XII Jornadas de Gestión de la Información.
Valor económico de la información: mercados, servicios y rentabilidad.
SEDIC, 18-19 de noviembre de 2010

MODELO/S DE COSTE PARA LA PRESERVACIÓN DE LOS DATOS CIENTÍFICOS EN LA e-CIENCIA

Pérez González, Lourdes. Biblioteca Concepción Arenal. Universidade de Santiago de Compostela, Campus Vida. Santiago de Compostela, mlourdes.perez@usc.es

Resumen: Dentro del marco de la e-ciencia, la preservación de los datos científicos es una tarea esencial para la consecución de ciberestructuras robustas y sostenibles en información. Su catalogación como bien económico (justificado por la necesidad de protección de la inversión en investigación, visibilidad de las instituciones, nuevas oportunidades de investigación en el futuro, materia prima y soporte para el conocimiento público) reclama modelos de coste que equilibren los altos recursos que la preservación digital absorbe o consume. Partiendo de los diferentes niveles de colección establecidos por el *National Science Foundation* (*NSF* 2005), colecciones de datos en investigación en curso, colecciones como recursos y colecciones para una comunidad científica, y en base al *OAIS Referente Model*, se dan a conocer los modelos de coste en curso (*Keeping Research Data Safe (I)* y *Keeping Research Data Safe (II)*, *JISC*, *LIFE*^{3}), estableciendo enlaces con guías al uso, recomendaciones y buenas prácticas (*Blue Ribbon Task Force on Sustainable Digital Preservation and Access*), establecidas y vigentes ya en la comunidad internacional.

Palabras clave: e-ciencia, preservación digital, modelos de coste, datos científicos, ciberinfraestructuras.

Abstract: Within the framework of the e-science, preservation of scientific data is an essential task for achieving robust and sustainable information cyberinfraestructure. Cataloguing as an economic good (justified by the need to protect investment in research, institutional visibility, new opportunities for future research, raw material and support for public knowledge) claims models that balance the high cost resources digital preservation consume. Based on the different levels of collection established by the National Science Foundation (NSF 2005): collections of data in ongoing research, resource collections and reference data collections for the scientific community and based on the *OAIS Reference Model*, will be announced ongoing cost models (*Keeping Research Data Safe (I)*, *Keeping Research Data Safe (II)*, *JISC Life*^{3}), establishing links to guides to use, recommendations and best practices (*Blue Ribbon Task Force on Sustainable Digital Preservation and Access*) based in the international community.

Keywords: e-science, digital preservation, cost model, cyberinfraestructure.

I) INTRODUCCIÓN.

Durante siglos las bibliotecas e instituciones afines han hecho acopio de la información para su uso futuro. Hoy se ha producido una mutación cualitativa e irreversible. La creación digital, las nuevas formas de comunicación en red y los modelos de consumo de la información *bit* está provocando que autores, editores e instituciones de investigación tengan que enfrentase a nuevas estrategias, políticas e infraestructurales, que permitan nuevos modelos de gestión.

Desde el punto de vista de la preservación, el principal problema es que al mismo ritmo en que el mundo se convierte en digital[i], se corre el riesgo de que este desaparezca. Esta desmaterialización de la información y los discursos lleva a que la creación y consumo de la misma esté en continua transformación, no sólo debido a la innovación tecnológica y la presión del mercado, sino a otros factores sociales y políticos, que lo van reconduciendo y modelando.

En cuanto a las instituciones académicas y las bibliotecas universitarias se nos dice que estamos ante un «cuarto paradigma» (Hey, 2009) de la investigación, el de la e-ciencia, en el que tenemos que construir nuevos modelos de adquisición, gestión, análisis y preservación de los datos científicos, ya que los mismos deben dar respaldo a los emergentes modelos de comunicación científica, investigación y docencia. Para lo cual, las instituciones académicas deben dar respuesta a las nuevas necesidades de la e-ciencia, constituyendo la creación y gestión de archivos digitales persistentes en el tiempo una pieza clave y determinante de estas nuevas infraestructuras. La gestión eficiente y efectiva de la información científica va a ser, sin lugar a duda, uno de los retos de las bibliotecas universitarias. Lo que obliga a plantearse la necesidad de confeccionar y establecer políticas sostenibles de preservación. Y el factor económico es aquí determinante, dado los altos costes de la misma.

Nuestro objetivo, en el presente trabajo, es presentar e introducir tres modelos de coste. Los tres vinculados al mundo anglosajón. Constituyen un punto de partida a la hora de planificar una política de preservación sostenible a nivel planetario.

a) El primero es el *Sustaining the digital investiment: Issues and Challenges of economically sustainable Digital preservation. Interim Report* (2008). con su actualización *Sustainable economics for a digital planet : ensuring Long-term Access to digital information: Final Report* (2010): el objetivo de ambos documentos es crear un modelo teórico en el que queden delimitadas las acciones necesarias para construir una preservación sostenible en el tiempo.

b) El segundo es el *Keeping research data safe: a cost model guidance for UK Universities* (2008) *(KRDS1)* y *Keeping Researcha safe 2* (2010) *(KRDS2*), conocido como el *modelo Beagrie*: se trata de una guía para la gestión económica aplicable a los proyectos de preservación digital desarrollados en instituciones académicas y de investigación británicas, consonante con la política de obligatoriedad de preservar los datos de la investigación financiados por el gobierno.

c) El tercero, y último, es el *LIFE[3]: a predictive costing tool for digital colections*, que, en esta tercera fase, aporta y ofrece una herramienta Web que permite la estimación y el control de los costes de un archivo digital.

