Consorci de Serveis Universitaris de Catalunya

Pla de treball i memòria econòmica per fer un repositori de dades de recerca al servei de les universitats de Catalunya

(Doc.ACO 19/10; 4/RepDades/F4/1909PlaTreballMemoEco; EcMp 26.09.19)

Resum executiu

Importància de les dades. En els darrers anys les dades de recerca han rebut l'atenció creixent de la comunitat científica, que considera que cal fer-les públiques. Des del 2016, totes les universitats de Catalunya tenen serveis de suport a la gestió de dades de recerca, però no disposen de la infraestructura adient per publicar les dades de recerca seguint els principis FAIR que la Comissió Europea demana.

Activitats fetes. El novembre del 2017, la Comissió Funcional de l'àrea de Ciència Oberta del CSUC (ACO) va acordar encarregar un informe que determinés els requeriments funcionals raonables que hauria de tenir un repositori de dades per complir amb els requisits FAIR. L'informe constata que la publicació de dades de recerca és una necessitat encara poc madura, que requereix desenvolupar bones pràctiques en la curació de les dades i que aquestes només es poden desenvolupar si es compta amb un repositori. L'informe es va aprovar en una reunió de la Comissió del març del 2019, també es va acordar elaborar un pla de treball i acompanyar el pla amb la memòria econòmica corresponent.

Pla de treball. Seguint les recomanacions de l'informe, aquest nou servei requeriria:

- Programari. Es proposa usar el programari de codi lliure Dataverse per la característica que té de permetre una construcció federada de repositoris.
- Servidors i emmagatzematge. Cal adoptar una solució adaptable a les necessitats reals (inicialment incertes però creixents). El CSUC ja disposa d'una infraestructura escalable.
- Preservació. Per tal que les dades siguin accessibles d'aquí uns anys cal gestionar-les sota un model de preservació Open Archival Information System (OAIS) i fer-ne còpies que han dipositar-se a diferents llocs.
- Curació de dades. Fer públiques les dades de recerca en forma FAIR requereix la seva preparació i descripció de manera determinada. Això només es pot fer amb els coneixements i feina d'un 'data curator' o 'data officer'.
- Identificadors persistents. Als conjunts de dades se'ls ha d'assignat identificadors persistents i es proposa usar DOIs com a tals.
- Formació i promoció. Cal dur a terme activitats de formació i de promoció de l'ús de la infraestructura que es creï.

Cost. Es considera la creació d'una infraestructura comuna, tal com la que tenen les institucions i països més avançats en aquesta matèria, amb un cost estimat de 300.000€ per un període de dos anys.

Calendari. El projecte tindria les fases següents: (1) Preparació: tres mesos, (2) Fase pilot: sis mesos, (3) Obertura del servei: quinze mesos (4) Avaluació: els tres mesos finals.

Àmbit de servei. El servei i la infraestructura s'adrecen a les necessitats de les universitats però podria cobrir una part de les necessitats dels centres CERCA. El servei hauria de desenvolupar-se en coordinació amb el d'altres institucions que a Catalunya facin la mateixa funció.

1. Antecedents, necessitats i recomanacions

En els darrers anys, les dades recollides, generades o utilitzades en el desenvolupament dels projectes de recerca han rebut l'atenció de la comunitat científica i dels òrgans gestors de la recerca. Es considera que fer-les públiques de manera FAIR (trobables, accessibles, interoperables i reutilitzables), beneficia la ciència i representa estalvis per a la recerca.

Totes les universitats de Catalunya tenen serveis de suport a la gestió de dades de recerca, però no disposen de les prestacions que la Comissió Europea (CE) demana per publicar les dades de recerca. La Comissió Funcional de l'àrea de Ciència Oberta del CSUC (ACO) va encarregar l'informe "FAIRxFAIR: requeriments factibles, assolibles i implementables per a un repositori de dades de recerca FAIR" que es va desenvolupar consultant experts tant dins de l'àmbit del CSUC com de fora, examinant experiències similars i estudiant els principals documents publicats sobre aquest tema.

