Livre blanc

Approches contemporaines en hébergement et gestion de données

HÉLÈNE TANGHE, SANAA AIT DAOUD, PAUL-OLIVIER GIBERT (DIGITAL & ETHICS)
SOROR SAHRI, SALIMA BENBERNOU (UNIVERSITÉ DE PARIS 5)
JÉRÔME TOTEL, LOÏC BERTIN (DATA4)
CRISTINEL DIACONU (CENTRE DE PHYSIQUE DES PARTICULES DE MARSEILLE, CNRS)
ITEBEDDINE GHORBEL (INSERM)
DANIELLE GELDWERTH-FENIGER (CNRS, UNIVERSITÉ DE PARIS 13)
LEILA ABIDI (USPC, IMAGERIES DU VIVANT)
CHRISTOPHE CÉRIN (UNIVERSITÉ DE PARIS 13).
PHILIPPE WERLE (UNIVERSITÉ DE PARIS 13)
MARIE LAFAILLE (USPC)

URL : http://lipn.univ-paris13.fr/~cerin/Livre_blanc_data_hosting.pdf

CONTACT : christophe.cerin@lipn.univ-paris13.fr

Relecture 2017, version au 18 décembre 2017

Préface

LIVRE BLANC, APPROCHES CONTEMPORAINES en hébergement et gestion de données, une parution de plus sur l'hébergement des données, ses opportunités et ses risques ? Et bien, non ! C'est avant tout une approche pédagogique du sujet qui a été retenue, afin que chacun, et tout particulièrement s'il n'est pas informaticien, puisse s'approprier ce vaste domaine qu'est celui de la donnée et surtout de son hébergement.

La donnée est là, partout présente et ce n'est pas un hasard si de plus en plus d'entreprises, d'institutions, et d'individus s'y intéressent. Encore faut-il savoir où la trouver, comment l'exploiter et où la stocker, notamment quand il s'agit de sujets sensibles tels que la santé ou la sécurité, sachant que, dès qu'il y a un problème de confidentialité supposé ou avéré, les médias s'en emparent et en font leurs choux gras, alors que l'immense champ des possibles que sont le cloud computing recèle n'est que très rarement traité.

Ce document est d'abord une excellente synthèse de l'état de l'art, conçue en mode collaboratif, dont le résultat est tout à fait probant. Il illustre la richesse du domaine et met en lumière toutes les facettes du sujet.

Document de vulgarisation, il s'adresse à tous ceux qui veulent en savoir plus sur le cloud computing et l'hébergement des données et traite sans complaisance des grandes questions qui se posent en matière d'hébergement et de gestion des données, et qui peuvent laisser place à tous les fantasmes tout particulièrement autour des données sensibles.

La sécurité et l'analyse des risques sont traitées avec une grande rigueur, tant ces questions peuvent encore freiner l'expansion du cloud computing, notamment dans le monde de la recherche, où les enjeux économiques peuvent être cruciaux sur le long terme. Aucune solution n'est parfaite, mais il faut très certainement prendre en compte les bénéfices d'un stockage en ligne sécurisé et ne pas avoir la mémoire courte : qui ne s'est jamais trouvé confronté à des difficultés d'échanges de données avant le développement des solutions de partage d'information en ligne ?

A l'heure où la technologie évolue toujours plus vite, où l'information circule très rapidement, les organisations, et en particulier les universités, doivent s'organiser et faire appel à des compétences multiples non seulement au niveau technique mais également dans tous les domaines fonctionnels car l'usage et le métier vont dicter les solutions à retenir en matière d'hébergement, de stockage et de partage de l'information, et l'avenir nous réserve probablement encore de nombreuses surprises !

Quel que soit le domaine d'activité, mais c'est encore probablement plus vrai dans celui de la recherche, c'est de la collaboration entre les équipes pluridisciplinaires que naissent les meilleures idées. La France est réputée pour la qualité de ses formations et nos chercheurs ne sont pas en reste : partageons nos savoir-faire pour faire émerger des solutions sur le plan international, et ne gardons que le meilleur des fantastiques opportunités que représente le cloud computing !

Dernier point mais par des moindres tant il est un marqueur important pour les rédacteurs de cet ouvrage : ce document se veut avant tout une base de travail, qui pourra être amendée, complétée, enrichie et actualisée par toutes les bonnes volontés qui voudront bien se l'approprier.

Bonne lecture à tous.

Françoise Farag
Présidente de Salvia Développement
Présidente du conseil d'institut de l'IUT de Villetaneuse

Préambule

LE PRÉSENT DOCUMENT a pour objectif de rentrer en douceur dans les problématiques liées à l'hébergement des données en particulier lorsque les technologies du cloud computing sont mises en oeuvre. Il a été rédigé de la manière la plus claire et pédagogique possible afin qu'il soit compréhensible par un public de néophytes, en particulier les non-informaticiens. Nous l'avons rendu auto-suffisant en ce qui concerne le vocabulaire en cherchant à expliciter le plus simplement possible les termes techniques. Il s'agit donc d'un point d'entrée dans le domaine de l'hébergement des données avec bien entendu ses a priori et ses choix rédactionnels.

Pour cette édition, il est destiné en premier lieu aux publics de l'enseignement supérieur et de la recherche pour qu'ils puissent se saisir des problématiques et des enjeux. Nous supposons en effet qu'une université ou un individu au sein de cette université a besoin d'éléments concrets pour définir une politique d'hébergement de données dans son cloud privé, ou s'imaginer les conséquences d'une telle politique dans son quotidien de chercheur. Nous lui proposons un panorama complet, un recueil d'informations objectives et factuelles qui sont proposées, discutées, commentées dans les écosystèmes sous-jacents, public ou privé. Une incise toute particulière est faite sur les sujets éthiques et juridiques.

Le document peut cependant être instructif pour toute personne n'appartenant pas au monde de l'enseignement supérieur et de la recherche mais désireuse de se faire une idée sur les problèmes et les solutions en matière d'hébergement et de gestion de données massives. Nous espérons aussi que le lecteur trouvera une approche qui constitue un tout, dont des éléments méthodologiques concernant les sujets de la migration et de l'adoption, dans notre quotidien, des technologies de cloud computing et des centres de données.

Un élément discriminant, parmi d'autres, pour décider si un individu ou une organisation a intérêt à basculer sur des technologies de type cloud, c'est-à-dire à externaliser les données, est de se demander si les données confiées à un tiers peuvent être valorisées en toute sécurité. C'est un peu la même question qu'avec votre argent. Si vous estimez que votre banque valorise bien votre argent, il n'y a aucune raison d'en changer. Si vous estimez que votre banque ne fait pas suffisamment pour investir, avec votre argent, dans l'économie réelle et les projets innovants, vous allez peut être en changer.

Le cloud est un tiers comme l'est une banque. De même, il suppose un choix de relation. Le problème est donc de savoir qui contrôle les données et avec quel type de contractualisation. Ce rapport qui, nous l'espérons, sera suivi par d'autres éditions, est organisé en quatre grandes parties :

- La partie 1 donne les éléments de vocabulaire, introduit l'hébergement des données et en particulier les données de recherche ;
- La partie 2 dresse un état de l'art des grandes questions qui se posent en matière d'hébergement et de gestion des données ; Une discussion particulière est faite sur les données de santé ;
- La partie 3 est une analyse situationnelle de solutions mises en œuvre pour l'hébergement des données. Nous donnons des exemples précis de projets en œuvre dans le monde académique principalement. Un focus important est fait sur les aspects de sécurité et d'analyse des risques ;
- La partie 4 élabore des éléments de politique, propose des lignes directrices et des recommandations à destination des individus et des décideurs.

Nous voudrions également remercier les personnes suivantes pour leurs relectures approfondies et leurs commentaires : Jean-Philippe Gouigoux (directeur technique de la société MGDIS et Microsoft Most Valuable Professional sur la spécialité Azure), Daniel Balvay (PARCC-HEGP et Université René Descartes), Laura Werle (Master 2 IPAG).

Le groupe PREDON (`https://predon.org`), qui s'occupe de préservation de données scientifiques, soutient activement ce projet de livre blanc sur l'hébergement des données. Plusieurs rédacteurs du livre blanc sont par ailleurs investis dans l'action PREDON qui sera présentée ultérieurement. Au sein du projet interdisciplinaire PREDON la préservation est discutée avec comme but l'échange de méthodes, pratiques et technologies utiles aux projets scientifiques qui nécessitent de la collecte et de l'analyse de données digitales. Le projet regroupe un large éventail de disciplines (physique des particules, astrophysique, écologie, informatique, sciences du vivant etc.) et a des contacts dans de grands centres de calcul nationaux comme le CC-IN2P3 (Lyon), le CINES (Montpellier) et le CDS (Strasbourg). PREDON est aussi une action au sein du GdR Madics (`http://madics.fr`).

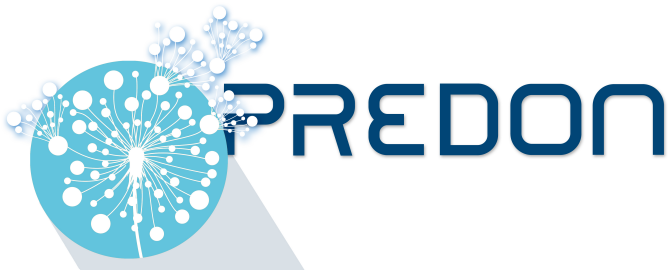


Table des matières

1. Les données de recherche et leur gestion

1.1 Qu'est ce qu'une donnée ?

1.1.1 Définition générale

Dans son acceptation générale, une donnée est définie de la manière suivante :

Definition 1 Une donnée est un ensemble de valeurs faisant référence à la représentation et au codage d'une information ou un savoir sous une forme adaptée à un usage. Une donnée n'est pas une information. Une donnée requiert une interprétation pour devenir une information.

Dans ce document nous nous concentrons sur les *données de recherche*, appelées aussi données scientifiques. Selon l'Organisation de Coopération et de Développement Économiques (OCDE), les données scientifiques sont « des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche ». Dans le rapport Gestion et partage des données [1], une donnée scientifique est définie comme « le résultat d'une expérience, généralement issue d'un instrument (robot, capteur, enregistrement audio ou vidéo, etc.), ou bien d'une observation humaine sur le monde naturel. Observées essentiellement dans les domaines où l'acquisition de l'information se fait sur le terrain (réalisation d'observations, de mesures, de comptages, etc.), elles contribuent aux avancées de la recherche en matière de connaissance et d'inventaires. »

Il existe différents types de données de la recherche qui diffèrent selon la manière dont les données sont produites et selon leur valeur supposée. Les données de recherche peuvent être des données brutes, des données traitées, des donnés dérivées, des données d'observation, des données expérimentales ou encore des données computationnelles ou de simulation [2, 3]. Souvent, un des

1. Rapport du groupe de travail sur la gestion et la partage des données, INRA, 2012
2. De l'open data à l'open research data : quelle(s) politiques(s) pour les données de la recherche, R. Gaillard, 2014
3. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century, NSF, 2005

points communs à tous ces types de données est leur volumétrie importante, la volumétrie étant une dimension clé du big data comme nous le verrons plus loin.

Ainsi une donnée doit être considérée selon une dimension ou encore un attribut qui la caractérise de manière plus précise. La donnée peut être massive ou pas, ouverte ou pas, personnelle ou pas. Souvent ces attributs renvoient également à des questions juridiques précises.

Par exemple le degré d'ouverture des données (publiques ou communautaires) pose des problèmes de risques potentiels de fuites d'information. La notion d'attractivité de la donnée confidentielle pour un tiers renvoie à la valeur de vente de l'information confidentielle et à un potentiel de nuisance pour la personne dont l'information a été divulguée (personne publique ou non / information sur l'intimité contenue dans la donnée). Il y a donc des risques d'utilisation politique ou économique non désirés par les individus.

1.1.2 Les données ouvertes

Definition 2 Les *données ouvertes (open data)* sont des données numériques d'origine publique ou privée qu'un organisme (collectivité, service public, entreprise) diffuse de manière structurée selon une méthode et une licence ouverte garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière. Les données de départ, avant leur diffusion, sont trop souvent fournies dans des formats inexploitables et non structurés.

Selon l'Open Government Data Group [4], les données ouvertes doivent satisfaire à huit principes. Elles doivent être :

- complètes (chaque jeu de données doit comporter toutes les données disponibles à l'exception des données sujettes à des limitations concernant la vie privée, la sécurité ou des privilèges d'accès) ;
- primaires (les données ouvertes sont des données brutes, prises directement à la source, aussi détaillées que possible et sans traitement ni modification) ; Les données ouvertes doivent contenir les données primaires mais elles ne s'y résument pas ;
- opportunes (les données doivent être rendues disponibles aussi vite que possible pour être à jour) ;
- accessibles (les données doivent être disponibles pour le plus grand nombre) ;
- exploitables (prêtes à être traitées par des outils informatiques) ;
- non discriminatoires (accessibles sans inscription) ;
- non propriétaires (disponibles dans des formats ouverts) et
- libres de droits.

La notion de données ouvertes s'inscrit dans le mouvement de *Science ouverte (open science)* qui considère la science comme un bien commun dont la diffusion est d'intérêt public et général. L'ouverture des données de recherche répond ainsi à cinq enjeux [5] :

- accélérer les découvertes scientifiques, les innovations et le retour sur investissement en recherche et développement ;
- encourager la collaboration scientifique et les possibilités de recherche interdisciplinaire ;
- éviter la duplication des expériences, favoriser la réutilisation des données et minimiser le risque de perte des données ;
- assurer l'intégrité et la reproductibilité de la recherche et

4. The 8 Principles of Open Government Data, 2007
5. Rendre public ses jeux de données, Cirad, 2015

- accéder librement à une masse de données ouvrant de nouveaux champs d'analyse non envisagés par le producteur des données.

Les données de la recherche sont liées à des contraintes juridiques qui font obstacle à leur ouverture libre. D'ailleurs, les huit principes cités ci-dessus reflètent deux groupes d'aspects accompagnant le processus d'ouverture de données dont les aspects juridiques. Le deuxième groupe concerne les aspects techniques.

R Vous pouvez retrouver de plus amples détails sur la question de données ouvertes dans deux documents : "Livre blanc sur les données ouvertes" rédigé dans le cadre du projet Infonomics Ressource Facility (IRF) [6] et dans le livre blanc récemment publié par le CNRS « Une science ouverte dans une république numérique [7] ».

1.1.3 Les métadonnées

Pas de données sans métadonnées ! En effet, les données n'ont de valeur que dans un contexte bien précis et l'ajout d'informations qui permettent de les identifier et de les localiser est essentielle.

Definition 3 Les métadonnées sont les données qui décrivent d'autres données (sémantiquement les données à propos des données).

Elles permettent à un individu ou un ordinateur de comprendre le sens et l'organisation des données et facilite leur intégration, leur collecte et leur partage. Dans le domaine médical, les métadonnées vont, par exemple, indiquer un certain nombre d'informations concernant une image (taille, dimensions, paramétrage des équipements utilisés pour capturer l'image, détails expérimentaux du sujet comme son âge, son poids, son sexe voire son nom). Pour un document texte, les métadonnées peuvent spécifier la longueur du document, le nom de l'auteur, la date de rédaction et un résumé qui peut inclure des avis ou des résultats cliniques. Les métadonnées sont à la base des techniques du web sémantique.

R Le terme de web sémantique [8] désigne une évolution du web qui permettrait aux données disponibles (contenus, liens) d'être plus facilement utilisables et interprétables automatiquement, par des agents logiciels. Dans le cadre de la recherche scientifique, le web sémantique permettrait par exemple de rendre plus efficace la recherche bibliographique. Vous pouvez vous reporter à l'article suivant qui reprend les problématiques soulevées lors de la formation sur les données de recherche organisée par la TGIR Huma-Num et qui traitent du web sémantique, des métadonnées et d'interoperabilité au sein des SHS [9] (Sciences Humaines et Sociales).

Les principaux éléments de vocabulaire de base ayant été posés, nous allons maintenant évoquer quelques aspects juridiques concernant les données puis nous présenterons le big data qui est très certainement l'univers actuel où l'on a affaire le plus à la gestion des données. Nous reviendrons ensuite sur le problème de l'hébergement des données pour introduire des problématiques d'infrastructure matérielle et logicielle. Le cloud est introduit comme une des infrastructures permettant de gérer les big data. Le chapitre se termine par une présentation des nouveaux métiers, des nouveaux acteurs et des rôles respectifs de chacun dans l'hébergement et la gestion des données à l'heure du big data.

6. Livre sur les données ouvertes, B. Meszaros et al., 2015.
7. Une science ouverte dans une république numérique, CNRS, 2016
8. The Semantic Web, Scientific American Magazine, 2001
9. Gérer les données de la recherche, de la création à l'interopérabilité, J. Demange, 2015

1.2 Quelques aspects juridiques concernant les données de recherche

Les données de recherche soulèvent de nombreuses questions juridiques. Le droit associé aux données de recherche dépend de la nature de ces données. Il y a des données couvertes par le droit d'auteur, d'autres par le droit *sui generis* c'est-à-dire un droit spécifique au producteur de base de données. Il y a aussi beaucoup de données qui sont couvertes par des réglementations. Ces réglementations peuvent être issues de directives ou de lois diverses sur des données brutes, comme par exemple la directive INSPIRE concernant des données géographiques [10]. Le but de ces directives est principalement d'assurer l'interopérabilité et l'harmonisation des données pour qu'elles soient utilisables partout dans le monde. Les réglementations peuvent aussi concerner des données à caractère personnel (les données SHS ou les données biomédicales), ou des données à caractère secret ou sensibles (les données de santé par exemple). Ces mêmes données peuvent aussi être appréhendées comme des informations publiques.

Lionel Maurel, juriste et bibliothécaire, souligne l'importance d'établir « un diagnostic juridique précis pour déterminer le statut de chaque couche composant les résultats d'un projet de recherche : logiciels, inventions, données de recherche, contenus numérisés, métadonnées associées, valorisation éditoriale par le biais d'articles, d'ouvrages, de sites Internet etc ». Chaque couche peut avoir un statut différent, ce qui conditionne aussi les licences à choisir pour encadrer leur mise à disposition et leur réutilisation [11].

Dans ce qui suit, nous détaillons tous ces points liés au droit associé aux données de recherche.

1.2.1 Propriété intellectuelle

Definition 4 La propriété intellectuelle peut être définie comme l'ensemble des droits exclusifs octroyés à l'auteur d'une œuvre intellectuelle (Code de la propriété intellectuelle).

Elle est constituée de deux branches :

- la propriété littéraire et artistique, dont droit d'auteur et droit des bases de données et
- la propriété industrielle, qui concerne les brevets, marques, dessins et modèles, noms de domaine.

Les bases de données sont définies par le Code de la propriété intellectuelle.

Definition 5 Les bases de données sont un recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement accessibles par des moyens électroniques ou par tout autre moyen (art. L.112-3 du code de la propriété intellectuelle).

La directive du 11 mars 1996 [12], transposée en France par la loi du 1er juillet 1998 [13] concernant la protection des bases de données, définit le régime de protection de ces bases de données. Les bases de données bénéficient d'une protection sous deux fondements possibles :

- au titre du droit d'auteur, sur la structure de la base de données (agencement, disposition), en tant qu'œuvre de l'esprit originale et/ou

10. La directive Inspire pour les néophytes, F. Merrien et M. Leobet, 2011
11. Le statut juridique des données de la recherche, L. Maurel, 2015
12. Directive 96/9/CE du 11 mars 1996 sur la protection des bases de données
13. Loi 98-536 du 1er juillet 1998

- au titre du droit *sui generis* du producteur de la base de données, dès lors que le producteur justifie d'un investissement financier, matériel et humain dans la base de données.

Exception aux droits d'auteur et du producteur de base de données dans le cadre de la recherche

La loi pour une République numérique [14] entrée en vigueur le 9 octobre 2016, introduit une modification au Code de la propriété intellectuelle prévoyant une exception aux droits d'auteur et du producteur de base de données pour la fouille et l'exploration de textes et de données dans le cadre de la recherche.

> Ainsi lorsqu'une œuvre a été divulguée, son auteur ne peut interdire les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données incluses ou associées aux écrits scientifiques pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale. Un décret fixe les conditions dans lesquelles l'exploration des textes et des données est mise en œuvre, ainsi que les modalités de conservation et de communication des fichiers produits au terme des activités de recherche pour lesquelles elles ont été produites ; ces fichiers constituent des données de la recherche (article L 122-5 du Code de la Propriété Intellectuelle).