Aquí daremos a conocer estos tres grandes modelos.

II) LOS MODELOS Y SU CONTEXTO.

Se entiende por «modelo de coste» una herramienta que facilita una toma de decisión informada.

Se espera que los costes totales en el mantenimiento de los archivos digitales (inicial, intermedio o final) sean elevados y permanentes en el tiempo, por lo cual es fundamental que en el marco y dinámica de la gestión se utilice una metodología que permita evaluar modelos y estrategias parciales en curso encaminadas a fortalecer así la estrategia global o general de preservación que se

plantee en relación a una institución y su escenario, misión, material y comunidad designada para su uso en el futuro.

Ahora bien, el desarrollar una política que asegure el máximo valor de los datos preservados en el tiempo para la comunidad científica y académica no es una tarea sencilla. Las cuestiones imbricadas en el proceso son de gran complejidad. Una de las primeras y primordiales necesidades es el uso de un lenguaje común. En este punto, no partimos de cero. Las comunidades científicas (metadatos), agencias internacionales (a través de estándares), universidades (políticas y planes) y bibliotecas (buenas prácticas, formación, eco a la comunidad de usuario, etc.), han logrado importantes avances en este campo. Nuestra primera tarea, por lo tanto, será intentar clarificar este lenguaje, y las constelación de categorías y conceptos a él asociados.

Comenzaremos por deslindar y definir los tres elementos que constituyen el marco común de estos modelos de coste, a saber:

1. Las ciberinfraestructuras.
2. Las entidades funcionales del *OAIS Reference Model*.
3. Los niveles de colección de datos establecidos por la *National Science Foundation*, *NSF* (2005).

1.-*Las ciberinfraestructuras y la necesidad de preservar los datos para el futuro.*

Por «ciberestructura» hay que entender una nueva forma de cultura científica que se sustenta en una infraestructura tecnológica de alto nivel, a través de la cual se intenta dar apoyo a nuevos mecanismos de colaboración, basados en el acceso a una ingente cantidad de datos, recursos de información interpretados y reutilizados por potentes herramientas de observación, visualización y simulación. Es un medio que permite el acceso y circulación del conocimiento distribuido, en el que colaboran y se comunican diferentes comunidades y disciplinas rompiendo fronteras culturales, geográficas y temporales. Su fin es crear una base de datos global, cuya sostenibilidad dependerá enteramente del trabajo y esfuerzo coordinado de la interoperabilidad, el consenso normativo en políticas de propiedad intelectual y la consideración y consolidación de la preservación de la ciencia como bien público.

2.- Las entidades funcionales del *OAIS Reference Model*.

El *OAIS Reference Model* es un modelo conceptual, en el que quedan establecidos la terminología, los componentes y una arquitectura en abstracto de lo que debe ser un archivo digital para la preservación en el tiempo.

Seis son las áreas funcionales que establece: dos de ellas, dedicadas a la planificación y gestión (en donde entraría la política, estrategias y planificación del mismo); y las cuatro restantes, relacionadas con el ciclo de vida del archivo y las actividades vinculadas al mismo.

a) Absorción de la información (*Ingest*).
b) Almacenamiento (*Archival storage*).
c) Gestión (Data management).
d) Acceso / Difusión (Access / Dissemination).
e) Planificación (Perserving Planning).
f) Gestión de archivo (*Administration*).

Este modelo conceptual es, en parte, la ruta común a seguir por el *Proyecto LIFE* y el *modelo Beagrie*, desmenuzando las actividades y procesos en variables de coste.

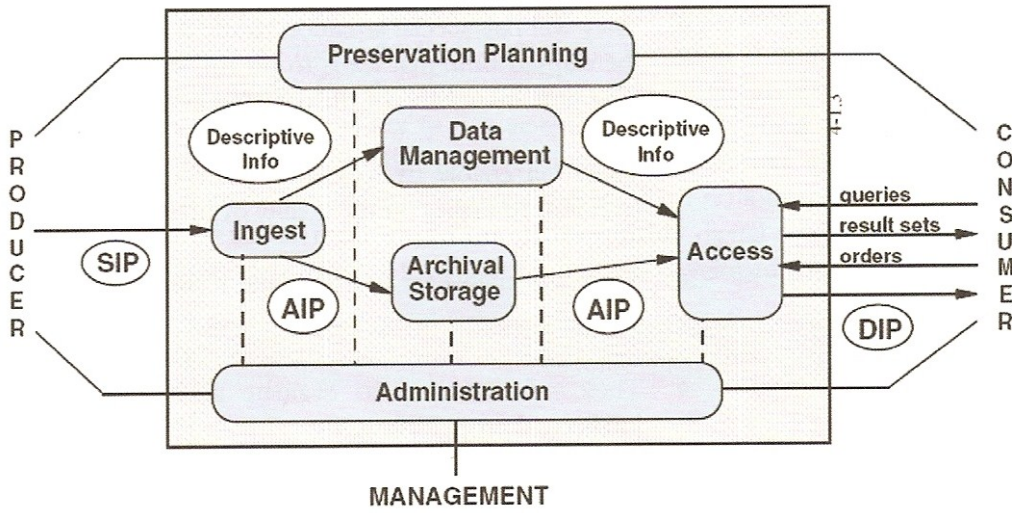


Figura 1. ***OAIS Reference Model. Consultative Committe for Space Data Systems.***

3.-Los niveles de colección establecidos por la NSF

Breve diccionario o nomenclatura de la NSF:

El término «dato», se refiere a cualquier información que puede ser almacenada en formato digital; aquí se incluyen texto, números, imágenes, vídeo, audio, software, algoritmos, ecuaciones, simulaciones, etc. Tales atos pueden ser generados desde la observación, la computación o a través de experimentación.

