Valutazione delle pratiche di Open Science attraverso l'IA: il caso di SCOAP^3

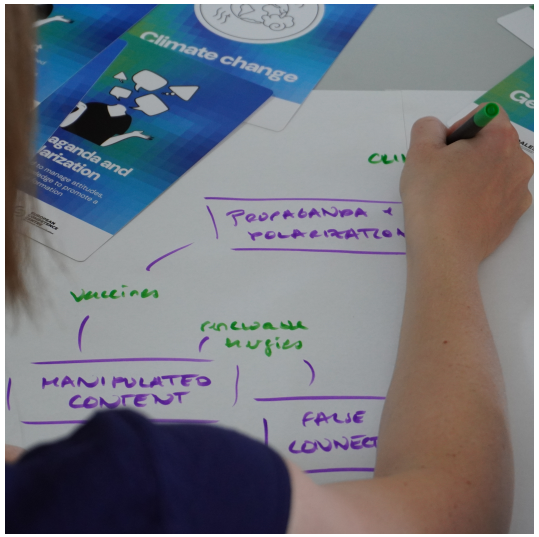
Jacopo Zuppa, 03/11/2026

SISSA Medialab









SISSA Medialab

- **Data curator** e **data analyst** presso Sissa Medialab S.r.l
- Società *in-house* della **Scuola Internazionale Superiore di Studi Avanzati** (SISSA)
- Pubblicazione di **riviste internazionali** di fisica e comunicazione della scienza
- Progettazione di **interventi educativi**, **masterclass** e **progetti europei** nell'ambito della comunicazione scientifica

Contenuto dell'intervento

- Progetto di tesi del **Master in Data Management and Curation** svolto nel corso di uno stage presso SISSA Medialab
- Analisi di possibili cambiamenti associati alle politiche di Open Science introdotte dallo **Sponsoring Consortium for Open Access Publishing in Particle Physics** (SCOAP^3) nel 2025
- Consorzio formato da migliaia di biblioteche, istituzioni ed enti finanziatori e coordinato dal **CERN**
- **Obiettivo**: convertire le maggiori riviste nell'ambito della fisica ad alte energie ad un modello ***open access***

MDMC
Master in Data Management and Curation

MASTER IN DATA MANAGEMENT AND CURATION

Building a FAIR-aligned data-processing pipeline for investigating research artifact availability in high-energy physics studies

Supervisor(s):
Dr. Maria Teresa LEO

Candidate:
Jacopo ZUPPA

2025–2026

SISSA

AREA SCIENCE PARK

Open Science Elements

- Serie di pratiche centrali ad un approccio ***open science***
- Al fine di incentivare un approccio OS, SCOAP^3 eroga **finanziamenti aggiuntivi** basati sull'adozione di queste pratiche
- Focus del progetto: verificare ***Dataset Linking*** e ***Software Linking*** su larga scala
- Requisito **ideale**: dati e codice apertamente disponibili
- Requisito **minimo**: inclusione di *Data Availability Statements* e *Code Availability Statements* nelle pubblicazioni

SCOAP³ Open Science Elements

2025–2027 · Requirements and evaluation criteria

ORCiD integration
Support all authors' identifiers throughout publishing; validate submitting authors' ORCiDs via API. Wider coverage earns extra credit.

ROR integration
Include persistent institutional identifiers in article affiliation metadata.

Public Peer Review
Offer transparent reviewing with publicly accessible reports made citable through DOIs.

Dataset linking
Require public data sharing or, at minimum, a data availability statement.

Dataset linking (2)
From 2026: reward annual growth in the share of articles linking public datasets through references or metadata.

Software linking
Require public software sharing or, at minimum, a software availability statement.

Software linking (2)
From 2026: reward annual growth in the share of articles linking public software through references or metadata.

Depositing of detailed metadata in Crossref
Deposit required metadata (e.g. identifiers, affiliations, licences, dates, references); richer deposits (e.g. abstracts, funders) earn extra credit.

Excellence in Accessibility
Websites and articles meet WCAG A/AA; AAA compliance earns extra credit.

Disclosure on SCOAP³ Community Values
Disclose diversity/equity/inclusion, sustainability, privacy, financial transparency, referee recognition/compensation and publication transparency. Disclosure is rewarded, not adherence.

