso.

Fornite manuali che includono la metodologia seguita, la lista delle abbreviazioni, la descrizione delle lacune nei dati, il setup del database ecc.

17 Seguite le convenzioni terminologiche

Seguire una convenzione terminologica precisa e coerente – uno schema generalmente condiviso per nominare i file di dati – rende molto più semplice per le generazioni di futuri ricercatori recuperare e comprendere le risorse e i dataset.

Consultate le politiche e le buone pratiche nella vostra disciplina o dominio di ricerca per trovare la convenzione terminologica più adatta.

Dichiarate chiaramente quali siano le buone pratiche per creare e applicare specifiche convenzioni relative alle terminologie dei file.

18 Adoperate formati comuni

La riusabilità dei dati aumenta utilizzando i formati standardizzati più diffusi nella vostra comunità.

Utilizzate gli attuali formati più diffusi insieme ai formati di archiviazione per condividere i vostri dati, per esempio Excel (xlsx) e CSV o GeoPackage Shapefiles.

Pubblicate i dati nei formati più comuni e che più si avvicinano a quelli di archiviazione quando i due sono diversi.

19 Mantenete l'integrità dei dati

I dati di ricerca che sono stati raccolti dovrebbero essere identici a quelli cui si accederà in un secondo momento. Per garantire il mantenimento dell'integrità, bisogna eseguire delle verifiche.

Implementate un metodo per il controllo della versione. È di fondamentale importanza garantire che ogni modifica alla versione originale di un dataset sia correttamente documentata.

Per riconoscere un file modificato, è essenziale registrare la provenienza – l'origine dei dati oltre a qualsiasi successivo cambiamento – e confrontare le copie con l'originale. Il controllo dell'integrità dei dati può avvenire tramite un codice di controllo tipo checksum, oppure dal confronto diretto fra due file. Fornite un meccanismo per evidenziare le diverse versioni, ad esempio aggiungendo come parametro di ricerca la versione dell'identificativo.

20 Licenza per il riuso

Per consentire il maggior riuso possibile, deve essere chiaramente indicato chi detiene i diritti sui (meta)dati e quale licenza si applica ad essi.

Prima di pubblicare la vostra ricerca, assicuratevi di conoscere chi detiene i diritti sui (meta)dati.

Comunicate in maniera trasparente e applicate licenze ai (meta)dati in un formato leggibile anche dal computer, possibilmente adottando licenze utilizzate in ambito internazionale o mappando le vostre licenze su quelle largamente utilizzate, come Creative Commons e Rights Management, per migliorare l'interoperabilità.

PARTHENOS è un progetto Horizon 2020 finanziato dalla Commissione europea. Le opinioni e le opinioni espresse in questa pubblicazione sono di esclusiva responsabilità dell'autore e non riflettono necessariamente le opinioni della Commissione europea.

 DOI: 10.5281/zenodo.3363243.
Design: Verbeeldingskr8.