El término «colección», se refiere no sólo a los datos almacenados, sino también a la infraestructura, las organizaciones y las personas necesarias para preservar el acceso a los datos.

La voz «colecciones digitales», se limita a aquella colección de datos a la que accedemos a través redes electrónicas.

La voz o expresión «colecciones digitales vivas en el tiempo», o *Long–lived data collection*, se refiere a aquella capaz de ser preservada en un tiempo suficiente para sobrevivir a un nuevo cambio tecnológico. Este concepto marca un proceso desde la creación de la información hasta el consumo de la misma por una comunidad de usuarios en donde aparecen diferentes actores y funcionalidades, tanto a nivel local (investigador, biblioteca, organización, etc.), como Estatal o internacional.

El lexema o expresión «universo de las colecciones de datos», apunta o alude a un sistema de datos digitales, colecciones de datos, software, hardware y comunicaciones, autores, gestores, usuarios, científicos de datos, agencias y centros de investigación que permiten la colección, análisis de datos, preservación (*curation*) y distribución de datos en la investigación y docencia en curso.

Partiendo del esquema conceptual de la *NSF*, distinguiremos tres categorías funcionales o tipologías de colecciones, base para los modelos de coste aquí presentados:

1ª) «Colecciones de datos de investigación», o *Research data collections*: los resultados de una investigación. Generalmente, son objeto de una preservación limitada en el tiempo, siendo las normas y formatos comunes en una comunidad dada, pudiendo o no adaptarse a los estándares y metadatos de la misma. A menudo, al ser una información novedosa, no existen sustitutos. Puede que no haya necesidad de preservar la misma más allá del proyecto de investigación. Estas colecciones por lo general cuentan con presupuestos de financiación bajos. La *NSF* y *JISC* (*Joint Information System Committee*, 2010) recomiendan políticas de selección y retención por parte de los autores, en algunos casos con medidas legales[ii].

2ª) «Colecciones de datos base de una comunidad» o *Resource or community data collections*: colecciones vinculadas a una disciplina. La propia comunidad científica establece ciertos parámetros, normas y estándares para su creación, acceso y uso. Los presupuestos vinculados a estas colecciones son de nivel intermedio y financiado a través de agencia gubernamentales. Debido al cambio en la financiación del proyecto y la vida de éste, es difícil en ocasiones establecer las políticas de retención. La *NSF* y el *JISC* recomiendan la colaboración entre autores e instituciones a fin de establecer cierta política de retención de las mismas.

3ª) «Colecciones de datos referenciales» o *Reference data collections*: colecciones fundamentales para una comunidad científica. Están vinculadas a un número de usuarios amplio, más allá de una institución y generalmente de alcance internacional. Las normas de creación son rígidas y con gran nivel de especialización. Los presupuestos de financiación son elevados y provienen de diferentes fuentes. La preservación de estas colecciones recae en diferentes organismos, no siempre estables.

Estos niveles de colección son elemento y premisa comúnn los modelos económicos a estudiar y deben ser los que marquen cualquier política de selección en un archivo digital sostenible. Tendrá, a su vez, por otra parte, importantes implicaciones para su financiación y sostenibilidad en el tiempo de dichas colecciones, el hecho de estar o no albergadas en ubicación física o virtual, centralizada o descentralizada.

II.1) Los informes del *Blue Ribbon Task Force on Sustainable Digital preservation* (primer modelo)

El organismo *Blue Ribbon Task Force on Sustainable Digital Preservation and Access, BRTF-SDPA*, como es sabido, es una fundación conjunta de la *National Science Foundation*, *Andrew W. Mellon Foundation*, *Library of Congress*, *Joint Information Systems Committee* del Reino Unido, *Council on Library and Information Resources*, *CLIR.* y el *National Archives and Records Administration, NARA.* Son dos los informes emitidos por él en punto de preservación*:* el primero, titulado *Sustaining the digital investment: Issues and chagenges of economic sustainable digital preservation*. Interim Report (2008), y, el segundo, con el título de *Sustainable Economics for a digital planet: Ensuring long-term access to digital information.* Final Report (2010).

El objetivo del *BRTF-SDPA* es establecer un cuádruple esquema o cuadro conceptual a partir del cuál se puedan construir marcos económicos, estrategias y buenas prácticas adaptables a los diferentes escenarios de la información digital.

Partiendo de que la información digital es fundamental para el desarrollo de la ciencia, la educación, la sociedad civil y el gobierno, el futuro de éstos dependerá en parte del acceso persistente a ésta en un futuro cara a su explotación y reutilización en diferentes campos. De cualquier manera no queda claro, pese a las políticas de retención obligatorias por parte muchos gobiernos, *qué información preservar, quién es el responsable de dicha preservación y quién pagará por ello*. De esta manera la *BRTF-SDPA* trata de crear un modelo económico sostenible que permita la generación de estrategias y planes coherentes de preservación. Advierte que *el problema es urgente* y que el *acceso a los datos del mañana va a depender de las decisiones que se tomen hoy.*

Cuatro son los escenarios o dominios emergentes de información digital que es necesarios identificar y pautar para la su sostenibilidad de ésta en el futuro:

1. La información académica (*Scholarly discurse*).
2. Datos científicos (*Research data).*
3. Información comercial y cultural (*Commercially owned cultural content*).
4. Información colectiva en la WEB (*Collectively produced Web content*).

