Ridurre l'impronta carbonica dei repository Open Access

L'esperienza di 4Science con DSpaceWay™

La transizione verso la Scienza Aperta digitale comporta un costo ambientale spesso sottovalutato: i repository istituzionali e le infrastrutture di archiviazione richiedono data center attivi 24/7, con un consumo energetico significativo e una crescente impronta carbonica. Questo intervento presenta l'esperienza concreta di 4Science nel ridurre l'impatto ambientale delle infrastrutture per la ricerca aperta, con particolare riferimento alla piattaforma DSpaceWay™.



Andrea Bollini — COO, 4Science

4SCIENCE
Share your knowledge

GenOA week 2026 Sessione 4: Costi ambientali della scienza digitale

Agenda

01

Il problema

Il costo ambientale nascosto della Scienza Aperta digitale

02

Crawler e AI

La duplice natura — problema e opportunità

03

La risposta architetturale

Infrastrutture moderne con DSpaceWay™

04

Risultati e prospettive

Impegno ambientale e prospettive future

Il costo ambientale della Scienza Aperta digitale

485 TWh

Consumo elettrico globale
dei data center (2025, IEA)

950 TWh

Proiezione entro il 2030
Pari a circa il 3% dell'energia globale

- I repository istituzionali attivi 24/7 contribuiscono a questa impronta ambientale
- Consumo idrico per il raffreddamento dei server spesso non rendicontato
- L'introduzione sempre più marcata dell'AI nei workflow della ricerca amplifica il problema se non gestita in modo sostenibile
- Ogni richiesta HTTP consuma energia: CPU, memoria, rete, storage

Dati tratti da "Electricity 2026: Analysis and forecast to 2030, Internation Energy Agency" https://www.iea.org/reports/electricity-2026/demand

4Science: chi siamo

Contributori attivi della community, 8 DSpace Committer alcuni coinvolti da oltre 20 anni nel progetto (da DSpace 1.2)

DSpace Certified Platinum Provider con il maggior contributo tecnico e unico contributore attivo Italiano di Dataverse

Oltre 100 progetti completati, clienti di tutte le dimensioni nel mondo

Hosting enterprise-grade su AWS dal 2016 (Europa e USA)
100% match con energie rinnovabili – Dublino e N.Virginia

~100 TB	50+
Storage	Clienti
~2,7M	~200
Item	Istanze DSpace

Dati riferiti ai soli clienti hosted / SaaS

Certificazioni: ISO 27001, 27017, 27018, 22301, 20000-1, 9001 • CSA Trusted Cloud • AWS Partner

La duplice natura — Il problema



- I crawler (motori di ricerca, aggregatori, bot AI) generano un volume di traffico sproporzionato
- Traffico "non umano" consuma risorse computazionali (potenzialmente) senza beneficio diretto per gli utenti
- Richieste ridondanti (favicon, asset statici, API ripetute) amplificano lo spreco energetico
- Bot malevoli e scanner automatici (wp-login.php) aggiungono carico inutile
- Il footprint ambientale è migliorato, MA il traffico crawler sta impattando negativamente il footprint totale

Più traffico = Più CPU = Più energia = Più emissioni

Il traffico dei crawler: i dati reali – maggio 2026

6,77M
richieste in un singolo burst

27M
richieste in una settimana senza attacchi

160k
req/min al picco (100x il normale)

381.256 log

in sole 2 ore di analisi dei log del server

333K+ richieste

generate da favicon.ico — completamente inutili

Sitemap e robots.txt

Asset statici (.js, .css, font) generano quota significativa del traffico

Dati reali dall'infrastruttura 4Science relativi ad un singolo repository in produzione — burst del 27 maggio 2026

Il traffico dei crawler: i dati reali – ottobre 2026

1,54G (224x)
richieste in un singolo burst

68M (2.5x)
richieste in una settimana senza attacchi

937k
req/min al picco (585x il normale)



16.2M log **(42x)**

in sole 2 ore di analisi dei log del server



Punte del 99.85% di 4xx

generate da URL inesistenti e tentativi ripetuti di utilizzo di risorse protette — completamente inutili



Richieste bloccate dal WAF

Dal 10.5% a giugno al 24.9% di fine ottobre (escludendo attacchi massivi)

Dati reali dall'infrastruttura 4Science relativi ad un singolo repository in produzione — burst del 12 ottobre 2026