De même, lorsqu'une base de données est mise à la disposition du public par le titulaire des droits

> celui-ci ne peut interdire les copies ou reproductions numériques de la base réalisées par une personne qui y a licitement accès, en vue de fouilles de textes et de données incluses ou associées aux écrits scientifiques dans un cadre de recherche, à l'exclusion de toute finalité commerciale. La conservation et la communication des copies techniques issues des traitements, au terme des activités de recherche pour lesquelles elles ont été produites, sont assurées par des organismes désignés par décret. Les autres copies ou reproductions sont détruites (article L 342-3 du Code de la Propriété Intellectuelle).

Les critères suivants doivent donc être remplis pour bénéficier de cette exception :

- l'accès à une source licite
- en vue de l'exploration / de la fouille de textes et de données incluses ou associées aux écrits scientifiques
- pour les besoins de la recherche publique/dans un cadre de recherche, avec un but non commercial.

Le texte ne donne pas de définition de la notion même d'exploration ou de fouille de textes et de données. Cependant, il est généralement considéré que la fouille de textes et de données ou le Text Data Mining (TDM) consiste à explorer, via des outils et techniques de fouille, des corpuses de textes et/ou de données multisources et multisupports - afin d'en déduire de nouvelles connaissances.

Des décrets d'application sont attendus pour janvier 2017 afin de préciser les conditions de mise en œuvre de ces exceptions.

14. Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique

1.2.2 Les données personnelles et leur anonymisation

L'article 2 de la loi « informatique et libertés » [15] définit la notion de *donnée personnelle*.

Definition 6 Constitue une donnée à caractère personnel « toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auquel peut avoir accès le responsable de traitement ou toute autre personne ».

Les données personnelles peuvent être *directement identifiantes* (nom, prénom,...) ou *indirectement identifiantes* (numéro d'inscription au répertoire de l'INSEE, numéro téléphone, adresse IP,...).

À l'inverse, n'est pas une donnée à caractère personnel une donnée, ou une combinaison de données qui n'est pas reliée à une personne physique et qui n'identifie pas une personne physique. Le caractère personnel dépend des moyens pouvant être mis en œuvre et conduisant à l'identification d'une personne. Le champ des données personnelles est donc *évolutif*, car il dépend de l'état de la technique et de ses performances. Il est également important de prendre en compte le *contexte du traitement* de l'information pour en déduire le caractère personnel : en effet, l'identification d'un individu peut également être déduite à partir des connaissances propres des personnes qui traitent les données.

Definition 7 L'anonymisation est une procédure permettant de casser le lien entre une/des donnée(s) et une personne physique, protégeant ainsi la vie privée des individus.

Le recours à l'anonymisation peut s'avérer nécessaire quand une entité n'est pas légitime à traiter des données personnelles pour les finalités qu'elle a déterminées. En effet, l'anonymisation permettrait de bénéficier de la richesse des données, tout en protégeant la vie privée des individus et en répondant aux exigences de protection des données à caractère personnel.

Cependant le concept d'anonymisation n'est pas simple à appliquer, car il s'avère difficile, voire impossible de déterminer si une donnée est toujours une donnée à caractère personnel, ou si elle a été anonymisée. En effet, un individu peut être identifié indirectement, par exemple grâce à d'autres sources de données qui, combinées, conduisent à son identification.

Pour évaluer l'anonymisation, il faudrait pouvoir répondre à plusieurs questions :

- Est-il raisonnablement possible qu'une personne physique soit identifiée à partir des données traitées et à partir d'autres données ?
- Quelle est la probabilité qu'une ré-identification soit effectuée ?
- Quelle est la probabilité que la ré-identification soit correcte ?

Si vous êtes intéressés par les différentes techniques d'anonymsation existantes, leur efficacité et leurs limites, vous pouvez vous referez au document rédigé par Benjamin Nguyen [16].

15. Article 2 de Loi 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés
16. Techniques d'anonymisation, B. Nguyen, Statistique et société, 2014

1.2.3 Les données de santé

Les données de santé sont considérées comme des données *sensibles* au sens de la loi « Informatique et Libertés ».

Definition 8 Les données de santé sont des « données à caractère personnel qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à la santé ou à la vie sexuelle de celles-ci » (art.8 loi « Informatique et Libertés »).

Il n'y a pas de définition légale de la donnée de santé ni en droit français, ni en droit européen, mais des indications apparaissent dans :

- l'article 1111-8 du Code de la Santé Publique (CSP) : les données « recueillies ou produites à l'occasion des activités de prévention, de diagnostic ou de soins » ;
- la jurisprudence Lindqvist Cour de Justice des Communautés Européennes ((CJCE) 6/11/2003, Bodil Lindqvist, C-101/01) éclaire quelque peu sur la définition des données de santé : la Cour de justice de l'UE considère dans cette affaire que l'indication du fait qu'une personne s'est blessée au pied et est en congé de maladie partiel constitue une donnée à caractère personnel relative à la santé au sens de l'article 8, paragraphe 1, de la directive 95/46/CE. La Cour donne à cette occasion une interprétation large à l'expression « données relatives à la santé », « de sorte qu'elle comprenne des informations concernant tous les aspects, tant physiques que psychiques, de la santé d'une personne » ;
- l'arrêt du Conseil d'État du 19 juillet 2010 N° 317182 apporte également quelques indications en considérant que, pour les élèves inscrits en classe spécialisée, « la seule mention du fait que l'élève est scolarisé dans une structure de soins ne donne pas assez d'informations sur la nature ou la gravité de l'affection dont il souffre pour être considérée comme une donnée de santé ». Par contre, si l'information permet d'identifier le type de handicap ou de déficience de l'élève, alors il s'agira d'une donnée de santé ;
- le règlement européen déclare que les « données concernant la santé regroupent toute information relative à la santé physique ou mentale d'une personne, ou à la prestation de services de santé à cette personne ».

Les données de santé sont considérées comme particulièrement sensibles, et font l'objet d'un encadrement particulier.

1.2.4 Licences logicielles

Pour terminer sur les problématiques juridiques, nous souhaiterions faire une incise sur les problématiques liées aux logiciels et plus particulièrement aux logiciels libres en citant une brochure [17] qui émane du Groupe Thématique Logiciel Libre (GTTL) du pôle de compétitivité SYSTEMATIC. Cette brochure aborde la protection du logiciel, le cadre juridique spécifique du logiciel libre, l'usage et l'exploitation du logiciel libre. Ce document est en lien indirect avec notre problématique mais il nous semble important de rappeler combien les logiciels libres jouent un rôle majeur dans notre quotidien. En particulier de nombreux intergiciels (logiciels servant d'intermédiaire de communication entre plusieurs applications) et outils de cloud et de big data sont des projets en logiciel libre. Ainsi l'appropriation de ces technologies par des communautés est plus aisée et, en contre

17. Fondamentaux juridiques - Collaboration industrielle et innovation ouverte

partie, l'adoption passe par les communautés qui en font des succès ou des échecs et non plus par les circuits, qui peuvent être verrouillés, des grands opérateurs.

1.3 Le big data

Beaucoup de chiffres circulent sur le volume de données produites, collectées et analysées quotidiennement. Tous s'accordent pour nous dire que la volumétrie va croître de manière exponentielle. Par exemple le cabinet IDC [18] estimait en 2011 déjà que le nombre de données produites et partagées sur Internet atteindrait les 8 zettaoctets en 2015. Rappelons que 1 zettaoctets = 1000 exaoctets = 1 million de pétaoctets = 1 milliard de téraoctets ! Ce « déluge » de données est aussi perceptible depuis plus d'une dizaine d'années dans le domaine de la recherche scientifique et il apparaît désormais clairement que le volume des données collectées durant ces prochaines années dépassera celui recueilli durant les siècles précédents [19].

Le big data caractérise les données de recherche de nombreux domaines scientifiques (génomique, astronomique, climatique, etc.) qui comprennent des volumétries de l'ordre du téraoctet, voire du pétaoctet en 2015 pour une seule expérience. Les big data sont définies de multiples manières dans la littérature. Nous reprenons celle du cabinet IDC [20] :

Definition 9 Le big data est présenté par nos collègues américains et défini comme « a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis ». (Cabinet IDC)

C'est à partir de 2001 que la communauté scientifique a commencé à expliquer les enjeux inhérents à la croissance des données comme étant tridimensionnels au travers de la règle dite « des 3V » (volume, vélocité et variété), puis « des 5V » (volume, vélocité, variété, véracité, valeur) ou plus car les spécialistes sont en désaccord sur le nombre de V à considérer et cela dépend des disciplines. Les définitions usuelles associées à ces dimensions sont les suivantes :

- Volume - Fait référence à la taille de l'ensemble des données créées par les différents acteurs ou dispositifs de l'écosystème ;
- Variété - Fait référence au nombre de systèmes individuels et/ou types de données distincts utilisés par les acteurs ; Les multiples formats de données au sein d'un même système posent problème, en pratique ;
- Vélocité - Représente la fréquence à laquelle les données sont à la fois générées, capturées, partagées et mises à jour ;
- Valeur - Fait référence à notre capacité à transformer en une valeur l'ensemble des données créées compte tenu des deux types de données précédentes : initiales et non-immuables ;
- Véracité - Fait référence au fait que les données initiales peuvent être modifiées involontairement à travers leurs cycles de vie, ou alors au fait que l'on gère des données non-immuables. L'ensemble affecte la qualité de l'information et donc la confiance en l'information.

Sur la Figure 1.1, d'autres termes que les cinq termes initiaux apparaissent et ces termes permettent de préciser certains buts, qui sont en lien avec les étapes du cycle de vie des données. À

18. International Data Corporation
19. The data deluge: an e-Science perspective
20. Big data Analytics: Future Architectures, Skills and Roadmaps for the CIO, IDC, 2011

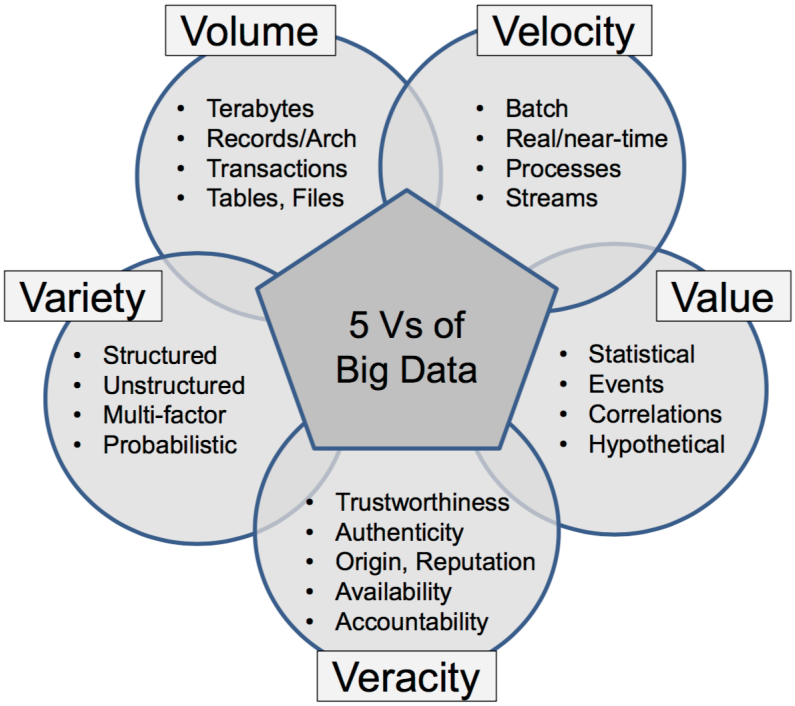


FIGURE 1.1 – Les 5V des big data

titre d'exemple, considérons les termes Terabytes, Records/Arch, Transactions, Tables, Files. Tous ces termes font référence aux problématiques du stockage, de la représentation (Tables, Fichiers) et au modèle d'accès (Transactionnel). *Dans un tel cadre, le but du big data est de fournir des moyens performants pour stocker, représenter et accéder aux données*.

La Figure 1.2 est un raffinement des notions du big data et elle est issue d'un rapport du NIST [21]. Cette figure en couleur, où chacune des cinq couleurs concerne soit une définition soit un défi ou but du big data, permet de lancer les discussions sur les grandes questions scientifiques et techniques qui se posent dans le domaine, notamment sur les questions des nouveaux modèles des données, les problématiques d'analyse et les problématiques d'architectures fonctionnelles et techniques ainsi que sur les outils. La densité de termes démontre que les questions enfouies derrière la terminologie big data sont à envisager sous de multiples angles d'approche. Dans la suite nous allons nous concentrer plus particulièrement sur les aspects d'architecture des systèmes parce qu'ils sont liés directement à notre sujet sur l'hébergement des données.

1.3.1 Gestion de données dans le cadre de la recherche

Aujourd'hui les dimensions du big data impliquent que la gestion des données, dans le cadre de la recherche, fasse de plus en plus partie intégrante du projet de recherche. La gestion des données requiert une organisation, une planification, et un suivi rigoureux tout au long de la vie du projet

21. National Institute of Standards and Technology

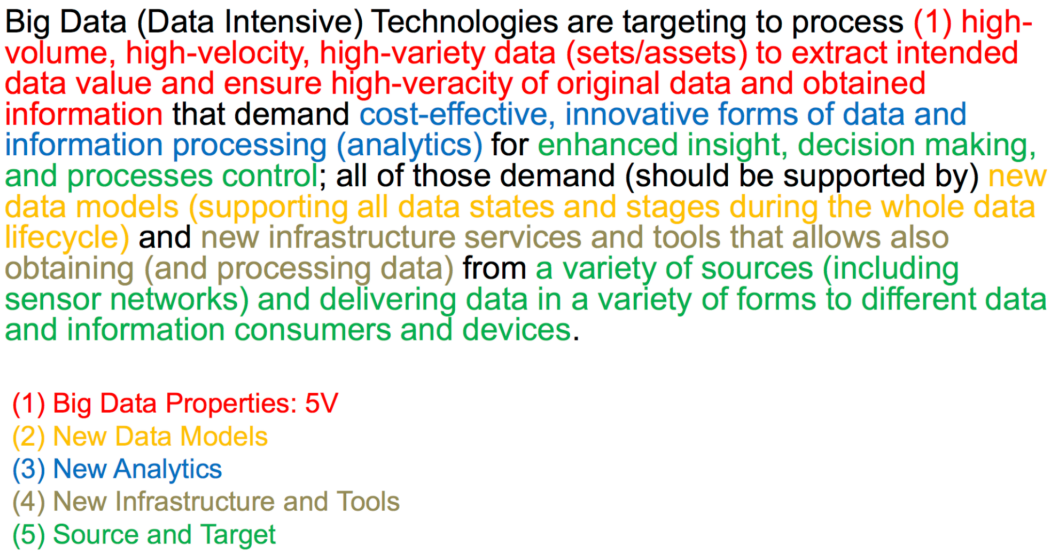


FIGURE 1.2 – Catégorisation des termes et buts du big data (d'après le NIST)

et au-delà, pour assurer leur pérennité, accessibilité et réutilisation. La gestion des données dans la recherche répond aux objectifs suivants [22] :

- elle accroît l'efficience de la recherche, en facilitant l'accès et l'analyse des données par le chercheur qui a conduit la recherche ou par tout nouveau chercheur ;
- elle assure la continuité de la recherche par la réutilisation des données, tout en évitant la duplication des efforts ;
- elle favorise la diffusion élargie et accroît l'impact : des données de recherche correctement formatées, décrites et identifiées gardent une valeur à long terme ;
- elle permet d'assurer l'intégrité de la recherche et la validation des résultats. Des données de recherche exactes et complètes permettent également la reconstruction des événements et des processus qui ont conduit aux résultats ;
- elle réduit le risque de perte et renforce la sécurité des données par l'utilisation de dispositifs de stockage robustes et adaptés. Cependant, notons dès à présent que ces problèmes ne sont pas que de nature technique. Nous examinerons plus loin les différents acteurs œuvrant à la sécurisation des systèmes d'information. L'enjeu est alors de faire travailler ensemble tous ces acteurs ;
- elle accompagne l'évolution actuelle de la publication : les revues scientifiques proposent de plus en plus que les données qui constituent la base d'une publication soient partagées et déposées dans un entrepôt de données accessible. De ce fait, la gestion des données de recherche facilite la soumission aux revues scientifiques d'articles s'appuyant sur des jeux de données documentés ;
- elle satisfait aux conditions de financement du projet par les bailleurs de fonds : ceux-ci s'intéressent de plus en plus à ce que les chercheurs font des données produites au cours d'un projet et conditionnent souvent le financement à l'ouverture de ces données pour qu'elles soient accessibles librement et gratuitement ;

22. Pourquoi gérer les données de la recherche ?, Cirad, 2015

- elle atteste de responsabilité : en gérant vos données de recherche et en les rendant disponibles, cela fait preuve d'une utilisation responsable du financement public de la recherche.

1.3.2 Plan de gestion et cycle de vie des données

Une bonne gestion des données nécessite l'élaboration d'un plan de gestion de données (Data Management Plan - DMP) et doit prendre en compte toutes les étapes du cycle de vie des données.

Definition 10 Le plan de gestion de données est un document formel explicitant la façon dont on obtient, documente, analyse et utilise les données à la fois au cours et à l'issue d'un projet de recherche. Il décrit la manière dont les données sont produites, décrites, stockées et diffusées.

Les opérateurs de la recherche, par exemple au niveau européen, demandent à ce que les chercheurs construisent un plan de gestion de données et l'intègrent à leur projet scientifique [23]. Il y aura bientôt très certainement une obligation également au niveau national à fournir ce type de document en réponse à chaque appel à projets.

R Vous pouvez vous inspirer des modèles de DMP préexistants comme celui produit par le Service Commun de Documentation (SCD) de Paris Descartes, le Bureau des archives et la Direction d'Appui à la Recherche et à l'Innovation (DARI) de Paris Diderot [24] ou élaborer votre propre plan de gestion des données à l'aide d'outils en ligne [25].

Il existe plusieurs représentations du cycle de vie des données de la recherche.

Definition 11 Schématiquement, le cycle de vie d'une donnée correspond à la période qui s'étend de la conception de la donnée à son utilisation, jusqu'à sa destruction ou sa conservation pour raison historique ou scientifique.

Selon le centre d'archives UK Data Archives [26] spécialisé dans les données de recherche en sciences sociales, le cycle de vie des données de recherche comporte six grandes étapes :

- la création des données
  - définition du référentiel de la recherche
  - mise en place d'un plan de gestion de données
  - localisation des données
  - collecte des données
  - description des données
- le traitement des données
  - saisie des données, numérisation, traduction, transcription
  - contrôle, validation, nettoyage des données
  - anonymisation et description des données
  - gestion et stockage des données
- l'analyse des données
  - interprétation
  - production des données dérivées
  - production des résultats de recherche

23. Le libre accès aux publications et aux données de recherche, Portail H2020
24. Réaliser un plan de gestion de données, A. Cartier et al. 2015
25. DMP Tool
26. Research Data Lifecycle, UK Data Archives

  - préparation des données pour la préservation
- la préservation des données
  - migration dans un format pérenne
  - création des métadonnées
  - documentation des données
  - archivage des données
- l'accessibilité et le partage des données
  - distribution/partage des données
  - contrôle des accès (ou non)
  - établissement des protections (copyright vs. copyleft)
  - promotion des données via une plateforme Internet ouverte
- la réutilisation des données
  - suivi et revues des recherches
  - nouvelles recherches à partir des données
  - croisement des données avec d'autres données issues d'autres domaines

Dans la sous-section 2.1.2, nous présenterons la vue purement fonctionnelle d'un modèle du cycle de vie des données dans les e-sciences (sciences utilisant le numérique comme le cloud et les big data).

La gestion des données doit prendre en compte l'intégralité du cycle de vie de la donnée. Il est important de noter que la préservation et l'accessibilité des données ne se décident pas au moment de l'archivage, mais qu'elles sont affaire d'anticipation. Toute personne souhaitant optimiser la gestion de ses données devrait donc prendre un temps de réflexion afin d'établir clairement comment naissent et meurent les données. La description des données par des métadonnées dès l'étape de création est primordiale pour pouvoir gérer efficacement les données tout au long de leur cycle de vie. L'anticipation est vraiment à encourager et inversement il doit être possible de changer les statuts des documents en fonction de l'évolution de l'environnement.