Aquí nos interesan solo los dos primeros para crear cartografías con los *Informes Beagrie* y el *Proyecto LIFE*.

Por otro lado, la *BRTF-SDPA* identifica una serie de barreras u obstáculos que imposibilitan llevar a cabo el acceso y la preservación digital de forma sostenible:

-Fondos insuficientes que impidan una preservación persistente, generalmente los fondos son «one time», únicos, vinculados a un proyecto de investigación, un presupuesto de obra, etc. Una preservación fiable no puede sufrir fisuras. La información digital perdida no es recuperable.

-Confusión y / o falta de coordinación entre de las partes interesadas en relación a roles y responsabilidades. En ocasiones los creadores y gestores de los datos a preservar no van a ser los beneficiarios, con lo que los costes no son asumidos por éstos últimos. Es un reto importante asignar los roles y las responsabilidades para establecer un modelo sostenible para la institución.

-Incentivos inadecuados para las instituciones, empresas o comunidades imposibilitan un modelo sostenible. La preservación requiere una planificación amplia en la que se de un compromiso de responsabilidad en base al acuerdo de normas, formatos, modelos de uso, interoperabilidad de sistemas. Lo que, a su vez, exige personal y equipos de trabajo muy especializados, para lo que no toda institución está preparada.

-La urgente necesidad de establecer modelos sostenibles de preservación digital, no siempre es bien entendida. A pesar de saber que el liderazgo y la competitividad de la institución, incluso su supervivencia, dependen de la disponibilidad de información, esto no está asumido por la misma ni queda adecuadamente institucionalizado.

-El temor a que el acceso a la información digital sea un reto demasiado grande a asumir, ya que requiere estrategias, políticas y planes estables por parte de la institución. No obstante, una buena hoja de ruta, una eficaz normativa sobre los datos, una «data bill», como elemento explicito en el modelo institucional, puede disipar tales miedos.

Si en su primer Informe, define el marco de la sustentabilidad de la información digital como: *el conjunto de medidas económicas, sociales, tecnológicas y políticas que estimulan el acopio de aquellos recursos importantes de información digital a través de los sistemas de preservación, y dan a su vez el apoyo indefinido a la permanencia indefinida en dichos sistemas , permitiendo el acceso y uso de recursos de información en el futuro.*

En su Informe final, profundiza en lo que esto conlleva:

1. El reconocimiento de los beneficios de la preservación por parte de los grupos de gestión[iii].

2. Incentivar a los responsables de la toma de decisiones para actuar en interés del bien público.

3. Incidir en el proceso de selección como elemento clave en la retención de materiales para la preservación en el tiempo.

4. Activar mecanismos que aseguren el funcionamiento y la localización eficiente de recursos en el proceso de la preservación digital. Para ello se establece como criterio prioritario la eficiencia tanto en el ámbito público como en el privado; la necesidad de trabajar con economías de escala y lograr economías de alcance.

5. Llevar a cabo una política y organización apropiadas para las actividades requeridas por la preservación digital. Se analizan escenarios dispares, las posibilidades de organismos públicos, privados y semi-públicos. En todos ellos se cree necesaria una cultura de responsabilidad y el compromiso con la misma, en línea de solventar problemas institucionales identificados en la Carta de la UNESCO (2003)[iv].

Por lo tanto, el trabajo del BRTF-SDPA tiene la virtud de hacernos reflexionar sobre la urgencia de plantear un modelo económico de información digital eficaz que permita la sostenibilidad, en el amplio mapa o campo que abarca y en el que opera la preservación digital (de alto nivel: e-ciencia; nivel medio: repositorios, servicios comerciales, etc., u otros más cercanos: proyectos de investigación, publicaciones académicas, etc.), así como la necesidad de su coordinación

a través de diferentes organismos, trabajando a nivel internacional, nacional, local, federado o temático.

Las organizaciones deben valorar su modelo de gestión económica teniendo en cuenta entornos, mapa de escenarios, sus recursos, actores, productores y servicios. Para ello, un modelo tal, requiere un conjunto mínimo de elementos o características:

a) Hay que contar con recursos suficientes que garanticen la sostenibilidad y que permitan o posibiliten el acceso de la información en el futuro.

b) Se prestará una atención especial al factor tiempo, tanto desde un punto de vista simple en relación al momento actual y la posibilidad de pérdida de información, como desde un punto de vista más amplio y complejo (de provisión), que va más allá de las posibilidades actuales de obtención y propiedad de la información, y contemple la disponibilidad de nuevas tecnologías que la posibiliten.

c) El modelo debe permitir examinar y tener en cuenta los efectos de las estrategias técnicas (en curso y futuras) de preservación en otras organizaciones.

d) Nos ha de permitir calcular el riesgo y las consecuencias económicas y técnicas de la pérdida de datos.

e) Posibilitará evaluar políticas alternativas, incluso aquellas relacionadas con cambios requeridos en las leyes de propiedad intelectual (derechos de autor, patentes de invención, *Science commons*, etc).

f) Nos permitirá evaluar las implicaciones de los cinco componentes que se han contemplado en la definición de sostenibilidad de la información digital, a todos los niveles en nuestras organizaciones, tanto individual como colectivo.

Siendo así, para resumir, que la sostenibilidad económica de la preservación digital requiere un equilibrio entre los recursos y resultados, la oferta y la demanda de información, y la percepción del valor y la voluntad de pago por ello.