La duplice natura — L'opportunità

- I crawler dei motori di ricerca indicizzano i contenuti e aumentano la visibilità della ricerca
- Gli aggregatori (OpenAIRE, BASE, CORE, Google Scholar) amplificano la scoperta e il riuso
- I tool AI abilitano l'accesso in linguaggio naturale ai repository e diventeranno sempre più lo strumento di lavoro dei ricercatori
- Più visibilità = più citazioni, più impatto, più giustificazione per il finanziamento
- L'Open Access raggiunge il suo scopo solo se i contenuti sono effettivamente trovabili e accessibili

La sfida: gestire questo traffico in modo efficiente, non bloccarlo

La necessità di infrastrutture moderne

Il modello tradizionale non è sostenibile

VM isolate (5-7 VM per installazione):

- ✗ Over-provisioning costante
- ✗ HA complessa e costosa
- ✗ Autoscaling lento (minuti)
- ✗ Manutenzione elevata
- ✗ Failure casuali

Plan S
Making full & immediate Open Access a reality

Richiede un uptime del 99.7% minimo

CDN — Caching intelligente per contenuti statici e crawler

WAF — Protezione da bot malevoli e attacchi DDoS

K8s — Alta Disponibilità — resilienza senza over-provisioning

AUTO — Autoscaling rapido per assorbire i picchi di traffico

DSpaceWay™: il repository come servizio sostenibile



SaaS turnkey

Le istituzioni si concentrano su ricerca, dati e impatto



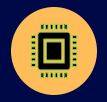
AWS Cloud

Hosting con focus sulla sostenibilità ambientale ed economica



Data center Dublino e N.Virginia (per clienti US)

Dublino - energia eolica, N.Virginia - energia rinnovabile



ARM/Graviton

Più efficienti energeticamente vs x86/AMD (fino a 60% di riduzione dei consumi)



Orchestrazione tramite k8s. Cluster condiviso, dati isolati

Risparmio energetico vs over-provisioning dedicato 24/7

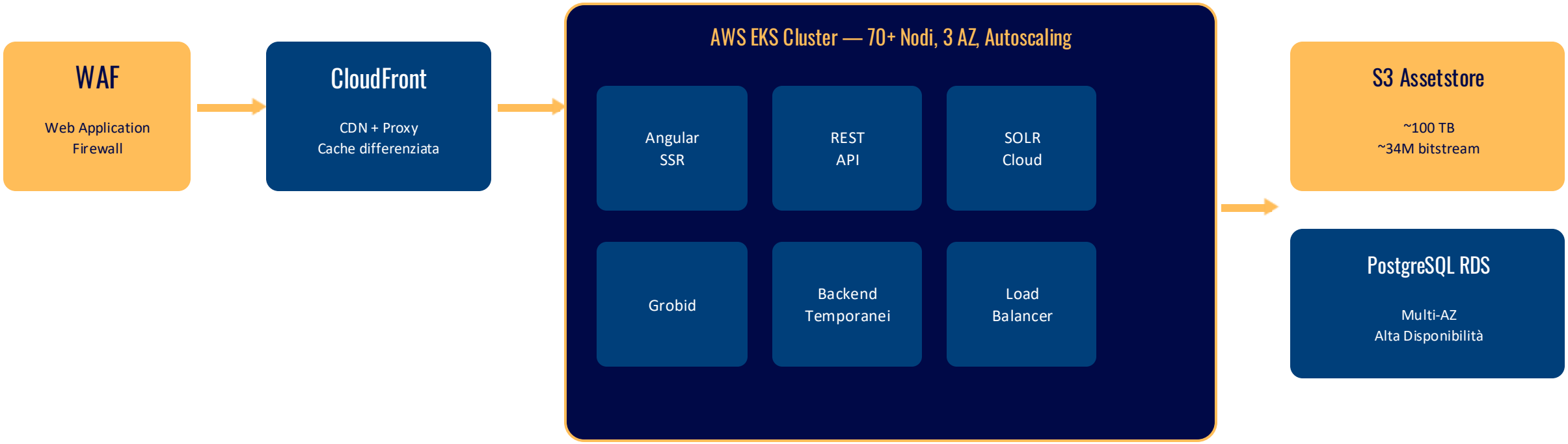
A partire da

299

EUR/USD al mese

Aggiornamenti automatici inclusi
Accessibile a istituzioni di tutte le dimensioni

Architettura Kubernetes: la soluzione 4Science



Servizi Serverless e Monitoring

IIIF
API Gateway + Lambda

GuardDuty
Security Monitoring

CloudWatch
Metrics & Logs

Datadog
Observability

I vantaggi di Kubernetes per la sostenibilità

Densità computazionale

Ogni nodo al ~100% di carico, eliminando l'over-provisioning delle VM tradizionali