La part principal de l'informe se centra en la descripció dels 25 requeriments funcionals que es consideren mínims per tal de garantir que la infraestructura creada compleixi amb els requeriments FAIR. Alhora, a banda dels requeriments tècnics, destaca dos aspectes fonamentals relacionats amb la gestió de les dades científiques:

- les pràctiques de la gestió de les dades de recerca no estan encara del tot definides però s'estan convertint en estratègiques i cal posicionar-s'hi des d'un inici per treure'n el màxim avantatge,
- aquesta gestió es basarà en desenvolupar expertesa i bones pràctiques en curació de dades i altres, i que aquests desenvolupaments només es poden dur a terme tenint en funcionament un repositori amb el que treballar.

L'informe conclou amb tres recomanacions:

- Crear de forma immediata i a Catalunya, amb programari ja existent i de codi lliure, un repositori on es puguin publicar les dades de recerca i que compleixi amb els requeriments FAIR de la CE (coneguts o que s'estableixin en el futur) que permeti desenvolupar expertesa i bones pràctiques en la gestió de dades de recerca.
- Promoure i facilitar la publicació de les dades de recerca en obert amb accions en què la Universitat faci públic el servei de gestió de dades de recerca existent. L'acció de difusió hauria de fer-se entre les diferents unitats que vehiculen la recerca i amb la participació activa de les oficines i serveis de recerca.
- Fer formació sobre els conceptes de Ciència Oberta i, concretament, sobre gestió de dades de recerca. Seguint les directrius de l'Expert Group on FAIR Data de la CE, aquesta formació per ser eficaç hauria de distingir entre usuaris avançats, investigadors joves i personal de suport de les universitats.

2. Beneficis de la creació d'un servei de repositori de dades de recerca (RDR)

De forma addicional a allò expressat a l'informe, cal remarcar que la publicació en obert de les dades de recerca comporta, entre d'altres, els següents beneficis:

- Obrir les dades pot derivar en noves relacions i col·laboracions.
- Els conjunts de dades són cada cop més citats en publicacions, per exemple gràcies als DOI, que els fan trobables de forma independent als articles.

És en aquest sentit que, com a teixit, és de gran interès disposar ben aviat d'una infraestructura comuna que abordi la gestió de les dades FAIR en un sentit ampli (infraestructura, formació, bones pràctiques...). Aquest servei ha de ser considerat estratègic, ja que:

- Fomentarà la visibilitat de la recerca produïda a Catalunya
- Millorarà els indicadors dels investigadors (més citacions a través dels conjunts de dades en obert)
- Fomentarà noves col·laboracions
- Com a conseqüència de tot plegat, facilitarà l'obtenció de finançament i d'altres recursos competitius, com per exemple projectes H2020.

3. Elements per posar en funcionament el servei de repositori

A continuació es descriuen els elements necessaris per posar en funcionament el servei objecte del present pla de treball, fonamentalment basat en un repositori de dades de recerca (RDR) que permeti publicar les dades de manera FAIR.

3.1 Programari

No cal desenvolupar cap programari propi ja que n'hi ha tant de comercials com de codi lliure (vegeu les opcions a l'annex 1 de l'informe). L'informe recomana usar-ne un de codi lliure; d'entre els existents es proposa usar Dataverse: les seves prestacions actuals i garanties de desenvolupament són similars a les d'altres programaris de codi lliure, però, per al projecte català, té l'avantatge de permetre fàcilment una construcció federal de repositoris. Això significa que les dades es poden visualitzar a nivell de cada institució participant o de forma col·lectiva i que cada institució pot organitzar com vulgui les seves comunitats. Aquest es considera un punt clau del servei, per obtenir un elevat grau d'acceptació per part de les universitats catalanes. Com exemples basats en programari Dataverse es poden veure els repositoris cooperatius DataverseNL (Països Baixos), eCiencia Datos (Consorcio Madroño, Espanya) o DataverseNO (Noruega), entre d'altres.

Els costos d'instal·lació de Dataverse són mínims, però cal preveure, a banda de les habituals despeses de manteniment, un esforç com –sobretot- despeses de confecció d'APIs que

permetin la interoperabilitat del repositori amb la resta de components de l'ecosistema d'informació de recerca (especialment amb els CRIS de les universitats, el Portal de la Recerca de Catalunya i les que pugui definir l'EOSC).