II.2) *Keeping Research Data Safe* (2008) y *Keeping Research Data Safe 2* (2010) (segundo modelo)

Los *Keeping Research Data Safe* (2008) y *Keeping Research Data Safe 2* (2010), son la aproximación por parte del *JISC* a un modelo de coste viable para la preservación de datos científicos en las instituciones británicas que reciben y reciban financiación pública; en el futuro, pretenden preparar una herramienta útil que funcione en el seno del modelo de financiación académico británico *Transparent Approach to Costing, TRAC*. Ambos estudios son deudores, y así lo reconocen, del Proyecto *LIFE* de la *British Library* y del *Cost Estimation Tool (CET)*, de la *NASA* (fundamental para el cálculo de los costes de absorción, almacenamiento y preservación de los datos aeroespaciales), así como de la *NSF* y de los niveles de colección de *Long-Lived data collectios* (2007), manteniendo las entidades funcionales del *OAIS Reference Model*.

En su primera fase, estudian los casos de tres archivos del Reino Unido: el de la *Cambridge University*, el modelo del *King's College* de Londres y los informes anuales de el *Archaeology Data Service* de la York University.

Partiendo de esta base, el *KRDS* establece punto de arranque, pautas y guía:

a) Una serie de costes variables y sus unidades de coste. Estas se ven incrementadas si se compara con los repositorios institucionales de publicaciones electrónicas. Las dos variables más importantes son la evolución tecnológica y el factor humano. No obstante, se tiene en cuenta que la investigación en preservación tiende a una automatización de los procesos de absorción de datos.

b) Un modelo de actividad basado en tres fases: pre-archivo, archivo y servicio del modelo *OAIS*.

c) La adaptación de las variables y unidades de coste al modelo *TRAC*.

A lo largo de ambos informes se hacen una serie de recomendaciones a las instituciones, antes de plantearse una estrategia de información, que sintetizamos:

1) Examinar el modelo de archivo y su estructura contemplando la posibilidad de que sea internacional, nacional y local. El *JISC* recomienda estructuras federadas para las colecciones locales o de investigación con un pre-archivo institucional. En el caso de colecciones relacionadas con una comunidad o disciplina, se deben establecer acuerdos con los archivos existentes a nivel nacional e internacional, si existieran.

2) Trabajar dentro del *High Education Information system* con economías de escala, compartiendo servicios mediante archivos temáticos, de forma centralizada y a nivel nacional.

3) Las organizaciones deben tener en cuenta que establecer archivos de datos es más costosos que los repositorios institucionales implantados en el tejido académico británico. La complejidad y diversidad de formatos a preservar dentro de estos requieren altas inversiones en tecnología y personal. Para ello se recomienda la colaboración entre instituciones a través de archivos centralizados.

4) Por otro lado, el *JISC* considera que el análisis de costes no es suficiente para una gestión económica. Aparte de ver qué es lo que puede o no costear la propia institución, es necesario que el análisis de coste vaya acompañado siempre de un cuadro de beneficios anticipados.

En este sentido, el *JSIC* ha explicitado un conjunto o serie de beneficios en relación a los costes de inversión, en el informe avanzado, *KRS2*, en donde se presentan en varios niveles y dimensiones según el área de impacto (para la investigación, para la institución y la propia gestión del archivo, a nivel privado y para el beneficio público) y que, básicamente, serían los siguientes:

a) Beneficios directos: nuevas oportunidades para la investigación; acceso a la comunicación científica como bien público, re-utilización de los datos, incremento de la productividad científica, fomento de la colaboración e impulso de redes en la investigación, mayor transferencia de resultados a la industria, verificación de los procesos de la ciencia, nuevos modelos de *peer-review*.

b) Beneficios indirectos: la no duplicidad de datos, no perder nuevas oportunidades de investigación, reducción de costes en una preservación futura, ofrecer datos a nuevas audiencias, servir como materia prima en nuevas metodologías, proteger los beneficios de las primeras inversiones

c) Beneficios a corto plazo: valorar a la investigación en curso, no tener pérdidas de información en etapas post-doctorales, posibilitar la reutilización de datos a corto plazo, asegurar el correcto almacenamiento de los datos de investigaciones en curso, dar disponibilidad a los datos y materiales anexos de los nuevos modelos de comunicación científica.

d) Beneficios a largo plazo: asegurar el valor futuro de estudiantes e investigadores, añadir valor a la formación de colecciones en el tiempo mediante el incremento de su masa crítica. alentar la

gestión de la preservación desde las primeras etapas de creación y almacenamiento a través de la creación de paquetes de datos consistentes. Para ello es fundamental la formación de los autores e investigadores, así como la creación de una e-estrategia de información *bit*[v] .

e) Beneficios privados o internos : beneficios para los agentes implicados en la financiación, beneficios para la comunidad científica (prestigio, carrera, difusión y visibilidad de trabajo, etc.), cumplimiento con la normativa de retención de datos, comercialización de la investigación (en este punto existen fricciones entre las patentes de invención, la transferencia de innovación y los *science commons*, de momento las estrategias apuntan al embargo de información en las primeras fases de la comercialización).

f) Beneficios públicos: materia prima para futuras investigaciones, motivación de nuevas investigaciones, avance de la ciencia a nivel planetario, impacto en el tejido social (de nuevo, se recomienda que las instituciones trabajen en los aspectos éticos de la información que deben quedar plasmados en la «carta» o «normativa» sobre la estrategia de retención de datos).

