Enjeux de l'archivage à long terme des données primaires de la recherche scientifique : implications pour le GIPDIR

Travail de Bachelor réalisé en vue de l'obtention du Bachelor HES

par :

Igor MILHIT

Conseiller au travail de Bachelor :

Alexandre BODER, Chargé d'enseignement HES

Genève, le 16 juillet 2012

Haute École de Gestion de Genève (HEG-GE)

Filière Information documentaire

Déclaration

Ce travail de Bachelor est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre de Bachelor en sciences HES en information documentaire. L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Bachelor, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de Bachelor, du juré et de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 16 juillet 2012

Igor Milhit

Remerciements

Je désire remercier ici toutes les personnes qui m'ont supporté durant les périodes de ma vie où je me suis montré particulièrement insupportable, et toutes celles qui, d'une manière ou d'une autre, m'ont permis de reprendre cette formation, sans quoi ce travail, bien entendu, n'aurait pas vu le jour.

Je remercie également l'ensemble des professeurs de la filière Information documentaire de la Haute École de Gestion de Genève, ainsi que mes camarades d'études, avec lesquels j'ai eu beaucoup de plaisir à travailler.

Plus directement en lien avec ce travail, je voudrais remercier :

- M. Jean-Blaise Claivaz pour m'avoir indiqué la bonne porte quand je cherchais un mandant pour mon sujet, alors que ce dernier était encore très vague et incertain ;
- M. Pierre-Yves Burgi, le directeur du service NTICE, ainsi que Jan Melichar, responsable du pôle « *Data Management* » de ce même service, qui m'ont accueilli avec enthousiasme et permis de développer une réflexion que j'ai trouvée passionnante ;
- M. Jean-Daniel Zeller qui m'a donné de bons conseils et indiqué des sources très utiles ;
- M. Didier Grange, archiviste de la Ville de Genève, qui a su recadrer mon travail au moyen d'une liste très pertinente de questions ;
- enfin, M. Alexandre Boder, mon conseiller, qui m'a prodigué un soutien précieux.

J'aimerais, pour terminer, mentionner Mme Jeanne Wagner pour ses encouragements et pour les relectures attentives.

Résumé

Le service des Nouvelles Technologies de l'Information, de la Communication et de l'Enseignement (NTICE) de l'Université de Genève développe un outil informatique, afin de répondre aux besoins de partage de fichiers et de centralisation des informations au sein d'un projet de recherche d'envergure et pluridisciplinaire. Cet outil doit être en mesure de supporter une grande diversité de formats, correspondant aux besoins et aux habitudes des chercheurs. C'est pourquoi ce projet s'intitule « Gestion Intégrée d'une Pluralité de Données Issues de la Recherche » (GIPDIR).

Pour le NTICE, cette centralisation des données primaires de la recherche constitue une étape préalable à leur archivage à long terme conforme à la norme OAIS. Pour que les données de la recherche puissent être gérées par un tel système d'archives, il est nécessaire que les données de la recherche se conforment à des exigences précises en terme de formats et de métadonnées. À partir de l'expérience de projet d'archivage des données primaires existants, nous établissons des recommandations, afin d'atteindre ces exigences, tout en limitant autant que faire se peut les coûts qu'elles impliquent, principalement pour documenter les données.

Les chercheurs peuvent en effet considérer l'effort supplémentaire induit par l'objectif d'archivage à long terme comme un obstacle à l'avancement de leur travail. Il s'agit donc de relever les arguments qui plaident en faveur de cet archivage, arguments parfois contraignants comme le sont les règlements des bailleurs de fonds, ainsi que les bénéfices que les chercheurs peuvent en escompter.

Table des matières

Index des tableaux

Index des figures

Introduction

Contexte

La recherche scientifique n'échappe pas aux grandes tendances de notre société, parmi lesquelles nous pouvons citer l'informatisation généralisée des activités humaines, l'interconnexion quasi permanente de tous les acteurs et l'explosion documentaire. Ces phénomènes trouvent leur lointaine origine dans le berceau de la modernité occidentale (humanisme, imprimerie, république des lettres, méthodologie scientifique), mais naissent véritablement, dans la configuration que nous connaissons actuellement, autour de la moitié du siècle dernier. Nous pouvons penser qu'à bien des égards, l'utilisation de ces technologies a atteint aujourd'hui un seuil qui impose une reconfiguration générale des méthodes de travail. Ces développements ouvrent de nouvelles possibilités, nous confrontent à de nouveaux défis, ou du moins, changent quantitativement la dimension des interactions auxquelles nous étions habitués.

Puisqu'une grande partie des activités humaines, et c'est encore plus marqué pour les activités intellectuelles, est réalisée ou gérée au moyen d'outils informatisés, nous assistons à une convergence des médias véhiculant de l'information. En effet, pour consulter ou éditer du texte, des images, des photographies, du son, des vidéos, des cartes géographiques, de la correspondance écrite et bien d'autres types d'information, nous utilisons un ordinateur. De plus, l'informatique a permis, grâce aux technologies de la communication, de constituer des réseaux de plus en plus vastes, de plus en plus interconnectés via le réseau des réseaux : Internet.

Cette convergence et cette interconnexion ont multiplié significativement les possibilités et les habitudes d'échange d'informations et de collaboration. Aussi, la recherche scientifique, qui est en partie caractérisée par le partage des méthodes et des résultats, multiplie des projets de recherche vastes et réunit des équipes constituées de nombreux chercheurs, parfois de disciplines très diverses.

Dans ce contexte, le service des Nouvelles Technologies de l'Information, de la Communication et de l'Enseignement (NTICE) de l'Université de Genève développe un système de Gestion Intégrée d'une Pluralité de Données Issues de la Recherche (GIPDIR). Cet outil répond à la demande d'un chercheur de l'Unité d'archéologie classique du département des Sciences de l'Antiquité, directeur du « Projet Crotone »[1].

1 http://www.unige.ch/rectorat/maison-histoire/Recherche/ProjetCrotone.html (consulté le 2 juillet 2012).

Ce projet réunit autour de Crotone, une ville de Calabre dont les traces archéologiques documentent un passé historique particulièrement riche, au moins quatre disciplines scientifiques différentes et plus de cinquante chercheurs géographiquement éloignés.

Le GIPDIR vise à offrir une souplesse et une liberté de fonctionnement aux chercheurs. En effet, ceux-ci ont besoin de pouvoir créer des espaces de travail correspondant aux diverses équipes de recherche constitutives du projet, sans devoir en faire la demande aux administrateurs du système. De même, le responsable du projet est en mesure de gérer les droits d'accès et d'écriture dont disposent les participants à tel ou tel espace de travail. Ces différents espaces de travail, aux droits d'accès et d'écriture particuliers, sont utilisés par les chercheurs participants au projet pour partager des fichiers informatiques qu'ils génèrent ou dont ils ont besoin pour leur activité et pour y accéder de manière centralisée. Ces fichiers sont d'une grande diversité, comprenant des fichiers de bureautique usuels, des formats d'affichage tels que le PDF[2], des données GPS[3], des photographies digitales de différents formats ou des formats de logiciels de traitement d'image. À cela s'ajoutent la correspondance électronique et les pièces jointes que s'échangent les chercheurs.

L'utilisation de cette plateforme, qui, après sa phase de projet pilote, doit être mise à disposition d'autres projets de recherche dans d'autres facultés – ce qui élargit d'autant la diversité des formats informatiques concernés – revient à centraliser les données d'un projet de recherche, qui sont d'habitude réparties sur les différents postes informatiques de chaque chercheur. Cette centralisation constitue en réalité un pas dans la direction de l'archivage à long terme des données de la recherche.

Objectifs

Dans ce but, le service NTICE a besoin d'éclaircir quels sont les enjeux de l'archivage à long terme des données primaires et quelles recommandations peuvent être proposées pour un futur système d'archives et déterminer des fonctionnalités et des exigences à intégrer au projet GIPDIR. Avant même de disposer de ces recommandations, le NTICE doit pouvoir recourir à des arguments en faveur de la conservation et de l'archivage des données, car celui-ci entraîne forcément un effort accru de la part des chercheurs qui, on le comprend aisément, veulent concentrer leur énergie dans l'avancement de leur recherche.

2 Portable Document Format.
3 Global Positioning System.

De ce point de vue, qu'il s'agisse des arguments ou des recommandations, un équilibre est à trouver entre les besoins légitimes de la recherche scientifique et les exigences propres à un système d'archives sur le long terme de données numériques. Cet arbitrage s'obtient au cours d'un processus de négociation qui aboutit à la signature du protocole de versement des lots de données dans le système ouvert d'archivage d'information[4]. Bien entendu, les divergences et les tensions qui existent entre les différents acteurs concernés par un système d'archives des données de la recherche scientifique n'empêchent pas de réunir vers un objectif commun les efforts et les énergies de ceux-ci. Ce travail tente de mettre en évidence que les méthodes des professionnels de la documentation et des archives peuvent aider à cette mise en commun.

Structure

Nous commençons par définir les notions centrales de « données primaires », de « *digital curation* » et de cycle de vie des données, ainsi que des éléments terminologiques du domaine archivistique. Puis, nous présentons les arguments principaux qui viennent étayer l'effort d'archivage des données de la recherche, qu'ils soient d'origine légale, réglementaire, scientifique et sociale.

Avant d'entrer dans le vif du sujet, c'est-à-dire des recommandations que nous proposons pour un futur système d'archives et pour le projet GIPDIR, nous mentionnons les tensions qui peuvent exister entre les divers objectifs des différents acteurs concernés par le sujet de l'archivage des données primaires qui sont les chercheurs eux-mêmes, les administrateurs du système d'archives, les institutions dans lesquelles se déroulent la recherche et les bailleurs de fonds.

Nous espérons parvenir à donner des pistes de résolution de ces tensions au moyen des recommandations, que nous élaborons sur la base de la description du paquet d'informations qui est celle du modèle OAIS. Cette description met en évidence le besoin de métadonnées et leurs différents types. Puis, nous explorons le sujet des formats de fichiers numériques.

À ce moment de notre réflexion, nous pensons pouvoir dégager plusieurs responsabilités qui incombent au système d'archives lui-même, à savoir la définition et la publication de la mission et des objectifs clairs, qui doivent s'appuyer sur une volonté non moins claire de la direction de l'institution. À cela s'ajoute un processus d'enquête

4 C'est-à-dire un OAIS. Le modèle OAIS est expliqué en partie au point 4.1, page 23.

auprès des producteurs des données primaires, de préférence avant toute création de données, c'est-à-dire au moment de la planification du projet de recherche, afin de prévoir autant que faire se peut, et le plus précisément possible, quels types et quels volumes de données seront concernés et de quelles particularités il faudra tenir compte pour leur gestion, leur conservation et leur archivage.

Enfin, une activité intimement liée à cette enquête est abordée à ce point du travail : l'évaluation. Celle-ci permet de définir les critères, le moment et sur qui repose la responsabilité de l'évaluation des lots de données afin de les sélectionner, s'il y a lieu, en vue de leur archivage ou au contraire, de leur destruction. Puis, nous terminons, avant de conclure, en résumant ce que nous considérons comme les exigences minimales qui devraient être intégrées au projet GIPDIR, pour permettre un archivage des données primaires satisfaisant.

1. Définitions

Comme annoncé dans l'introduction, ce travail débute par la définition de quelques notions, soit parce qu'elles sont centrales, comme le sont les données primaires, soit parce qu'elles appartiennent à la terminologie propre à l'archivistique, qui est le domaine de l'information documentaire particulièrement concerné par le sujet qui nous occupe dans ces pages.