1.4 Hébergement des données et cloud computing

Le cloud offre aujourd'hui un concept opérationnel qui permet d'accéder, via le réseau et à la demande, à des ressources informatiques de stockage et de calcul virtualisées et mutualisées. Ce concept est aujourd'hui de plus en plus en vogue dans les contextes du big data et de l'open data notamment pour y déposer des données y compris dans les usages les plus simples comme le stockage de fichiers en ligne pour les particuliers. Les services de type Dropbox [27] se sont en effet multipliés dans un passé récent et sont utilisés par « monsieur tout le monde ».

Afin de mieux tirer parti de l'approche cloud, de nouvelles solutions de gestion et de traitement de données sont apparues. En effet, Les systèmes de gestion de données traditionnels (bases de données relationnelles) ne peuvent pas répondre aux défis de variétés et d'échelle requis par les big data. Les évolutions logicielles ont donc suivi assez naturellement les évolutions matérielles. Il s'agit tout particulièrement du développement de nouvelles bases de données adaptées aux données volumineuses et non-structurées, que l'on range principalement sous le vocable de NoSQL [28], et la

27. Dropbox, service de stockage et de partage de copies de fichiers locaux en ligne
28. Système de gestion de bases de données NoSQL

mise au point de modes de calcul à haute performance, par exemple le paradigme MapReduce [29]. Ces deux notions ont clairement marqué une évolution majeure pour accompagner le phénomène des big data du coté des langages et modes de programmation.

1.4.1 Infrastructure informatique de type cloud et données

Pour l'informaticien, une *infrastructure* (on parle aussi de *plateforme*) comprend des processeurs, du stockage et du réseau. Dans ce document nous allons parler plus spécifiquement d'infrastructures de type *cloud*. La définition du cloud computing donnée par le NIST (National Institute for Standards and Technology), est la suivante :

Definition 12 Le cloud computing est un modèle informatique qui permet un « accès à la demande et de manière aisée à un ensemble de ressources configurables (du réseau, des processeurs, du stockage, des applications et des services) qui peuvent être approvisionnées et libérées avec un effort minimal quant aux interactions avec le fournisseur. »

Un cloud du point de vue de l'informaticien est donc constitué de trois ingrédients :
- un ERP (Entreprise Resource Planning) pour gérer la relation cliente et le catalogue des applications que l'utilisateur peut déployer dans le cloud ;
- un modèle de déploiement (comment et sur quels types de processeurs les applications sont-elles déployées) ;
- des nœuds de stockage et de calcul c'est-à-dire une infrastructure matérielle.

D'un point de vue d'ensemble, le cloud est composé de *cinq caractéristiques essentielles*, et il repose sur *trois modèles de services* (SaaS, PaaS et IaaS, voir la Figure 1.4), et *trois modèles de déploiement* (cloud privé, cloud public et cloud hybride).

Les cinq caractéristiques essentielles du cloud, mises en avant par le NIST, sont :
- Ressources en libre-service et à la demande (on-demand self-service)
  Le cloud est dit « orienté service » c'est-à-dire qu'il arbore une architecture logicielle s'appuyant sur un ensemble de services simples développés en s'inspirant des processus métier de l'entreprise ou de l'université par exemple. Dans l'approche service il y a aussi une séparation claire entre la notion de contrat et celle d'implémentation. Le cloud s'affranchit des couches physique et matérielle (plus de gestion de serveurs et de logiciels système), plus d'installations/configurations de logiciels sur les PC, et pour les développeurs, un déploiement instantané sans gestion de plateformes hétérogènes. Le cloud est basé sur le principe de virtualisation (cf. 14 page 16) qui permet de réduire voire de supprimer la dépendance entre matériel et logiciel.
- Accès via le réseau étendu (broad network access)
  Les services sont accessibles via l'usage de protocoles et standards venant du monde Internet.
- Mutualisation des ressources (resource pooling)
  Les applications partagent un ensemble de ressources permettant d'atteindre de larges économies d'échelle.
- Approvisionnement rapide et ajustable (rapid elasticity)
  Le cloud permet de gagner en flexibilité. Les infrastructures sont allouées selon les besoins et ajustables à la hausse comme à la baisse.

29. Paradigme MapReduce

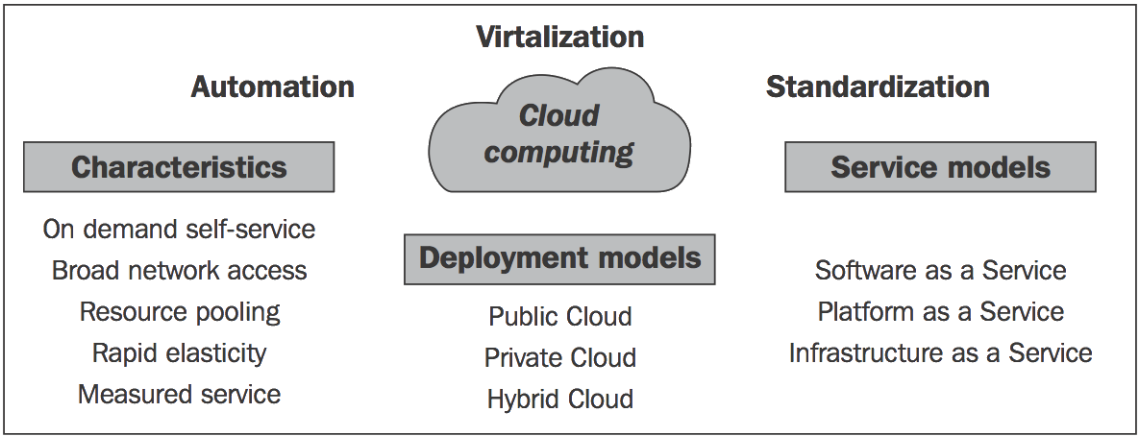


FIGURE 1.3 – Caractéristiques, modèles de services et de déploiement du cloud computing.

- Ressources et services pouvant être mesurés et contrôlés (measured service)
  Le cloud est aussi orienté « à l'usage » c'est-à-dire qu'il s'appuie sur des outils dits de « reporting » permettant de suivre la consommation avec une facturation souple à l'usage.

Le cloud computing a ouvert des possibilités de collaboration entre les individus et les entreprises, quelles que soient leurs localisations géographiques. Le cloud peut être public, souvent par l'intermédiaire d'un réseau distribué de ressources informatiques qui peuvent ne pas être dans les locaux de l'organisme d'accueil. Le cloud peut être privé quand un organisme ne l'ouvre pas sur l'extérieur et enfin il peut être hybride quand il permet de prendre des ressources à la fois dans un cloud public et privé.

1.4.2 Modèles de services proposés par le cloud computing

Selon le NIST, il existe trois modèles de services qui peuvent être offerts en cloud computing : SaaS, PaaS et IaaS (Figure 1.4).

Le logiciel en tant que service ou *Software as a Service* (SaaS) est un modèle d'exploitation des logiciels dans les clouds dans lequel ceux-ci sont installés sur des serveurs distants plutôt que sur la machine de l'utilisateur. Les clients ne paient pas de licence d'utilisation pour une version, mais utilisent généralement gratuitement le service en ligne ou payent un abonnement récurrent.

Une plateforme en tant que service, de l'anglais *Platform as a Service* (PaaS), est une autre forme de cloud computing où le fournisseur de cloud met à disposition des organisations un environnement d'exécution rapidement disponible, en leur laissant la maîtrise des applications qu'elles peuvent installer, configurer et utiliser elles-mêmes. A titre d'exemple la pile logicielle LAMP permettant de construire des serveurs de sites web est un PaaS. LAMP regroupe le système d'exploitation Linux, le serveur Web Apache, un serveur de base de données (MySQL ou MariaDB) et, à l'origine, PHP, Perl ou Python. Cloud9 IDE est un environnement de développement intégré en ligne (IDE) qui peut être vu aussi comme un service applicatif pour développer en ligne des applications dans de multiples langages de programmation.

L'infrastructure en tant que service ou, en anglais, *Infrastructure as a Service* (IaaS) est encore une autre forme de cloud où l'organisation achète du service chez un fournisseur de cloud. Ce service

peut être une machine nue ou une machine munie d'un système d'exploitation (OS) tel que Windows ou Linux. Cela peut représenter pour certaines directions des systèmes d'information (DSI) un moyen de réaliser des économies, principalement en transformant des investissements en contrats de location et aussi en évitant de s'occuper directement du refroidissement, de la redondance électrique, du contrôle d'accès physique. Nous en reparlerons au moment de la présentation des data centers.

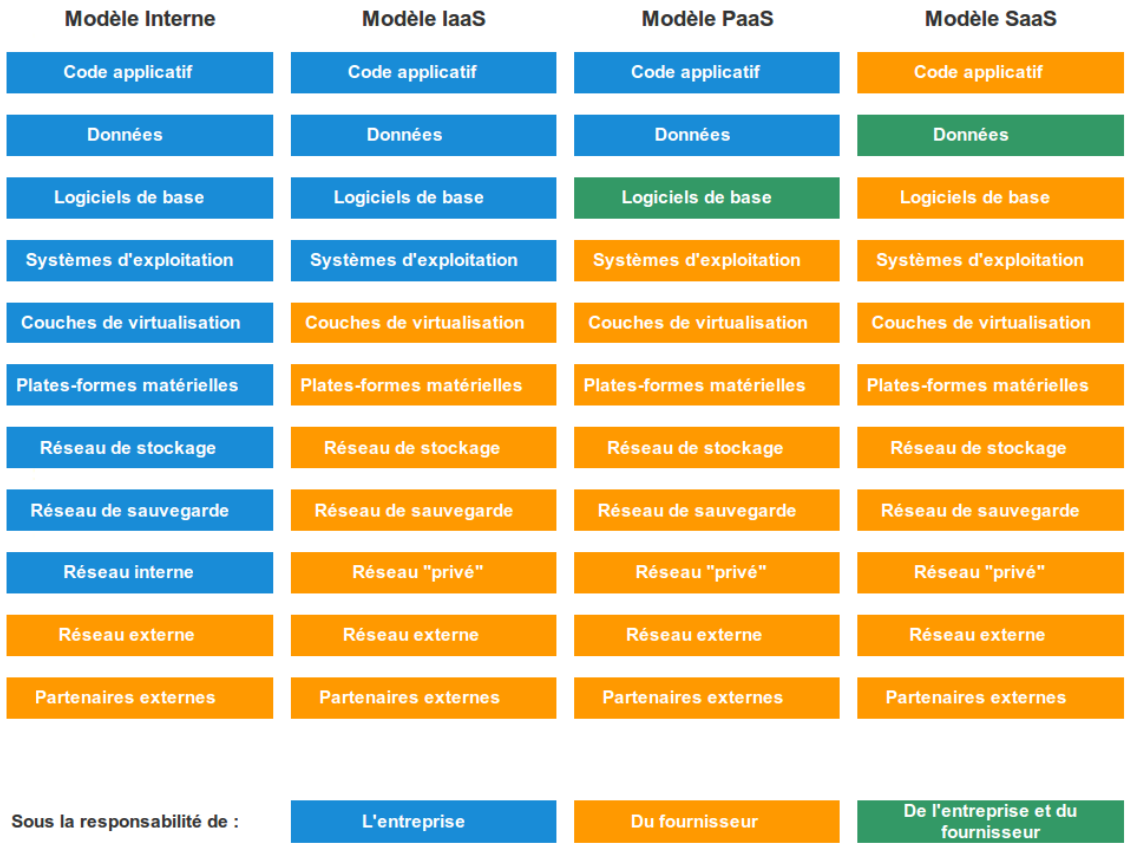


FIGURE 1.4 – Les différents services proposés par le cloud computing. Tiré du site d'Auranext.

1.4.3 Fonctionnement général du cloud computing

D'un point de vue des modèles économiques, le cloud computing est essentiellement une offre commerciale d'abonnement à des services externes. Les fournisseurs de cloud ont par exemple créé un système de paiement « pay-as-you-go » permettant de payer les ressources utilisées à la demande. Ainsi, un client peut louer un serveur sur une courte période et aux dimensions voulues en termes de nombre de processeurs, espace disque et bande passante réseau. Cela sous entend que le fournisseur mesure votre consommation en ressources pour vous la re-facturer, ce qui peut être conflictuel si la méthode est intrusive.

L'utilisateur du cloud est *au centre des préoccupations* et l'effort minimal que l'utilisateur doit faire pour l'utiliser est obtenu grâce à une automatisation tous azimuts comme par exemple lors du déploiement des services.

Comme nous l'avons déjà mentionné, le cloud vise à mutualiser à l'extrême les ressources matérielles et logicielles et à procurer un environnement unique et cohérent. Il est basé sur ce qu'on appelle une architecture *multi-tenant* (ou multi-entité).

Definition 13 L'architecture multi-tenant est un principe permettant de mettre à disposition un logiciel à plusieurs organisations clientes à partir d'une seule installation.

Ce concept peut être vu comme des locataires qui vivraient chacun dans un appartement d'un même immeuble. Cette organisation présente l'avantage de partager un certain nombre de ressources telles que le chauffage, la distribution d'eau, d'électricité, un service de gestion et d'entretien par un syndic... contrairement à des personnes habitant chacune dans leur propre maison. La mutualisation qui découle de ce type d'infrastructure permet au fournisseur de ne mettre à jour qu'un seul logiciel et pas N copies situées sur des lieux géographiques différents. L'économie qui en résulte est immédiate.

La question du service support doit également être la clé de la réussite d'un Cloud. On pourrait penser, au moyen d'arguments comptables, que puisqu'il s'agit de ne mettre à jour une seule copie plutôt que N, éventuellement dispersées sur plusieurs sites, nous pourrions diminuer le nombre d'informaticiens. C'est un raisonnement du passé. Avec le cloud il n'y a plus d'informaticien qui passe re-formater les disques durs. Les métiers évoluent. Il conviendrait mieux de redéployer les informaticiens pour ne pas se retrouver avec des outils non fonctionnels. Un service support aurait donc avantage à diffuser sa compétence, à renvoyer aux services compétents, ou encore à gérer les sorties de route.

Le cloud computing gère les services par le biais de la *virtualisation*. Cloud et virtualisation sont deux concepts bien différents ; la virtualisation est une technologie, le cloud est plutôt une approche conjuguant de multiples technologies, dont la virtualisation.

Definition 14 La virtualisation consiste à faire fonctionner un ou plusieurs systèmes d'exploitation / applications comme un simple logiciel, sur un ou plusieurs ordinateurs - serveurs / système d'exploitation, au lieu de ne pouvoir en installer qu'un seul par ordinateur.

Ainsi le fournisseur de services dans un cloud fait une économie sur le matériel nécessaire pour servir plusieurs clients. Les systèmes d'exploitation sont isolés les uns des autres par des techniques propres à la virtualisation, autrement dit il n'y a pas de mélange entre les environnements des clients, la virtualisation est donc « étanche ». L'idée est d'offrir à l'utilisateur un bac à sable dans lequel il peut travailler mais sans perturber l'utilisateur d'à côté. Depuis quelques années, le multiplexage de systèmes d'exploitation c'est-à-dire le fait d'exécuter plusieurs systèmes d'exploitation sur la même machine, est remplacé par un multiplexage de plusieurs applications dans des conteneurs isolés les uns des autres mais partageant les bibliothèques du système d'exploitation hôte. Nous obtenons de la sorte un usage plus économe en ressources car il est possible d'exécuter plusieurs dizaines de conteneurs voire une centaine sur un processeur alors qu'il n'est possible de faire exécuter que quelques systèmes d'exploitation complets sur le même processeur via la virtualisation classique. Nous sommes donc désormais loin des pratiques d'il y a quelques années durant lesquelles la règle « un serveur pour une application » était en vigueur.

1.4.4 Cloud, sécurisation et droit de propriété des données

Au delà de la mise en œuvre technique du déploiement dans le cloud, il peut souvent devenir difficile de savoir comment la propriété de l'information hébergée, utilisée et/ou consultée dans le cloud par les organisations informatiques multi-tenant est définie à cause de la mutualisation. En

outre, les incertitudes sur la sécurité et la confidentialité de l'information retenue dans le cloud pose problème ; les cas de violation de sécurité rencontrés par des entreprises établies ont soulevé de nouvelles inquiétudes parmi les entreprises du secteur public et dans le secteur privé ainsi qu'entre les individus.

Il est de plus en plus admis que les utilisateurs de systèmes de cloud doivent prendre conscience des risques potentiels vis à vis des questions de vie privée lorsque l'information est détenue et/ou est traitée dans les systèmes de cloud publics ou partagés. Ces préoccupations retardent la migration et l'adoption des technologies de cloud malgré leurs avantages indéniables. Cependant, selon le cabinet Gartner, les dépenses en services de cloud et en Infrastructure as a Service (IaaS) ont des taux de croissance annuelle estimés à 17,7% et 41,3% de 2011 à 2016, respectivement.

Le cloud correspond plus à une organisation virtuelle des services qu'à une organisation physique. Les systèmes de cloud sont généralement hébergés dans des centres de données.

Definition 15 Un centre de données ou data center est un site physique sur lequel se trouvent regroupés des équipements constituants du système d'information d'une organisation.

Il comprend un contrôle sur l'environnement (climatisation, système de prévention et de lutte contre l'incendie, etc.), une alimentation électrique d'urgence et redondante, ainsi qu'une sécurité physique élevée comme par exemple un grillage autour des serveurs, un contrôle d'accès biométrique, une surveillance avec des rayons laser.

Pour une description détaillée des infrastructures d'un data center, nous vous renvoyons à la Partie 3 de ce document qui présente un data center francilien.

Concernant les données, il est reconnu que le cloud computing peut conduire à différentes formes d'emprisonnement et les menaces suivantes ont été identifiées :

- Emprisonnement dans une plateforme/infrastructure : la migration d'un fournisseur de cloud utilisant une plateforme à un fournisseur de cloud utilisant une plateforme différente peut s'avérer très complexe ;
- Verrouillage des données : puisque le cloud est encore nouveau, les normes liées à la propriété, à savoir qui possède réellement les données une fois déposées dans le cloud, ne sont pas encore suffisamment développées, ce qui peut rendre complexe la migration des données si les utilisateurs de cloud computing décident de déplacer les données hors de la plateforme du fournisseur choisi initialement ;
- Emprisonnement par les outils : si des outils intégrés à un certain cloud pour gérer un environnement de cloud computing ne sont pas compatibles avec d'autres clouds, ces outils ne peuvent gérer que des données ou des applications qui vivent dans le seul cloud du fournisseur.
- Les inconvénients attribués aux clouds mis en place par des prestataires privés (e.g., Google, Amazon, Microsoft, Apple) sont : (1) la soumission au cadre légal en vigueur dans le pays qui accueille les données ; (2) les temps et coût de transfert (à relier avec les débits en vigueur et les garanties offertes sur ces débits) ; (3) la dépendance au prestataire de service ; (4) la sécurisation des données et (5) la pérennité du service.

Pour parer à toutes les éventualités, les grands comptes étudient à la loupe les contrats qui les lient aux fournisseurs. . . , et prennent aussi de bonnes assurances !

1.4.5 Adoption du cloud

En octobre 2015 paraissait un rapport du CIGREF (Club Informatique des GRandes Entreprises Françaises) intitulé « La réalité du cloud dans les grandes entreprises ». Ce rapport note que trois ans après avoir analysé les « fondamentaux du cloud computing » et la question de « la protection des données dans le cloud », le CIGREF s'interroge maintenant sur les raisons pour lesquelles ses adhérents passent au cloud. Il en ressort principalement que

- la pression des directions métiers,
- le cloud comme vecteur d'innovation et d'agilité,
- la pression des fournisseurs,
- la simplification des infrastructures,
- la réduction des coûts,

se révèlent comme les principaux moteurs d'adoption de solutions de cloud.

Pour autant, l'introduction du cloud comporte aussi sa part d'ombre et de difficultés dans sa mise en œuvre ; l'étude s'est donc aussi intéressée :

- à la gouvernance à mettre en place,
- aux problématiques de mise en œuvre,
- à la difficulté d'évaluer le niveau de sécurité et d'assurer la conformité à la réglementation d'une offre cloud,
- aux difficultés de contractualisation avec les fournisseurs,
- aux problèmes d'adaptation de la réglementation à cette innovation que représente le cloud.