3.2 Servidors i emmagatzematge

Tots els experts assenyalen que les necessitats en servidors per a l'emmagatzematge de dades és un factor clau del cost, i en conseqüència la sostenibilitat, d'aquest tipus de serveis. Això és degut no només a les dimensions que l'espai ha de tenir sinó, sobretot, per l'acumulació de les dades en el temps (al menys durant els 10 anys inicials) i per la necessitat de duplicar-lo o triplicar-lo a efectes de preservació.

A més, cal tenir en compte la incertesa actual: avui dia es fa molt difícil saber de forma precisa les necessitats d'emmagatzematge perquè la dinàmica actual no és de fer públiques les dades finals dels projectes de recerca i, per tant, no hi ha cap substrat real a partir del qual poder fer projeccions futures.

Essent un cost rellevant, es fa aconsellable implementar el servei sobre una infraestructura d'emmagatzematge fàcilment escalable a les necessitats reals del servei, conforme aquest vagi evolucionant. D'acord amb les previsions realitzades a l'informe, es considera oportú posar a disposició del projecte una infraestructura de fins a 100 TB, que haurien de ser suficients per als dos primers anys del projecte (vegeu les diferents projeccions a l'annex 1).

L'informe proposa concentrar-se en els conjunts de dades de no més de 10GB i donar així un servei "estàndard" per a conjunts de dades que requereixin d'una capacitat moderada d'emmagatzematge. També es donaria solució a conjunts més grans (amb un llindar a l'entorn dels 10TB, segons disponibilitat), establint uns costos addicionals segons la capacitat.

Els costos d'emmagatzematge es preveuen baixos en un estadi inicial fins que la pràctica de publicar les dades de la recerca es generalitzi, però creixents, conforme la comunitat científica vagi incorporant-se al servei. Un cop el servei es trobi en maduresa, el cost acumulat de l'emmagatzematge es compensarà per diversos factors:

- Les depeses d'emmagatzematge (i.e. la tecnologia) tendeixen a disminuir i avui es disposa de sistemes de baix cost per a dades que, tot i haver d'estar disponibles en línia, no requereixen d'una velocitat d'accés rellevant.
- El creixement es pot fer 'just-in-case', sense que calgui anticipar despesa si no és necessari.
- La possibilitat d'emmagatzemar les dades a diversos tipus de discs en funció que siguin més o menys consultats

3.3 Preservació

L'objectiu final dels repositoris de dades de recerca és garantir la reutilització de les dades, i aquesta depèn de forma molt important de la seva preservació. La preservació de fitxers de qualsevol tipus té bàsicament dos components:

- de curació de dades per tal de complir els requisits del model Open Archival Information System (OAIS)
- mecànics (entre d'altres, comprovació de la integritat de les dades i manteniment de múltiples còpies geogràficament disperses).

El model de preservació OAIS té la finalitat de permetre l'accés a llarg termini a la informació en format digital, obviant els possibles canvis tecnològics i, per això, entre d'altres recomanacions defineix les metadades que calen per a una adequada conservació de la informació. L'aplicació d'aquest model requereix la intervenció humana (vegeu, "Curació de dades", més avall), però aquesta es veu facilitada per programaris que sistematitzen el procés.

La integritat de les dades es comprova amb eines de checksum, però cal a més reforçar-la disposant de diverses còpies a diferents llocs en prevenció d'accidents i desastres naturals. Aquestes còpies no han de facilitar un accés immediat a les dades, sinó que han de servir de rèplica en cas de danys als fitxers originals; això n'abarateix el cost. En aquesta línia, es preveu disposar de dues còpies a llocs allunyats de l'RDR: per això podrien establir-se acords de col·laboració mútua amb entitats similars al CSUC En ambdós casos, caldrà assumir els costos que se'n derivin.