1.1. Données primaires et système d'archives

La recherche scientifique produit une très large documentation qui peut être répartie en quatre grandes catégories : les documents administratifs (par exemple la description du projet ou le dossier de requête de subsides), les données brutes, les données analysées et enfin les résultats de la recherche, c'est-à-dire dire les publications sur lesquelles elle débouche (Arovelius, 2010 : 11-12). Ces publications peuvent être considérées comme les données secondaires, et les trois premières comme les données primaires (Keller-Marxer, Buetikofer, 2008 : 5). Nous ajoutons à cette définition toute la documentation et les métadonnées qui décrivent ces données, leur processus d'acquisition, de création et de traitement, ainsi que les méthodes et les éventuels outils nécessaires à leur interprétation. Enfin, est intégrée à la notion de données primaires toute la correspondance échangée entre les chercheurs du projet, voire entre ceux-ci et des intervenants extérieurs, ainsi que les pièces jointes qui les accompagnent. C'est à cet ensemble très large que nous nous référons lorsque que, dans le présent travail, nous employons l'expression « données primaires » ou « données de la recherche ».

Ces données primaires sont intégrées, en vue de leur conservation, de leur archivage et de leur diffusion, dans un « système d'archives ». Par cette expression nous nous référons au modèle de « *data repository* » particulier et exigeant que constitue l'*Open Archival Information System* (OAIS). Ce modèle repose sur la définition abstraite, mais parfaitement claire du « paquet d'informations » contenant les données que le système d'archives a pour objectif de conserver et d'archiver (l'objet-donnée), ainsi que les informations nécessaires à cette conservation et cet archivage, mais également à la diffusion de ces données (l'information de représentation et l'information de pérennisation)[5].

5 Comme annoncé plus haut, une explication de ces éléments constitutifs du modèle OAIS se trouve au point 4.1, page 23.

Précisons encore que le projet GIPDIR n'est pas un système d'archives au sens de cette définition. En effet, cet outil se situe du côté du producteur de données, c'est-à-dire avant le paquet d'informations à verser. De plus, le GIPDIR a pour objectif premier de faciliter le partage des données primaires entre les chercheur d'un projet et non pas l'archivage à long terme de ces données. Toutefois, en vue de cet archivage, les données de recherche doivent être prises en charge avant même leur création, aussi le GIPDIR est-il bien concerné par la démarche qui nous occupe dans ce travail.

1.2. Digital Curation et cycle de vie

Dans la littérature et sur les sites web que nous avons consultés, se retrouve très régulièrement la notion de « *digital curation* », qui est particulièrement difficile à traduire en français. Si cette expression revient aussi souvent dans l'univers anglophone, et est reprise parfois telle quelle dans la documentation francophone ou germanophone[6], ce n'est pas uniquement un effet de mode ou une stratégie marketing. Le terme « *curation* », puisque c'est bien lui qui nous pose problème, regroupe dans son acception anglaise les notions de gestion, de conservation et de mise à disposition des collections, des documents ou des données. Il est à relever que le terme lui-même est souvent absent des dictionnaires anglophones qui renvoient au verbe « *to curate* » ou au substantif « *curator* », avec les acceptions liées soit au domaine juridique des curatelles, soit à la conservation au sens muséal du terme.

Et c'est bien d'une extension de cette dernière notion dont il est question ici. Pourtant, il semble tout à fait contre-indiqué de traduire « *digital curation* » par « conservateur de contenu numérique ». Le *Grand dictionnaire terminologique* proposé par l'*Office québécois de la langue française*[7] propose « éditeur de contenu » ou « organisateur de contenu » comme équivalent à « *digital curator* », mais à la lecture de la définition, nous ne pouvons que constater que cette notion, bien que proche de celle qui nous occupe, n'est pas tout à fait adaptée. Selon ce dictionnaire, l'éditeur de contenu « cherche, trie de façon sélective, collectionne, organise, commente et partage des contenus Web sur une thématique donnée, au sein d'une communauté. »[8]

6 Le projet « *Digital Curation* » de l'École polytechnique fédérale de Zürich utilise l'expression allemande « *Digitaler Datenerhalt* », dans laquelle « *Erhalt* » réunit l'idée d'acquisition et celle de conservation. http://www.library.ethz.ch/de/About-us/Projects/Digital-Curation (consulté le 10 juillet 2012).

7 http://gdt.oqlf.gouv.qc.ca/index.aspx (consulté le 4 juillet 2012).

8 http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26507021 (consulté le 4 juillet 2012).

The Digital Curation Centre (DCC)[9] offre une explication détaillée de ce qu'il faut entendre par « *digital curation* ». Dans le domaine des données de la recherche, de manière similaire à celui du *Records Management* ou de la gestion électronique des documents, la notion recouvre l'ensemble des actions en lien avec les données au cours de leur cycle de vie[10]. Elle réunit autour des données les différents acteurs concernés, depuis les responsables du projet de recherche et les chercheurs eux-mêmes, autrement dit les producteurs des données, jusqu'au public qui accède à ces données, en passant par les administrateurs du système d'archives. Ce sont les trois acteurs principaux du modèle OAIS : les producteurs, les administrateurs et le public cible.

Le cycle de vie débute par la conceptualisation (*conceptualize*), c'est-à-dire la planification de la création ou du relevé des données, mais également les différentes possibilités de stockage qui peuvent répondre aux besoins du projet de recherche, ainsi que de l'archivage à long terme. Puis, intervient la création (*create*) des objets numériques, ainsi que la collecte et la constitution des métadonnées qui doivent les accompagner[11] pour se conformer aux exigences du système d'archives. Ces données doivent bien entendu être accessibles aux chercheurs qui les utilisent ou de manière publique (*access and use*). C'est à ce moment qu'interviennent l'évaluation et la sélection (*appraisal and select*), afin de déterminer quelles sont les données qui doivent être archivées à long terme et lesquelles doivent être détruites (*dispose*). Les données sélectionnées pour la conservation à long terme sont acquises par le système d'archives (*ingest*), qui entreprend des actions pour les pérenniser et en préserver l'authenticité (*preservation action*). Si les données versées n'atteignent pas les standards exigés pour l'archivage à long terme, elles sont à nouveau évaluées et sélectionnées (*reappraisal*). Ensuite, elles peuvent êtres conservées (*store*), mises à disposition afin d'être consultées et réutilisées (*access and reuse*). Enfin, le système d'archives peut transformer ces données, afin d'en améliorer la conservation à long terme, principalement par le biais de migration d'un format à un autre (*transform*).[12]

Ce qui doit être retenu de la notion de « *digital curation* », c'est qu'elle couvre l'entier du cycle de vie des documents, et qu'elle engage la responsabilité de l'ensemble des

9 http://www.dcc.ac.uk/ (consulté le 4 juillet 2012).
10 Il est fait mention d'un modèle du cycle de vie des données primaires au point 4.4.2, à la page 32.
11 Voir le point 4.2, à la page 25.
12 Voir la figure 2 : DDC Lifecycle Model, à la page 32.

acteurs d'un système d'archives tel que le modélise la norme OAIS. L'expression « *digital curation* » met en évidence une certaine convergence des intérêts de ces acteurs, pour autant qu'ils parviennent à collaborer.

1.3. Fonds, série, dossier, document

L'archivage à long terme des données primaires de la recherche scientifique fait partie du domaine archivistique, aussi ne semble-t-il pas inutile de préciser brièvement les notions fondamentales de fonds, série, dossier et document. Le fonds regroupe l'ensemble de la production documentaire d'un producteur, qui peut-être une personne physique ou une organisation. Le fonds est constitué sur un principe fondamental en archivistique : le principe de provenance. Dans les archives, en effet, les documents ne sont pas classés comme une collection en bibliothèque, qui se base sur des critères thématiques, mais bien selon le contexte de leur production, puisque que les archives ne documentent pas seulement les informations qu'elles contiennent, mais également l'activité qui est à l'origine de leur production.

Il est possible de rencontrer plusieurs niveaux de fonds et de sous-fonds. Dans le contexte de notre travail, l'ensemble des archives des données primaires de la recherche d'une université correspond au fonds du niveau le plus élevé. S'ajoute un niveau inférieur correspondant aux différentes disciplines ou facultés.

Les fonds sont eux-mêmes constitués par des séries, qui rassemblent la documentation liée à une activité, et non plus à un producteur. Il n'est pas toujours évident de tracer une frontière tout à fait claire entre le fonds et la série. Par exemple, un projet de recherche est-il une série, une activité particulière d'un producteur comme une faculté, ou est-il en réalité un producteur, c'est-à-dire un fonds ?

La série se divise en dossiers qui sont autant d'affaires prenant place à l'intérieur d'une activité. Ces dossiers, enfin, sont constitués des documents eux-mêmes, c'est-à-dire de la plus petite unité archivistique qui correspond, dans le monde numérique, à un fichier particulier, avec son format et son extension.

1.4. Description

Les différentes strates de la hiérarchie documentaire des archives, désignées par l'expression « unité de description » doivent toutes, selon l'orthodoxie archivistique être décrites. Cette description est une « [p]résentation intellectuelle et matérielle [...] faite pour en donner une identification exacte et unique, en expliquer le contexte d’origine et

en permettre l'exploitation administrative ou historique. »[13] À bien des égards, elle recoupe la documentation sur les données, c'est-à-dire les métadonnées exigées par le modèle OAIS pour la conservation et l'archivage de données numériques. Mais, l'univers digital impose des éléments de description supplémentaires, comme l'information de représentation[14].

La description archivistique a d'autres fonctions que de simplement offrir un accès aux données au moyen des outils de recherche qu'elle rend possible. Par le fait même de décrire les unités de description dans leur imbrication hiérarchique, elle donne une image du contexte de la création des données primaires et des informations sur leurs producteurs. La description rend également visible les relations que les différents groupes de données entretiennent entre elles, et donc la cohérence de l'ensemble. Dans ce sens, elle participe au renforcement de l'intégrité des données (Ballegooie, Duff, 2006 : 8-9).

1.5. Évaluation[15]

Dans l'univers des archives, l'évaluation consiste à déterminer dans un ensemble de documents le sort de ceux-ci : certains documents sont détruits rapidement après leur création, d'autres sont conservés jusqu'à une échéance précise, puis sont détruits. Enfin, il existe une catégorie de documents qui, lorsque la période de leur utilisation est terminée, sont archivés sur le long terme. Les critères de cette évaluation reposent d'une part sur l'existence ou non d'exigences légales ou réglementaires qui déterminent la destruction, la conservation à terme ou l'archivage sur le long terme. D'autre part, lorsque la « valeur primaire » des documents, c'est-à-dire la raison pour laquelle ils ont été créés, disparaît, l'évaluation détermine si ces documents possèdent une « valeur secondaire », par exemple un intérêt historique, auquel cas ils sont conservés.

Dans le contexte de l'archivage des données primaires, l'évaluation pose des questions particulières. Tout d'abord du point de vue de la pertinence de ce processus. Le coût de l'espace de stockage disponible ne semble plus constituer un argument aussi central, car il ne cesse de diminuer, du moins à l'achat (Komorowski, sans date). Ces coûts restent toutefois non négligeables en ce qui concerne la maintenance, la nécessité des sauvegardes et des redondances, et surtout les actions nécessaires

13 Dictionnaire de terminologie archivistique, Direction des Archives de France, 2002.
14 Voir le point 4.1, à la page 23.
15 Ce point s'inspire principalement de Whyte, Wilson, 2010.

pour assurer l'intégrité et l'authenticité des données conservées, ainsi que des éventuelles migrations de formats pour garantir la pérennité des données sur le long terme.

Dans certaines disciplines scientifiques, les volumes nécessaires atteignent des niveaux tels – il peut s'agir de plusieurs pétaoctets par année – que les arguments pour une évaluation et une sélection des données primaires à archiver ne se fondent pas sur la notion de volume. Par contre, les ressources qui doivent être engagées pour la documentation de ces données, pour la récolte, la constitution et la gestion des métadonnées de qualité sont importantes, et ce sont des ressources qui ne sont alors plus disponibles pour l'avancement de la recherche elle-même. L'évaluation permet de ce point de vue d'éviter de gaspiller de l'énergie et du temps à documenter des données dont l'archivage ne présente pas un intérêt significatif.