Source: scoap3.org/journals-2025-2027/open-science-elements/

Soluzione tecnica proposta

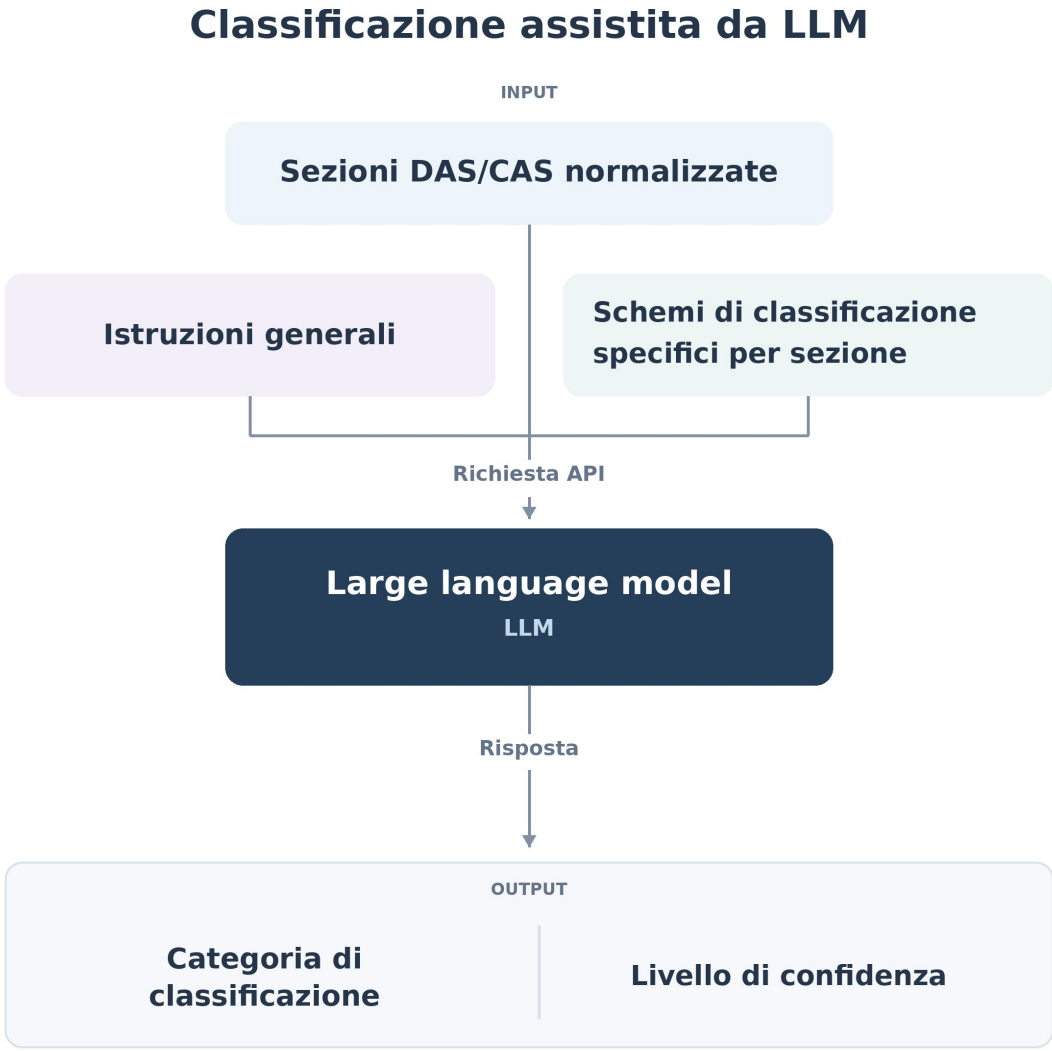
- Software per **verificare la presenza** di DAS/CAS e **classificarne il contenuto** attraverso l'IA
- Pipeline **modulare** di elaborazione dati
- **Input**: PDF delle pubblicazioni scientifiche
- **Output**: sezioni DAS/CAS e categorizzazione dei contenuti

Pipeline di elaborazione delle pubblicazioni



Meccanismo di classificazione

- La classificazione è basata sull'utilizzo di un ***large language model***
- Al modello vengono fornite le **sezioni DAS/CAS**, delle **istruzioni generali** e degli **schemi di classificazione**
- Il modello risponde con una **categoria di classificazione** e un **livello di confidenza**



Categorie di classificazione ed esempi reali

Condizioni di accesso tipiche ai dati e al codice

Nessun dato / nessun codice
Nessun dato o codice associato allo studio.

Accesso aperto
Dati o codice disponibili pubblicamente.

Accesso ristretto
Dati o codice disponibili su richiesta.

Accesso parziale
Solo una parte è disponibile pubblicamente.

Nessun accesso
Dati o codice non disponibili.

Non chiaro
Disponibilità ambigua o non specificata.