Dans le monde académique, une étude commandée en 2016 par l'ALECSO et l'ITU intitulée « Guidelines to improve the use of cloud computing in education in the Arab countries » [30] dresse à la fois un panorama d'expériences ayant abouti à l'adoption des technologies de cloud computing dans le monde de l'éducation et de la recherche, mais aussi propose une démarche et des recommandations techniques pour avancer sur ces questions. L'ensemble de l'ouvrage est plutôt destiné, d'une part aux décideurs, et d'autre part à un public d'informaticiens pour implémenter les recommandations. Cette étude est donc d'une portée générale en matière de stratégies de migration vers le cloud et contient de nombreuses références à des projets accomplis.

1.4.6 Stockage et archivage dans le cloud

Le stockage d'information est assuré par un dispositif informatique. Les technologies actuelles pour stocker sont les *mémoires de masse* (disque dur, carte SSD ou microSD) ou les *mémoires à accès rapide* comme la mémoire RAM (Random Access Memory) des ordinateurs portables. Quand on écrit dans un support de masse, l'information y reste même après une coupure de courant. Quand on écrit dans une RAM et qu'il y a une coupure de courant, l'information est perdue. L'avantage des mémoires RAM est que les temps d'accès et les débits de lecture et d'écriture sont bien supérieurs à ceux d'un support de masse. Cependant, pour un prix équivalent, un support de masse offre une capacité de stockage bien supérieure à une mémoire RAM.

Un disque dur classique est un ensemble de plateaux munis d'une surface magnétique qui, telle une bande, enregistre des informations grâce aux déplacements de têtes de lecture-écriture fixées à des bras. L'ensemble de ces bras s'appelle le peigne. Grâce à la rotation rapide du disque (de 5400 à 15000 tours/minute) la tête de lecture flotte au-dessus de la surface du disque magnétique à une distance comparable à l'épaisseur d'une mouche entre le train d'atterrissage d'un A320 et le

30. Guidelines to improve the use of cloud computing in education in the Arab countries, 2016

tarmac. Cette technologie atteint ses limites et exploite les propriétés d'un mécanisme peu intuitif appelé effet tunnel en mécanique quantique. Son temps de réponse se mesure en milliseconde ou 10-3 seconde (environ 15 ms).

La carte SSD, a contrario, est un mécanisme de mémoire de masse électronique basé sur des composants électroniques capables de conserver l'information sans alimentation électrique mais dont les temps de réponses en lecture et écriture sont bien meilleurs que ceux du disque dur mais moins bons que ceux de la mémoire vive ou à accès rapide RAM. Son temps de réponse se mesure en milliseconde (environ 0,1 ms). Celui de la mémoire RAM est de 10 nanoseconde (10 10-9 seconde).

En terme de coût, nous avons le classement simple suivant : 1 Go en disque dur mécanique < 1 Go en mémoire de masse SSD < 1 Go en mémoire vive RAM.

Ces technologies évoluent et la baisse du coût de leur fabrication y contribue grandement. Nous parlons actuellement de mémoire NVRAM (Non Volatile RAM) qui devrait être un standard de nos machines dans un proche avenir, et qui promet des temps d'accès 100 fois supérieurs à ceux des SSD. Cette technologie a pour objectif de garder à la fois les avantages des RAM et ceux des supports de masse. La technologie NVRAM actuelle la plus connue est la mémoire flash. Les mémoires NVRAM impactent potentiellement l'organisation de la mémoire sur un système informatique. L'organisation est hiérarchique avec en haut de la hiérarchie les mémoires à accès très rapides et en bas de la hiérarchie les mémoires « lentes » mais de grandes capacités. Les NVRAM suppriment un niveau dans la hiérarchie. C'est un phénomène qui a toujours existé : les disquettes $5"^{1/4}$ ont été remplacées par d'autres supports qui eux mêmes sont supplantés un jour ou l'autre. Le CD-ROM par exemple est en train de disparaître.

R Une présentation technique de la technologie NVRAM et de son intégration dans les systèmes informatiques est disponible en ligne [31]. Il s'agit d'une présentation faite dans le cadre de l'atelier international High Performance Data Intensive Computing (HPDIC) à Phoenix, Arizona. Cette présentation s'adresse plutôt à un public d'informaticiens.

Dans le passé, on *archivait* un document quand il était en fin de cycle de vie. L'archivage aujourd'hui est mis en œuvre dès la création des documents qui sont enrichis tout au long du cycle de vie du document. Il convient cependant de distinguer l'archivage du stockage et de la sauvegarde.

Definition 16 Le stockage concerne les actions, outils et méthodes permettant d'entreposer des contenus électroniques. La sauvegarde concerne l'ensemble des actions, outils et méthodes destinés à dupliquer des contenus électroniques d'origine dans un but de sécurisation des contenus et pour éviter leur perte. Enfin, l'archivage électronique concerne l'ensemble des actions, outils et méthodes mis en œuvre pour rassembler, identifier, sélectionner, classer et conserver des contenus électroniques à très long terme.

Dans le cloud, il est possible de louer de l'espace de stockage. Prenons ici l'exemple du service S3 d'Amazon. Nous ne parlons pas ici de la location de type Dropbox car nous voulons discuter du service de stockage comme un service parmi d'autres services et non pas comme un service isolé et qui ne sert qu'à stocker des photos par exemple. Il faut cependant reconnaître que nombre de ces services utilisent en sous main le service S3.

Le service Amazon S3 [32] a été lancé en 2006. Il est généralement couplé avec le service EC2

31. Future server platforms: persistent memory for data-intensive applications, S. Kannan and A. Gavrilovska, 2014
32. Service Amazon S3

(Elastic Compute cloud) qui lui, concerne la location de services de calcul (sur les fichiers gérés par S3 [33]). Le service S3 se présente sous la forme d'un éventail de classes de stockage conçues pour différents cas d'utilisation. L'utilisateur paie en fonction de la classe de stockage choisie. Ce modèle économique est également appliqué pour S2. Avec S3 l'utilisateur peut choisir la classe qu'il souhaite en fonction de ses besoins. Il dispose de certaines subtilités comme la possibilité d'utiliser « le stockage à redondance réduite (RRS) qui est une option de stockage d'Amazon S3 qui permet aux clients de réduire leurs coûts en stockant les données non critiques et reproductibles à des niveaux de redondance inférieurs à ceux du stockage standard Amazon S3. »

Les métiers de bibliothécaires (voir 1.5.4) et de documentalistes sont particulièrement concernés par l'archivage. L'informaticien quant à lui construit des outils pour faciliter toutes les tâches que nous venons de mentionner parce que de plus en plus de supports à archiver se présentent sous la forme électronique.

Pour le grand public ou les professionnels, Amazon propose depuis 2012 un service d'archivage sur le long terme et de sauvegarde. Il s'agit du service Glacier [34]. Amazon Glacier est adapté aux données qui sont accédées rarement et pour lesquelles un temps de récupération de 3 à 5 heures est acceptable. L'offre tarifaire est également intéressante puisqu'on trouve une offre à partir de 0,007 USD par gigaoctet et par mois. Amazon affirme que le service est sécurisé à la fois pour le transfert des données et aussi via un chiffrement automatique sur le site hébergeur.

Un des avantages des services Amazon dont nous venons de parler tient dans la possibilité de les interfacer de manière simple via une API (Application Programming Interface) normalisée de type REST (REpresentational State Transfer [35]) par exemple. Quand les interfaces sont normalisées, des écosystèmes émergent plus facilement. Cela explique par exemple que beaucoup d'offres logicielles de stockage utilisent en sous main les infrastructures d'Amazon. REST est un style architectural qui s'applique tout autant à la réalisation d'applications pour un utilisateur humain qu'à la réalisation d'architectures orientées services destinées à la communication entre machines.

Enfin, il convient de mentionner des initiatives de type Owncloud [36] et Seafile [37] qui offrent des solutions complètes (côté serveur et côté client) pour la sauvegarde de fichiers. Ils remplacent avantageusement les services « à la Dropbox ». Il s'agit de deux solutions open source de type cloud (SaaS) qu'une université peut déployer pour ses usages. L'université de Strasbourg utilise Seafile ; Owncloud est utilisé à Reims avec le cluster de calcul. Les aspects collaboratifs, de synchronisation, de partage et de cryptage de fichiers sont particulièrement bien étudiés dans ces solutions.

1.5 Nouveaux métiers, nouveaux acteurs et rôles

C'est entre 2010 et 2013 que l'on retrouve en France de nombreux rapports concernant le marché de l'emploi autour du cloud et des big data. Sur le dernier semestre de 2013, plusieurs études montrent que la recherche d'expertise dans le domaine du big data en France se précise [38]. Une étude

33. Amazon S3 and Amazon EC2
34. Amazon Glacier
35. Representational State Transfer
36. Owncloud
37. SeaFile
38. HTML5 et big data, les compétences IT les plus en vogue, V. Juhan, 2013

américaine montre aussi qu'au niveau international les besoins sont sur les technologies du Web (HTML5) ou encore dans la gestion de données via les technologies MongoDB et Hadoop. Le mot clé PaaS, qui nous vient du cloud, arrive en 9e position des mots clés les plus cités.

1.5.1 Syntec numérique

Le Syntec numérique, 1er syndicat professionnel de l'écosystème numérique français, a publié fin 2013 les résultats d'un Contrat d'Étude Prospective (CEP) qui identifie clairement de nouveaux métiers dont analyste big data [39]. Cette étude analyse les besoins en compétences et en recrutement de la filière numérique à l'horizon 2018 et dresse la cartographie de l'offre de formation initiale et continue dans les différents domaines du numérique.

1.5.2 Étude de l'APEC

L'APEC (Association Pour l'Emploi des Cadres) relève dans l'étude emploi-cadre 2013-2014 [40] que « le marché de l'emploi cadre affiche un fort dynamisme dans les activités informatiques et télécommunications. Les recrutements de cadres y ont ainsi augmenté de 13 % en 2012 par rapport à 2011, alors qu'ils ont reculé de 1 % pour l'ensemble des entreprises ». De plus, les prévisions de croissance des effectifs sont bonnes et « la grande majorité des recrutements de cadres dans le secteur concernent le cœur de métier des entreprises, c'est-à-dire la fonction informatique. À titre d'exemple, les métiers concernés peuvent être des ingénieurs informaticiens, des directeurs informatiques, des développeurs, des ingénieurs d'exploitation. » Cela veut dire que les postes « cœur métier » sont beaucoup plus recherchés que les postes de commercial, étude R&D, administration, finance… Ils représentent 77% des recrutements de cadre du secteur informatique d'après l'étude de l'APEC.

1.5.3 Identifications par le pôle Systematic et par l'OPIIEC

Le livre blanc sur le big data de Stéfane Fermigier [41] démontre que cette thématique est largement dominée par les outils open source ou logiciels libres. L'observatoire des métiers de l'informatique propose une étude [42] sur les compétences et formations open source en France et note, en synthèse, que « l'évolution du marché vers le cloud computing ne fait que renforcer ce modèle de développement. Les besoins en compétences sont ainsi très importants et l'offre de formations sur le territoire français est peu ou pas connue des professionnels comme des éventuels apprenants ». Le rapport de synthèse globale de cette étude note que :

> La quasi-totalité des étudiants est confrontée à l'open source, a minima dans un objectif pédagogique. La majorité des établissements ne fait pas de distinction entre les enseignements open source et les outils du marché. Pour autant, certaines formations ont développé des approches de gestion de projet open source, incluant par exemple les contenus suivants :
> - partenariat avec un acteur majeur de l'open source ;
> - projets comportant des jalons avec des versions intermédiaires, intégrant leurs propres développements mais aussi ceux de la communauté et des contributeurs (dans la promotion ou de l'extérieur) ;

39. Contrat d'Études Prospectives du secteur professionnel du Numérique, Syntec numérique, 2013
40. le marché de l'emploi cadre dans l'informatique et les télécommunications, Apec, 2013
41. big data & open source : une convergence inévitable ?, S. Fermigier, 2012
42. Compétence et formation open source en France, OPIIEC, 2013

- utilisation des outils de gestion de projet open source : GitHub, blog, gestion des contributions.

Enfin l'étude montre que peu d'établissements en France sont aujourd'hui présentés comme spécialisés autour du logiciel libre. Les perspectives d'évolution des établissements de formation en informatique vont, toujours d'après l'étude de l'observatoire, « vers un renforcement du poids de la formation autour des infrastructures, des réseaux et du cloud computing ».

L'observatoire pense même qu'il y aura un « renforcement du nombre de diplômés issus de l'apprentissage et une baisse du nombre de diplômés issus des formations initiales ». L'élément le plus important pour introduire les enjeux de la formation figure dans la synthèse de l'étude et qui propose les recommandations suivantes :

- développer un module de formation à destination des entreprises sur les stratégies d'innovation et le développement de nouveaux usages via le logiciel libre ;
- développer des modules de formation et de perfectionnement à destination des acteurs du secteur IT (Informatique et Télécommunications) sur les technologies associées aux segments en forte croissance (e-commerce, mobilité, datacenter, cloud computing, big data…) : Linux, Apache, Java, Tomcat, NoSQL, Hadoop, cloud OpenStack ;
- mettre en place un groupe de réflexion sur une stratégie de certification des compétences autour des technologies clés de l'open source.

1.5.4 Émergence de nouveaux métiers

Les nouveaux métiers apparus ces dernières années sont le data scientist [43] qui est le résultat de l'évolution de plusieurs métiers comme le data miner et le data analyst. Le data manager [44] quant à lui organise les données qu'il recueille pour faciliter la recherche d'information et permettre aux entreprises de définir des axes stratégiques.

Le site Talents du numérique [45] par le passé connu sous le nom de PassInformatique est un point d'entrée à consulter si vous êtes à la recherche d'informations sur les métiers de l'informatique. Ce site comporte des fiches métiers et donne la liste des formations, à tous les niveaux, pour atteindre ces métiers. Les compétences requises y sont également clairement indiquées.

Concernant la gestion des données de recherche, dans le rapport de Swan & Brown [46] sur « les compétences, le rôle et la structure des parcours professionnels des chercheurs et curateurs de données », en plus des métiers de créateur de données (la plupart du temps des scientifiques), d'analyste de données (data scientist) et d'administrateur de données (data manager), on mentionne la nécessité des qualifications des *bibliothécaires* pour les aspects organisationnels, de description, de sélection et de préservation des données. Il s'agit tout particulièrement des qualifications et compétences nécessaires pour améliorer la qualité des activités de curation des données. La curation

43. Métier du Data scientist
44. Métier du Data manager
45. Talents du numérique
46. The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs, A. Swan and S. Brown, 2008

est issue des pays anglo-saxons et apparaît en France en 2011. Le « curator » est la personne qui sélectionne les œuvres d'art en vue d'une exposition en fonction d'un certain type de public. Le « content curator » est la personne qui partage sur Internet ses découvertes de données. Les compétences associées aux activités de curation des données s'appuient sur des scenarii ou les bibliothécaires consultent des équipes de recherche et étudient leurs activités, en particulier celles liées à l'administration des données [47].

Definition 17 Le rôle des bibliothécaires, actuellement considérés comme les spécialistes de la science de l'information, se résume à (1) analyser les statistiques de consultation et de documentation en vue d'évaluer les besoins des chercheurs, (2) être un acteur de l'indexation des données : « médiateur de la valorisation des données » et (3) constituer une cellule d'appui pour la gestion des données de la recherche en élaborant un plan de gestion de données.

Les bibliothécaires vont tout particulièrement contribuer à identifier les besoins des chercheurs et les assister sur le volet « métadonnées ». Aussi, la question de l'ouverture des données de recherche offre une opportunité unique à ces professionnels de la documentation : celle de remodeler, à l'échelle des établissements de recherche, leur(s) lien(s) avec la communauté des chercheurs [48].

Dans l'article [49], les auteurs présentent un tableau de synthèse sur le rôle des bibliothécaires dans la gestion des données de la recherche.

Le rôle des bibliothécaires a été souligné dans le rapport « Large-scale data sharing in the life sciences » [50], ainsi que dans le rapport de l'INRA sur la gestion et le partage des données de la recherche : « les documentalistes disposent d'un socle de compétences et d'outils acquis dans la gestion des publications qui est transposable à la gestion des données, mais il est important de souligner qu'ils n'ont généralement pas une connaissance intime de la nature des recherches, et des jeux de données associés. Ils ne sont donc pas en mesure de définir les métadonnées correspondantes [51]. »

R Dans l'article [52] les auteurs proposent une architecture de gestion de données à grande échelle, basée sur les rôles (data role) pour permettre le partage et la curation des données scientifiques sur le climat et la science de la terre.

1.6 Hébergement des données et liens à la cybersécurité

Il est quasi immédiat de relier les problématiques d'hébergement avec les problématiques de sécurité. On peut également parler de cybersécurité qui est définie par l'ITU [53] comme suit :

47. Le rôle des Bibliothèques dans la conservation des données, l'accès et la préservation : conclusions d'une enquête, IFLA, 2012
48. De l'open data à l'open research data : quelle(s) politiques(s) pour les données de la recherche, R. Gaillard, 2014
49. Upskilling Liaison Librarians for Research Data Management, A.Cox et al., 2012
50. Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models, Lord et al., 2005
51. Rapport du groupe de travail sur la gestion et le partage des données, INRA, 2012
52. Publication and Curation of Shared Scientific Climate and Earth Sciences Data, M. Humphrey et al. 2009
53. Définitions et termes relatifs à l'instauration de la confiance et de la sécurité dans l'utilisation des technologies de l'information et de la communication, UIT

Definition 18 On entend par cybersécurité l'ensemble des outils, politiques, concepts de sécurité, mécanismes de sécurité, lignes directrices, méthodes de gestion des risques, actions, formations, bonnes pratiques, garanties et technologies qui peuvent être utilisés pour protéger le cyberenvironnement et les actifs des organisations et des utilisateurs. Les actifs des organisations et des utilisateurs comprennent les dispositifs informatiques connectés, le personnel, l'infrastructure, les applications, les services, les systèmes de télécommunication, et la totalité des informations transmises et/ou stockées dans le cyberenvironnement.

La cybersécurité cherche à garantir que les propriétés de sécurité des actifs des organisations et des utilisateurs sont assurées et maintenues par rapport aux risques affectant le cyberenvironnement. Les objectifs généraux en matière de sécurité sont les suivants :

- Disponibilité : le système doit fonctionner sans faille durant les plages d'utilisation prévues et garantir l'accès aux services et ressources installées avec le temps de réponse attendu.
- Intégrité, qui peut englober l'authenticité et la non-répudiation : les données doivent être celles que l'on attend, et ne doivent pas être altérées de façon fortuite, illicite ou malveillante.
- Confidentialité : seules les personnes autorisées ont accès aux informations qui leur sont destinées. Tout accès indésirable doit être empêché.

Definition 19 La cyberstratégie quant à elle englobe l'ensemble des pratiques civiles et militaires, publiques et privées, intérieures et extérieures visant à aménager et à utiliser le cyberespace, l'ensemble des cyber environnements, afin de répondre aux objectifs fixés par l'autorité politique pour assurer la prospérité et la sécurité des citoyens.

La cyberstratégie en France est par exemple étudiée dans le cadre de la chaire Castex [54]. La chaire Castex développe la recherche fondamentale et appliquée en géopolitique du cyberespace dans le but de nourrir la réflexion stratégique liée aux enjeux du cyberespace dans les domaines politique, économique, militaire et règlementaire. Elle a également vocation à être une plateforme de ressources et d'échanges où convergent les chercheurs, les acteurs publics et privés concernés qui souhaitent étudier, partager, comprendre et sensibiliser leur public aux enjeux du cyberespace.

Face à l'émergence de l'open data, du big data, des objets connectés, des solutions de cloud, face à la multiplication de plus en plus préoccupante des « cyber-incidents » et la prolifération des données personnelles et des données confidentielles, face aux algorithmes qui les sondent et les auscultent pour en tirer les meilleurs profits sans le consentement de nos concitoyens, pillant ainsi notre patrimoine scientifique, médical et technique, il est impératif de s'emparer des stratégies, méthodologies et outils afin d'implémenter une gouvernance, une organisation, une conformité aux standards et des cadres règlementaires nationaux, européens et internationaux. Le dernier rapport «Panorama de la cybercriminalité 2015» réalisé par le CLUSIF (Club de la Sécurité de l'Information Français) montre clairement les insuffisances de conception de dispositifs même récents dont la sécurité a été compromise.