Enfin, malgré la qualité des outils de recherche et des métadonnées rendant ceux-ci possible, l'accroissement non-maîtrisé des volumes conservés, archivés et mis à disposition, dégrade le rapport entre l'information pertinente et le bruit dans les résultats de recherche.

Néanmoins, il est particulièrement difficile de prendre aujourd'hui des décisions quant au choix des données primaires qui doivent être conservées sur le long terme, car nous ne pouvons pas nous faire une idée correcte des questionnements et des nouvelles possibilités de traitement des données qui seront ceux de la recherche de demain (Archaeology Data Service, 2012). C'est pourquoi, entre autres raisons, il est indispensable que les critères qui déterminent l'évaluation et la sélection des données soient, d'une part, établis avec le concours de la communauté scientifique et, d'autre part, transparents et objectifs, afin d'en permettre le contrôle, mais également afin d'informer la communauté scientifique future des choix qui ont été faits, et sur quelle base ils ont été effectués.

2. Arguments en faveur de l'archivage

Ce chapitre a pour objectif de réunir les différents arguments qui peuvent soutenir l'effort d'archivage des données primaires numériques. Dans un premier temps, nous relevons des éléments légaux et réglementaires qui sont susceptibles d'avoir un impact sur le mode de fonctionnement de la recherche, notamment du point de vue de la gestion des données. Ces éléments proviennent de différents acteurs. Les prescriptions légales sont bien entendu du ressort des pouvoirs publics, avec les particularités de ceux-ci dans le contexte helvétique et sa structure fédérale. Les règlements et les directives peuvent être élaborés soit par les bailleurs de fonds – nous avons analysé la réglementation produite par le Fonds national suisse de la recherche scientifique (FNS) –, soit par les universités elles-mêmes.

Dans un deuxième temps, nous rappelons que la déontologie et les principes de la science soutiennent en bonne partie la volonté de conserver, d'archiver et de mettre à disposition les données primaires. En effet, la science et la quête sinon de la vérité du moins d'une connaissance objective reposent notamment sur un degré d'ouverture suffisant pour permettre la critique, la confirmation ou l'invalidation des résultats de la recherche.

Précisons que, tout comme dans le sujet de l'ouverture des données publiques[16], il existe des limites de nature légale[17] et éthique qui restreignent une large diffusion des données de la recherche. Dans un certains nombre de cas, ces données sont constituées de données personnelles, qui ne peuvent être accessibles à toutes et à tous, du moins pas sans avoir été anonymisées. Des restrictions imposées par les droits de la propriété intellectuelle peuvent également exister, bien que ce soit en partie contradictoire avec les buts et les méthodes du progrès de la science.

2.1. Lois, règlements et directives

2.1.1. Lois fédérales

Dans certains secteurs, le cadre juridique est à la fois précis et contraignant, ce qui offre une assise solide aux efforts d'archivage. Nous pensons par exemple aux prescriptions légales en matière de comptabilité instituant un délai de conservation de

16 Open Data.

17 RS 235.1 Loi fédérale du 19 juin 1992 sur la protection des données (LPD) au niveau fédéral, ou RSG A 2 08 Loi sur l'information du public, l'accès aux documents et la protection des données personnelles (LIPAD), au niveau cantonal (Genève).

dix ans[18]. Un équivalent dans le domaine des archives scientifiques n'existe pas encore en Suisse. Toutefois, le terrain n'est pas totalement en friche et certains points de repère sont d'ores et déjà disponibles.

La Loi fédérale sur l'archivage (LAr) permet de définir deux éléments. D'une part, elle spécifie ce qui ressort de sa compétence, à savoir les organismes et les institutions qui dépendent de la Confédération, comme les Écoles polytechniques fédérales et certains instituts de recherche[19]. Les université et les hautes écoles, quant à elles, sont régies par des lois cantonales. D'autre part, la loi précise quels sont les documents qui doivent être conservés et archivés : « [t]ous les documents de la Confédération qui ont une valeur juridique, politique, économique, historique, sociale ou culturelle [...] »[20], ce qui concerne certainement les données produites par et pour la recherche scientifique.

Toujours au niveau de la législation fédérale, l'article 11*a*, alinéa 1 de la Loi sur l'encouragement de la recherche et de l'innovation (LERI) déclare que « [l]es institutions chargées d'encourager la recherche veillent à ce que les recherches qu'elles soutiennent soient menées selon les règles de bonne pratique scientifique. »[21] Le Fonds national suisse de la recherche scientifique (FNS)[22] reprend l'expression « bonne pratique scientifique » à l'article 32, alinéa 3 de son règlement des subsides (FNS, 2008 : 34), dans le chapitre qui définit les droits et les devoirs des bénéficiaires de son aide. Ce qui a son importance étant donné la situation prépondérante du FNS comme bailleur de fonds de la recherche scientifique à but non commercial en Suisse.

2.1.2. Bonne pratique et intégrité

La « bonne pratique scientifique » est une notion peu précise, bien qu'elle se réfère à une conception de la science communément acceptée. Nous pouvons assurément considérer que « l'intégrité scientifique » en fait partie. Or, elle est mentionnée dans le règlement des subsides du FNS à l'article 11, alinéa 3, article qui établit les cas dans lesquels le FNS n'entre pas en matière lors d'une requête de subsides (FNS, 2008 : 27). Pour une définition claire de l'intégrité scientifique, le FNS renvoie à une publication des Académies suisses des sciences (Salathé, 2008).

Ces *Principes de bases et procédures* apportent des arguments utiles à la question de

18 RS 220 Art. 962, C. Durée de conservation.
19 RS 152.1 Art. 1, al. 1 But et champ d'application.
20 RS 152.1 Art. 2, al. 1 Principes.
21 RS 420.1 Art. 11*a*, alinéa 1.
22 http://www.snf.ch/F/pages/default.aspx (consulté le 16 juin 2012).

la conservation et de l'archivage des données primaires. Dans la partie B, « principes de bases », l'article 3.1 stipule que :

> « *[p]our permettre la supervision de la recherche, la reproduction des essais et l'analyse ultérieure des données selon d'autres points de vue, il convient de documenter toutes les données (y inclus les données brutes) d'une manière claire, complète et précise. Les données et matériaux doivent être conservés de sorte que soient exclus tout dommage, toute perte ou toute manipulation. Il en va ainsi non seulement pour les données manuscrites, mais aussi pour les données électroniques. Il est nécessaire de documenter les incidents particuliers, tels que par exemple la perte de données et les écarts du plan de recherche initial. À la conclusion du projet, la direction du projet est responsable de la conservation des données et matériaux pendant une durée définie en fonction de la spécialité. Elle doit veiller à leur durabilité et à leur protection. »* *(Salathé, 2008 : 17)*

Nous avons jugé nécessaire de citer l'article en entier, car il précise clairement ce qui en terme d'intégrité scientifique est nécessaire, non seulement du point de vue de la conservation des données primaires (au sens large), mais également du point de vue de la documentation de ces données, de leur création, saisie et manipulation. Nous pouvons noter encore que la responsabilité de la conservation est imputée à la direction du projet de recherche. Le fait qu'elle soit en mesure de s'adresser à un système d'archives est assurément un soulagement pour cette direction et un pas supplémentaire décisif vers la pérennisation des données primaires.

Dans la partie « Mémorandum », au point 6, le texte des Académies suisses des sciences « approuv[e] les dispositions qui existent déjà dans certaines universités et facultés » et encourage les universités et les hautes écoles à élaborer « un règlement contraignant dans le but de garantir l'intégrité scientifique » (Salathé, 2008 : 11). L'Université de Genève dispose d'une directive à ce sujet, mise à jour en 2011 et publiée sur son site Internet le 12 avril 2012 (Université de Genève, 2012). Celle-ci aborde la question des données primaires, qu'elle désigne par les termes « données de bases » au point 2.6. Ces données, qu'elles soient sur support papier ou numérique, doivent être documentées « de manière claire, complète et précise », afin d'exclure les dommages, les pertes et les manipulations ciblées. Il est judicieusement précisé que la documentation doit permettre de distinguer les données brutes, les données traitées et les résultats. En plus d'aborder la question des droits d'accès à ces données pendant et après le projet de recherche, le règlement établit que la responsabilité de la conservation des données incombe au responsable du projet de

recherche.

On le voit, la directive relative à l'intégrité de la recherche de l'Université de Genève est très proche des *Principes de bases et procédures* des Académies suisses des sciences en matière d'intégrité. Du point de vue de la question de la conservation et de l'archivage des données de la recherche, la notion, relativement vague, de « bonne pratique scientifique » de la Loi sur l'encouragement de la recherche et de l'innovation, se clarifie avec les précisions apportées par les deux directives qui ont été abordées ci-dessus.

La conservation des données primaires, clairement documentées, permet bien entendu de soutenir une défense en cas de procédure pour comportement incorrect. Selon l'article 4, alinéa 2 du règlement des subsides du FNS, « la personne incriminée », doit être en mesure de « fournir des pièces justificatives » dans le but de démontrer sa bonne foi (FNS, 2009 : 2), ce qui ne saurait être possible si ces pièces (les données primaires) n'ont pas été conservées et documentées. De plus, dans la liste des comportements qui « [c]onstituent des infractions à l'intégrité dans le domaine scientifique » des directives de l'Université de Genève, figure non seulement « la falsification intentionnelle de données de base » ou la « dissimulation de données », mais également la « suppression de données de base consignées, avant l'expiration du délai de conservation prescrit ou après avoir pris connaissance du désir de tiers de les consulter » (Université de Genève, 2012 : 3.2.1).

Les prescriptions en matière d'intégrité du travail de la science et des bonnes pratiques de la recherche qui appuient l'effort de conservation et d'archivage des données primaires ne se limitent fort heureusement pas aux cas des comportements incorrects. Les directives de l'Université de Genève reviennent plusieurs fois sur le sujet de l'accessibilité des données primaires en cherchant à exprimer un équilibre entre une gestion relativement restrictive des droits d'accès (Université de Genève, 2012 : 2.6) et une mise à disposition plus large, afin de rendre possible la « répétition et/ou la vérification des expériences » (Université de Genève, 2012 : 2.7). Le règlement du FNS quant à lui spécifie à l'article 44, alinéa 1, lettre b, que les bénéficiaires des subsides du FNS doivent « [m]ettre à disposition d'autres chercheurs les données recueillies durant les travaux de recherche soutenus par le FNS et les déposer dans des fichiers scientifiques reconnus, conformément aux prescriptions du FNS » (FNS, 2008 : 39).

2.1.3. Résumé

Pour résumer ce survol des aspects légaux et réglementaires, nous pouvons retenir qu'il est vraisemblable que les données primaires puissent être considérées comme des documents à archiver au sens de la Loi fédérale sur les archives. Cette loi n'a pas d'impact sur les projets de recherche menés dans les universités et les hautes écoles qui dépendent des législations cantonales. Et en l'occurrence, nous n'avons pas trouvé de mentions d'archivage et de conservation des données primaires dans la loi cantonale de Genève.

Par contre, l'université elle-même, ainsi que les bailleurs de fonds, le FNS par exemple, prescrivent, voire exigent que les projets de recherche se conforment à des bonnes pratiques, ce qui induit la documentation, la conservation et l'archivage des données primaires, et ceci afin de répondre à plusieurs besoins : piloter la recherche elle-même, assurer la reproductibilité des résultats, démontrer l'intégrité de la démarche scientifique en cas de suspicion, permettre une analyse ultérieure des données selon des angles nouveaux.

2.1.4. Le fonctionnement de la recherche

L'angle légal et réglementaire a été abordé afin de relever les exigences administratives, externes au fonctionnement intrinsèque de la science. Ces exigences ou directives ne naissent toutefois pas du néant, elles s'appuient sur une conception de la science qui est largement admise. La conservation et l'archivage des données scientifiques découlent de cette conception et contribuent à renforcer la qualité de la recherche scientifique.