Non corretto
Contenuto non pertinente alla sezione.

DAS · DATI

Accesso aperto

“This article has associated data in a data repository. Available at https://doi.org/10.25835/2ovdfuq6.”

I dati sono dichiarati disponibili in un repository.

DOI: 10.1007/JHEP12(2025)117

CAS · CODICE

Accesso ristretto

“The code used in this paper is available from the corresponding author on reasonable request.”

Per ottenere il codice occorre contattare l’autore.

DOI: 10.1007/JHEP09(2025)116

DAS · DATI

Accesso parziale

“The main data supporting this study’s findings are available within the article. More data are available from the authors upon reasonable request.”

Parte dei dati è nell’articolo; altri dati sono su richiesta.

DOI: 10.1103/PhysRevD.111.074504

CAS · CODICE

Nessun accesso

“[…] The code analysed during the current study is not publicly available because it is still in use on other works, and we are improving it to make the code more efficient.”

Codice non condiviso perché ancora usato e in corso di ottimizzazione.

DOI: 10.1140/epjc/s10052-025-14224-8

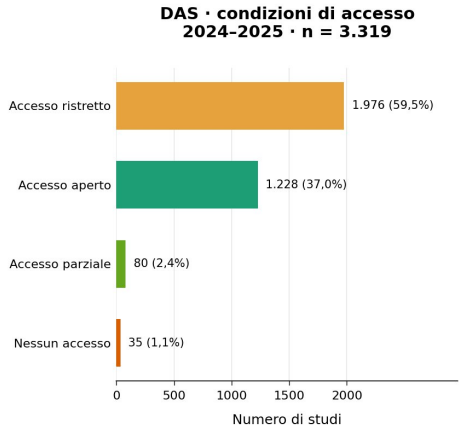
Esempi dalle registries del progetto · 2 DAS e 2 CAS · […] indica un’omissione

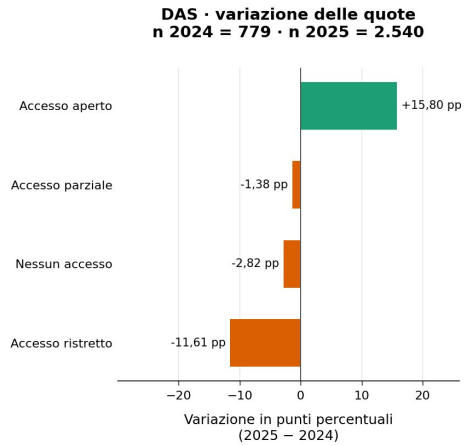
Analisi e risultati – cambiamenti tra 2024 e 2025

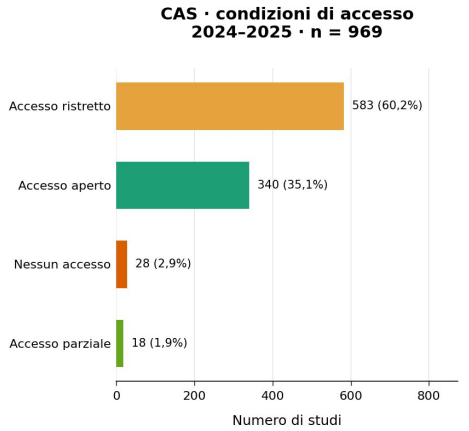
- Analizzato un campione di **17.076 articoli** (n = 8.442 nel 2024, n = 8.634 nel 2025) da **11 riviste** di fisica delle alte energie
- Aumento netto di **DAS** (+ 56,2 pp) e **CAS** (+ 30,9 pp) nel 2025
- Tra gli studi con dati associati, la categoria di accesso più comune è risultata essere l'**accesso ristretto**, sia per i dati (n = 1,976, 59,5%) che per il codice (n = 583, 60,2%)
- MA aumento visibile di dati (+15,8 pp) e codice (+16,2 pp) dichiarati come **apertamente disponibili** nel 2025 negli studi con artefatti associati

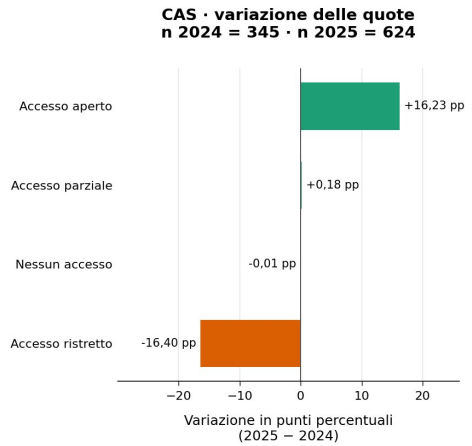
Condizioni di accesso dichiarate

Studi con dati o codice associati e condizioni di accesso chiare









Esclusi: nessun dato/codice, non chiaro e non corretto.
Per i CAS si considera il codice prodotto dagli autori.