El *JISC* es consciente de la necesidad de trabajar en proyectos que permitan ir adaptando modelos de coste a diferentes instituciones y escenarios de información científica. Los mayores retos, en este momento se centran, siguiendo el modelo funcional del *OAIS Reference Model*, en la absorción de datos. La creación de paquetes robustos y ricos en metadatos, sólo provenir de las comunidades de investigación conocedoras de los sistemas y aplicaciones de creación, recuperación y difusión de datos cada vez más mediados por la e-ciencia. Los profesionales de la información, en este caso las bibliotecas universitarias, con todo, deben incidir en la selección de colecciones y la formación de la comunidad en relación a formatos requeridos, empaquetamientos de datos, modelos éticos y motivación de retención de información.

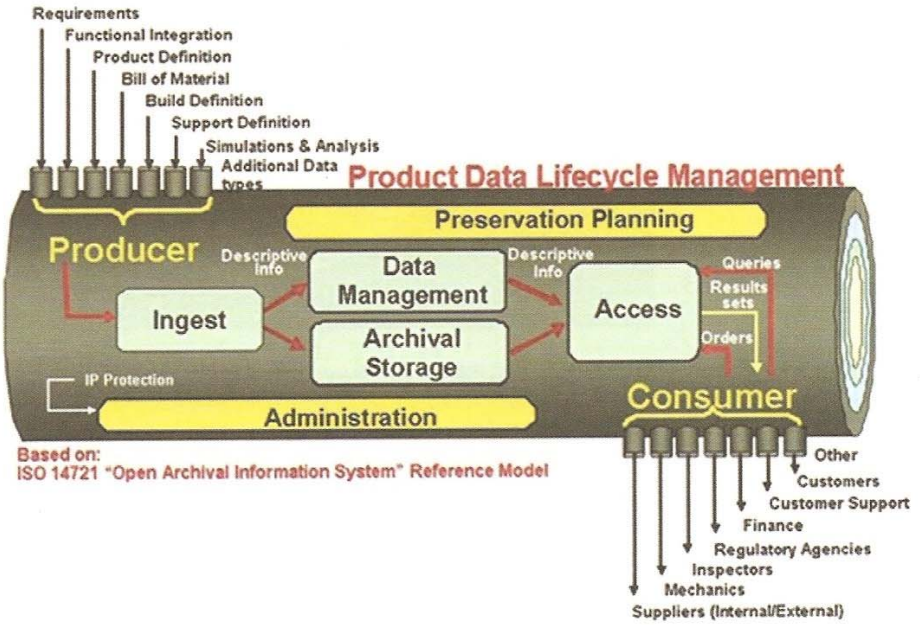


Figura 2. ***OAIS Reference Model. Consultative Committe for Space Data Systems.***

Ambos informes, *KRDS1* y *KRDS2*, advierten de la importancia estratégica de construir modelos propios, aprendiendo de otros existentes y en curso, de las propias necesidades de información de la comunidad científica y de adaptar *OAIS Reference Model* a la gestión del archivo. El Informe avanzado del *KRDS* termina con ocho recomendaciones, entre las que aquí destacamos dos, la necesidad de investigar futuros desarrollos en la fase de pre-archivo trabajando desde la

perspectiva de la comunidad científica y que el equipo *JISC / Beargie* prosiga con la investigación a fin de implementar el proceso logrando alguna herramienta práctica eficaz.

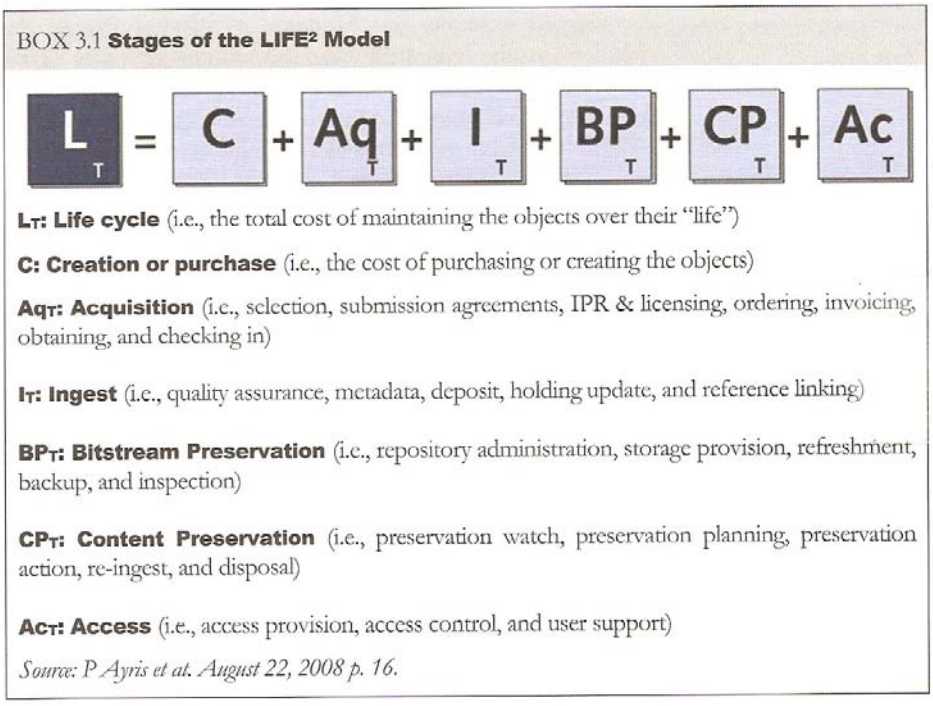
II.3) *LIFE*[3]

El proyecto *LIFE* (*Life Cycle Information for E-Literature*) es la continuación de el concepto del ciclo de vida de la información en las colecciones físicas desarrollado por la *British Library* en la década de los noventa, retomado para su adaptación al mundo de la información digital Su principal objetivo es lograr una herramienta útil que permita a las instituciones encargadas de la preservación digital estimar los altos costes de la preservación adaptándose a requerimientos funcionales, validación de formatos, arquitecturas de archivo en curso, modelos de gestión, etc.