Autoscaling rapido

Aumento delle risorse in secondi (non minuti come VM) per rispondere ai picchi senza sprechi

Riduzione manutenzione

Componenti non sane rimpiazzate automaticamente, IaC by design con approccio GitOps

Maggiore sicurezza

Nodi e componenti applicative aggiornati automaticamente e periodicamente

Osservabilità

Metriche individuali per componente, ottimizzazione delle risorse allocate

Risultato netto

Meno hardware per stesso livello di servizio = meno energia = meno emissioni

CDN e caching intelligente

Il problema

Angular Transfer State non scala: ogni thread NodeJS mantiene cache indipendente — meno hit, alto consumo di memoria, altamente inefficiente

La soluzione

CDN (CloudFront) come standard industriale con caching differenziato per tipo di client

Utenti umani

Vogliono contenuto aggiornato → cache breve

Crawler

Soddisfatti con contenuto di giorni/settimane → cache lunga

- Contenuti statici (favicon, .js, .css, font, immagini) cachati a livello CDN
- Sitemap e robots.txt serviti dalla CDN senza colpire l'applicazione
- Disabilitazione transfer state inefficiente (PR condivisa con la community DSpace)
- Risultato: drastica riduzione delle richieste che raggiungono l'applicativo

WAF e protezione dal traffico malevolo



Richiesta ▶ WAF ▶ CloudFront ▶ Applicazione

Blocco bot malevoli

Scanner automatici e richieste wp-login.php, path WordPress bloccate prima di raggiungere l'applicazione

Protezione DDoS

Protezione DDoS integrata a livello infrastrutturale per assorbire attacchi volumetrici

Fail fast per 404

Richieste a risorse inesistenti gestite immediatamente senza caricare Angular

GuardDuty

Amazon GuardDuty per monitoraggio continuo della sicurezza dell'infrastruttura

Riduzione del carico computazionale = Riduzione del consumo energetico

Off-load del batch processing

Il problema nativo DSpace

- ✗ Script CLI pesanti (cronjob) con alto consumo CPU/memoria
- ✗ Processi eseguiti sullo stesso JVM della REST API
- ✗ Competizione per risorse tra batch e richieste utente
- ✗ Over-provisioning necessario per gestire i picchi batch

La soluzione 4Science

- ✓ Disaccoppiamento scheduling/executor
- ✓ REST API schedula → pod temporaneo esegue
- ✓ Pod creati on-demand, distrutti al completamento
- ✓ REST API resta reattiva e performante

Flusso di esecuzione

REST API schedula → K8s crea pod temporaneo → JVM dedicato esegue batch → Pod distrutto al completamento

AI come opportunità

Server MCP (Model Context Protocol) open-source per l'ecosistema DSpace
Fornisce un accesso efficiente ai tool AI al repository riducendo gli sprechi sia lato repository che lato LLM

Linguaggio naturale

Connette assistenti AI ai repository DSpace tramite conversazione

Dual-runtime

Locale (stdio per Claude, Cursor) e cloud (AWS Lambda)

Ricerca documenti

Ricerca e navigazione del repository via AI

Gestione item

Creazione item, aggiornamento metadati, workflow submission

Efficienza operativa

Meno operazioni manuali = meno risorse sprecate

Workflow agentici

DSpace diventa sistema abilitato per workflow AI autonomi

GitHub: https://github.com/4Science/dspace-mcp

Risultati concreti e contribuzioni

1M

req/min

Picco gestito
(600x il traffico normale)

0

interruzioni

Servizio continuo
durante il picco

299

EUR/mese

DSpaceWay SaaS
accessibile a tutti

Contribuzioni open-source condivise con la community

SSR

Internal routing — riduzione milioni di richieste interne

Cache

Ottimizzazione caching e disabilitazione transfer state

Crawler

Performance fine-tuning per gestione traffico crawler

S3

Presigned URL per download diretto da S3 assetstore

DSpace-MCP

Server MCP open source per integrazione con AI

In roadmap per la contribuzione: off-load processi • SOLR Cloud • K8s leader election

Attenzione al nostro impatto: si inizia con la consapevolezza

4SCIENCE
Share your knowledge

Carbon emissions summary Info

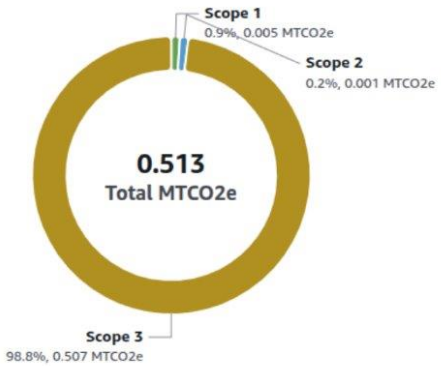
Provides your estimated carbon emissions calculated using both the market-based (MBM) and location-based (LBM) methods and the estimated savings from using AWS. The savings represent the estimated carbon emission reduction due to AWS's carbon-free energy purchases. The values in this page are dynamically updated based on the filters selected.