Le retard n'est pas technique, même si les progrès restent à faire. Il est dans l'organisation, la structuration, la sensibilisation, la formation et l'accompagnement, nécessaires à tout l'univers technologique, à tout ensemble d'outils techniques. Si les textes dépeignant cet état de fait ne manquent pas, les méthodologies, démarches et cadres règlementaires permettant d'y palier et de mettre en œuvre une politique de sécurité à la hauteur des enjeux sont disponibles et attendent d'être mis en œuvre, d'être conjugués au terrain de l'enseignement, de la recherche, de l'administration, d'être articulés avec les différents métiers de nos établissements. Le dernier rapport « Menaces

54. Chaire Castex de Cyberstratégie

Informatiques et Pratiques de Sécurité en France - Édition 2016 » réalisé par le CLUSIF montre que dans leurs pratiques, les entreprises et structures publiques ont une approche encore trop techniciste de la sécurité du patrimoine informationnel. À ce propos nous pouvons citer également Bruce Schneier (cryptologue, spécialiste en sécurité informatique) qui affirme : « Si vous pensez que seule la technologie peut résoudre vos problèmes de sécurité, alors vous n'avez rien compris à la technologie, ni à vos problèmes. »

1.6.1 Rapport de la CNIL

Le rapport d'activité 2015 analyse « La protection des données personnelles au cœur de la cybersécurité » fait le constat suivant :

> L'année 2015 fut marquée par de nombreux changements dans l'écosystème du numérique et de la cybersécurité : le cloud computing, les objets connectés et le big data ont pris de l'ampleur ; le paysage légal a évolué avec la loi de programmation militaire et la loi relative au renseignement ; le nombre de cyberattaques a encore progressé ; les violations de données se sont multipliées (Uber, Anthem, Ashley Madison…) et les données concernées sont souvent comptées en dizaines de millions. Comment, dans ce contexte, instaurer la confiance des partenaires et des internautes pour accompagner l'innovation numérique ? Les efforts en matière de sécurité devront être non seulement poursuivis, mais aussi adaptés. Le besoin d'intégrer la cybersécurité et la protection de la vie privée se fait ressentir. Le respect de la vie privée est au cœur du développement du numérique. Il en est de même de la prise en compte de la sécurité des systèmes d'information, permettant d'assurer notamment la résilience des infrastructures. Les notions de sécurité et de protection de la vie privée sont aujourd'hui indissociables. De même qu'il n'est plus envisageable de développer un service sans prendre en compte la dimension sécurité, la confiance dans le monde du numérique passe notamment, comme l'a rappelé le Premier Ministre lors de la présentation de la stratégie nationale pour la sécurité du numérique, par la prise en compte et le respect des notions de vie privée et de protection des données à caractère personnel.

1.6.2 ANSSI - Espace de confiance numérique

L'État au travers de l'Agence Nationale de la Sécurité des Systèmes d'Information [55] (ANSSI) se doit de promouvoir un espace de confiance pour les services en ligne. Comme elle l'indique sur son site, l'ANSSI participe ainsi avec les ministères et les autres services du Premier ministre à la création d'un écosystème de confiance pérenne et lisible pour les acteurs économiques. L'ANSSI intervient au sein de cet espace à travers la réglementation, ses labels (certification des produits de sécurité et qualification des produits de sécurité et des prestataires de service de confiance) et le soutien qu'elle peut apporter aux acteurs économiques pour l'évaluation de leurs solutions.

En particulier, l'ANSSI participe à la mise en œuvre du

- référentiel général de sécurité ;
- règlement européen sur l'identification électronique et les services de confiance pour les transactions électroniques au sein du marché intérieur (eIDAS), applicable depuis le 1er juillet 2016.

55. Agence Nationale de la Sécurité des Systèmes d'Information

Le référentiel général de sécurité (RGS)

Le site Internet de l'ANSSI précise que le référentiel général de sécurité [56] (RGS) est le cadre règlementaire permettant d'instaurer la confiance dans les échanges au sein de l'administration et avec les citoyens. Il a pour objet le renforcement de la confiance des usagers dans les services électroniques mis à disposition par les autorités administratives. Il comprend les règles permettant aux autorités administratives de garantir aux citoyens et aux autres administrations un niveau de sécurité de leurs systèmes d'information adapté aux enjeux et risques liés à la cyber sécurité. Dans le cadre du développement des téléservices et des échanges électroniques entre l'administration et les usagers, les autorités administratives doivent garantir la sécurité de leurs systèmes d'information en charge de la mise en œuvre de ces services.

Champ d'application et destinataires

L'ANSSI insiste sur le fait que le Référentiel général de sécurité s'impose spécifiquement aux systèmes d'information mis en œuvre par les autorités administratives dans leurs relations entre elles et dans leurs relations avec les usagers. Indirectement, le RGS s'adresse à l'ensemble des prestataires de services qui assistent les autorités administratives dans la sécurisation des échanges électroniques qu'elles mettent en œuvre, ainsi qu'aux industriels dont l'activité est de proposer des produits de sécurité. De façon générale, pour tout autre organisme souhaitant organiser la gestion de la sécurisation de ses systèmes d'information et de ses échanges électroniques, le Référentiel général de sécurité se présente comme un guide de bonnes pratiques conformes à l'état de l'art.

Contenu

Enfin l'ANSSI stipule que le Référentiel général de sécurité propose d'une part une méthodologie orientée autour de la responsabilisation des autorités vis-à-vis de leurs systèmes d'information à travers la démarche d'homologation et d'autre part des règles et bonnes pratiques que doivent mettre en œuvre les administrations lorsqu'elles recourent à des prestations spécifiques : certification et horodatage électroniques, audit de sécurité.

Il intègre les principes et règles liés à :

- la description des étapes de la mise en conformité ;
- la cryptologie et à la protection des échanges électroniques ;
- la gestion des accusés d'enregistrement et des accusés de réception ;
- la qualification des produits de sécurité et des prestataires de services de confiance ;
- la validation des certificats par l'État.

Le règlement eIDAS vu plus haut (voir 1.6.2) s'applique aux échanges entre l'administration et le public (citoyens, entreprises) et ne s'applique pas aux systèmes fermés, sans impact direct sur les tiers. Le périmètre fonctionnel du règlement eIDAS n'est pas identique à celui du RGS. Le RGS couvre notamment la délivrance de certificats d'authentification de personnes ou de machines, et la délivrance de certificats de confidentialité, deux services non couverts par le règlement. Le règlement n'induit pas d'obligation pour les administrations de recourir à des moyens d'identification électronique notifiés ou à des services de confiance qualifiés au titre du règlement eIDAS.

1.6.3 Politique de sécurité pour l'État

Universités, laboratoires, grandes écoles, grands organismes de recherche et le ministère de l'Enseignement supérieur et Recherche (MESR) ont à mettre en conformité leur système d'information

56. Référentiel général de sécurité

au regard des différents cadres règlementaires comme le Référentiel général de sécurité, la protection du potentiel scientifique et technique (PPST), la politique de sécurité des systèmes d'information (PSSI) d'État et la loi « Informatique et Libertés ». La PSSI de l'État (PSSIE) est un processus d'amélioration continue de la sécurité du SI qui doit être pleinement opérationnel d'ici 2017. Tous les ans, l'État demande à ses différents ministères de produire un rapport d'avancement sur la mise en place de cette politique de sécurité.

R L'Agence Nationale de la Sécurité des Systèmes d'Information publie sur son site une page référençant les textes réglementaires décrivant les exigences en matière de sécurité en fonction du niveau de sensibilité de l'information à protéger.

1.6.4 Politique de sécurité pour la recherche et la santé

Les grands organismes de recherche CNRS et INSERM ont décliné leur propre version de PSSI, voisine et adaptée de celle de l'État. Pour le milieu des praticiens hospitaliers et leurs outils numériques toujours plus nombreux dans le SI, le ministère des Affaires sociales et de la Santé au travers de son Agence des Systèmes d'Information Partagés de Santé a élaboré, avec l'aide et l'expertise de l'Agence Nationale de la Sécurité des Systèmes d'Information (cf. ANSSI), une politique de sécurité adaptée qui tient compte des spécificités et des exigences en matière de sécurité que requièrent les données de santé des patients.

Le ministère des Affaires sociales et de la Santé au travers de son Agence des Systèmes d'Information Partagés de Santé propose sur son site un espace « Services Pro » qui référence :

- la Politique Générale de Sécurité des Systèmes d'Information de Santé (PGSSI-S) ;
- une rubrique « référentiels » ;
- les hébergeurs de données de santé agréés et la procédure d'agrément ;
- un espace pour les produits de certification ;
- une rubrique « repères juridiques ».

1.6.5 Guides d'externalisation de la CNIL et de l'ANSSI

La CNIL propose un guide des services du cloud computing [57] et un guide des 7 étapes clés pour garantir la confidentialité des données [58] dont il convient de surveiller les évolutions avec l'adoption du « Privacy Shield » et du Règlement Général Européen de Protection des Données dont la mise en œuvre pratique est en cours d'élaboration.

L'ANSSI a mis en ligne également un guide des bonnes pratiques de l'externalisation [59] qui « propose une démarche pour apprécier les risques et fixer les exigences qu'appelle votre contexte, afin de garantir la sécurité de votre système d'information et des données qu'il traite. Il s'appuie pour cela sur le plan d'assurance sécurité et fournit des clauses types permettant d'obtenir du prestataire des engagements contractuels. »

1.6.6 Stratégie nationale pour la sécurité du numérique

Comme indiqué sur le site de l'ANSSI, la stratégie nationale pour la sécurité du numérique [60], dévoilée le 16 octobre 2015 par Monsieur le Premier Ministre Manuel Valls, est destinée à accompagner la transition numérique de la société française. Elle a fait l'objet de travaux interministériels

57. Cloud computing : les conseils de la CNIL pour les entreprises qui utilisent ces nouveaux services, CNIL, 2012
58. Cloud computing : les spet étapes clés pour garantir la confidentialité des données, CNIL, 2013
59. Externalisation et sécurité des systèmes d'information : un guide pour maîtriser les risques, ANSSI, 2010
60. Stratégie nationale pour la sécurité du numérique, site de l'ANSSI, documents FR, EN, DE

coordonnés par l'ANSSI. Elle répond aux nouveaux enjeux nés des évolutions des usages numériques et des menaces qui y sont liées avec cinq objectifs :

- garantir la souveraineté nationale
- apporter une réponse forte contre les actes de cyber malveillance
- informer le grand public
- faire de la sécurité numérique un avantage concurrentiel pour les entreprises françaises
- renforcer la voix de la France à l'international.

Avec la Stratégie nationale pour la sécurité du numérique, l'État s'engage au bénéfice de la sécurité des systèmes d'information pour aller, par une réponse collective, vers la confiance numérique propice à la stabilité de l'État, au développement économique et à la protection des citoyens.

1.6.7 Rapport d'activité 2015 de l'ANSSI

L'année 2015 a été une année charnière pour l'ANSSI avec de nombreux temps forts ponctués d'incidents de cybersécurité graves, dont celui très spectaculaire de TV5 Monde. L'agence a vu son rôle et son expertise dans la sécurité du numérique confortés. Le rapport d'activité 2015 de l'agence décrit ses grandes actions et les axes stratégiques qui les ont guidés dont la mise en place de la *stratégie nationale pour la sécurité du numérique* (voir la sous-section 1.6.6).

1.6.8 Réunir tous les acteurs et les impliquer dans la sécurité des systèmes d'information

Recommandations du Haut Fonctionnaire de Défense et de Sécurité

La Politique de Sécurité des Systèmes d'Information de l'État (PSSIE) s'impose à tous les ministères et toutes les entités sous leurs responsabilités (MESR et universités). La PSSIE a son agenda et un plan d'actions de mise en conformité des systèmes d'information de l'État. Elle devra être totalement mise en œuvre en janvier 2017.

Cette mise en conformité a été évaluée dans l'ensemble des ministères. En avril 2015, cette évaluation a été demandée aux présidents d'universités (Autorité Qualifiée de la SSI) assistés de leur Responsable SSI à la demande du haut fonctionnaire de défense et de sécurité (HFDS) du MESR. Devenue annuelle, cette évaluation de la PSSIE dans nos établissements amène le HFDS du MESR à conclure que la sécurité des systèmes d'information demeure trop technique, trop du domaine exclusif de la DSI (Direction des systèmes d'Information). La prise en compte des contraintes techniques nécessaires à la SSI doit aussi s'accompagner d'un engagement afin d'associer les ressources humaines et financières sur toute la durée du projet, d'organiser et de suivre le fonctionnement du SI (Système d'Information) en associant en permanence tous les acteurs. La SSI doit devenir un axe majeur de la gouvernance et des métiers.

1.6.9 Rapport au Sénat - Sécurité numérique et risques

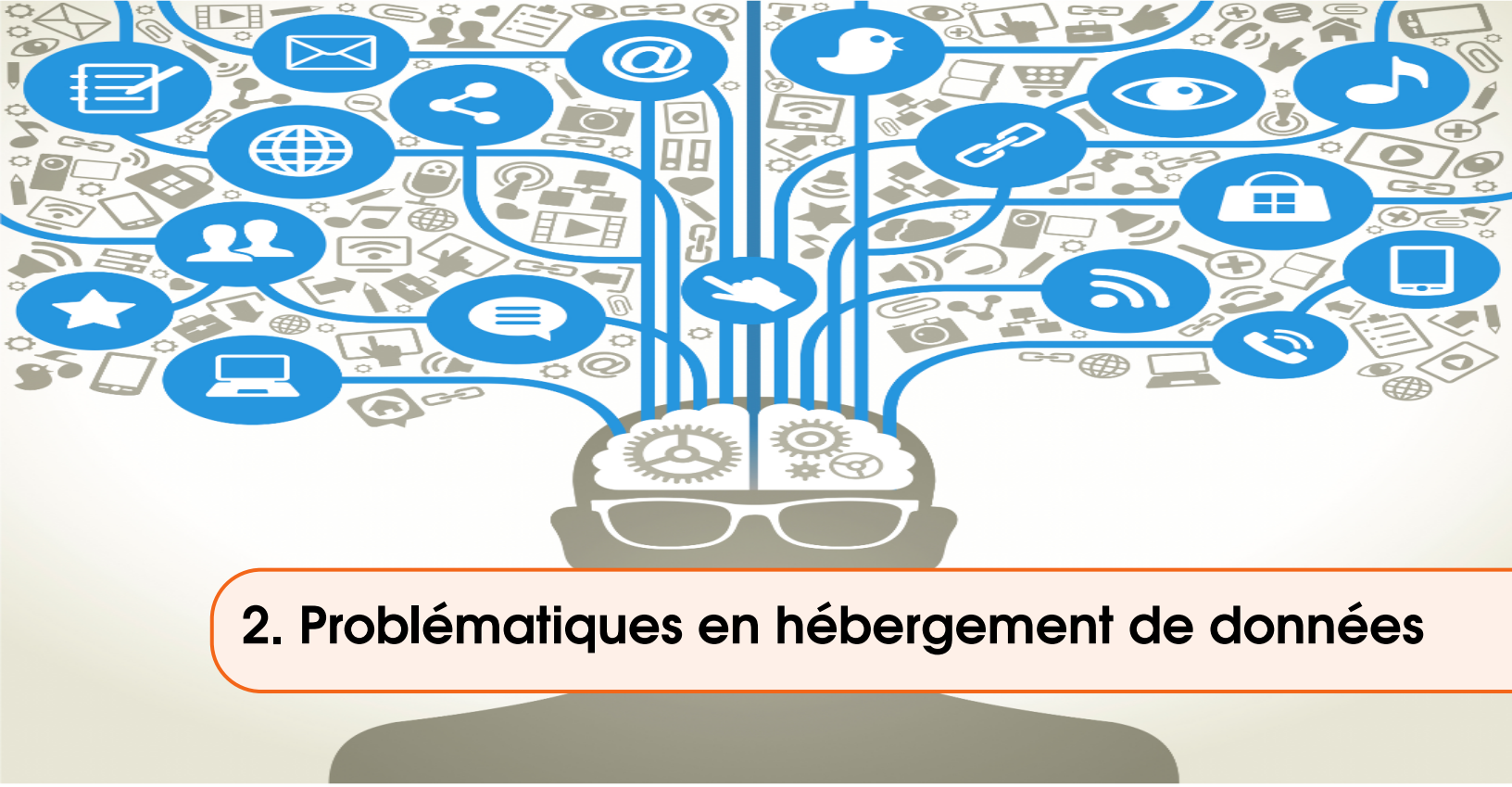
Les rapporteurs de ce travail étaient Mme Anne-Yvonne Le Dain, députée, et M. Bruno Sido, sénateur. Une phrase clé de ce rapport qui nous sert d'ultime citation est la suivante :

III. SE DONNER LES MOYENS DE LA SÉCURITÉ NUMÉRIQUE PAR UNE MEILLEURE COOPÉRATION ENTRE LES ACTEURS.

— Créer un lieu d'échange sur le numérique réunissant ingénieurs, politiques et administratifs pour développer une culture du numérique au sein de la sphère politique et administrative.

— Instituer une coopération entre les industriels, la communauté de défense et le monde académique pour élaborer et appliquer une stratégie nationale de cybersécurité à moyen et long

termes pour faire face aux attaques.
— Élargir les pouvoirs de l'ANSSI en lui donnant un pouvoir de régulation et d'injonction.
— Encourager, sur tout le territoire national, le développement d'acteurs de confiance spécialistes de la sécurité informatique (retombées économiques possibles).



2. Problématiques en hébergement de données

2.1 Hébergement et liens au cloud et aux centres de données

2.1.1 Introduction

Vous utilisez très certainement déjà des services de cloud hébergés dans des centres de données. À titre personnel votre messagerie est hébergée dans un centre de données ; vous utilisez aussi les services d'édition en ligne, vous gérez votre agenda en ligne et vous déposez vos photos… dans le cloud. Tous ces services sont les prémices d'une industrialisation des tâches du chercheur. Nous pouvons maintenant les examiner sous un angle plus large. À noter que la liste des grandes questions que nous abordons dans ce document est bien entendu non exhaustive.

La science des données (en anglais « data science ») est une nouvelle discipline qui s'appuie sur l'informatique et les mathématiques, en particulier les statistiques afin d'extraire de l'information des données. Le terme big data caractérise plutôt les données en tant qu'objet d'étude et moins les méthodes scientifiques d'extraction des connaissances. En science des données, les chercheurs s'appuient sur la fouille de données (en anglais « data mining »), les statistiques, le traitement du signal, l'apprentissage et la visualisation des données. Chacune de ces disciplines produit et échange des données. Les sites de production peuvent être géographiquement distants, ce qui implique une circulation et un hébergement des données entre les sites.

2.1.2 Vue fonctionnelle du cycle de vie des données

D'un point de vue architectural et fonctionnel, c'est-à-dire pour l'informaticien la vue haute de l'architecture informatique décrivant les grandes fonctions du système nécessaires aux personnes travaillant dans les e-Sciences, le modèle est par exemple celui du NIST présenté à la Figure 2.1. Nous défendons ici que ce modèle est le modèle qui va s'imposer et se généraliser auprès de toute personne travaillant en e-Sciences (à savoir toutes les sciences qui utilisent le numérique comme le cloud et les big data).