Dicho proyecto se remonta, en su primera fase, fase 1, al bienio 2005-2006, cuando se fijó una metodología que permitiera entender el ciclo de vida de la información bit. Los estudios se realizaron en el *Voluntarily Deposited Electonic Publications* (*VDEP*) de la *British Library* y el *e-repository* de la *Cambridge University Library* (*UCL*). A partir de aquí se fue construyendo un modelo que permitiera estimar los costes de la preservación de los objetos digitales en el ciclo *LIFE*.

Su segunda fase, fase 2, data del bienio 2007-2008 y se extendió a nuevos proyectos de preservación: *Newspaper Digitalization Project* de la *British Library*, el estudio del repositorio *SHERPA-LEAP* y lservicios de preservación del *SHERPA-DP*.

En la actualidad se halla en su tercera fase, fase avanzada, fase 3, *LIFE3*, con el resultado de una herramienta web, a través de *Hojas Excel* y mediante el uso de plantillas extensible y adaptable a los diferentes escenarios de la preservación digital. Dicha herramienta tiene de fondo, y ello es fundamental, el modelo LIFE derivado de la fase 2, en el que se puede observar la adaptación del *OAIS Reference Model* a los escenarios de las instituciones académica (caso adquisición como actividad funcional).



BOX 3.1 **Stages of the LIFE² Model**

$$L_T = C + Aq_T + I_T + BP_T + CP_T + Ac_T$$

L_T: Life cycle (i.e., the total cost of maintaining the objects over their "life")

C: Creation or purchase (i.e., the cost of purchasing or creating the objects)

Aq_T: Acquisition (i.e., selection, submission agreements, IPR & licensing, ordering, invoicing, obtaining, and checking in)

I_T: Ingest (i.e., quality assurance, metadata, deposit, holding update, and reference linking)

BP_T: Bitstream Preservation (i.e., repository administration, storage provision, refreshment, backup, and inspection)

CP_T: Content Preservation (i.e., preservation watch, preservation planning, preservation action, re-ingest, and disposal)

Ac_T: Access (i.e., access provision, access control, and user support)

Source: P Ayris et al. August 22, 2008 p. 16.

Figura 3. ***LIFE Model.***

El objetivo es obtener una visión sinóptica inmediata en la que requerimientos funcionales y costes queden identificados y cuantificados por la institución, en relación al tamaño y propósito de su archivo. Las plantillas de dicha herramienta permiten la selección de los modelos por tipología de la organización, tipo de contenidos (formatos, número de registros, arquitectura del archivo, personal, etc.) y, pese a que la herramienta viene en estimaciones basadas en los modelos estudiados, se permite la modificación del mismo a través del comando *«Refine Organizational profile»*.

Es fundamental entender el *Workflow* del Proyecto a fin de adaptar el archivo al modelo LIFE:

a) *Creación / Adquisición – Ingest* : en donde se identifican los mayores costes y la posibilidad de reducción en base a modelos de sostenibilidad.

b) *Bitstream preservation*: coste por unidad de almacenamiento *MG*, deflación de las tecnologías, contratación externa de posibles servicios, etc.

c) *Content Preservation*: esta fase es quizás la más dura y menos desarrollada por tratarse de las estrategias y políticas propias. No obstante, se da la posibilidad de trabajar con tres modelos de migración y emulación, tomando como ejemplos los datos de la *Danish National Library and Archives* para la migración y de la la *Koninkligke Bibliotheek, KB*, de Holanda para la emulación.

d) *Access***:** en donde se incide en las estimaciones de costes directos de personal y costes estimados sobre diferentes tecnologías y aplicaciones de acceso).

Se supone que la aplicación de esta herramienta estará abierta a todas las instituciones.

III. CONCLUSIONES.

Cada vez más las instituciones académicas tienen que enfrentarse a nuevos retos de archivo de la información digital generada por sus comunidades científicas. La gestión de estos nuevos archivos requiere altos costes para las mismas.

Hemos presentado aquí tres modelos. Los tres convergen en sus metas: difundir el valor de los datos científicos en la nueva infraestructura de información global, entender las posibles demandas de las comunidades científicas en el futuro, trabajar en una visión de sostenibilidad por parte de las organizaciones implicadas en la financiación, creación, retención y acceso a los mismos, y, finalmente, utilizar herramientas adaptadas a modelos funcionales capaces de soportar la interoperabilidad con diferentes arquitecturas existentes y por venir. Son modelos vivos.

Son los retos de un futuro a compartir.

Referencias bibliográficas :

American Council of Learned Societies Commission on Cyberinfraestructures for the Humanities and Social Science (2004). Our Cultural Commonwealth. The report of the American Council of Learned Societies on Cyberinfraestructure for the Humanities and Social Science: Consultado en: 01-09-2010. http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf

Bargrie, N.; Chuszcz, J. Y Lavoie, B. (2008). Keeping Research Data Safe: a cost model and guidance for UK Universities (Joint Information System Committee 2008). Consultado en: 01-09-2010. http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Fink, J. Lynn; Bourne, Philip E. "Reinventing Scholarly Communication for the Electronic Age". En: *CT Watch Quarterly* , 2007, Aug. v. [3], n. [3]. Consultado en: 01-09-2010. http://www.ctwatch.org/quarterly/articles/2007/08/reinventing-scholarly-communication-for-the-electronic-age/index.html.