3.4 Curació de les dades

L'informe assenyala com element clau per tenir conjunts de dades públics de qualitat que aquests hagin estat 'curats' o 'editats' seguint estàndards, criteris i protocols que els facin FAIR: tanmateix, en l'estat actual de la gestió de dades de recerca, aquests estan encara en bona part en construcció i molts es definiran en base a l'experiència i les bones pràctiques acumulades en el procés de publicar les dades. En qualsevol cas:

- la curació de dades implicarà diferents activitats tant pel que fa a la selecció dels conjunts de dades com al seu enriquiment amb metadades i ontologies
- la preparació de les dades de recerca per a la seva publicació requerirà complir amb uns requisits generals i amb d'altres propis de les diferents disciplines científiques.

La curació de les dades és una tasca integral en la que han d'intervenir-hi els mateixos investigadors, però també personal de suport especialitzat. Cal tenir en compte que un dels objectius de l'RDR és justament el poder desenvolupar expertesa i bones pràctiques en aquest àmbit.

Conseqüentment, la creació i generalització de l'RDR requereix una persona experta dedicada íntegrament a importar bones pràctiques estrangeres, a facilitar l'intercanvi de les que es desenvolupin aquí i a fixar les que es consensuïn com a protocols propis. En aquest sentit, el servei preveu la contractació d'un tècnic especialista ('data curator' o 'data officer'), que amb, idealment, experiència en la gestió de dades de recerca.

3.5 Identificadors persistents

La interoperabilitat dels registres de les dades de l'RDR depèn dels estàndards amb que estiguin tractades i dels identificadors persistents que portin associades. El primer requeriment de l'informe indica que el repositori de dades hauria d'assignar DOIs com identificadors persistents. Aquests s'han d'assignar a cada fitxer de dades així com a les seves versions.

La principal organització sense ànim de lucre que proporciona DOIs per a les dades de recerca és DataCite. Les institucions que utilitzen el repositori Dataverse poden passar a formar part del Global Dataverse Community Consortium (GDCC) que es dedica a proporcionar un espai de col·laboració perquè s'aprofitin les economies d'escala en diferents casos, per exemple, en la provisió de DOIs de DataCite (per a més detall sobre el cost del DOI, veure l'annex 2).

3.6 Formació i promoció

La gestió de dades de recerca és una activitat nova que encara no està incorporada als processos habituals de la pràctica de la recerca. Cal, doncs, tal com recomana l'informe, proporcionar formació sobre els conceptes de Ciència Oberta i, concretament, sobre gestió de dades de recerca.

Aquesta formació ha de seguir les directrius del grup d'experts sobre dades FAIR de la CE i fer-se de forma diferenciada per a tots els implicats en la gestió de la recerca (usuaris avançats, investigadors joves i personal de suport de les universitats).

Per dur-la a terme, el servei preveu cursos presencials o telemàtics i l'elaboració de materials formatius. Finalment, es preveu promoure tant els serveis de suport a la gestió de dades de recerca que ja s'estan donant des de les universitats com l'RDR que es proposa crear.

4. Taula de requeriments i costos

En resum, la posada en funcionament del servei proposat d'RDR precisa de:

- Programari. Unes 700h de tècnic informàtic per al desplegament i la configuració inicials, i unes 300 h/any per a manteniment
- Emmagatzematge. Capacitat inicial per a 100TB. Es disposa de la capacitat per créixer fins a 250 TB per al 2n any, si bé aquesta no es reflecteix als costos presentats (ja que seria un cost a assumir en el cas que fos necessari)
- Curació de dades. Un tècnic especialitzat a temps complet.
- Identificadors persistents. Fer-se membre de l'organització DataCite i comprar un mínim de 10.000 DOIs.
- Preservació. Un programari que faciliti seguir el model OAIS i acords per a còpies remotes. Aquests recursos no caldrien fins el 2n any del projecte.
- Recursos de formació, promoció i generals

La següent taula mostra les despeses associades als recursos enumerats anteriorment:

Despeses				
	Inicials	**1r any**	**2n any**	**Total**
Programari	14.000,00 €	14.000,00 €	8.400,00 €	36.400,00 €
Emmagatzematge		5.000,00 €	15.000,00 €	20.000,00 €
Curació de dades		60.000,00 €	60.000,00 €	120.000,00 €
Identificadors	10.000,00 €	2.000,00 €	2.000,00 €	14.000,00 €
Preservació		15.000,00 €	15.000,00 €	30.000,00 €
Promoció i formació		15.000,00 €	15.000,00 €	30.000,00 €
Despeses generals		10.000,00 €	10.000,00 €	20.000,00 €
	24.000,00 €	**121.000,00 €**	**125.400,00 €**	**270.400,00 €**

Si l'RDR es volgués desenvolupar per a cada universitat (en comptes de fer-lo cooperatiu), el cost de cada component variaria i seria en cada cas:

- Programari. Lleugerament inferior, però per a cadascuna de les universitats
- Emmagatzematge. De forma agregada, majors, tant per economies d'escala com per percentatge d'ocupació (una solució compartida és molt menys sensible a variacions d'ocupació respecte a solucions individuals, i cal tenir en compte que per aquest projecte les incerteses són elevades)
- Curació de dades. Un tècnic especialitzat a mitja jornada per a cada universitat
- Identificadors persistents. De forma agregada, el mateix.
- Preservació. Les mateixes de programari per a cadascuna de les universitats, i similars (de forma agregada) per a les de còpies remotes.
- Promoció i formació. La meitat, però per a cada universitat.
- Altres. De forma agregada, similars.

El cost de cada repositori individual per 2 anys podria aproximar-se a uns 125.000€; és a dir, si fossin 5 les universitats que desenvolupessin un RDR, el cost per al sistema seria d'uns 625.000€. Per al cas de 8 universitats, aquest cost s'acostaria a 1.000.000 €.

5. Calendari de treball i termini d'execució

El repositori pot estar operatiu per treballar-hi de forma molt ràpida perquè els primers passos (instal·lació del programari Dataverse, contractació d'un tècnic i afiliació a DataCite) poden fer-se en un termini curt. Els elements que requeriran més temps són els d'establir mecanismes d'interoperabilitat entre l'RDR i els altres elements de l'ecosistema de recerca, l'establiment de bones pràctiques de curació de dades i la promoció de la publicació de dades de recerca entre els investigadors.

El projecte tindria les fases següents:

- Preparació (tres mesos): consistiria en la instal·lació del programari Dataverse, la contractació d'un tècnic i l'afiliació a DataCite
- Fase pilot (sis mesos): consistiria en la publicació de dades de recerca d'investigadors o grups de recerca preparats o motivats per fer-ho i l'establiment, a partir d'aquestes primeres experiències, d'uns primers protocols d'actuació
- Obertura del servei (quinze mesos): amb les bases prèviament establertes de les fases anteriors, caldria estendre el servei al màxim nombre possible d'investigadors
- Avaluació (els tres mesos finals): per analitzar el funcionament del servei i planificar la seva evolució de cara al seu futur.

L'informe FAIRxFAIR recull l'opinió dels experts que les dades de recerca es faran públiques a diferents repositoris en funció de la gran dimensió d'algunes i de l'existència de repositoris de dades ja existents i consolidats en algunes disciplines. Això vol dir que un RDR com el descrit pot no ser el millor lloc on publicar determinades dades de recerca produïdes a una universitat, però també significa que aquest pot servir també per fer públiques dades de recerca produïdes a centres de recerca o a organitzacions no universitàries que fan recerca. Alhora, un RDR com el descrit hauria de desenvolupar-se en coordinació amb el d'altres institucions que a Catalunya estiguin treballant en iniciatives similars (probablement, complementàries).

Annexos:

Annex 1. Càlcul suposat de projectes i datasets pel repositori de dades

Per tal de calcular les necessitats del repositori de dades caldria tenir dades sobre la quantitat de projectes que haurà d'emmagatzemar així com de l'ocupació de les dades de cada projecte. Degut a que la publicació de dades de recerca és encara un fenomen incipient, no es disposa de xifres que puguin extrapolar-se. Aquí es fa un càlcul basat en les dues suposicions: (A) en base els projectes europeus, i (B), en base al nombre d'investigadors.