Le fonctionnement de la science vise un objectif majeur qui est de constituer une connaissance objective sur laquelle il est raisonnable de s'appuyer. Pour assurer cet objectif, il est nécessaire que toute affirmation soit soutenue par une argumentation et un raisonnement ouvert, c'est-à-dire explicite et partagé. Cette publicité permet à la communauté scientifique de vérifier la validité du raisonnement et la véracité des faits avancés. Dans les cas où un désaccord intervient, cette même publicité permet une critique argumentée et c'est au moyen de ce genre de conversation ouverte et publique que le progrès de la connaissance est rendu possible.

L'effort nécessaire pour parvenir à un tel degré d'ouverture passe par l'établissement et le suivi de procédures, propres à chaque discipline, ainsi que par la documentation de tous les aspects d'un projet de recherche, que ce soit la gestion du projet, le relevé de

mesures, la constitution de données, la mise en place et le déroulement d'expériences.

La nécessité de documenter les processus de la recherche étant établit, nous pouvons souligner les bénéfices de l'archivage à long terme des données primaires de la recherche. Pour ce faire nous nous inspirons notamment du manuel pour la gestion et l'archivage des données scientifique publié par le Conseil international des archives (Arovelius et al., 2010 : 30).

La documentation des données primaires augmente la confiance que l'on peut accorder à une démarche scientifique, puisque, comme il a été indiqué plus haut, cette documentation permet de vérifier que les résultats effectifs correspondent bien à ce qui est affirmé, qu'ils sont bien le produit d'une méthode précise et documentée. Plus encore, cette documentation assure que la recherche est bien en accord avec les normes éthiques en vigueur. Aussi, conserver et archiver à long terme cette documentation contribue à une recherche de bonne qualité.

La conservation à long terme des données primaires, à condition qu'elles soient documentées avec précision, rend possible la réutilisation de ces données dans le futur. Cette réutilisation est souhaitable pour diverses raisons. La constitution d'un ensemble de données a un coût – financier, en terme de temps d'acquisition, etc. –, qui peut atteindre des niveaux élevés dans certaines disciplines et il est donc économiquement intéressant d'avoir accès à des lots de données déjà existants, lorsque le projet de recherche le permet. Dans d'autres situations, il n'est tout simplement pas possible de reproduire les données primaires, puisque le « terrain » duquel les données ont été extraites est voué à disparaître (par exemple des témoins), ou a été détruit par le fait même du projet de recherche (une fouille archéologique). Il faut également garder à l'esprit que de nombreux projets de recherche, si ce n'est la totalité d'entre-eux, ne sont pas en mesure d'épuiser toutes les questions qu'il est possible de poser à un ensemble de données, ni toutes les conclusions que l'on peut en retirer. Ceci est d'autant plus vrai que nous n'avons aucun moyen de connaître aujourd'hui les interrogations de la recherche de demain. Il nous est donc très difficile de prétendre pouvoir évaluer quelles données pourraient être détruites, dans le but de diminuer les coûts et les efforts induits par leur conservation et leur archivage. Enfin, la conservation et l'archivage des données de la recherche fourniront la matière et les sources à l'histoire des sciences et à la philosophie des sciences.

Les chercheurs peuvent aussi trouver un avantage supplémentaire dans l'existence de

système d'archives numériques. Parmi les diverses fonctions de tels systèmes se trouve l'attribution d'un identifiant pérenne, comme par exemple un *Digital Object Identifier* (DOI)[23], pour chaque lot de données. Cet identifiant permet donc de se référer de manière non ambiguë à un ensemble de données lors de la rédaction d'un article scientifique.

2.1.5. Le public contribuable

Un dernier argument trouve son origine dans un raisonnement proche de celui qui sous-tend la démarche des données publiques ouvertes. Les données produites par la recherche font la plupart du temps partie de projets qui ont été financés, en partie ou intégralement, par des fonds publics et donc par les citoyennes et les citoyens. C'est pourquoi il ne semble pas absurde d'imaginer que ces données puissent, à terme, s'élever dans le domaine public et être accessibles en consultation au plus grand nombre. Des auteurs n'hésitent pas à l'affirmer : « It is widely accepted that publicly funded research data is a public good and that it should be made available for sharing and reuse »[24] (Caplan, 2006 : 15).

Ce n'est pas uniquement une question d'origine du financement, mais une question bien plus large. La science, comme bien d'autres activités humaines, est mue par des mobiles individuels honorables à des degrés divers, mais d'un point de vue plus général, elle a comme mission d'améliorer le sort de tous. C'est d'autant plus vrai dans le contexte d'une société qui affirme ses idéaux démocratiques. En conséquences, les produits de la science, y compris les données primaires de la recherche, doivent être conservés et archivés, afin d'en permettre la diffusion.

23 Voir la note 34, à la page 24.

24 « Il est largement admis que les données produites par une recherche financée par des fonds publics sont un bien commun et qu'elles doivent être partagées et accessibles pour une utilisation ultérieure » (notre traduction).

3. Les tensions entre les différents acteurs et objectifs

Dans ce chapitre, nous nous appuyons notamment sur la confrontation entre les informations que nous avons pu récolter dans la littérature sur l'archivage numérique à long terme des données primaires, et les points qui nous ont été signalés au cours des entretiens avec deux chercheurs, à savoir le chercheur à l'origine de la demande de développement du projet GIPDIR et un chercheur en sciences informatiques. À ces deux entretiens s'ajoutent les conversations que nous avons pu avoir avec la direction du NTICE, mais également avec l'archiviste de la ville de Genève.

Les arguments[25] à l'origine de la tendance de l'archivage des données primaires sont de natures diverses, allant d'exigences légales, d'ailleurs à préciser et à développer, à des considérations sociétales ou politiques, en passant par les directives et règlements des universités et la méthodologie scientifique elle-même. Or, ces différents angles d'approche ont des objectifs, et donc des implications, qui peuvent parfois être contradictoires ou entrer en conflit les uns avec les autres. Aussi est-ce le but de ce chapitre d'explorer ces tensions, de trouver les points où des compromis semblent possibles et où des convergences apparaissent.

Les recommandations de bonnes pratiques et les directives produites par les Académies des sciences, le FNS ou l'Université de Genève insistent sur la notion d'intégrité de la recherche. Il s'agit de garantir la qualité et l'authenticité des publications scientifiques, et de lutter contre les fraudes et les manipulations éventuelles (Keller-Marxer, Buetikofer, 2008 : 7-8). Pour le permettre, les données doivent être conservées intégralement, sans procéder à une évaluation et une sélection. Les formats dans lesquels ces données ont été produites et analysées doivent être respectés, même s'ils sont binaires ou propriétaires. En effet, si des données ont été éliminées ou modifiées, il semble difficile de vérifier si celles-ci correspondent bien aux affirmations contenues dans les publications scientifiques qui en découlent. Toutefois, cette intransigeance doit être relativisée, car d'un point de vue plus pragmatique, une modification des « *bits streams*[26] » qui n'altère pas le contenu des données, ainsi que leur représentation, est possible.

Malgré tout, les exigences liées à permettre la vérification de l'intégrité de la recherche

25 Il est bien question dans ce travail des arguments pour l'archivage à long terme des données primaires, et non pas des causes ou des origines de cette tendance, analyse qui nous intrigue passablement, mais qui s'écarte de notre sujet.

26 La séquence de bits qui constitue un fichier informatique.

et l'archivage restent difficile à concilier. Bien qu'il soit question dans ce travail de formats et de supports numériques, un délai de conservation de trois, cinq ou dix ans ne peut pas être considéré comme équivalent à un « long terme ». Dans le monde analogique, du moins dans celui du papier, le long terme suppose des fourchettes temporelles supérieures au siècle, ce qui pose de nombreuses questions, par exemple du point de vue de la construction des bâtiments dans lesquelles doivent être conservées des archives. S'il est vrai que l'expérience humaine du « long terme » des données numériques est forcément limitée, le modèle OAIS le définit comme « étant suffisamment long pour être soumis à l'impact des changements technologiques, y compris à la prise en compte de nouveaux supports et nouveaux formats de données ou à des changements de la communauté d'utilisateurs. Le long terme peut se poursuivre indéfiniment » (Organisation internationale de normalisation, 2003 : 1-1).

L'absence d'évaluation est également un critère qui différencie la conservation pour assurer l'intégrité scientifique de l'archivage des données primaires. L'évaluation est une action importante dans le cycle de vie des données primaires, qui a le principal avantage d'éliminer en partie le bruit qui, sinon, serait généré par les requêtes formulées dans les outils de recherche. Enfin, il n'est raisonnable d'imaginer pouvoir réutiliser des formats binaires ou propriétaires que dans un délai très court, car il est indispensable de pouvoir disposer des logiciels capables de représenter les données, voire du matériel informatique nécessaire à leur exécution. Or, il n'est pas possible de compter sur un avenir stable dans ce domaine.

Nous pensons qu'il est important d'avoir à l'esprit que la conservation dans l'objectif d'assurer l'intégrité de la recherche et l'archivage à long terme des données primaires ne sont pas parfaitement compatibles, et qu'il semble difficile de vouloir assurer ces deux buts au moyen d'une procédure unifiée.

D'autres objectifs de l'archivage à long terme des données primaires sont mis en évidence par les arguments énumérés au chapitre 2. Il s'agit principalement d'en donner l'accès à un public ciblé, dont la définition est susceptible d'évoluer dans le temps, et de rendre possible leur réutilisation. Bien que ces buts ne soient pas aisément accessibles, ils ont néanmoins l'avantage de rejoindre certains besoins de la recherche elle-même, comme bien entendu la critique de la méthode et des résultats par les pairs, qui concourt à la qualité de la science. Le fait que les données primaires, documentées, soient conservées améliore les conditions de cette critique. Les chercheurs sont bien entendu les plus concernés par l'éventuelle réutilisation des

données primaires. Enfin, les chercheurs tirent profit de la possibilité de pouvoir citer un lot de données précis, comme ils citent des articles scientifiques, mais aussi du fait que les données qu'ils ont produites soient citées par d'autres auteurs. Ces fonctionnalités ne peuvent être assurées par le système d'archives qu'au prix d'un effort de documentation, même si tout doit être entrepris pour le maintenir à un niveau le plus réduit possible.

Par ailleurs, la recherche a également besoin pour fonctionner de partager et d'échanger, au sein d'un projet de recherche, de grandes quantités d'informations, de données et de documents. Si l'archivage des données primaires est en mesure de répondre aux besoins précédents, ce n'est pas le cas pour ces besoins d'échange. C'est bien à ce besoin que répond le projet GIPDIR. Lors d'un entretien, le chercheur qui est à l'origine de la demande du développement de ce projet, a insisté sur certains éléments pour que le projet GIPDIR reste adapté aux besoins du fonctionnement quotidien de la recherche. L'objectif principal est de faciliter l'échange de fichier au moyen d'une plateforme centralisée. Celle-ci doit être simple à utiliser et ne pas demander un apprentissage supplémentaire de la part des chercheurs, qui ont besoin de pouvoir avancer dans leur travail. Elle doit supporter tous les formats qu'utilisent les chercheurs pour leur activité et ne pas les forcer à recourir à d'autres outils logiciels pour se conformer à des exigences de conservation.

L'outil GIPDIR a également pour objectif de donner au responsable du projet de recherche la plus grande autonomie et liberté possibles du point de vue de la structuration des répertoires de fichiers. Celle-ci doit refléter les préoccupations et les représentations intellectuelles des chercheurs eux-mêmes et ne pas être contrainte par l'effort de normalisation des documentalistes. Enfin, la documentation, c'est-à-dire les métadonnées, doivent être récoltées le plus possible de manière automatique et, du moins, rester très basiques et non obligatoires. Les chercheurs n'ont pas toujours les ressources nécessaires, en temps et en personne, pour documenter des données, d'autant qu'une part non négligeable d'entre elles sont susceptibles d'être détruites, ce qui doit être vérifié par la définition des critères d'évaluation lors de l'enquête[27].