Cette figure présente un modèle du cycle de vie des données et donc de la circulation et l'hébergement des données dans les e-Sciences. Comme nous l'avons introduit tout au début du

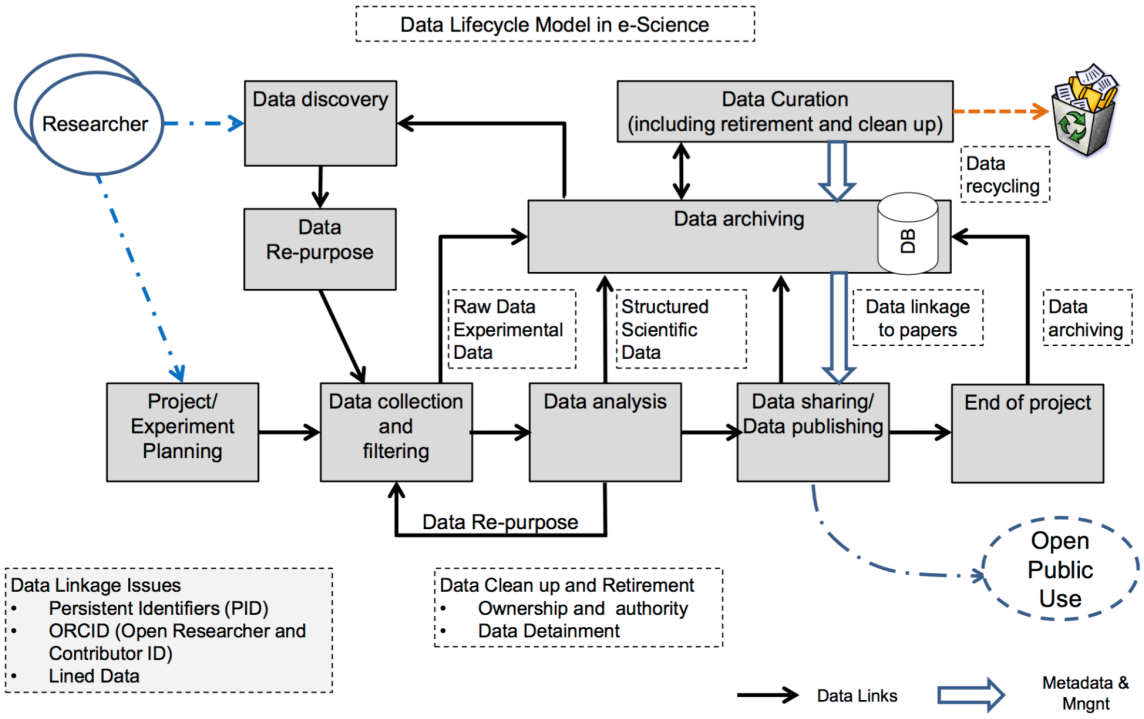


FIGURE 2.1 – Vue fonctionnelle du cycle de vie des données. D'après le NIST.

document, le cycle de vie d'une donnée explicite où les données naissent, où elles sont transférées, hébergées et où elles meurent. Sur la Figure 2.1 nous remarquons distinctement le cycle qui débute à la boîte Data discovery et s'achève à la boîte Data archiving ou Data recycling.

Il est important de noter qu'à chaque boîte correspond un ensemble d'outils informatiques et que ces outils définissent la vue technique, concrète de l'ensemble. Avec le cloud computing, il est possible à tout scientifique de composer sa vue technique en y intégrer les outils qu'il souhaite, à la demande, sans intervention d'un administrateur système.

Cette possibilité offerte par le cloud computing complexifie de fait toutes les problématiques de gestion et d'hébergement de données tout en offrant une grande souplesse d'utilisation. Par exemple faut-il traiter la sécurité au niveau des boîtes simples ou au niveau d'un groupe de boîtes ? Concernant la législation, laquelle fait référence ? Il faut en effet imaginer que le service rendu par une boîte puisse être localisé sur un site géographiquement différent du service rendu par une autre boîte.

L'objectif de l'informaticien en « Système » est de permettre que les scientifiques travaillant en e-Sciences puissent insérer leur discipline dans la vue fonctionnelle de la Figure 2.1. Il doit donc favoriser l'émergence de grands patrons de conception, c'est-à-dire d'arrangements et de manières d'organiser des services de gestion de données, configurables pour une discipline.

La vue d'esprit de la Figure 2.1, une fois implémentée dans un système réel, permet à un individu de déployer ses outils favoris à la demande. Au niveau des équipes scientifiques, cette même vue est un cadre cohérent permettant de réfléchir aux interactions entre équipes et aux flux d'information.

Dans le cadre de grands projets comme le LHC [61], cette vue est déjà implémentée mais pas à travers des systèmes de clouds et de data centers. Ces architectures pourraient pourtant servir à unifier les systèmes actuels, à mutualiser les ressources et pour lesquels on pourrait implémenter des politiques d'élasticité (le fait d'approvisionner le système avec des ressources externes lors de pics d'activité). Dans le cadre de petits projets, la vue de la Figure 2.1 permet aussi à un seul individu de maîtriser l'ensemble des processus métiers (analyse, archivage, curation...) à partir du moment où cet individu reconnaît les apports de cette approche de gestion des données.

Pour contrôler les échanges de données entre les boites de la Figure 2.1, l'informaticien développe également des modèles de programmation pour la gestion des cycles de vie des données. Pour l'informaticien Gilles Fedak [62], un système parfait de gestion du cycle de vie des données devrait :

- capturer les étapes et les propriétés essentielles du cycle de vie : création, destruction, fautes, réplication, vérification d'erreurs...
- permettre aux systèmes existants d'exposer leur cycle de vie intrinsèque des données ;
- raisonner sur des jeux de données répartis sur des ensembles d'infrastructures et des systèmes hétérogènes et
- simplifier la programmation d'applications « Data Life Cycle ».

Dans ce travail, les questions de provenance des données sont centrales afin d'estimer la qualité du jeu de données et de garder une trace des conditions d'acquisition et de transformation des données.

R À titre informatif, le site du UK Data Archive [63] est un autre point d'entrée sur la question du cycle de vie. Ce site permet également de déposer ses propres données. Enfin comme lecture complémentaire nous vous conseillons le numéro spécial en date de novembre 2015 de Computing Edge de IEEE Computer Society [64]. Ce numéro évoque les big data dans différents champs disciplinaires, le médical compris.

2.1.3 Exemples de démarches scientifiques impactées par le cloud et les big data

Après avoir examiné comment les nouveaux paradigmes du cloud et des big data influent sur le quotidien du chercheur, regardons comment le chercheur impacte ces univers à travers quelques exemples. Concernant les démarches scientifiques, les considérations architecturales présentées dans la partie précédente appellent à plus de créativité de la part du chercheur. Celui-ci est maintenant en capacité d'élaborer de nouvelles approches et méthodes scientifiques.

Le premier exemple concerne la gestion des hypothèses. Une fois énoncée, une hypothèse peut être discutée par exemple dans le cadre d'une démarche expérimentale. Dans l'article de Gonçalves et Porto [65], les hypothèses sont vues comme des données, au même titre qu'une donnée de base de données et ceci afin d'obtenir un cadre unifié entre hypothèses et données observées.

Rappelons d'abord le rôle des hypothèses dans le cadre d'une démarche scientifique expérimentale. Bien entendu l'étude des pratiques des chercheurs révèle une si grande diversité de démarches

61. Projet Le Grand collisionneur de hadrons (LHC)
62. ActiveData, un modèle de programmation pour la gestion des cycles de vie des données, Simonet et al., 2014
63. Research Data Lifecycle, UK Data Archives
64. big data, Computing Edge, IEEE Computer Society, 2015
65. Managing Scientific Hypotheses as Data with Support for Predictive Analytics, B. Gonçalves and F. Porto, 2015

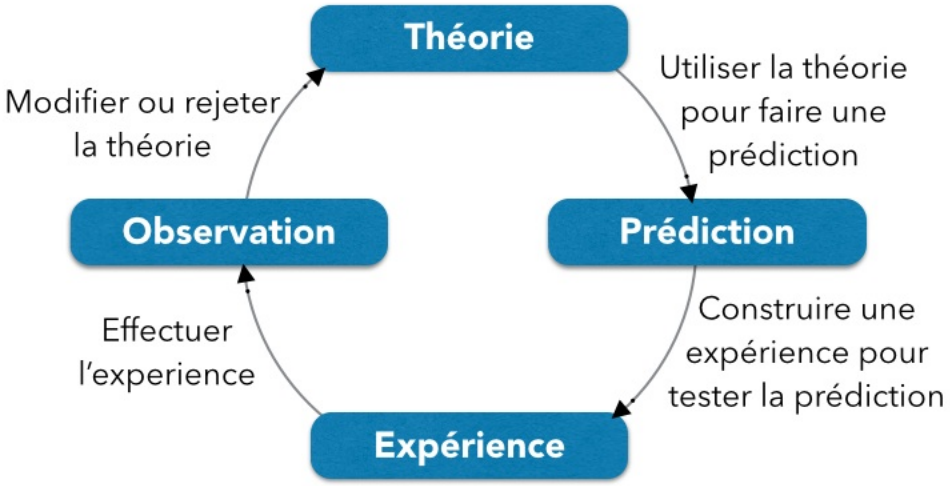


FIGURE 2.2 – Méthode scientifique expérimentale

et de disciplines scientifiques que l'idée d'une définition universelle des hypothèses est fausse. La démarche scientifique est rappelée à la Figure 2.2 et la formulation d'hypothèses apparaît lorsqu'il s'agit de faire une prédiction. Ici nous supposons qu'il y a une théorie à partir de laquelle nous formulons des hypothèses.

Dans l'article de Gonçalves et Porto, il s'agit d'intégrer, dans un même cadre de réflexion, les données (observées) et les théories (simulées). Pour cela les auteurs supposent que l'on est capable d'extraire des hypothèses une simulation computationnelle (en première approximation des équations que l'on injecte dans un simulateur). Les résultats qui sortent de la simulation sont testés et confrontés aux données observées issues de l'expérimentation. Si une validation intervient alors le chercheur peut publier de nouveaux résultats, sinon il lui faut retourner à l'étape de formulation des hypothèses.

Un des points clés du travail tient dans la gestion des données de simulation qui par essence sont incertaines contrairement aux données brutes telles qu'on les a en physique des particules ou en astronomie et telles qu'on les traite sur les grands clusters de calcul. Le caractère incertain peut venir de deux sources : non complétude (données manquantes) et multiplicité (données inconsistantes). Un autre point clé est la façon dont on considère une unité de données élémentaire. En gestion de données simulée les chercheurs s'intéressent plus au contenu prédictif d'une donnée qu'à sa simple dimension. Les technologies et outils clés sont alors les bases de données probabilistes [66, 67] et la statistique Bayésienne [68].

Enfin l'avantage de la méthode de Gonçalves et Porto tient aussi dans le fait que le nombre de données issues des hypothèses est d'un ordre de grandeur plus petit qu'avec la méthode plus

66. Bases de données probabilistes
67. Probabilistic Database, D. Suciu et al. 2011
68. Introduction to Bayesian Statistics, 2nd Edition, W. M. Bolstad, 2007

traditionnelle centrée sur une dimension du problème à partir des données brutes. En conclusion de ce premier exemple nous pouvons dire qu'ici la stratégie a été de réduire le volume des données traitées dans le processus expérimental.

Le deuxième exemple qui montre l'impact du cloud computing sur l'hébergement des données concerne la maîtrise énergétique des centres de données. Différentes techniques ont été mises au point ces dernières années pour consolider les serveurs du cloud. Il s'agit de regrouper des applications sur certaines machines et d'en éteindre d'autres. Des solutions exactes ou approchées à ce problème existent (elles relèvent de la discipline informatique de l'Optimisation Combinatoire) et elles vous disent quelles machines éteindre, à quel moment et où déplacer les applications qui migrent. La solution considère que tous les serveurs sont égaux par ailleurs. En fait il s'agit d'une solution globale qui vous garantit que tout le centre de données consommera le moins possible.

Imaginons la situation suivante. Chez un hébergeur, dans un centre de données, les serveurs appartiennent à des clients différents de sorte que le centre de données abrite une collection de mini centres de données, potentiellement gérés par des technologies de type cloud. Il y a isolation physique et logicielle entre les mini centres de données. En fait, le fait de minimiser l'énergie consommée par chacun des mini centres de données ne constitue pas, la plupart du temps, la solution optimale du problème d'optimisation énergétique lorsque l'on considère tous les serveurs d'un coup.

La question de savoir si un centre de données (constitué de mini centres de données) est énergiquement efficace est donc discutable et discutée. Ce qui est sûr, c'est que si l'on autorise les données à migrer sur n'importe quel serveur alors la théorie de l'Optimisation Combinatoire expliquera comment atteindre l'énergie minimale pour exécuter toutes les applications hébergées.

Peut-être même que le cloud et le centre de données le plus énergétiquement efficace serait le cloud constitué de vos téléphones mobiles, de vos tablettes et des autres dispositifs connectés à la maison qui sont connus pour consommer beaucoup moins qu'un ordinateur PC ou un portable. Ainsi il pourrait y avoir mutualisation des services selon une nouvelle modalité. Votre tablette rendrait les services que vous connaissez mais aussi abriterait des services et des données pour d'autres personnes. Des chercheurs appellent ce type de cloud, un *cloud volontaire* [69].

Des travaux récents, pour utiliser des dispositifs légers en lieu et place de grandes infrastructures, ont été conduits et touchent au déploiement de moteurs de recherche en texte intégral [70], au déploiement sur des téléphones mobiles de bases de données transactionnelles [71] et aux problématiques de l'infrastructure d'archivage sur Raspberry Pi [72].

2.2 Hébergement et liens à l'éthique et au juridique

2.2.1 La question de l'hébergement de données personnelles

L'externalisation du centre de données consiste à sous-traiter des ressources informatiques à un prestataire externe. Les données à caractère personnel doivent bénéficier de certaines garanties lorsque leur hébergement est externalisé.

69. Energy-aware service provisioning in volunteers clouds, Y. Ngoko et al. 2015
70. FindAll: A Local Search Engine for Mobile Phones, A. Balasubramanian et al. 2012
71. One DBMS for all: the Brawny Few and the Wimpy Crouwd, T. Mühlbauer et al. 2014
72. Scaling Down Distributed Infrastructure on Wimpy Machines for Personal Web Archiving, J. Lin, 2015

Le responsable de traitement

Selon l'article 3-I de la loi « Informatique et Libertés », le responsable de traitement est « sauf désignation expresse par les dispositions législatives ou réglementaires relatives à ce traitement, la personne, l'autorité publique, le service ou l'organisme qui détermine ses finalités et ses moyens ». Le responsable de traitement est tenu de respecter les exigences prévues à l'article 34 Loi « Informatique et Libertés », en prenant « toutes précautions utiles, au regard de la nature des données et des risques présentés par le traitement, pour préserver la sécurité des données et, notamment, empêcher qu'elles soient déformées, endommagées, ou que des tiers non-autorisés y aient accès ».

Dans la pratique, il s'agit de la personne en charge notamment de :

- veiller au bon respect des principes de la protection des données à caractère personnel ;
- informer les personnes de l'existence de leurs droits d'accès, de rectification et d'opposition ;
- procéder à l'accomplissement des formalités avec le CIL (Correspondant Informatique et Libertés) si l'organisme en dispose d'un ou à défaut auprès de la CNIL.

Le sous-traitant

Selon l'article 35 de la loi « Informatique et Libertés » , le sous-traitant est la personne traitant des données à caractère personnel pour le compte du responsable de traitement. La loi prévoit que celui-ci :

- n'agit que sur instruction du responsable de traitement et
- présente des garanties suffisantes pour assurer la mise en œuvre des exigences de sécurité et de confidentialité qui incombent au responsable de traitement prévues à l'article 34 de la loi « Informatique et Libertés » (sans que cela ne décharge ce dernier de veiller au respect de ces mesures).

Obligation d'une convention

L'article 35 de la loi « Informatique et Libertés » prévoit qu'un contrat lie le sous-traitant au responsable du traitement. Il doit comporter l'indication des obligations incombant au sous-traitant en matière de protection de la sécurité et de la confidentialité des données et prévoir que le sous-traitant n'agit que sur instruction du responsable du traitement.

La CNIL préconise de prévoir certains éléments essentiels dans le cadre des contrats de sous-traitance [73] :

- une clause spécifique couvrant la confidentialité des données personnelles [74]
- des dispositions (audits de sécurité…) afin de s'assurer de l'effectivité des garanties en matière de protection des données et notamment :
  - le chiffrement des données selon leur sensibilité ou à défaut l'existence de procédures garantissant que la société de prestation n'a pas accès aux données qui lui sont confiées ;
  - le chiffrement de la liaison de données (connexion de type https par exemple) ;
  - des garanties en matière de protection du réseau, traçabilité (journaux, audits), gestion des habilitations, authentification, etc.

73. La sécurité des données personnelles, CNIL, 2010
74. Un modèle de clause est disponible sur le site de la CNIL

- les conditions de restitution des données et de leur destruction au terme de l'hébergement.

R Il existe des exigences particulières en cas de transfert de données personnelles depuis le territoire européen vers un territoire où les dispositions de la directive 95/46/CE ne s'appliquent pas, c'est à dire hors Union européenne et Espace économique européen.

Le projet de règlement européen

La Directive européenne 95/46/CE de 1995, texte de référence en matière de protection des données à caractère personnel au niveau européen, a été récemment remplacée par un règlement général sur la protection des données. Le projet a été déposé par la Commission européenne le 25 janvier 2012 et a été approuvé par le Parlement européen le 12 mars 2014. Il a été définitivement adopté le 14 avril 2016 par le Parlement européen et le Conseil de l'Union européenne et publié le 27 avril de la même année.

Le projet de règlement vise à moderniser le cadre législatif devenu obsolète, à clarifier l'application de certains principes, à éliminer l'actuelle fragmentation des législations des États membres dans le domaine. La réforme prend la forme d'un règlement, permettant ainsi au texte une applicabilité directe et obligatoire dans tous ses éléments dans tous les états membres de l'Union européenne.

R Ce texte de référence est disponible sur le site du Journal officiel de l'Union européenne en plusieurs langues [75].

Le texte prévoit la suppression des formalités administratives préalables pour une logique de responsabilité des responsables de traitement via un système d'analyse de risques en interne. Ce nouveau système rendrait obligatoire la tenue d'une documentation régulièrement mise à jour sur l'ensemble des mesures prises pour la conformité des traitements de données personnelles, par le responsable de traitement et par le sous-traitant.

Le responsable du traitement ou, le cas échéant, le sous-traitant devra réaliser une analyse du risque en ce qui concerne les répercussions potentielles du traitement des données en question sur les droits et les libertés des personnes concernées, tout en évaluant si les traitements sont susceptibles de présenter des risques spécifiques.

En effet, le projet de règlement européen portant sur la protection des données personnelles prévoit également une version étendue de la notion de responsable de traitement qui est désormais définie comme :

Definition 20 « la personne physique ou morale, l'autorité publique, le service ou tout autre organisme qui, seul ou conjointement avec d'autres, détermine les finalités, les conditions et les moyens du traitement de données à caractère personnel ».

75. Journal officiel de l'Union européenne, L 119, 4 mai 2016

Cette nouvelle définition vise à instaurer un système de co-responsabilité entre tous les acteurs ayant décidé, de manière autonome et concertée, de la création d'un traitement de données à caractère personnel, notamment les sous-traitants.

Ce système de co-responsabilité implique que « lorsqu'un responsable du traitement définit, conjointement avec d'autres, les finalités, conditions et moyens du traitement de données à caractère personnel, les responsables conjoints du traitement définissent, par voie d'accord, leurs obligations respectives afin de se conformer aux exigences du présent règlement, en ce qui concerne notamment les procédures et mécanismes régissant l'exercice des droits de la personne concernée ».

De plus, dans le cas où un sous-traitant traiterait des données d'une autre manière que celle prévue dans les instructions du responsable du traitement alors il pourra également être considéré comme responsable conjoint du traitement.

2.2.2 L'hébergement de données de santé

L'activité d'hébergement de données de santé est encadrée par la loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé [76] , qui modifie l'article L. 1111-8 du code de la santé publique. Cet article prévoit l'organisation du dépôt, de la conservation et de la restitution des données de santé à caractère personnel dans des conditions de nature à garantir leur confidentialité et leur sécurité.

À des fins d'hébergement des données de santé, cet article insiste entre autres sur l'information et la non-opposition de la personne concernée, le mode d'établissement d'accès et de transmissions des données, les garanties de sécurité des données, le choix des systèmes d'hébergement et définit le cadre d'utilisation des données.