El projecte canadenc Portage ha presentat recentment xifres dels repositoris de dades canadencs. Tal com fa l'informe FAIRxFAIR, també ells distingeixen entre datasets de dades molt grans i datasets de dades de la 'cua llarga'. Els que tenen recollits d'aquests darrers són uns 1.500 que ocupen menys de 2 TB. Segons això, la mitjana d'ocupació d'un dataset dels que ens ocupem seria de 1,33 GB.

A. Comptabilització a partir dels projectes (de dalt a baix)

Durant l'any 2018, les universitats catalanes han participat en un total de 92 projectes europeus H2020 (segons dades de RIS3-MCAT). D'aquests 92 projectes, 43 han estat coordinats per universitats catalanes. Considerem que, a priori, els coordinadors del projecte s'encarreguen de gestionar les dades.

Cal tenir en compte que:

- alguns projectes no generaran dades,
- que hi haurà projectes no europeus que voldran publicar les dades,
- que els projectes poden tenir una quantitat variable de datasets, i
- que els datasets poden tenir un pes variable.

De cara a tenir un rang d'ocupació probable, i basant-se en els 43 projectes europeus coordinats en un any per alguna universitat catalana, fem els supòsits següents:

- Projectes que no generaran dades. Es fan tres suposicions: no en generaran un 10%, un 20% i un 30%.

Hipòtesis 1	Hipòtesis 2	Hipòtesis 3
-10%	-20%	-30%
38,7	34,4	30,1

- Projectes no europeus que voldran publicar les dades al repositori. Es fan tres suposicions: seran el 50, el 100 i el 200% dels projectes europeus.

	Hipòtesis 1	Hipòtesis 2	Hipòtesis 3
	+50%	+100%	+200%
38,7	58,1	77,4	116,1
34,4	51,6	68,8	103,2
30,1	45,2	60,2	90,3

- Datasets generats per cada projecte. Es fan tres suposicions: es considera que el 50% dels casos tindrà 1 dataset, el 40% en tindrà 3 i el 10% 5. Així, el nombre de datasets que es generaran estarà entre el rang 99 (valor mínim) i 255 (valor màxim).

	Hipòtesis 1	Hipòtesis 2	Hipòtesis 3
38,7	127,71	170,28	255,42
34,4	113,52	151,36	227,04
30,1	99,33	132,44	198,66

- Pes dels datasets. Es fan tres supòsits: es considera que el 50% dels casos el dataset pesarà 1GB, el 30% 2 GB, el 15% 5 GB i el 5% 10 GB.

- A partir de les dades anteriors, l'espai d'emmagatzematge necessari està dins un rang d'entre 233 GBs (valor mínim) i 600 GBs (valor màxim).

	Hipòtesis 1	Hipòtesis 2	Hipòtesis 3
38,7	300,12	400,16	600,24
34,4	266,77	355,70	533,54
30,1	233,43	311,23	466,85

B. Comptabilització a partir dels investigadors (de baix a dalt)

Al juliol de 2019, hi ha 11.982 investigadors actius (segons dades del Portal de la Recerca de Catalunya). Es té en compte que:

- La UAB ofereix 200 GB d'espai de disc per a dades en brut a cada membre del cos de PDI i aquests no demanen més espai de disc que l'ofert.
- Aquesta no demanda indica que no es consumeix tot l'espai ofert. Se suposa que l'espai usat pot ser de 100GB/PDI.
- L'espai necessari per publicar les dades és molt menor que el que cal per gestionar-les. Se suposa que l'espai necessari per publicar les dades finals és del 10% respecte l'anterior.

De cara a tenir un rang d'ocupació probable, i basant-se en el nombre d'investigadors del PRC i de l'espai de disc ofert per la UAB, l'espai necessari per publicar dades és de 119.820 GB (117 TB).

C. Comptabilització a partir de la mitjana de Portage

De cara a tenir un rang d'ocupació probable, i basant-se en la mitjana de pes del dataset segons Portage (1,33 GB) i el rang de datasets generats per projecte definits en la opció A (entre 99 i 255), l'espai necessari per publicar dades és de 131,67 GB (valor mínim) i 339,15 GB (valor màxim).

