

Compartir dades de recerca: algunes reflexions

Taula rodona: *Dades de recerca*
(Institut d'Estudis Catalans, 24 d'abril de 2014)



Jordi Ardanuy

jordi_ardanuy@ub.edu



Projecte Genoma Humà: acords de les Bermudes

- ▶ Alliberament automàtic de conjunts de seqüències més grans d'1 KB (preferentment dins de les 24 hores següents)
- ▶ Publicació immediata de les seqüències finalitzades
- ▶ Domini públic de la seqüència per a facilitar el lliure accés tant per a recerca com desenvolupament

La informació obtinguda del PGH és de lliure accés per a qualsevol persona i es pot consultar de manera gratuïta

1996

Per a què?

- ▶ Reutilització de les dades
 - Nous usos per a les dades
 - Promoure la innovació i els usos comercials
 - Redueix el cost de la duplicació de la recopilació de dades
 - Conduex a noves col·laboracions entre els usuaris i els creadors de dades
- ▶ Reconeixement per als investigadors (output)
- ▶ Recursos importants per a la formació i l'educació
- ▶ Maximitza la transparència i la rendició de comptes en el cas de recerca amb fons públics.
- ▶ **Revisió ex-post (escrutini resultats de recerca)**
 - Encoratja la millora i validació dels mètodes de recerca

Diferents realitats



Study Group for Data Preservation and
Long Term Analysis in High Energy Physics



THE UK'S LARGEST COLLECTION OF DIGITAL RESEARCH DATA
IN THE SOCIAL SCIENCES AND HUMANITIES

Preservació en HEP: un punt de partida

«Les **dades** dels experiments de física d'alta energia es recullen amb un significatiu esforç financer i humà i són majoritàriament úniques. Tanmateix, fins fa poc no existia una estratègia coherent per a la preservació de les dades i la reutilització; i, molts conjunts de dades importants i complexos, **simplement s'han perdut**» (2012)

R. Kogler; D. M. South and M. Steder, on behalf of the ICFA DPHEP Study Group Data preservation in High Energy Physics. «Data preservation in High Energy Physics». *Journal of Physics: Conference Series* 368 (2012) 012026 doi:10.1088/1742-6596/368/1/012026

Particularitats de la HEP

- ▶ **Gran volum de dades prèvies al processat (Raw data)**
- ▶ Diferents capes de dades processades de complexitat descendent
- ▶ **Programari específic de simulació, reconstrucció i anàlisi**
- ▶ Diferents nivells de documentació
- ▶ **Gran quantitat de recursos humans implicats amb col·laboracions a diferents nivells**
- ▶ **Experiments únics de llarga durada**

Cal preservar:

- tot el programari de simulació i reconstrucció
- les dades sense processar

Particularitats repositoris CC.SS

▶ Heterogeneïtat d'aproximacions i disciplines

- Recerques quantitatives i qualitatives
- Contextos de recopilació de dades molt diferents
- Varietat de metodologies de recopilació de dades
- Reaprofitament de dades des d'altres disciplines

▶ Recopilació de dades sobre individus

- Consentiments
- Transformar a formes anònimes
- Accés a les dades amb diferents privilegis

Algunes implicacions

- ▶ Cal preveure un pla per a compartir dades junt amb el pla de recerca i publicació de resultats
- ▶ Increment de les tasques a desenvolupar
- ▶ Possibilitat de reconeixement fins i tot quan hi ha resultats negatius, dificultats de publicació d'articles, etc
- ▶ Possibilitat de participació de documentalistes de les institucions
 - Formació del personal de recerca
 - Participació en el procés de definició del pla per a compartir dades i col·laboració en tot el procés
 - Facilitar suport en les qüestions ètiques i legals
 - Facilitar suport en el procés de compartir dades en recerques completades (actuals o antigues)