> Toute personne qui héberge des données de santé à caractère personnel recueillies à l'occasion d'activités de prévention, de diagnostic, de soins ou de suivi social et médico-social, pour le compte de personnes physiques ou morales à l'origine de la production ou du recueil desdites données ou pour le compte du patient lui-même, doit être agréée à cet effet. Cet hébergement, quel qu'en soit le support, papier ou électronique, est réalisé après que la personne prise en charge en a été dûment informée et sauf opposition pour un motif légitime. Les traitements de données de santé à caractère personnel que nécessite l'hébergement prévu au premier alinéa, quel qu'en soit le support, papier ou informatique, doivent être réalisés dans le respect des dispositions de la loi n° 78-17 du 6 janvier 1978 [77] relative à l'informatique, aux fichiers et aux libertés. La prestation d'hébergement, quel qu'en soit le support, fait l'objet d'un contrat.
>
> Les conditions d'agrément des hébergeurs des données, quel qu'en soit le support, sont fixées par décret en Conseil d'État pris après avis de la Commission nationale de l'informatique et des libertés et des conseils de l'ordre des professions de santé. Ce décret mentionne les informations qui doivent être fournies à l'appui de la demande d'agrément, notamment les modèles de contrats prévus au deuxième alinéa et les dispositions prises pour garantir la sécurité des données traitées en application de l'article 34 de la loi

76. Loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé
77. Loi n° 78-17 du 6 janvier 1978

n° 78-17 du 6 janvier 1978 précitée [78], en particulier les mécanismes de contrôle et de sécurité dans le domaine informatique ainsi que les procédures de contrôle interne. Les dispositions de l'article L.4113-6 [79] s'appliquent aux contrats prévus à l'alinéa précédent.

L'agrément peut être retiré, dans les conditions prévues par l'article 24 de la loi n° 2000-321 du 12 avril 2000 [80] relative aux droits des citoyens dans leurs relations avec les administrations, en cas de violation des prescriptions législatives ou réglementaires relatives à cette activité ou des prescriptions fixées par l'agrément.

Seules peuvent accéder aux données ayant fait l'objet d'un hébergement les personnes physiques ou morales à l'origine de la production de soins ou de leur recueil et qui sont désignées par les personnes concernées. L'accès aux données ayant fait l'objet d'un hébergement s'effectue selon les modalités fixées dans le contrat, dans le respect des articles L. 1110-4 [81] et L. 1111-7 [82].

Les hébergeurs tiennent les données de santé à caractère personnel qui ont été déposées auprès d'eux à la disposition de ceux qui les leur ont confiées. Ils ne peuvent les utiliser à d'autres fins. Ils ne peuvent les transmettre à d'autres personnes que celles qui les leur ont confiées.

Lorsqu'il est mis fin à l'hébergement, l'hébergeur restitue les données aux personnes qui les lui ont confiées, sans en garder de copie.

Les hébergeurs de données de santé à caractère personnel et les personnes placées sous leur autorité qui ont accès aux données déposées sont astreintes au secret professionnel dans les conditions et sous les peines prévues à l'article 226-13 du code pénal [83].

Les hébergeurs de données de santé à caractère personnel ou qui proposent cette prestation d'hébergement sont soumis, dans les conditions prévues aux articles L. 1421-2 [84] et L. 1421-3 [85], au contrôle de l'Inspection générale des affaires sociales et des agents mentionnés aux articles L. 1421-1 [86] et L. 1435-7 [87]. Les agents chargés du contrôle peuvent être assistés par des experts désignés par le ministre chargé de la santé.

Tout acte de cession à titre onéreux de données de santé identifiantes, directement ou indirectement, y compris avec l'accord de la personne concernée, est interdit sous peine des sanctions prévues à l'article 226-21 du code pénal [88].»

Il est à noter que ces traitements de données de santé à caractère personnel que nécessite l'hébergement doivent être réalisés dans le respect des dispositions de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. De même, la loi prévoit une harmonisation des dispositions relatives à l'hébergement de données de santé avec celles relatives aux archives

78. Article 34 de la loi n° 78-17 du 6 janvier 1978
79. Article L.4113-6 du CSP
80. Article 24 de la loi n° 2000-321 du 12 avril 2000
81. Article L. 1110-4 du CSP
82. Article L. 1111-7 du CSP
83. Article 226-13 du code pénal
84. Article L. 1421-2 du CSP
85. Article L. 1421-3 du CSP
86. Article L. 1421-1 du CSP
87. Article L. 1435-7 du CSP
88. Article 226-21 du code pénal

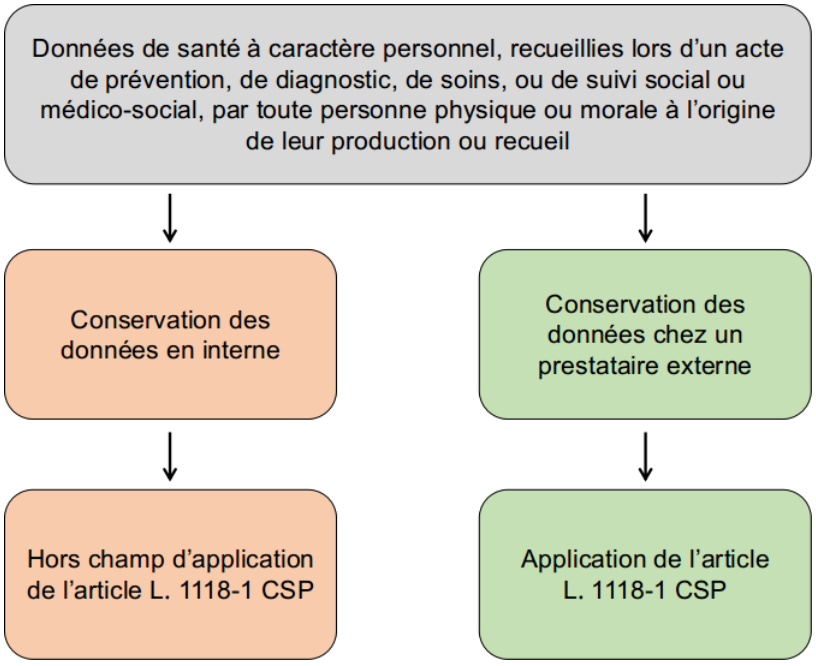


FIGURE 2.3 – Données personnelles concernées par l'application de l'article L. 1111-8 CSP

publiques (article L. 212-4 du code du patrimoine), qui devront également respecter l'article L1118-1 du CSP.

Données concernées

Les données concernées par l'application de l'article L. 1111-8 du CSP sont (Figure 2.3) :

- les données de santé à caractère personnel
- recueillies lors d'un acte de prévention, de diagnostic, de soins ou de suivi social ou médico-social
- par toute personne physique ou morale à l'origine de leur production ou recueil
- et hébergées chez un prestataire externe.

Information et non-opposition de la personne concernée

L'article L. 1111-8 du CSP dispense le déposant de données de santé auprès d'un hébergeur de données de santé de recueillir le consentement exprès de la personne concernée, sous réserve de la délivrance d'une information préalable à celle-ci. Cette dernière pourra toujours s'y opposer si elle le souhaite.

2.2.3 Hébergement de données de santé dans le cadre de la recherche

L'identification des frontières du champ d'application de l'agrément n'est pas toujours évidente et constitue un sujet de débats.

A priori, l'article L. 1111-8 du CSP prévoit un champ d'application délimité, couvrant les cas où des données de santé recueillies ou produites à l'occasion des activités de prévention, de diagnostic, de soins ou de suivi social ou médico-social sont déposées chez un tiers par les personnes physiques ou morales à l'origine de leur recueil ou production. Néanmoins, tant l'ASIP Santé que le Comité

d'Agrément des Hébergeurs (CAH) ont démontré une tendance à l'interprétation large de ce champ d'application.

Pour l'ASIP Santé, la question de l'application des dispositions de l'article L. 1111-8 du CSP doit se poser dès lors que les données concernées sont recueillies ou produites à l'occasion des activités pré-citées (bases de prévention, de diagnostic ou de soins que ce soit pour des bases de données de recherche, d'assurance, bases médico-sociales, ...). D'ailleurs le Comité d'agrément a été saisi de dossiers de candidatures présentés par des sociétés spécialisées dans la conduite de recherches biomédicales pour le compte d'établissements de soins ou de laboratoires pharmaceutiques. Cela pose la question de l'application de l'article L. 1111-8 à de telles bases de données.

Cependant aucune recommandation formelle ne précise si la recherche est incluse dans le champ des activités concernées par l'agrément, même s'il semble que les termes de « prévention, diagnostic, de soins ou de suivi social ou médico-social » ne doivent pas être interprétés de manière restrictive.

2.2.4 Procédure d'agrément hébergement de données de santé

La loi de modernisation de notre système de santé [89] (art 204-I-5°-c) prévoit de remplacer la procédure d'agrément des hébergeurs de données de santé par une procédure de certification, avec une évaluation de conformité technique réalisée par un organisme certificateur accrédité par le Comité Français d'Accréditation (COFRAC). L'idée est de simplifier la procédure d'agrément actuellement en place et d'améliorer les délais. La nouvelle procédure sera définie par voie d'ordonnance par le Gouvernement, et précisée par un décret. Tant que cette nouvelle procédure n'est pas en vigueur, les hébergeurs sont tenus de respecter la procédure d'agrément actuelle prévue par le décret n° 2006-6 du 4 janvier 2006 [90] relatif à l'hébergement de données de santé à caractère personnel et modifiant le code de la santé publique, qui continue de s'appliquer.

L'instruction d'une demande d'agrément se fonde sur le dépôt d'un dossier conforme au référentiel de constitution des dossiers.

Référentiel

Le référentiel de constitution de dossier se compose de six formulaires standards à renseigner (P1 à P6) et deux formulaires d'engagement (C1 et C2) à signer par le candidat :

- Formulaire de présentation détaillée du candidat
- Formulaire de présentation détaillée d'un sous-traitant
- Formulaire de description des clauses d'un modèle de contrat
- Formulaire de présentation du service d'hébergement
- Formulaire de présentation des résultats de l'analyse des risques
- Formulaire de description des dispositions de sécurité
- Formulaire d'engagement à la fourniture d'un rapport d'auto-évaluation annuel
- Formulaire de prise de connaissance des dispositions de contrôle

Le médecin de l'hébergeur

Une des exigences du décret n° 2006-6 du 4 janvier 2006 dans son article R. 1111-9-6 est la présence d'un médecin dans l'organisation candidate à l'agrément. Le médecin de l'hébergeur doit

89. Loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé
90. Décret n° 2006-6 du 4 janvier 2006 relatif à l'hébergement de données de santé à caractère personnel

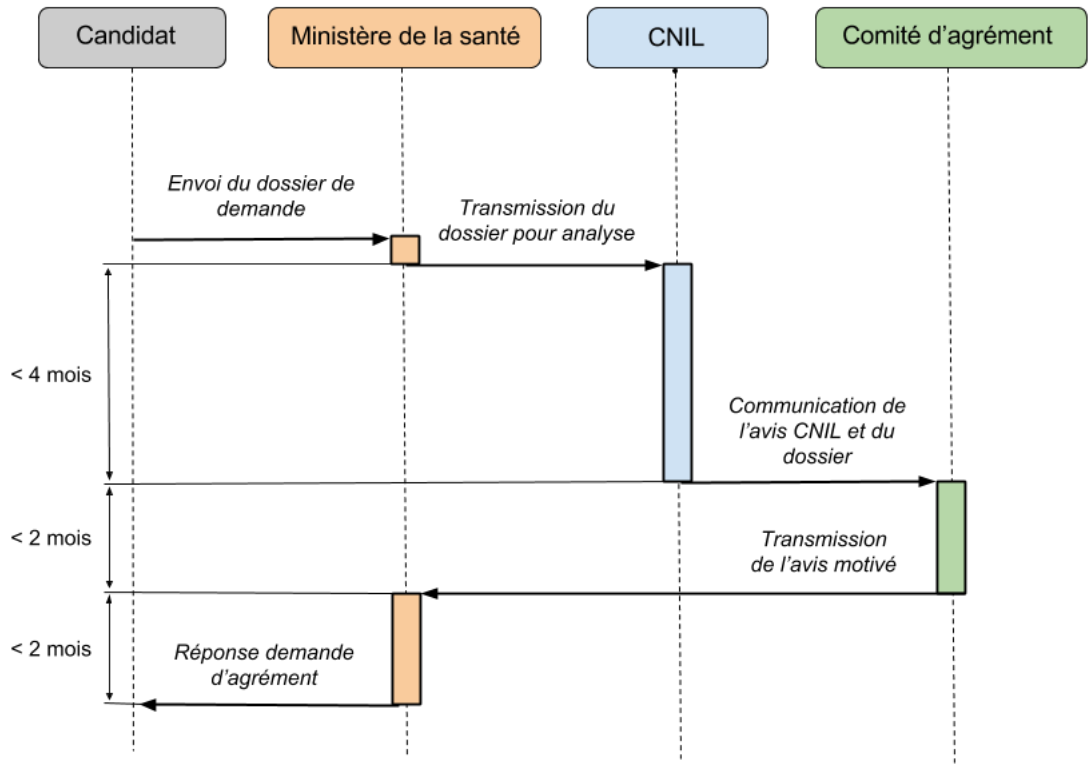


FIGURE 2.4 – Procédure actuelle d'agrément pour hébergeurs données de santé. Tiré de : Le rôle de l'Agence des systèmes d'information partagés de santé dans la procédure

être inscrit à l'Ordre des Médecins. Il est contractuellement lié à l'hébergeur, mais pas forcément salarié.

Ses missions ne sont pas expressément prévues par la loi, mais il peut par exemple veiller à la confidentialité des données, au respect des droits des personnes, aux conditions d'accès des données, recevoir les demandes liées aux données…

Déroulement de la procédure

Étape 1 : réception du dossier de demande d'agrément par le secrétariat du comité d'agrément assuré par l'ASIP Santé

- Accusé de réception
- Transmission du dossier à la CNIL

Étape 2 : instruction du dossier par l'ASIP Santé et la CNIL

- Instruction par les chargés d'analyse de l'ASIP Santé sous trois angles :
  - — volet éthique et juridique : l'examen de la demande suivant des considérations de garanties d'ordre éthique et déontologique en relation avec la pratique et les finalités médicales de l'hébergement de données de santé à caractère personnel et le respect des droits du patient ;
  - — volet sécurité et technique : examen des résultats de l'analyse du dossier sur les garanties apportées en terme de politique de sécurité des systèmes d'information et de confidentialité des données de santé, en considérant les aspects techniques mais également organisationnels ;
  - — volet économique et financier, exprimant une analyse de la demande sur des considérations en relation avec le modèle économique et la structure financière du candidat. Les rapports d'instruction sont présentés lors du comité d'instruction interne.

  Les rapports d'instruction sont présentés lors du comité d'instruction interne et validés par le responsable du comité. Le dossier est ensuite présenté à un des membres du Comité d'agrément « rapporteur » du dossier.
- Instruction par la CNIL en parallèle (délai de deux mois renouvelable une fois), et transmission de l'avis au comité d'agrément.

Étape 3 : avis du comité d'agrément des hébergeurs

- Réunion du comité d'agrément dans un délai d'un mois suivant la réception de l'avis de la CNIL (délai renouvelable une fois), qui se prononce sur tous les aspects du dossier, en particulier sur les garanties d'ordre éthique, déontologique, technique, financier et économique qu'offre le candidat, et rend un avis.
- Transmission de l'avis au ministre en charge de la santé.

Étape 4 : décision du ministre en charge de la santé

- Décision dans un délai de 2 mois suivant la réception de l'avis du Comité d'agrément pour prendre sa décision. À l'issue de ce délai, le silence vaut décision de rejet.
- L'agrément est délivré pour une durée de trois ans, et les demandes de renouvellement de l'agrément doivent être déposées au plus tard six mois avant le terme de la période d'agrément.

2.3 Préservation des données

2.3.1 Introduction

Definition 21 Pour les bibliothécaires et les archivistes, la conservation numérique est un processus formel pour assurer que l'information numérique authentifiée reste accessible et utilisable. Ce processus comprend la planification et l'allocation des ressources digitales via l'application de méthodes et de technologies de conservation.

Dans les domaines scientifiques, la préservation des données a été souvent négligé dans le passé. En effet, au milieu du XX^{e} siècle, l'entrée dans l'ère digitale n'a pas fondamentalement changé au début la démarche scientifique. Les communautés scientifiques ont profité pleinement de ce nouvel outil pour avancer vers des expériences plus précises, rendant les lots de données obsolètes assez rapidement. Une certaine manière de prévoir l'analyse des données, avec la focalisation sur le résultat recherché, ainsi que l'évolution rapide de la technologie, qui n'a fait que s'intensifier, n'ont pas encouragé une réflexion approfondie sur la conservation numérique dans le domaine de la recherche. Avec l'exception notable de l'astrophysique, les communautés scientifiques n'ont pas pris en compte la préservation et la mise à disposition des données digitales dans la planification des expériences.

La donne change au début des années 2000 pour plusieurs raisons : la capacité à collecter des données augmente de manière spectaculaire. En même temps, dans plusieurs disciplines, la complexité des dispositifs expérimentaux a également évolué rapidement, pour donner naissance à des flux de données qui excèdent la capacité d'analyse immédiate : on collecte plus de données que prévu, avec un potentiel scientifique à plus long terme. De plus, les scientifiques ont en commun des moyens pour accéder à des expérimentations plus vastes (« big science ») comme le « Large Hadron Collider » (LHC), ce qui rend ces expérimentations difficilement reproductibles, et par conséquent uniques. Ceci est encore plus vrai pour des données digitales contenant des informations temporelles, comme par exemple les données d'observation de la terre ou des écosystèmes, ou encore le suivi de la position des objets stellaires. La perte de lots de données serait donc irréparable.

La physique des hautes énergies a été précurseur dans la collecte et le calcul massif des données digitales, ainsi que pour la mise en place de projets complexes, avec une durée de vie au delà de dix ans. Malgré ce contexte, les problèmes de préservation des données ont été posés relativement tard, à l'aube du démarrage du LHC. L'exemple de l'astrophysique, organisé en observatoire virtuel depuis plusieurs décennies, devient une source d'inspiration ainsi qu'une invitation à une démarche plus large, interdisciplinaire.

Dans la communauté de la physique des hautes énergies il est d'usage de fonctionner comme cela est décrit à la Figure 2.5 où l'on observe le chevauchement d'un même cycle (préparation - prise de données - analyse) entre différents projets d'une même expérience. La coordination des phases est donc vitale.

Comme le souligne le journal du CNRS en 2016, c'est dans ce contexte qu'est né en 2012 le projet interdisciplinaire PREDON, composante du Groupe de Recherche MADICS. Sous l'impulsion de Cristinel Diaconu, directeur de recherche au CNRS, le diagnostic est le suivant : « L'explosion du volume de données issues des expériences menées au CERN nous a conduits à mener une réflexion autour de leur préservation. Nous avons dans un premier temps structuré notre communauté autour d'une organisation internationale nommée Data Preservation and Long Term Analysis in High Energy Physics. Par la suite, nous nous sommes aperçus que nombre de disciplines étaient confrontées à la même préoccupation. C'est de ce constat que nous est venue l'idée de former une communauté interdisciplinaire autour de la question de la préservation des données scientifiques. »

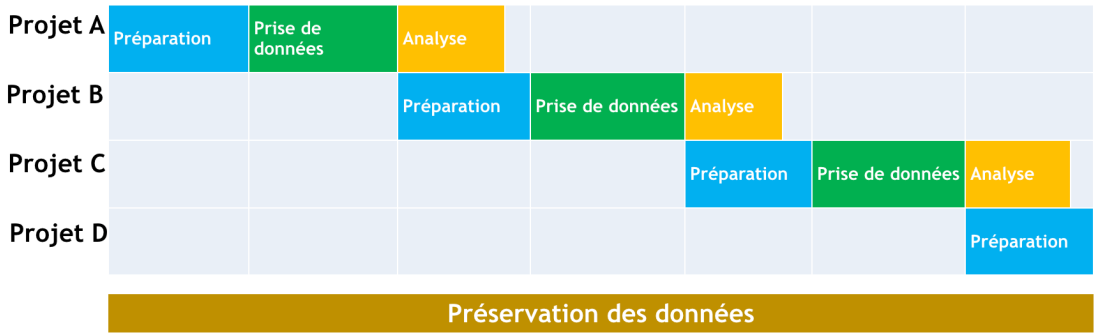


FIGURE 2.5 – Les cycles d'une expérience de physique des hautes énergies

Au sein du forum, les participants échangent donc autour de questions telles que « Comment conserver les données à long terme ? », « Comment garantir qu'on saura les lire dans dix ou vingt ans ? » ou encore « Comment permettre à la prochaine génération de chercheurs de comprendre les données archivées ? » afin d'adapter les stratégies de préservation à leur propre secteur. Ces questions deviennent d'autant plus importantes que la présentation d'un Data Management Plan (le plan de gestion des données que nous avons introduit plus haut dans ce document) est d'ores et déjà demandée dans le cadre d'un projet pilote du programme européen pour la recherche et l'innovation Horizon 2020 et qu'elle a vocation à se généraliser.