Projecció a 3 anys

Un aspecte crític dels repositoris de dades de recerca és que les necessitats d'emmagatzematge s'acumulen, almenys durant el període que s'ha decidit conservar-les (10 anys com a mínim). Això fa que, de cara un càlcul de necessitats, a les dades calculades sota els supòsits precedents per un any calgui sumar-hi les dels següents.

S'ha fet una projecció a 3 anys agafant els valors màxims que es basa en el supòsit que cada any les dades a publicar augmenten en un 50%. Segons això, les necessitats són:

	Any 1	Any 2	Any 3
Opció A (segons projectes)			
Dades generades	600 GB	900 GB	1.350 GB
Dades acumulades	600 GB	1.500 GB	**2.850 GB**
Opció B (segons investigadors)			
Dades generades	11.982 GB	17.973 GB	26.959 GB
Dades acumulades	11.982 GB	29.955 GB	**56.914 GB**
Opció C (segons Portage)			
Dades generades	339,15 GB	508,75 GB	763, 12 GB
Dades acumulades	339,15 GB	847,87 GB	**1.610,99 GB**

Si apliquem les tarifes del CSUC per aquest 2019 a aquesta projecció a 3 anys trobem que el rang de preu per emmagatzemar dades mitjançant espai de disc SAS va de 12.926,81€ (valor mínim) a 456.691,62€ (valor màxim):

	Any 1	Any 2	Any 3	Despesa total
Opció A (segons projectes)				
Dades generades	2.772,00 €	4.158,00 €	6.237,00 €	
Dades acumulades	2.772,00 €	6.930,00 €	13.167,00 €	**22.869,00 €**
Opció B (segons investigadors)				
Dades generades	55.356,84 €	83.035,26 €	124.550,58 €	
Dades acumulades	55.356,84 €	138.392,10 €	262.942,68 €	**456.691,62 €**
Opció C (segons Portage)				
Dades generades	1.566,87 €	2.350,43 €	3.525,61 €	
Dades acumulades	1.566,87 €	3.917,16 €	7.442,77 €	**12.926,81 €**

Annex 2. Assignació del DOI com a identificador persistent

El primer requeriment de l'informe indica que el repositori ha d'assignar DOIs com a identificador persistent per les dades. Tenint en compte això, la principal organització sense ànim de lucre que proporciona DOIs per a les dades és DataCite.

DataCite permet contractar els seus serveis[1] directament però, les diferents institucions que utilitzen el programari de Dataverse, com és el cas que ens pertoca, poden proveir-se de DOIs a través del Global Dataverse Community Consortium[2] (GDCC). Aquest consorci té com objectiu oferir un espai de col·laboració entre les diferents institucions que usen Dataverse per a que puguin aprofitar-se d'economies d'escala en diferents contextos. Per exemple, en la provisió i ús de DOIs de DataCite.

Formar part d'aquest consorci comporta uns costos mínims associats de $1.400 anuals que inclouen tant la quota per formar part del consorci com la provisió de 10.000 DOIs. En el cas que se'n necessitin més, hi ha un cost extra de $300 i caldria establir d'altres acords[3] amb el consorci.

Descripció	Cost/any
Quota del GDCC	$500
Quota per usar 10.000 DOIs amb un compte de DataCite	$900
Cost extra per l'excés de DOIs	$300

Existeixen dos supòsits: que el CSUC gestioni el compte de DataCite o que cada universitat gestioni el seu propi compte. Segons això, el rang de costos per cada supòsit són els següents:

	Quota GDCC	Quota DataCite	Cost extra	Total
Supòsit gestió CSUC	500	900	300	**$1.400-1.700**
Supòsit gestió universitats	500*10= 5.000	900*10= 9.000	300*10= 3.000	**$14.000-17.000**

[1] https://datacite.org/assets/DataCitePriceList2019.pdf
[2] http://dataversecommunity.global/
[3] Aquesta tarifa extra no té un preu establert perquè, generalment, els membres no superen els 10.000 DOIs anuals. Segons informen, seria necessari entrar en una "fase pilot" i valorar-ne els costos.