Applicazioni - Open Science

- Elaborazione, verifica e classificazione automatica di **volumi elevati** di pubblicazioni scientifiche
- Analisi delle **condizioni di accesso** più comuni dichiarate per dati e codice
- Studio degli **andamenti annuali** di metriche legate a pratiche di Open Science
- Arricchimento dei **metadati esistenti** delle pubblicazioni
- Ridefinizione delle **linee guida** sul contenuto di DAS/CAS

Monitoraggio e linee guida

Un processo ripetibile per seguire cambiamenti nel tempo

17.076 articoli
Corpus SCOAP³

→

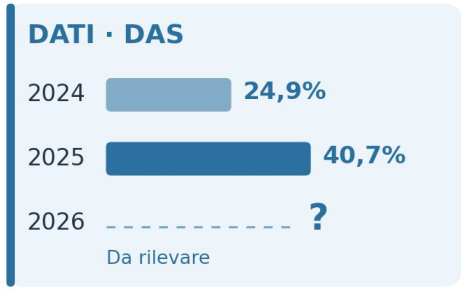
DAS e CAS
Analisi dei testi

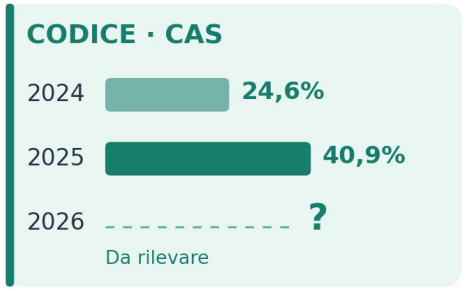
→

Indicatori
Confronto annuale

Dati e codice dichiarati ad accesso aperto

Studi con artefatti associati e condizioni di accesso chiare





Dai casi problematici a istruzioni più chiare

UN DAS REALE

“The data that support the findings of this article are openly available.”

Dati dichiarati aperti, senza link o DOI ai dati nel DAS.

DOI dell’articolo: 10.1103/bc5h-8jm3

PROPOSTA PER LE LINEE GUIDA

Quali artefatti?
Indicare i dati disponibili

Quali condizioni?
Esplicitare come accedere

Dove trovarli?
Inserire un link o DOI ai dati

Applicazioni - classificazione con LLM

- Il meccanismo di classificazione: schemi sviluppati da un esperto + large language model è risultato **altamente efficace**
- Accuratezza rispetto ad **annotazioni manuali:**
  - DAS: **98,8**% su 500 sezioni
  - CAS: **95,8**% su 500 sezioni
- Il modello utilizzato ha dimostrato la capacità di comprendere e applicare regole di classificazione **relativamente complesse**
- Meccanismo potenzialmente adattabile
  - Ad altri **contesti disciplinari** in cui vengono utilizzati DAS e CAS
  - Ad altri **contesti di monitoraggio di requisiti** in testi e pubblicazioni

Dove riutilizzare il metodo?

Possibili estensioni · Esempi illustrativi da validare

RICERCA BIOMEDICA

Condizioni di accesso ai dati

"I dati individuali sono accessibili previa approvazione del comitato di accesso."

POSSIBILE OUTPUT

Accesso controllato

PSICOLOGIA

Dichiarazioni di preregistrazione

"Lo studio è stato preregistrato nel registro indicato, con identificativo ABC-123."

POSSIBILE OUTPUT

Dichiarazione e ID presenti

MONITORAGGIO ISTITUZIONALE

Riferimenti ai finanziamenti

"Questo lavoro è stato finanziato dal programma di ricerca XYZ."

POSSIBILE OUTPUT

Finanziatore presente; codice del grant assente

Testo + schema esperto → classificazione → validazione

Testi inventati a scopo illustrativo; gli output riguardano quanto dichiarato.

Grazie dell’attenzione!

SISSA Medialab

Per **quesiti e/o chiarimenti** potete rivolgervi a me personalmente, o contattarmi per mail all’indirizzo **jacopo.zuppa@sissamedialab.it**

Slides: 10.5281/zenodo.23692096

Software: https://github.com/Appianibus/MDMC_Internship_Project_Zuppa

Dati: https://zenodo.org/records/20135209

Tesi: https://zenodo.org/records/22661028