On le voit, des tensions existent entre les objectifs des différents acteurs concernés par la recherche et leur archivage, à savoir les bailleurs de fonds, les institutions telles que les université et les hautes écoles, les chercheurs, voire les citoyennes et les citoyens

27 Voir le point 4.5, à la page 34.

dans leur ensemble[28]. Ces diverses tensions peuvent être arbitrées en tenant compte des coûts induits, par exemple pour la constitution des métadonnées, et les bénéfices que peut en retirer la recherche. De même, une prise de conscience de la part des bailleurs de fonds ou des universités, que les coûts de l'archivage des données primaires ne se résument pas aux coûts de stockage, est nécessaire. Enfin, et cet aspect est abordé plus en détail au chapitre suivant[29], une enquête dès la planification du projet de recherche est en mesure d'établir les directives et des procédures fixant un niveau minimal de documentation et facilitant le renseignement des métadonnées. Ces éléments font également partie de la négociation entre l'administration du système d'archives et les producteurs, en vue de fixer le protocole de versement, qui définit les conditions dans lesquelles un paquet d'informations à verser est accepté par le système d'archives.

28 Les citoyennes et les citoyens ont certainement besoin d'une documentation différente que la communauté scientifique qui a produit les données.

29 Notamment au point 4.4.2, à la page 32.

4. Recommandations

Ce chapitre a pour objectif de mettre en évidence quels sont les points à clarifier, les démarches à entreprendre et les informations à renseigner, afin de concourir à un système d'archives numériques des données primaires, qui parvienne à remplir sa mission. Celle-ci est d'assurer une conservation pérenne des données de la recherche et leur accessibilité sur le long terme, ce qui suppose la possibilité d'utiliser des outils de recherche pour les sélectionner, de disposer de la documentation nécessaire à leur compréhension et à leur réutilisation.

Aussi avons-nous jugé utile de commencer par brièvement décrire le concept de « paquet d'informations » selon le modèle OAIS. Cette description met en lumière l'importance des métadonnées dans la constitution d'un paquet d'informations, et souligne que le besoin en métadonnées est déterminé par les exigences de la conservation et de la diffusion de l'information que doit assurer le système d'archives. Enfin, nous expliquons que ces métadonnées ne peuvent être exclusivement récoltées de façon automatique et que les producteurs doivent renseigner les métadonnées dont ils sont les seuls à maîtriser le contenu, et ce le plus tôt possible dans le cycle de vie des données.

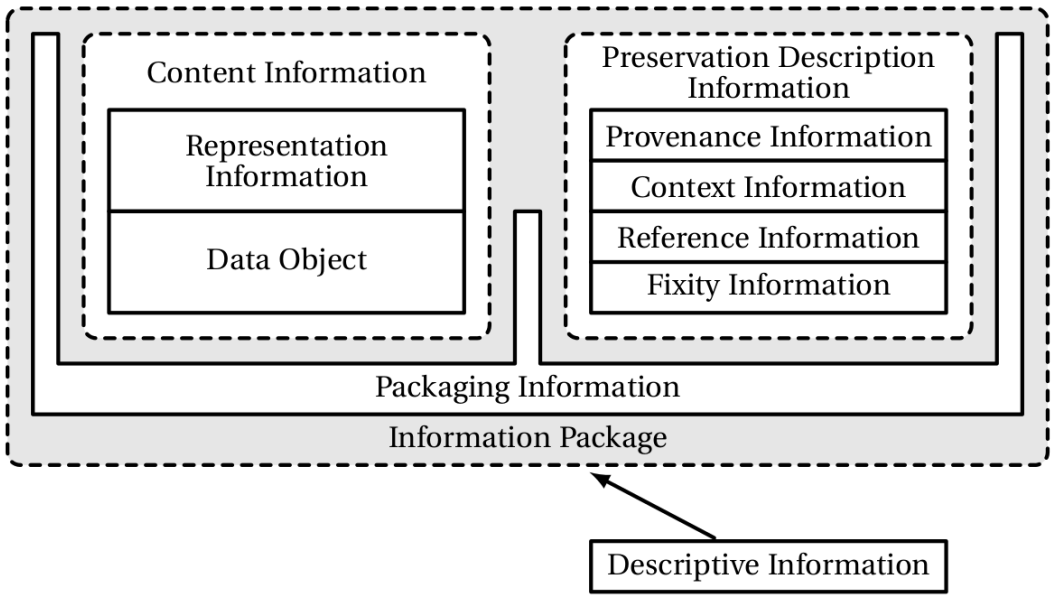
Puis nous abordons la question des formats dans lesquels les données primaires ainsi que les métadonnées associées sont versées dans le système d'archives en vue de leur conservation et de leur diffusion. De l'exploration des thèmes des métadonnées et des formats, nous aboutissons au constat que ces questions ne peuvent trouver de réponses sans clarifier plus précisément les missions et les objectifs du système d'archives. Enfin, une enquête réalisée par les gestionnaires du système d'archives et les chercheurs, au moment de la planification de la recherche est indispensable, afin de déterminer les responsabilités de chacun des acteurs et les procédures qui doivent être mises en œuvre.

Nous avons conscience que ces directives restent à un niveau très général. Ce point de vue nous a semblé toutefois adapté à la situation. En effet, le système d'archives des données primaires est, à notre connaissance, encore à l'état de projet.

4.1. Le paquet d'informations selon le modèle OAIS[30]

Le modèle OAIS (Open Archival Information System), développé par le *Consultative Committee for Space Data Systems*[31] (CCSDS) dans le contexte de l'archivage des données des sciences spatiales, est rapidement devenu une norme internationale[32] (Organisation internationale de normalisation, 2003). Il propose deux modèles – de l'information et des fonctions d'un système d'archives –, qui permettent d'utiliser une terminologie largement répandue dans le domaine des archives numériques.

Figure 1: Modèle du paquet d'informations (IP)



(Source : Ball, 2010 : 7)

L'information est décrite comme un « paquet d'informations » constitué de plusieurs éléments. Ce paquet contient deux parties principales. La première partie, appelée « contenu d'information » (*content information*), est elle-même constituée de deux éléments : l'objet-donnée (*data object*), à savoir l'information que le système d'archives a pour mission de conserver, et l'information de représentation (*representation information*). Cette dernière est « essentially the technical information (or metadata)

30 Ce point se réfère à Organisation internationale de normalisation, 2003 et Ball, 2010 : 6-7-8.

31 http://public.ccsds.org (consulté le 25 juin 2012).

32 ISO 14721:2003. Le CCSDS a publié sur son site, le 14 juin 2012, une nouvelle version de ce modèle qui peut être téléchargée à l'URL suivante : http://public.ccsds.org/publications/archive/650x0m2.pdf (consulté le 25 juin 2012).

needed to render the bit sequences into something meaningful » (Day, 2005 : 21)[33]. Il peut s'agir d'informations liées au domaine de l'informatique (une description de l'encodage ASCII, les spécifications d'un format informatique, le système d'exploitation, le logiciel et sa version utilisés lors de la création des données) ou d'informations en rapport avec les données (le fait que la suite de nombres conservés représentent, par exemple, des mesures de températures exprimées dans telle unité).

L'information de représentation constitue à son tour un objet-donnée qui doit être rendu plus intelligible, selon les cas, par une nouvelle information de représentation. Le niveau de récursivité et de détail de l'information de représentation dépend en grande partie du public que vise le système d'archives du point de vue de la diffusion, puisque c'est bien l'objectif d'un OAIS : archiver et conserver de l'information, sur le long terme, dans le but qu'elle soit mise à disposition d'un public, qui doit être en mesure de la comprendre et de l'utiliser.

Au contenu d'information s'ajoute l'information de pérennisation (*preservation description information*), qui rassemble les informations nécessaires pour assurer la conservation du contenu d'information sur le long terme. Ces informations sont subdivisées en quatre éléments. Premièrement, l'information de provenance (*provenance information*) relève l'histoire des données à conserver, à savoir leur origine et/ou leur source (qui en a été le créateur ou le collectionneur), les modifications dont elles ont été l'objet, et qui en a été le responsable. Deuxièmement, l'information de contexte (*context information*) décrit les relations qu'entretiennent les données conservées avec leur environnement, c'est-à-dire leur rapport avec d'autres ensembles d'informations, ainsi que les raisons de leur création, de leur récolte ou de leur saisie. Troisièmement, afin de pouvoir identifier de manière univoque le paquet d'informations, une information d'identification (*reference information*) est requise, comme par exemple un numéro ISBN ou un *Digital Object Idenfifier* (DOI)[34]. Quatrièmement, l'information d'intégrité (*fixity information*) assure que les données n'ont pas été modifiée entre le moment de leur conservation et celui de leur consultation, ce qui peut être réalisé grâce à la technique de la somme de contrôle

33 L'information de représentation est « essentiellement toute l'information technique (ou les métadonnées) indispensable pour transformer les séquences de bits en quelque chose d'intelligible » (notre traduction).

34 Il existe un consortium international de bibliothèque qui propose des identifiants persistant aux systèmes d'archives participant au consortium : *DataCite* http://www.datacite.org (consulté le 9 juillet 2012). À noter qu'en mai 2012, la norme ISO 26324:12 sur le *Digital Object Identifier System* a été publiée.

(*checksum*).

Le contenu d'information et l'information de pérennisation constituent un paquet d'informations en étant réunis par l'information d'empaquetage (*packaging information*). Celle-ci décrit comment, de manière physique ou logique, les éléments constitutifs du paquet d'informations sont regroupés sur un support ou dans un format donné.

Enfin, puisque la conservation de ce paquet d'informations n'a de sens que dans un contexte plus large et dans le but d'être consulté, ce qui suppose qu'il soit recherché et trouvé, il est nécessaire de le doter d'une information de description (*descriptive information*), c'est-à-dire, le plus souvent, d'une notice de catalogue. Notons que certaines informations contenues dans l'information de pérennisation se retrouvent dans l'information de description, comme par exemple l'information d'identification et une partie de l'information de provenance et de contexte.

Ce modèle de paquet d'informations se décline en trois paquets d'information différents, correspondant aux trois fonctions fondamentales du système d'archives lui-même. En premier lieu, selon un ordre chronologique, vient le paquet d'informations à verser (*submission information package*), que l'on abrège par l'acronyme SIP. Il s'agit de l'information créée ou rassemblée par le producteur et qu'il verse au système d'archives. La fonction d'acquisition de l'OAIS traite les paquets versés et en constitue des paquets d'informations archivés (AIP pour *archival information package*). Bien que dans leur structure les SIP et les AIP ne diffèrent pas, lors du passage du SIP vers l'AIP, s'ajoute l'information de description. C'est également à ce moment que des changements de formats peuvent avoir lieu, ainsi qu'une normalisation des métadonnées. Enfin, lorsque les données archivées sont consultées, l'OAIS délivre un paquet d'informations diffusé (DIP pour *dissemination information package*). Les formats lors de la diffusion peuvent différer de l'AIP et d'un DIP à l'autre, selon les usages (consultation ou réutilisation).

4.2. Les métadonnées

De cette description conceptuelle du paquet d'informations, nous pouvons différencier deux types d'informations : les données à conserver elles-mêmes (l'objet-donnée) et les données sur ce noyau d'information, les métadonnées. Ces dernières constituent une part importante du paquet d'informations, sans lesquelles il n'est tout simplement pas envisageable de conserver durablement de l'information. Non seulement elles permettent d'entreprendre des actions de pérennisation et rendent possible de

retrouver, d'accéder et de réutiliser des données archivées, mais elles documentent également le contexte de leur création, les organisations et les personnes productrices des données archivées, leur histoire. Plus encore, les métadonnées sont garantes de l'intégrité et l'authenticité de l'ensemble de données dans lequel un paquet d'informations particulier s'insère (Ballegooie, Duff, 2006 : 8-9).

