Pla de treball 2021 pel repositori de dades de recerca

(Doc.CO21/04) (4 RDM\Repositori de dades\F4 RDC\Dataverse\2021 PlaTreball2021_desenvolupat.docx,02.03.21)

1. Antecedents

Entre el setembre i desembre de 2020, s'ha posat en funcionament de forma pilot el repositori de dades de recerca i aquest ja publica dades que es poden veure en portals internacionals. En data de 29.01.21 es va fer una reunió dels tècnics participants en la fase pilot i es va validar el repositori com a apte per publicar-hi dades de recerca de manera FAIR. En data de 17.02.21 aquests resultats es van presentar als vice-rectors de recerca de les universitats membres del CSUC que formen la comissió de ciència oberta i es va acordar l'inici de l'ús del repositori per a la publicació de dades. En aquesta mateixa reunió es va exposar de forma sumària el Pla de Treball per a l'any 2021 que de forma més detallada es desenvolupa en aquest document.

El programari de codi obert emprat és Dataverse, desenvolupat per Harvard i sostingut per una comunitat internacional (dins l'àmbit europeu se'n destaca el de Noruega o els Països Baixos). Aquest programari pensat específicament per a dades de recerca disposa d'una estructura federada que permet la personalització de cada instància (d'estructura, logotip, descripcions), assigna el DOI com a identificador persistent i es registra automàticament a DataCite, permet la inclusió de dades restringides o tancades, transforma en el moment del dipòsit alguns formats de fitxer a formats de preservació i ofereix un versionat automàtic de les dades, entre moltes altres funcionalitats.

2. Pla de treball per al 2021

A continuació es detallen les principals accions a dur a terme al 2021 per semestres.

Primer semestre

- **Elaborar i signar un conveni amb cada institució**

S'ha encarregat l'elaboració d'un conveni a l'assessoria jurídica Legalment especialitzada en dret a la informació i de les noves tecnologies, protecció de dades de caràcter personal, propietat intel·lectual, administració electrònica, comerç electrònic i règim jurídic de les telecomunicacions.

Aquest conveni ha de vehicular la relació entre el CSUC i la institució que utilitzi el repositori, així com determinar els drets i deures de cadascuna de les parts.

- **Tenir unes normes generals del repositori i pròpies d'instància**

Associat al conveni que descriurà els drets i deures de cada membre que usi el repositori de dades, cal disposar d'unes normes generals del repositori i d'altres pròpies d'instància.

Aquestes normes generals hauran de determinar qui i com pot dipositar en aquest repositori, quin tipus de dades i amb quines mides, així com que cada institució haurà de garantir que compleix amb les normes i processos establerts i avalar la publicació de dades, entre d'altres.

- **Aprovar el procés de curació de dades (amb accions mínimes a fer)**

La xarxa Data Curation Network de curadors de dades dels Estats Units ha desenvolupat una checklist amb els processos que cal que un curador de dades porti a terme per tal de millorar la qualitat d'un dataset.

Per tal de potenciar cadascun dels principis FAIR i millorar la qualitat dels datasets dipositats al repositori de dades es considera que cal avalar el document "REVISAT: Criteris orientadors per curar conjunts de dades de manera FAIR" que és una traducció al català de la proposta americana. A la vegada, que caldrà determinar quines accions seran obligatòries a fer per totes i cadascuna de les instàncies que participen al repositori a fi i efecte que el repositori assoleixi uns mínims de qualitat i pugui optar a rebre un segell de qualitat.

- **Definir el nom del repositori**

Durant la fase de desenvolupament i pilot, el repositori de dades s'ha anomenat de diferents maneres fet que dificulta la identificació del servei.

Per aquest motiu cal que, abans que sigui difós de manera extensa, ja disposi d'un nom i una imatge gràfica que l'identifiqui.

- **Registrar el repositori a re3data**

Per millorar la descobribilitat del repositori, a més a més dels cercadors habituals és necessari registrar-lo en el portal re3data: Registry of Research Data Repository.

Es tracta d'un dels principals catàlegs destinats específicament a recollir repositoris per a dades de recerca. A més a més, la Comissió Europea recomana als investigadors l'ús d'aquest portal dins les seves directrius "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020".

- **Interoperabilitat**

El repositori ha de permetre intercanviar informació amb d'altres sistemes, i concretament amb d'altres repositoris que disposin les institucions que hi participen.

D'aquesta manera, cal que les metadades s'*exportin* en diferents formats però, concretament en Dublin Core perquè puguin ser ingestades en els sistemes d'informació i/o repositoris institucionals de les universitats i centres de recerca a través del protocol OAI-PMH. Actualment s'està fent una prova pilot amb la Universitat de Lleida.