Estimated MBM emissions
0.513 MTCO2e

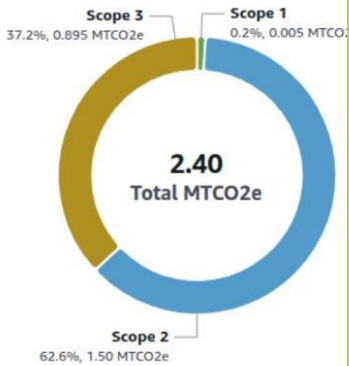
Estimated LBM emissions
2.40 MTCO2e

Estimated emissions savings
1.89 MTCO2e

Emissions by scope (MBM) Info



Emissions by scope (LBM) Info

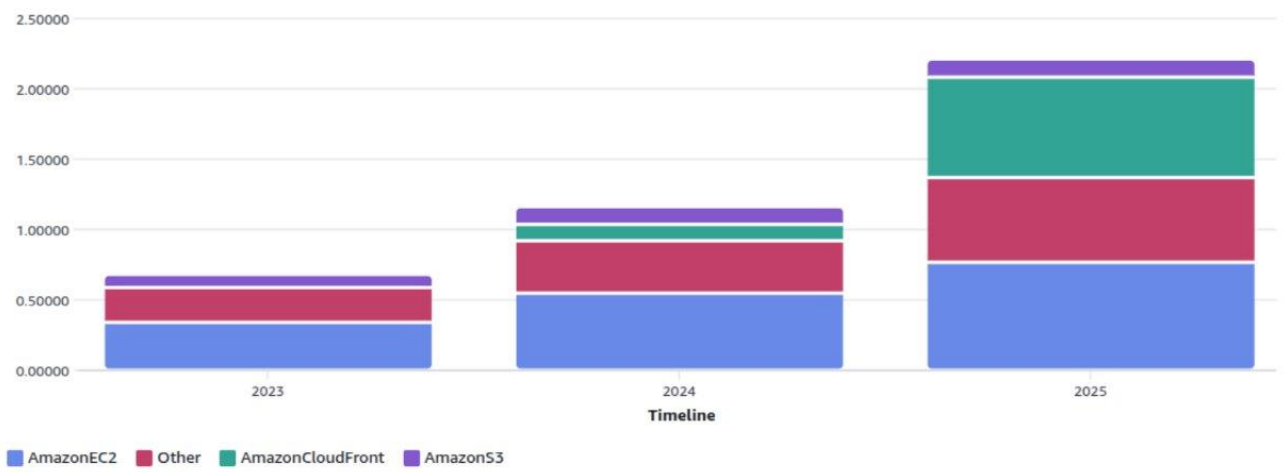


Water withdrawals summary Info

Provides the estimated water withdrawals associated with your usage of AWS. The values in this page are dynamically updated based on the filters selected.

Estimated total water withdrawals from using AWS services
4.06 m^3

Water withdrawals



Lavoro remoto e open source come scelte sostenibili

Lavoro remoto

- Pratica aziendale consolidata in 4Science
- Meno pendolarismo = meno emissioni di gas serra
- Meno spazi ufficio = meno energia per climatizzazione
- Riduzione impatto sulle infrastrutture stradali

Open Source

- Riutilizzabile, estensibile, mantenuto collaborativamente
- Ciclo di vita esteso, meno spreco di risorse
- Evita ri-sviluppo ripetuto → meno consumo computazionale
- 4Science contribuisce codice sostanziale alle comunità globali

Supporto alla Sustainable Libraries Initiative: donazione e visione condivisa per biblioteche sostenibili

Conclusioni

Hosting green + CDN/WAF + Autoscaling K8s + Open Source + AI = Repository sostenibili

Un modello replicabile per l'intero ecosistema della ricerca aperta

"Combinare innovazione tecnologica con responsabilità ambientale e sociale"

Contatti

4Science S.p.A.
www.4science.com
info@4science.com

Andrea Bollini: 0000-0002-9029-1854

Sedi

US Kansas City, USA

IT Milano, Italia

IT Roma, Italia

IT Ravenna, Italia

Certificazioni: ISO 27001, 27017, 27018, 22301, 20000-1, 9001 • CSA Trusted Cloud • AWS Partner