Les métadonnées se déclinent en différents types, selon ce qu'elles décrivent. Il y a les métadonnées techniques, encore appelées de gestion ou administratives. Elles concernent toutes les métadonnées nécessaires à la gestion des données archivées : des informations sur les processus de création, sur les formats de stockage, sur leur provenance et sur les droits de propriété intellectuelle ou les restriction de diffusion. Ensuite, on trouve les métadonnées de structure qui documentent l'affichage et les possibilité de navigation entre les données (par exemple une table des matières, ou une structure de liens hypertextes internes). Enfin, les métadonnées de description qui sont les éléments indispensables à la constitution d'outil de recherche. C'est bien entendu dans ce dernier groupe de métadonnées que l'univers archivistique a le plus à apporter. Cette répartition des métadonnées se retrouve dans le *Metadata Encoding Transmission Standard* (METS), qui est conforme au modèle OAIS (Day, 2005 : 8).

Il existe des ensembles de métadonnées utiles pour rassembler les différents groupes d'informations constitutifs d'un paquet d'informations du modèle OAIS. Les informations de pérennisation peuvent être encodées au moyen du schéma XML PREMIS, proposé par *Preservation Metadata Implementation Stategies*[35]. Relevons que d'un projet d'archivage à l'autre, les besoins en information de préservation ne sont pas les mêmes, en fonction d'un certain nombre de facteurs, comme par exemple le type de public visé (Caplan, 2006 : 12). Les informations de paquetage peuvent elles être intégrées dans le schéma XML du METS (Day, 2005 : 24-25-26).

Quant aux informations descriptives, elles doivent se conformer aux normes archivistiques existantes, comme ISAD(G) (Conseil international des archives, 2000). La norme permet de déterminer un ensemble d'informations minimales. Elles se répartissent dans la « zone d'identification » et la « zone de contexte » et concernent les informations suivantes :

- la référence : il s'agit de pouvoir identifier avec précision l'unité de

35 Le développement de cette initiative est maintenu par la *Libray of Congress* et est accessible à l'URL suivante : http://www.loc.gov/standards/premis/ (consulté le 26 juin 2012).

description[36] ;

- l'intitulé de cette unité de description ;
- les dates : les dates de création des documents ou les dates indiquant la période de rassemblement des documents ;
- le niveau de description : l'unité décrite est-elle un fonds (par exemple un projet de recherche), une série (une équipe particulière de ce projet de recherche), un dossier (telle campagne de prises de vues), un document (telle photographie)[37] ;
- l'importance matérielle de l'unité de description, afin de déterminer le volume (nombre de séries, dossiers, documents) et les supports, ou formats qui la compose ;
- le ou les producteurs de l'unité de description, qui doit être indiqué selon une forme normalisée, à savoir ISAAR(CPF)[38].

La constitution des métadonnées a un coût non négligeable en termes de ressources techniques et humaines, mais également en temps. Il s'agit donc de trouver un bon équilibre entre les besoins de pérennisation et de description, et les ressources disponibles pour cette activité. Aussi, la question de l'automatisation de la recherche, de la saisie et de la gestion des métadonnées est cruciale. Elle est possible grâce à l'analyse des objets numériques eux-mêmes ou au recourt à des bases de données rassemblant des informations de représentation ou de la documentation sur les formats. (Day, 2005 : 26-27). La *National Library of New Zealand* a développé un outil libre d'extraction de métadonnées, le *Metadata Extraction Tool*[39], qui réunit les différentes données dans un fichier XML. De plus, les données liées à des événements internes au système d'archives (vérification des sommes de contrôle, par exemple) peuvent être relevées et intégrées aux informations de pérennisation de manière automatique. (Caplan, 2006 : 15)

Un autre point fondamental au sujet des métadonnées réside dans le fait que pour une part importante de celles-ci, le producteur est le seul a être en mesure de les renseigner. Le producteur doit donc dès la création des données, et même avant, lors de la planification du projet de recherche, documenter le contexte des données, selon quelles procédures ces données sont produites et comment elles doivent être interprétées. Ces métadonnées doivent être constituées en fonction du public cible défini par le système d'archives. Or, ce public cible n'est pas forcément homogène et

36 Voir le point 1.3, à la page 8.
37 Ce sont là des indications qui ne correspondent pas forcément à une réalité du terrain.
38 ISAAR(CPF) normalise la description des producteurs d'archives, qu'ils soient des collectivités, des personnes ou des familles (Conseil international des archives, 2004).
39 http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-extraction-tool/ (consulté le 26 juin 2012), Apache License V2.0.

peut être constitué à la fois de diverses communautés scientifiques correspondant à diverses disciplines, y compris les historiens des sciences, mais également un public plus large. En conséquence, le degré explicatif des métadonnées diffère d'un public à l'autre et n'est pas trivial à établir. Surtout, le niveau de précision ou de vulgarisation exigé doit être explicité dans le protocole de versement qui est négocié entre l'administration de l'OAIS et le producteur, dans l'idéal avant que les données soient produites, c'est-à-dire au moment de la planification du projet de recherche[40].

4.3. Les formats

L'expression très large de « recherche scientifique » recouvre un nombre élevé de disciplines scientifiques. Celles-ci utilisent un ensemble relativement réduit de formats informatiques communs et un ensemble très étendu de formats propres à leurs méthodes et besoins. Or, un système d'archivage qui se veut ouvert et pérenne, c'est-à-dire conforme au modèle OAIS, est amené à préférer des formats dont la documentation est publique et qui permettent la meilleure interopérabilité possible, c'est-à-dire offrant une certaine indépendance par rapport au système d'exploitation ou au logiciel utilisé.

Tableau 1 : Liste de formats pour l'archivage numérique

Domaine d'utilisation	Formats compatibles	Remarques
Texte (non structuré)	"Texte uniquement" ("plain text")	UTF-8 UTF-16 ISO-8859-1 ISO 8859-15 US-ASCII
Bureautique	PDF/A	Correspond à PDF 1.4 avec restriction
Tableurs	CSV	comma separated value
Bases de données relationnelles	SIARD RDB DATA	
Format de fichier graphique Bitmap	TIFF	
Audio	WAVE	

(Source : Archives fédérales suisse, 2007 : 2)

À l'observation des listes de formats informatiques acceptés par des services d'archives officiels, nous sommes frappés par le nombre très réduit des formats proposés, du moins en regard des besoins de l'archivage des données primaires de la

40 Ce qui est plus longuement détaillé au point 4.4.2, page 32.

recherche. Les Archives fédérales suisses (AFS) proposent une fiche d'information intitulée « Formats de fichiers adaptés à l'archivage », dans l'introduction de laquelle il est expliqué que le nombre de formats acceptés est volontairement réduit, afin de garantir une meilleure pérennité sur le long terme des archives numériques (Archives fédérales suisses, 2007 : 1). Ce qui donne le tableau « liste de formats pour l'archivage » ci-dessus[41] (tableau 1).

Dans le domaine de l'archivage numérique des données de la recherche, il existe d'autres listes de formats plus étendues. Par exemple, l'*Archaeology Data Service* (un système ouvert d'archives numériques britannique), qui propose un ensemble très complet de guides, aborde la question des formats. À chaque type de donnée correspond un guide particulier : pour les documents et les fichiers textes, les bases de données et les tableurs, les images bitmap, les images vectorielles, les vidéos et l'audio (Archaeology Data Service, 2011). Pour ne prendre que l'exemple des images bitmap (*raster images*), alors que les directives des archives officielles de la Confédération helvétique ou du Canton de Genève n'acceptent que le TIFF, le guide de l'ADS décrit non seulement une liste de formats qui couvre la plupart des situations que peuvent rencontrer des chercheurs, mais explique également les choix induits par la création des images. Si l'ADS préfère naturellement le TIFF, il entre en matière sur d'autres formats, dans le cas d'utilisations particulières et selon certaines conditions (Archaeology Data Service, 2012).

Le cas des photographies numériques est exemplaire. Lors de la prise de vue, l'appareil, s'il est de qualité, offre deux choix de formats : habituellement le JPG (donc avec une compression) et le RAW, équivalent à un master. Le format RAW, en plus du fait qu'il est bien plus volumineux que le JPG, n'est ni un format de diffusion, ni un format d'archivage, notamment parce que la majorité des fabricants ne proposent que leur propre version propriétaire[42]. À partir d'un format RAW, il est possible d'obtenir des fichiers TIFF, des JPEG2000 ou des DNG, mais ce type de conversion n'est pas trivial, à la fois en terme de compétences techniques et de logiciels nécessaires. Le JPG, lui, subit une compression avec perte de qualité dès la prise de vue, et il peut sembler à première vue peu pertinent de le convertir en JPEG2000 (c'est-à-dire avec une compression sans perte) ou en TIFF, puisque la qualité n'en sera pas améliorée.

41 Les archives de l'État de Genève ont publié un tableau très similaire, dans le même objectif (République et Canton de Genève, Archives d'État, 2011).

42 Il existe bien le DNG, qui est une version ouverte du format RAW, mais il n'est pas répandu, ni au niveau des appareils, ni à celui des logiciels de traitement de l'image.

Toutefois, le format JPEG2000 offre d'autres avantages qui ne doivent pas être dédaignés : c'est un format reconnu pour la conservation à long terme, et son algorithme de compression permet d'offrir différentes résolutions de bonne qualité, par exemple lors de la consultation. Aussi est-il important d'encourager l'utilisation du JPEG2000, malgré les difficultés. Cette brève explication de la situation des formats dans le domaine de la photographie numérique mets en évidence que les questions de formats ne se résolvent pas de manière simple.

Les guides de l'ADS ne se limitent pas à ces types de données. Les domaines des photographies aériennes, des données géophysiques, des études marines, des scanning laser (à savoir des enregistrements précis d'objets tridimensionnels réels), ainsi que des données de systèmes d'information géographiques (GIS), de la conception assistée par ordinateur (CAD) et, enfin, des réalités virtuelles sont également traités, à chaque fois par un guide particulier.

Ce bref survol permet simplement de prendre conscience que les prescriptions des archivistes sont trop restrictives et que pour établir des bonnes pratiques en terme de formats, il est indispensable de bien connaître les méthodes de travail propre à chaque discipline. Il n'est en effet pas constructif de se limiter à n'accepter que les formats se conformant le mieux aux exigences de la conservation à long terme (formats standards et ouverts, si possible sous forme textuelle) et, ainsi, exclure une grande partie des travaux des chercheurs. D'autant que dans certains cas, il n'est actuellement pas possible de se passer de certains formats binaires, voire de formats propriétaires. Ce ne sont, bien entendu pas des formats adaptés à la conservation, mais s'il n'est d'autre possibilité, que faire ?

Pour ces raisons, il apparaît que la question des formats, tout comme celle des métadonnées, doivent être explicitement abordés par le protocole de versement qui est négocié entre l'administration de l'OAIS et le producteur, si possible lors de la planification du projet de recherche[43]. Or, le protocole de versement ne peut être en mesure de régler ces questions sans avoir au préalable procédé à une enquête sur les processus de création documentaire induits par un projet de recherche, et ce lors de la planification.

4.4. Responsabilités du système d'archives

Nous avons opté pour une approche relativement indirecte dans l'espoir d'atteindre

43 Voir le point 4.4.2, page 32.

cette section avec quelques arguments. Au vu de l'information nécessaire du point de vue du modèle OAIS pour qu'un objet-donnée puisse être conservé sur le long terme, rester intelligible, être accessible au moyen d'une recherche et grâce au réseau Internet, et être réutilisable par un nouveau projet de recherche ; au vu de la complexité des métadonnées à constituer, saisir et gérer, automatiquement, par les administrateurs du système d'archives et par les producteurs des données, et du coût induit par cette activité ; au vu, enfin, de l'immense diversité des formats informatiques concernés par l'archivage des données primaires produites par différentes disciplines scientifiques, nous pensons que le lecteur sera convaincu qu'il n'est pas possible d'affronter ces défis sans s'appuyer sur une vision claire et explicite et sans s'attaquer à ces questions au moment de la planification d'un projet de recherche, comme il a déjà été écrit plus haut.