Fry, Jenny; Lockeyr, Suzanne and Oppenheim, Charles (2008). Identifyng benefits arising from the curation and open sharing of research data produced by UK Higher Education and Research Institutes. Consultado en: 01-09-2010. http://ie-repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf

Hey, Tony; Tansley, Stewart and Tolle, Kristin. *The fourth paradigm: data-intensive scientific discovery. Redmond, Washington : Microsoft Research, 2009,* ISBN [978-0-9825442-0-4].

Joint Information System Committee (2010). Keeping Research Data Safe. Consultado en: 01-09-2010. http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Blue Ribbon Task Force on sustainable digital preservation and Access/ National Science Foundation; The Andrew W. Mellon Foundation; Library of Congress; Joint Information Systems Committee; Council on Library and Information Resources; National Archives and Records Administration (2008). Sustaining the digital investment : isues and changelles of economically sustainable digital preservation. Consultado en: 01-09-2010. http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

Blue Ribbon Task Force on sustainable digital preservation and Access/ National Science Foundation; The Andrew W. Mellon Foundation; Library of Congress; Joint Information Systems Committee; Council on Library and Information Resources; National Archives and Records Administration (2010). Sustainable economics for a digital planet : Ensuring long-tern Access for digital information. Consultado en: 01-09-2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

National Science Board, NSB (2005). Cyberinfraestructure Vision for 21st Century. Consultado en: 01-09-2010. http://www.arl.org/bm~doc/ci_vision_march07.pdf

National Science Board, NSB (2005). Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. Consultado en: 01-09-2010. http://www.nsf.gov/pubs/2005/nsb0540/

National Science Board, NSB (2010). Grant Proposal Guide. Consultado en: 01-09-2010. http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gpg

UNESCO; Biblioteca Nacional de Australia (2003). Directrices para la preservación del patrimonio digital. Consultado en: 01-09-2010. http://unesdoc.unesco.org/images/0013/001300/130071s.pdf

[i] Según el trabajo de Lyman y Hal: «la producción mundial de formato impreso, imagen, óptico, magnético y su contenido requieren alrededor de 1,5 millones de gigabytes. El equivalente a 250 megabytes por persona para cada hombre, mujer o niño en el planeta». En : **Lyman, Peter; Hal R. Varian**. How much information? 2003. Consultado en: 01-09-2010. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/ Para consultar informes actualizados. Se recomienda ver los estudios en curso de *How Much Information? (HMI) del Global Information Center*. Consulatado 01-09-2010. http://hmi.ucsd.edu/howmuchinfo.php

[ii] La *NSF*, a través de *Grant Proposal Guide /NSF 2010* (p. 25 y 51) establece la obligación de describir los futuros usos de los datos en la investigación .

Para las estrategias tomadas en las diferentes disciplinas científicas, ver : *Appendix c*: *Current Policies on Data Sharing and Archiving (NSF, 2005).*

[iii] Para la comunidad académica y los datos científicos, ver los Informes de la *NSB (2005)*; los Informes *KRDS 1*, *KRDS2* y *Our Cultural Commonwealth (2004)*

[iv] *Directrices para la preservación del patrimonio digital* (2003, p. 45-46). «La solvencia de los programas de preservación digital y la confianza que se puede depositar en ellos son asuntos muy importantes para las diversas partes interesadas, pues los productores, usuarios, inversores y la comunidad en general tienen gran interés en que los elementos del patrimonio digital estén sujetos a disposiciones en las que se puede confiar./ Quienes puedan ser responsables de los programas también tienen interés en evaluar lo que pueden ofrecer y los riesgos que supondrá aceptar esa responsabilidad. Es de esperar que los programas de preservación que ofrecen fiabilidad a largo plazo presenten las siguientes características:/ • Responsabilidad: un compromiso fundamental con la preservación de los objetos digitales considerados./ • Viabilidad de la organización, incluida la perspectiva de un mandato permanente; un estatuto jurídico que la defina como una organización que va a desempeñar un papel permanente de preservación; y una capacidad demostrada para reunir recursos, infraestructura y equipos de trabajo que puedan hacerse cargo de la complejidad de la preservación digital. • Durabilidad financiera: una perspectiva verosímil de que la organización será capaz de seguir suministrando los necesarios recursos a largo plazo, con un modelo de gestión sostenible para llevar a cabo su mandato de preservación digital./ • Adecuación tecnológica y de procedimiento: la utilización de sistemas y procedimientos apropiados para hacer lo que requieren la gestión y la preservación de los recursos digitales./ • Seguridad del sistema de muy alta calidad./ • Transparencia de los procedimientos, con una clara atribución de las responsabilidades y mecanismos para evaluar los resultados obtenidos e informar sobre ello».

[v] Existen en este momento varias iniciativas paradigmáticas. Dos ejemplos a tener en cuenta a la hora de establecer e-estrategias y buenas prácticas podrían ser: el Proyecto de la Universidad de Oxford, *Embedding Institutional Data Curation Services in Research (EIDCSR)* en: http://eidcsr.oucs.ox.ac.uk/ (Consultado 01-09-2010) y la guía *From data to Wisdom : Pathways to Succesful Data Management for Australian Science (2006)*