D'altra banda, per aquelles institucions que ja tinguin dipositades dades de recerca en els seus repositoris institucionals i vulguin traspassar i disposar de tota la producció de dades de recerca institucionals des d'un únic lloc cal que el repositori pugui *importar* tot aquest conjunt dades, incorporant l'identificador persistent original. Actualment s'està fent una prova pilot amb la Universitat de Barcelona que ha de permetre aquest semestre incorporar els fitxers que ara té al seu repositori alhora que es troba una metadada identificadora de la ruta prèvia (és a dir, del Handle designat al dataset).

- **Dipositar datasets de mida gran**

Des de l'inici del projecte, s'ha indicat que es permetria dipositar per defecte fins a 10GB per dataset (tot i que aquest límit podria ampliar-se properament). De totes maneres, en alguns casos, existeixen casos en els que, o bé per la singularitat del dataset o per raons de justificació de finançament, cal poder dipositar datasets amb una mida superior.

Per aquest motiu, el repositori ha de permetre disposar d'un sistema robust per dipositar fitxers de mida gran o molt gran (fins a 100GB o la mida que s'acabi determinant). Actualment, s'ha fet una prova pilot amb la Universitat Internacional de Catalunya però cal procedimentar tota aquesta tasca.

- **Autenticació**

Les institucions participants al pilot del repositori han manifestat la necessitat de disposar d'un mètode que permeti als usuaris procedir a una única autenticació (single sign-on). Atès que el sistema UNIFICAT està estès entre les universitats i els centres de recerca i s'ofereix des del CSUC cal prioritzar aquesta funcionalitat.

A més a més, caldrà valorar la necessitat d'implantació d'altres sistemes comuns com l'autenticació a través d'ORCID, Google, GitHub, entre d'altres.

Segon semestre

Per al segon semestre es marca un objectiu específic (la incorporació d'un curador de dades) i tres de genèrics. Els objectius genèrics s'hauran d'assolir-se en anys successius, però per al 2021 es vol avançar en alguns aspectes. De forma general, aquests objectius suposaran:

- En preservació, definir quins processos cal dur a terme per garantir la reutilització futura dels datasets i participar en alguna xarxa de preservació
- En qualitat, recollir la informació sobre els requisits de CoreTrustSeal de cara a certificar el repositori de confiança
- En interoperabilitat, tenir determinats els circuits d'exportació i importació de metadades amb d'altres sistemes (repositoris institucionals, CRIS, l'EOSC...) de manera que els conjunts de dades del repositori siguin trobables a nivell català juntament amb altres productes de recerca com articles o projectes

- **Preservació**

S'entén per preservació digital "l'aplicació de tècniques i mètodes que permeten garantir que la informació emmagatzemada digitalment en qualsevol tipus de format, programa, maquinària o sistema, continuïn sent accessibles en el futur"[1].

Per oferir unes prestacions mitjanes-altes de preservació cal disposar de diferents còpies geogràficament separades. Per aquest motiu, cal començar a estudiar les opcions de col·laboració amb d'altres consorcis nacionals i internacionals.

Entre les diferents opcions, hi hauria la possibilitat de contractar la conservació d'alguna còpia en servidors externs dins la Unió Europea (per complir amb la legislació GDRP) o bé es podria arribar a l'acord amb d'altres consorcis de conservar còpies mútuament.

- **Certificació del segell CoreTrustSeal**

Per tal d'oferir confiança i seguretat als investigadors, cal disposar d'un segell de certificació del repositori com el de CoreTrustSeal.

Diferents repositoris amb el programari Dataverse ja compten amb aquest segell, tot i que indiquen que la seva obtenció ha estat dificultosa. De totes maneres, el *Global Dataverse Community Consortium* ofereix als seus membres, com el CSUC, suport per a l'obtenció d'aquest segell.

[1] Viquipèdia 2021 https://ca.wikipedia.org/wiki/Preservaci%C3%B3_digital

- **Interoperabilitat**

A més a més del que s'ha indicat pel primer trimestre, el repositori també ha de permetre exportar les metadades d'altres sistemes, com podria ser als CRIS, al Portal de la Recerca de Catalunya o a l'EOSC.

També caldrà que pugui recollir metadades d'altres repositoris de dades (per exemple, recollir tota la producció d'investigadors catalans que es troben dipositats en repositoris disciplinaris).

- **Tenir un data curator**

Per millorar la qualitat dels datasets dipositats al repositori es considera clau que hi hagi una persona encarregada a curar les dades, tal i com està emergint en l'àmbit internacional la figura del *data curator* o *data stewardship*.

Aquesta plaça ja ha estat incorporada als pressupostos del CSUC per al 2021 però cal iniciar el procés de selecció i contractació.