4.4.1. Mission et objectifs

Le système d'archives doit définir précisément quels sont ses buts et objectifs. Il doit pouvoir s'appuyer non seulement sur des arguments éthiques ou de méthodologie scientifique, dont la valeur n'est pas mise en doute, mais également sur une volonté claire de l'institution (l'université) dans laquelle il s'inscrit[44]. Il doit rédiger, publier et communiquer, notamment aux différentes communautés de chercheurs concernés par le système d'archives, un document qui déclare qui est responsable du système d'archives, c'est-à-dire qui est en charge de sa gestion, dans quel(s) but(s) les données primaires sont conservées (intégrité de la recherche, réutilisation, archivage à long terme) et quels sont les publics cibles visés par le système d'archives. Ce document constitue sa mission. Son existence et sa publication sont considérées par *The Data Seal of Approval* (DSA)[45] comme un des seize critères pour estimer le degré de pérennité et d'ouverture d'un système d'archives (Ball, 2010 : 31).

Le système d'archives doit également pouvoir définir quelles sont les informations nécessaires pour assurer l'archivage et la conservation des données, ainsi que pour en permettre la compréhension et la réutilisation future, ce qui dépend entre autres des publics visés. De ces informations, quelle partie peut être constituées par le système d'archives lui-même ? À l'inverse, quelle partie de ces informations doit impérativement être produite par les chercheurs, puisque que pour bon nombre d'entre elles – par exemple les procédures d'acquisition de certaines données, la signification de

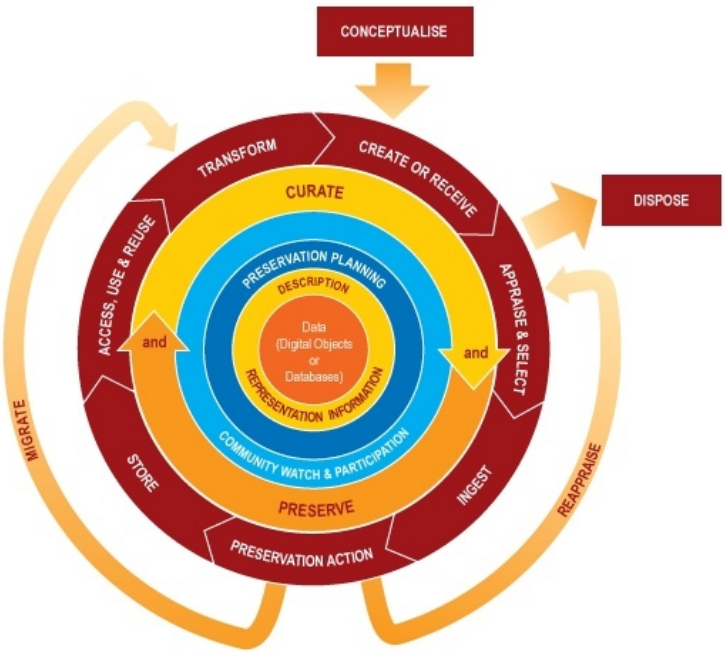
44 Voir le chapitre 1.
45 http://www.datasealofapproval.org/ (consulté le 27 juin 2012).

certaines suites de nombres –, ils sont les seules personnes en mesure de le faire ? Enfin, il s'agit d'évaluer dans quelle mesure les processus de travail des producteurs doivent être modifiés pour produire l'information nécessaire à la constitution d'un paquet d'informations à verser (SIP) conforme aux exigences du système d'archives. Il est essentiel de faire en sorte que les efforts à fournir dans ce but restent les plus limités possible et n'entravent pas inconsidérément le travail de la recherche (Ball, 2010 : 9).

4.4.2. Enquête

Pour répondre à ces questions, dont les réponses peuvent différer d'un projet de recherche à l'autre, les administrateurs du système d'archives doivent procéder à une enquête en collaboration avec les chercheurs, au moment de la planification de la recherche. Cette enquête a également pour objectif que les chercheurs disposent d'une image plus complète des données dont ils ont besoin, des données qu'ils vont produire, selon quels processus et l'influence que leur conservation à long terme exerce sur leurs propres méthodes de travail.

Figure 2: DDC Lifecycle Model



(Source : DDC Lifecycle Model)

Le *Data Asset Framework* (DAF)[46], qui est un outil développé pour évaluer les données de la recherche, ainsi que leur gestion, relève que cette enquête doit être planifiée

46 http://www.data-audit.eu/ (consulté le 27 juin 2012).

avec attention pour s'assurer la disponibilité des chercheurs concernés (Ball, 2010 : 30). L'enquête peut s'appuyer sur des modèles de cycle de vie des données, comme le *DDC Lifecycle Model* (figure 2)[47], qui permettent de vérifier si les différentes étapes nécessaires à la conservation sur le long terme des données sont prévues (Ball, 2010 : 9). Par exemple, l'utilisation d'un modèle de cycle de vie des données met rapidement en évidence qui est responsable de la constitution de telles métadonnées et à quel moment. Pour autant, les différents modèles de cycle de vie existants ne correspondent pas toujours à la réalité du terrain, car celui-ci peut être très différent d'une discipline à l'autre. Autrement dit, ces modèles ne permettent pas de se passer d'une analyse détaillée des méthodes de fonctionnement de chaque projet, tel qu'il est planifié (Ball, 2010 : 14).

Il est important d'établir, dès la planification, avec un certain degré de précision de quels types de données un projet de recherche a besoin. De ces données, quelle part est déjà existante et quelle part doit être produite. Sont-elles toutes d'origine numérique, ou existe-t-il des données récoltées ou créées sous forme analogique ? Dans le cas où il est prévu de les numériser, quelles procédures cette numérisation doit-elle suivre ? Pour les données existantes, il s'agit de préciser si leur provenance peut être clairement déterminée, si elles peuvent être citées et quels sont les droits de propriété qui y sont attachés, lorsque c'est le cas. Pour les données que le projet doit produire pour se propres besoins, il est nécessaire d'évaluer leur volume, de décrire leur procédures de création, la manière de les documenter (en accord avec les standards de chaque discipline). Une attention doit être apportée également aux questions de protection des données, ce qui peut signifier un processus d'anonymisation ou des restrictions à la diffusion. C'est également à cette étape de la planification que la question des droits de propriété intellectuelle des données produites doit être résolue.

Bien entendu, ce travail de planification doit également faire l'inventaire des formats qui interviennent dans la création des données, mais aussi pour le traitement de celles-ci. Cet inventaire doit permettre de donner des conseils quant aux formats à privilégier, lorsqu'il y a le choix, ainsi que les méthodes permettant de les obtenir. Il est encore une fois important de trouver un équilibre entre les exigences de la conservation à long terme des données et les habitudes de travail des chercheurs.

47 http://www.dcc.ac.uk/resources/curation-lifecycle-model (consulté le 27 juin 2012) d'où la figure 2 est extraite. Voir aussi le point 1.2, à la page 6.

Les entretiens entre les gestionnaires du système d'archives et les chercheurs qui planifient leur projet doivent également déterminer s'il y a lieu d'imposer des règles minimales de nommage des fichiers pour l'ensemble du système d'archives, ou s'il est possible d'envisager une convention adaptée à chaque contexte (projet de recherche, équipe de travail). Quelle que soit la réponse à cette question, des règles minimales de nommage sont nécessaires pour la constitution du SIP. Aussi, plus tôt sont-elles mises en œuvre dans l'histoire des données, plus simple sera la création des paquets à verser. Il n'est pas inutile de conseiller d'éviter les espaces (que l'on remplace par des *underscore*) et de s'en tenir aux caractères alphanumériques simples (sans accent, ni cédille). Le nom lui-même devrait contenir un élément commun à tout le projet de recherche[48], et être significatif et compréhensible le plus largement possible.

Nous avons proposé plus haut que le système d'archives publie sa mission et nous voulons conseiller encore de publier des recommandations et des directives claires, par exemple sur son site web, et les diffuser de manière active auprès des chercheurs de son institution. Ces directives doivent reprendre succinctement les questions qui sont abordées par l'enquête dont il est question ci-dessus et surtout encourager les chercheurs à prendre contact avec les gestionnaires du système d'archives dès la planification du projet, principalement pour atténuer tant que possible les coûts induits par les exigences de conservation, tout en permettant de tendre vers elles.

4.5. Évaluation et sélection[49]

En s'appuyant à la fois sur la mission et sur l'enquête, des critères d'évaluation, au sens archivistique du terme[50], doivent être définis par l'administration du système d'archives, avec l'active collaboration de la communauté scientifique, voire, lorsque cela est possible, avec le public visé par le système d'archives. Ces critères, les plus objectifs possibles, forment une politique d'évaluation qui doit être publique et explicite, afin que ce processus soit transparent et vérifiable. En effet, l'appréciation du degré d'importance d'un lot de données du point de vue de leur conservation et de leur archivage ne doit pas être laissé à la subjectivité d'un individu particulier. De plus, le public qui consulte les données primaires archivées doit pouvoir prendre connaissance des raisons qui ont participé à la décision de conservation et d'archivage de ces

48 Il serait d'ailleurs intéressant de savoir si, dans une certaine mesure, dans le contexte du projet GIPDIR, ces éléments identifiants ne pourraient pas être gérés par le système informatique lui-même.

49 Ce point s'inspire principalement de Whyte, Wilson, 2010.

50 Voir le point 1.5, à la page 9.

données.

Ces critères sont définis en partie lors de la conception du système d'archives lui-même, mais également lors de la planification du projet de recherche, ceci parce qu'ils doivent à la fois se conformer aux éventuelles exigences légales et réglementaires, à la mission et aux objectifs du système d'archives, ainsi qu'aux particularités de chaque projet de recherche et des standards des différentes disciplines scientifiques.

Un critère particulièrement difficile à établir est celui lié à la notion archivistique de valeur secondaire. Dans le cas des données primaires de la recherche, cette valeur secondaire recouvre la valeur scientifique, historique et sociale de ces données. La valeur sociale est constituée par les préoccupations d'une société, que certains voudraient mesurer quantitativement par le nombre de financements de projets de recherche dans un domaine particulier. Or, cette mesure ne nous dit que très peu de choses sur les interrogations qui seront celles de la société de demain. De même, concernant la valeur historique, nous ne sommes pas en mesure, aujourd'hui, de déterminer quelles sont les données qui seront utilisées comme sources pour l'histoire et l'histoire des sciences dans le futur. Par contre, la valeur scientifique peut reposer sur des critères plus objectifs et stables dans le temps.

L'évaluation peut néanmoins recourir à des critères plus solides. Par exemple, est-ce que tel lot de données est unique, représente la seule source pour démontrer telle hypothèse ? Si d'autres lots similaires existent, il s'agit de déterminer lequel de ces lots est le plus complet ou représentatif, le mieux documenté ou le moins sujet au risque d'altération ou de destruction. Est-ce que ces données ont été produites au cours d'un événement qui n'est pas reproductible (une éruption volcanique), auquel cas leur conservation est indispensable ?

Si les données primaires ne sont pas suffisamment documentées, si la qualité des métadonnées n'est pas satisfaisante, il est nécessaire de juger si cette documentation et ces métadonnées peuvent être améliorées. Si ce n'est pas le cas, conserver des données qui seront très difficilement consultables ou réutilisables n'a que peu de sens. De même, le format dans lequel ces données sont enregistrées est un critère de sélection, car les formats ouverts, indépendants des logiciels, des systèmes d'exploitation et du *hardware*, offrent de meilleures garanties de pérennité.

Toujours du point de vue de la réutilisation possible des données, il est important que leur intégrité et leur authenticité puissent être assurées. Si ce n'est pas le cas, il n'est

pas souhaitable de les archiver à long terme. De même, lorsque des restrictions importantes limitent leur diffusion future, comme des droits de propriété intellectuelle, liées à l'éthique ou à la protection de la sphère privée, la conservation peut être remise en question. Il s'agit par exemple de déterminer si une anonymisation peut être réalisée ou non, ou si les droits de la propriété intellectuelle n'interdisent pas à un public, même réduit, de consulter ou de réutiliser ces données.

Enfin, une évaluation du coût d'archivage sur le long terme doit être faite. Ce coût peut être comparé à celui de la création des données. Il est important de déterminer si ce coût peut-être supporté par le système d'archives sur le long terme et donc de vérifier dans quelle mesure le financement du système d'archives est garanti dans le temps.

L'évaluation, bien qu'elle soit chronologiquement située dans le *DDC Lifecycle Model* (figure 2) entre la création des données et leur versement dans le système d'archives, puis à nouveau effectuée par le système d'archives, est à notre sens un processus qui ne se réduit pas à deux moments ponctuels. En effet, non seulement les critères doivent être en partie définis par le travail commun des administrateurs du système d'archives et les producteurs des données, mais une évaluation *a priori* peut aussi être réalisée au moment de la planification du projet de recherche, dans le même temps que l'enquête, dont il est question au point 4.4.2[51], lorsqu'il s'agit d'estimer le volume et la nature des données qui seront produites par un projet de recherche.

Cette évaluation relève de la responsabilité des administrateurs du système qui en fixent les règles, non sans tenir compte de l'avis et de l'expertise des chercheurs. Les principes généraux du processus d'évaluation doivent figurer dans les directives que publie le système d'archives. Mais des critères et des règles plus précis et adaptés à chaque projet de recherche doivent donc être abordés lors de l'enquête préalable et être explicités dans le protocole de versement qui est négocié au moment de la planification du projet de recherche.

4.6. Exigences minimales

Dans l'espoir de parvenir à résumer et à clarifier ce qui a été vu jusqu'ici, nous présentons des recommandations pour un niveau minimal de métadonnées, que le système d'archives devrait demander aux chercheurs.

Tout d'abord, il est important d'insister sur le fait qu'il n'est pas raisonnable de vouloir

51 Page 32.

que les données soient décrites au niveau du fichier ou du document. Dans la plupart des cas, une description d'un ensemble de fichiers, c'est-à-dire en termes archivistiques d'un dossier, voire d'une série, est parfaitement satisfaisant. Néanmoins, la granularité de la description doit être cohérente avec celle de l'accès offert aux données. Le consortium *DataCite*[52] préconise une description minimal pour chaque enregistrement d'un *DOI*[53].

Par description, nous entendons les métadonnées et la documentation dont sont responsables les chercheurs. Elle doit renseigner sur la provenance des données (qui les a produites, collectés), selon quelles procédures elles ont été produites, quels traitements leur ont été appliqués, quelle est leur signification, comment doit-on les lire, et quelles limites à la diffusion sont à prendre en compte (protection des données et droits de propriété intellectuelle). Il est possible de s'appuyer sur les champs minimaux indiqués par la norme ISAD(G)[54], ainsi que les métadonnées obligatoir du schéma *Dublin Core* proposées par le consortium *DataCite*, à savoir l' « *Identifier* » lui-même, le « *Title* », le « *Creator* », la « *PublicationYear* » et le « *Publisher* » (DATACITE, 2011 : 5-6).

Dans le contexte de ce travail, c'est-à-dire en tenant compte du projet GIPDIR, il est envisagé qu'un nombre important de métadonnées peuvent être récoltées de manière automatique, notamment toutes les données techniques (format, spécification du format[55], version de logiciel, taille, date de création).

À partir des métadonnées constituées par les chercheurs et de manière automatique, ainsi que des directives en matière de formats qui doivent être mises en œuvre dès la planification du projet, il doit être réalisable d'assister la constitution de paquets d'informations à verser (SIP) conformes au protocole de versement sur l'exemple de l'outil développé par les Archives fédérales suisses[56].

52 *DataCite* http://www.datacite.org (consulté le 9 juillet 2012).
53 Voir la note 34, page 24.
54 Voir le point 4.2, page 25 et suivante.
55 Au moyen par exemple de bases de données comme PRONOM : https://www.nationalarchives.gov.uk/PRONOM/Default.aspx (consulté le 27 juin 2012).
56 http://www.bar.admin.ch/dienstleistungen/00823/01559/index.html?lang=fr (consulté le 27 juin 2012).

Conclusion

Au terme de ce travail, nous pensons que le chantier en vue de l'élaboration de véritables recommandations a démarré, mais qu'il y a encore beaucoup d'inconnues à résoudre. Néanmoins, nous espérons que les arguments donnés au chapitre 2 permettront d'appuyer les démarches qui iront dans le sens de l'archivage à long terme des données primaires numériques. Nous pensons d'ailleurs qu'il est souhaitable que la législation, notamment au niveau cantonal, prévoie cet archivage, afin d'offrir aux systèmes d'archives une base claire et solide.

Du point de vue des recommandations, nous avons indiqué l'importance pour un système d'archives de définir et de publier sa mission et ses objectifs, afin d'améliorer sensiblement la compréhension, l'adhésion et l'implication des différents acteurs, principalement des chercheurs. En effet, ces derniers sont responsables d'une part importante, en ce qui concerne la description des données, des métadonnées nécessaires.

L'importance de celles-ci a été mise en évidence par la description du modèle de paquet d'informations proposé par l'OAIS. À cela s'ajoute l'épineuse question des formats informatiques. Ces deux problématiques, métadonnées et formats, ne peuvent être résolues sans avoir, au préalable, procédé à une enquête en collaboration avec les chercheurs, afin d'évaluer, en termes de procédures de création documentaire, de formats et de volumes, les données éventuellement existantes et celles qui seront produites par le projet de recherche.

Nous avons voulu montrer que de situer chronologiquement cette enquête, dès la planification du projet de recherche et en collaboration avec les principaux intéressés, concourt à mettre en œuvre des directives et des procédures détaillées en vue de se conformer aux exigences du système d'archives, tout en respectant les besoins des chercheurs. C'est également dès ce moment du cycle de vie des données que doit intervenir la définition des critères et des procédures d'évaluation et de sélection des données primaires.

Bibliographie

ARCHAEOLOGY DATA SERVICE. 2011. Guides to Good Practice. In : *Site de The Archaeology Data Service*. http://guides.archaeologydataservice.ac.uk/g2gp/Main (consulté le 27 juin 2012).

ARCHAEOLOGY DATA SERVICE. 2012. Guidelines for Depositors. In. *Site de The Archaeology Data Service*. http://archaeologydataservice.ac.uk/advice/guidelinesForDepositors (consulté le 27 juin 2012)

ARCHIVES FÉDÉRALES SUISSES. 2007. Normes et standards pour l'archivage de documents numériques. In : *Site des Archives fédérales suisses*. Mis en ligne le 24 avril 2008. http://www.bar.admin.ch/themen/00876/00877/index.html?lang=fr&download=NHzLpZeg7t,lnp6I0NTU042l2Z6ln1ae2IZn4Z2qZpnO2Yuq2Z6gpJCDdIJ9hGym162epYbg2c_JjKbNoKSn6A-- (consulté le 27 juin 2012)

AROVELIUS, Renata, et al. 2010. *Management and Preservation of Scientific Records and Data*. International Council on Archives : Paris.

BALL, Alex. 2010. *Review of the State of the Art of the Digital Curation Research Data*. University of Bath : Bath.

BALLEGOOIE, Marlene (van), DUFF, Wendy. 2006. *DCC Digital Curation Manual Instalment on Archival Metadata*. HATII, University of Glasgow : Glasgow.

CAPLAN, Priscilla. 2006. *DCC Digital Curation Manual Instalment on Preservation Metadata*. HATII, University of Glasgow : Galsgow.

CONSEIL INTERNATIONAL DES ARCHIVES. 2000. *ISAD(G): Norme Générale Et Internationale De Description Archivistique: Adoptée Par La Commission Sur Les Normes De Description, Stockholm, Suède, 19-22 Septembre 1999*. CIA : Ottawa.

CONSEIL INTERNATIONAL DES ARCHIVES. 2004. *ISAAR (CPF): Norme Internationale Sur Les Notices D'autorité Utilisées Pour Les Archives Relatives Aux Collectivités, Aux Personnes Ou Aux Familles*. CIA : Paris.

DATACITE. 2011. DataCite Metadata Schema for the Publication and Citation of Research Data. In : *Site du DataCite Metadata Schema Repository* [en ligne]. http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

(consulté le 10 septembre 2012).

DAY, Michael. 2005. *DCC Digital Curation Manual Instalment on Metadata*. HATII, University of Glasgow : Glasgow.

FONDS NATIONAL SUISSE DE LA RECHERCHE SCIENTIFIQUE (FNS). 2008. *Règlement des subsides : règlements du Fonds national suisse de la recherche scientifique relatifs aux octrois de subsides*. Fonds national suisse : Berne.

DIGITAL CURATION CENTRE. What is digital curation ? In : *Site du Digital Curation Centre* [en ligne]. http://www.dcc.ac.uk/digital-curation/what-digital-curation (consulté le 4 juillet 2012).

FONDS NATIONAL SUISSE DE LA RECHERCHE SCIENTIFIQUE (FNS). 2009. *Règlement du Conseil de la recherche sur la gestion du comportement incorrect des requérant-e-s et des bénéficiaires de subsides dans le contexte scientifique*. Fonds national suisse : Berne.

KELLER-MARXER, Peter, BUETIKOFER, Niklaus. 2008*. Élaboration d'un modèle pour un archivage à long terme centralisé des données scientifiques primaires et secondaires en Suisse : situation, besoins, desiderata, modèles, protagonistes, conditions cadre et contexte européen*. Berne, iKeep.

KOMOROWSKI, Matthew. [Sans date]. A History of Storage Cost. In : *Site Mkomo.com* [en ligne]. http://www.mkomo.com/cost-per-gigabyte (consulté le 6 juillet 2012)

ORGANISATION INTERNATIONALE DE NORMALISATION, 2003. *Systèmes de transfert des informations et données spatiales - système ouvert d'archivage d'information - modèle de référence = Space data and information transfer systems - open archival information system - reference model*. Genève: ISO. ISO international standard, 14721.

RÉPUBLIQUE ET CANTON DE GENÈVE, ARCHIVES D'ÉTAT. 2011. Formats de fichiers adaptés à l'archivage électronique à moyen et long terme. In : *Site des Archives d'État de la République et Canton de Genève* [en ligne]. http://etat.geneve.ch/dt/SilverpeasWebFileServer/20111019_formats_archivage_etatge.pdf?ComponentId=kmelia66&SourceFile=1319206910719.pdf&MimeType=application/pdf&Directory=Attachment/Images/ (consulté le 27 juin 2012).

SALATHÉ, Michelle. 2008. *L'intégrité dans la recherche scientifique : principes de bases et procédures*. Berne : Académies suisses des sciences.

UNIVERSITÉ DE GENÈVE. 2012. Directive relative à l'intégrité dans le domaine de la recherche scientifique et à la procédure à suivre en cas de manquement à l'intégrité (Chercheurs-euses). In : *Site de l'Université de Genève*. Mis en ligne le 12 avril 2012. https://memento.unige.ch/doc/0003 (consulté le 18 juin 2012).

WHYTE, Angus, WILSON, Andrew. 2010. How to Appraise and Select Research Data for Curation : a Digital Curation Centre and Australian National Data Service 'working level' guide. In : *Site du Digital Curation Centre* [en ligne]. http://www.dcc.ac.uk/resources/how-guides/appraise-select-data (consulté le 5 juillet 2012).