Sur le plan national, outre la communauté de la physique des hautes énergies et de l'astronomie, les sciences humaines et sociales se sont dotées d'une très grande infrastructure de recherche (TGIR) appelée Huma-Num, pour gérer la diffusion et la préservation de leurs données numériques. Ce service est proposé en partenariat avec le Centre Informatique National de l'Enseignement Supérieur (CINES), qui fournit les outils et l'expertise nécessaires à l'archivage. La TGIR Huma-Num développe un dispositif technologique qui permet d'accompagner les différentes étapes du cycle de vie des données numériques. Ainsi, elle met à disposition un ensemble de services pour le stockage, le traitement, l'exposition, le signalement, la diffusion et la conservation sur le long terme des données numériques de la recherche en sciences humaines et sociales.

Nous allons maintenant discuter, à un niveau général, du contour, du dimensionnement d'un service de préservation en précisant des éléments de vocabulaire.

2.3.2 Facteurs poussant vers plus de préservation

Pour identifier les facteurs poussant vers plus de préservation nous avons besoin de définir quelques notions.

Definition 22 Les cyber-infrastructures peuvent être définies comme l'ensemble coordonné de technologies de l'information et de systèmes (y compris les experts et les organisations) permettant le travail, les loisirs, la recherche, l'éducation à l'ère de l'information digitale. Elles intègrent des fonctions avancées d'acquisition (data acquisition), de stockage (data storage), de gestion (data management), d'intégration (data integration), de fouille (data mining), de visualisation (data visualization) des données, ou d'autres services de traitement informatique ou informationnel.

Le développement de cyber-infrastructures de données est grandement affecté à la fois par les

cas d'utilisation actuels et projetés, et par notre besoin de rechercher, analyser, modéliser, fouiller, et visualiser des données numériques. Plus largement, le monde des cyber-infrastructures de données est influencé par les tendances de la technologie, l'économie, la politique et le droit. Quatre tendances significatives reflètent l'environnement dans lequel les cyber-infrastructures de données évoluent :

- *Volumétrie des données.*
  Il y a plus de données numériques créées que de volume de stockage pour les accueillir. Certains articles [91] notent qu'en 2007, au point de croisement, le volume de données a été estimé à environ 264 exaoctets (264×10^{18} octets). C'est près d'un million de fois la quantité de données numériques hébergées en 2008 par la Bibliothèque du Congrès américain considérée comme la bibliothèque la plus grande du monde.
- *Politiques et règlements.*
  De plus en plus de politiques et de règlements nationaux exigent l'accès, la gestion et la conservation des données numériques. Aux États-Unis, la loi Sarbanes-Oxley de 2002 promeut une gestion responsable et appropriée à la préservation. Une loi de 1996 précise la responsabilité des organismes publics pour conserver de manière digitale les informations financières et autres. La Health Insurance Portability assure la confidentialité des dossiers médicaux numériques. À cette époque déjà, sur le front de la recherche, certains personnels du National Institutes of Health sont tenus de soumettre des copies numériques de leurs publications au service PubMed Central [92].
- *Coûts de stockage de plus en plus bas.*
  Des disques SSD de plusieurs teraoctets sont maintenant disponibles, par exemple chez le constructeur SCANDISK [93]. Western Digital a annoncé le 20 septembre 2016 un prototype [94] de disque de capacité 1 To en technologie SDXC, un format pour le grand public proposé en 2009.
- *Commercialisation des services de stockage et de données numériques.*
  L'introduction de Amazon Simple Storage Solutions en 2006 est un exemple parmi d'autres. Ce service que le grand public peut utiliser en ligne est plus connu maintenant sous le nom de Amazon S3. À ce jour il y a même un accès gratuit à S3 qui inclut 5 Go de stockage, 20000 demandes GET (récupérer des données) et 2000 demandes PUT (déposer des données). Un intérêt du service S3 est qu'il peut être couplé à beaucoup de services Amazon, par exemple au service EC2 qui permet de calculer sur les données. Là aussi, il y a un mode gratuit qui inclut 750 heures par mois d'utilisation d'instances t2.micro Linux et Windows durant un an. Pour rester dans le cadre du niveau gratuit, vous devez uniquement utiliser des instances EC2 Micro [95].
- *Virtualisation des logiciels et des services.*
  Un facteur très important dans la préservation à long terme des données est la conservation des capacités de lecture et décodage, et donc la préservation des logiciels de lecture. En effet, les données scientifiques répondent à des impératifs de vitesse et efficacité qui vont souvent au déla du simple fichier sur disque, et nécessitent du traitement par des logiciels complexes ainsi que la gestion d'ensembles de données annexes (métadonnées) nécessaires au décodage et à l'exploitation de ces données. La démocratisation de ces moyens logiciels (notamment via

91. Got Data? a Guide to Data Preservation in the information age
92. PubMed Central
93. SCANDISK
94. Premier disque SDXC de 1To, Western Digital
95. Types d'instances Amazon EC2

des machines virtuelles) facilite la préservation des écosystèmes de calcul ad-hoc, typiques des expériences scientifiques.

2.3.3 Dimensionnement

Une question clé est de savoir qui est chargé de faciliter la préservation des données numériques de valeur. Il est notable que le terme « valeur » signifie différentes choses pour différentes personnes. Il est aussi notable que faciliter la préservation des données implique l'hébergement de plusieurs copies des mêmes données, la migration des données d'une génération de supports de stockage à une autre pour assurer la pérennité, et enfin la protection de son intégrité et son authenticité.

Pour construire une cyber-infrastructure de données il est nécessaire de distinguer les différents usages des données ainsi que les différents scénario de préservation. La Figure 2.6 décrit un modèle pour tout le spectre de la préservation. La flèche à gauche de la pyramide dénote des propriétés sur les collections de données. La flèche à droite de la pyramide dénote des propriétés des infrastructures.

Au sommet de la pyramide se trouvent les données de valeur pour la société en général et dont les opérateurs sont principalement des institutions d'intérêt public (tels que les organismes gouvernementaux, les bibliothèques, les musées et les universités).

Au milieu de la pyramide se trouvent les données de la valeur propre à une communauté spécifique. On y trouve par exemple les enregistrements numériques à partir de votre hôpital local, les données de recherche scientifique conservées dans des dépôts communautaires...

En bas de la pyramide se trouvent les données de tout un chacun : photos, documents texte... Il y a ici la nécessité de disposer de sites primaires, supplémentaires pour la sauvegarde individuelle et privée des collections d'un usager.

La création d'une pyramide de données économiquement viable doit également être complétée par de la recherche en continue entraînant le développement de solutions qui répondent aux défis techniques de gestion des données et de préservation. Cette démarche de recherche pourra alors conduire à la possibilité d'utiliser et de créer de nouvelles connaissances à partir des données stockées. Par exemple, le processus de fouille sur les données d'exploitation dépend de la façon dont elles sont organisées, sur le niveau d'information supplémentaire (métadonnées) associé avec ces données. Il est alors important, à ce moment de l'analyse, de travailler avec les communautés.

L'évaluation des archives à garder se réfère au processus d'identification des dossiers et autres documents à conserver par la détermination de leur valeur. Plusieurs facteurs sont généralement considérés lors de la prise de cette décision. C'est un processus difficile et critique parce que les enregistrements sélectionnés vont façonner la compréhension des chercheurs de cet organisme pour les dossiers, ou fonds.

L'évaluation des archives peut être effectuée une fois ou lors des différentes étapes d'acquisition et de traitement. Une évaluation au niveau macroscopique, c'est à dire une analyse fonctionnelle à un niveau élevé des enregistrements, peut être effectuée avant même que les dossiers ont été acquis afin de déterminer les enregistrements à acquérir. Une évaluation itérative peut aussi être effectuée alors que les enregistrements sont en cours de traitement.

À noter également l'organisation et le suivi des grandes masses de données, en particulier en ce qui concerne les données scientifiques. En effet, l'organisation sans faille de type « bibliothèque », pour indexer et sauvegarder de manière professionnelle les lots de données issues d'une expérimentation, reste un facteur essentiel pour la pérennité des données. Néanmoins, la connexion avec les communautés scientifiques ainsi que l'organisation de la mise à disposition, sont indispensables pour s'assurer de l'utilité ou non de sauvegarder sur le long terme des données de « mauvaise qualité ».

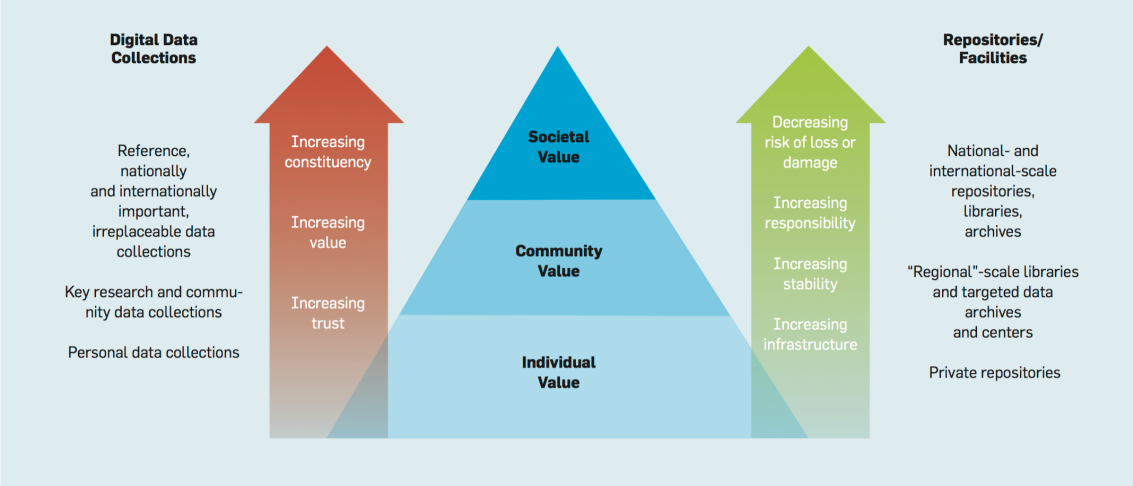


FIGURE 2.6 – Champ lexical de la préservation. D'après F. Berman, 2008. [97]

2.3.4 Initiatives de la communauté à l'échelle internationale

Pour normaliser la pratique de la conservation numérique et afin de fournir un ensemble de recommandations pour la mise en œuvre d'un programme de préservation, le modèle de référence Open Archival Information System [96] (OAIS) a été développé. OAIS concerne tous les aspects techniques du cycle de vie d'un objet numérique : ingestion, stockage d'archives, gestion des données, administration, accès et planification de la conservation. Le modèle aborde également les questions de métadonnées et recommande que cinq types de métadonnées soient attachés à un objet numérique : référence (identification) des informations, la provenance (y compris l'histoire de la conservation), le contexte, la fixité (indicateurs d'authenticité), et la représentation (mise en forme, structure de fichier).

International Research on Permanent Authentic Records in Electronic Systems [98] (InterPARES) est une initiative de recherche collaborative dirigée par l'Université de British-Columbia qui se concentre sur le traitement des questions de conservation à long terme des documents numériques authentiques. La recherche est menée par des groupes de discussion de diverses institutions en Amérique du Nord, en Europe, en Asie et en Australie, avec pour objectif de bâtir des théories et des méthodologies qui constitueront la base de stratégies, normes, politiques et procédures nécessaires pour assurer la fiabilité et l'exactitude des documents numériques au fil du temps.

2.3.5 Outils et méthodologies spécifiques

Digital Repository Audit Method Based On Risk Assessment [99] (DRAMBORA), présenté par le Centre curation numérique (DCC) et DigitalPreservationEurope (DPE) en 2007, propose une méthodologie et une boîte à outils pour l'évaluation numérique des risques référentiel. L'outil permet soit de procéder à l'évaluation en interne (auto-évaluation) soit d'externaliser le processus.

Le processus DRAMBORA est organisé en six étapes et se concentre sur la définition du mandat, l'identification des risques ainsi que l'évaluation de la probabilité et l'impact potentiel des risques.

97. Got data? A guide to data preservation in the information age, Communication of the ACM, 2008
96. Open Archival Information System
98. International Research on Permanent Authentic Records in Electronic Systems
99. Digital Repository Audit Method Based On Risk Assessment

L'auditeur est tenu de décrire et de documenter le rôle, les objectifs, les politiques, les activités de l'entité examinée, afin d'identifier et d'évaluer les risques associés à ces activités et de définir des mesures appropriées pour les gérer.

En 2002, le projet Preservation and Long-term Access through Networked Services (PLANETS), faisant partie du « EU Framework Programmes for Research and Technological Development », commença à adresser les défis de la préservation. Le but premier de PLANETS a été de construire des services et des outils pratiques pour aider à assurer la préservation sur le long terme de biens culturels et scientifiques. Le projet s'est terminé en 2010 et il s'est présenté sous la forme de l'Open Planets Foundation, puis à partir de 2014 sous la forme de l'Open Preservation Foundation [100] pour s'aligner sur les directions de la fondation.

En France, *le Centre Informatique National de l'Enseignement Supérieur* [101] (CINES) poursuit des actions de lobbying auprès des acteurs du marché du stockage afin que leurs formats de fichiers soient encore lisibles dans plusieurs années. Une équipe d'ingénieurs est engagée dans une course permanente contre l'obsolescence en veillant à ce que les logiciels ou lecteurs matériels puissent en permanence accéder aux données. La plateforme FACILE [102] recense la palette de formats actuellement pris en charge par le CINES.

Pour répondre à ce besoin d'échange et de préservation des données, la communauté astronomique internationale s'est structurée autour de l'Observatoire virtuel, un ensemble de services qui permet de retrouver l'information utile parmi toutes les données astronomiques ouvertes aux chercheurs grâce à un répertoire et à des standards partagés.

Le Centre de Données astronomiques de Strasbourg [103] (CDS) est un centre de données voué à la collecte et à la distribution dans le monde entier de données astronomiques. Il héberge la base de référence mondiale pour l'identification d'objets astronomiques et ses missions consistent :

- à rassembler des informations utiles concernant les objets astronomiques, sous forme informatisée ;
- à distribuer ces informations dans la communauté astronomique internationale ;
- à conduire des recherches utilisant ces données.

Le CDS a été créé en 1972 par l'Institut National d'Astronomie et de Géophysique (INAG), devenu depuis l'Institut National des Sciences de l'Univers (INSU), en accord avec l'Université Louis Pasteur, devenue depuis l'Université de Strasbourg.

Les services principaux du CDS sont Simbad, la base de données de référence pour l'identification et la bibliographie des objets astronomiques (hors système solaire), Vizier, qui collecte les catalogues astronomiques et les tables publiées dans les journaux académiques, et Aladin, un atlas interactif du ciel qui permet de visualiser des images astronomiques provenant des observatoires sol et spatiaux ou fournies par l'utilisateur, et des données provenant des services du CDS ou d'autres bases de données telles que celle de la NASA et de l'IPAC (NED).

En conclusion, les personnels du CDS standardisent les documents en formats pérennes pour leur conservation et qu'ils leur attribuent les métadonnées reconnues par la discipline. Ces métadonnées peuvent être présentes dans les documents reçus (fichiers Fits, Flexible Image Transport System) ou ajoutées par des documentalistes du CDS, montrant par la sorte la pluridisciplinarité du travail. Ces données sont ensuite dupliquées sur huit sites miroirs situés dans différents pays.

100. Open Preservation Foundation
101. Centre Informatique National de l'Enseignement Supérieur
102. plateforme FACILE
103. Centre de données astronomiques de Strasbourg

Ces exemples montrent le besoin d'une démarche structurée et permanente au sein de chaque projet et programme scientifiques, comme illustré à la Figure 2.5, mais surtout le besoin d'une structuration au sein des communautés et des organismes de recherche. En effet, la préservation des données scientifiques ne peut pas incomber aux seuls projets scientifiques car la structuration temporelle ne permet pas une vision, à plus long terme, sur le sort des données. Les succès évoqués plus haut s'appuient sur des structures spécialisées (centre de données dédié et spécialisé, observatoires virtuels), en connexion étroite avec les communautés scientifiques, suivant des politiques définies au niveau national et surtout international.

2.4 Éléments méthodologiques pour la gestion des risques

Dans le chapitre précédent nous avons introduit les principaux acteurs nationaux qui œuvrent à la sécurité au sens large. Nous allons maintenant donner des motivations et expliquer comment gérer les risques afin de protéger l'information. Nous allons surtout introduire une méthodologie. Celle-ci sera déclinée dans le prochain chapitre consacré aux études de cas. Ainsi nous nous limitons ici à une présentation générale en explicitant le vocabulaire lié à la gestion des risques.

2.4.1 Expression de besoin de sécurité pour protéger l'information

Il a longtemps été admis, enseigné et pratiqué qu'on ne pouvait pas développer un logiciel sans un cahier des charges qui décrit les fonctionnalités attendues du logiciel.

Definition 23 La maîtrise d'ouvrage (MOA) est un ensemble de personnes qui passe commande pour un outil avec des fonctionnalités métiers attendues (directeur de laboratoire et équipes de recherche, présidence de l'université et agence comptable, PDG d'une entreprise et commerciaux ou juristes ou architectes...).

Definition 24 La maîtrise d'œuvre (MOE) sont les personnes qui développent ou maintiennent en exploitation technique l'outil (développeur, ingénieur du data center,...).

Quand la maîtrise d'œuvre développe et/ou installe l'outil, elle doit respecter le cahier des charges de cet outil.

Definition 25 Le cahier des charges décrit les fonctionnalités attendues et c'est un engagement contractuel entre l'organisme (MOE) qui développe et/ou installe le nouvel outil et l'organisme (MOA) qui finance le développement et/ou l'installation de cet outil.

Nous verrons que la loi française impose aux services de l'État une analyse des risques sur l'information, les données à caractère personnel et l'emploi de technologies conformes à des standards de sécurité. Parmi les informations à sécuriser, les législations françaises et européennes imposent que les métiers (MOA) identifient les informations à caractère personnel qu'ils manipulent. La MOA est désignée par la loi « Informatique et Libertés » comme étant le *responsable de traitement* ou RT. De par la loi et la nature de plus en plus sophistiquée des données, celles à caractère personnel ne sont pas identifiables par les techniciens (MOE) mais bien uniquement par le métier (MOA), que l'on nomme le *responsable de traitement*. Ces fonctionnalités de sécurisation des données à caractère personnel attendues par les métiers forment un sous-ensemble distinct du cahier des charges de la sécurité de l'information et participent donc du cahier des charges de l'outil. On parle alors de « privacy by design ».

2.4.2 Conception sécurisée ou Security by design

La mise en œuvre de la sécurité doit passer par une prise en compte dès l'élaboration du projet et doit suivre tout le cycle de vie de la donnée. Cela passe par la mise en œuvre d'un dialogue entre les métiers et la direction des systèmes d'information, et par la compréhension mutuelle des enjeux associés à ces développements. Idéalement, cette réflexion sera menée le plus tôt possible, dès la conception des projets.

L'analyse de risque à la conception d'un logiciel ou d'un système d'information, le respect des bonnes pratiques lors du développement de nouveaux logiciels, lors de l'installation et l'exploitation d'un logiciel, d'un système et d'une infrastructure réseau sont les règles à suivre pour intégrer la sécurité à la conception que l'on appelle « Security by Design ».

2.4.3 Conception respectant la vie privée ou Privacy by Design

La création d'un nouveau télé service à destination des citoyens ou entre système d'information de services publics (ex : échanges entre des hôpitaux et un data center universitaire) doit respecter le Référentiel général de sécurité (RGS) de l'État. Cela implique la mise en œuvre de solutions techniques éprouvées suite à une analyse de risque et une analyse des besoins de sécurité spécifiques aux données à caractère personnel manipulées. La création de ce nouveau télé service donnera lieu à une démarche avec le Correspondant Informatique et Libertés de l'établissement ou du grand organisme de recherche (CNRS, INSERM,...).

2.4.4 Une démarche de protection

Le patrimoine informationnel est constitué de processus métiers et d'ensembles d'informations appelés *biens essentiels*. La production, la transformation, le stockage de l'information pour réaliser les processus métier sont réalisés par les matériels, logiciels informatiques et les réseaux de télécommunication. Ils sont appelés *biens supports*. Lorsque les métiers ont peur pour des informations, des biens essentiels, on parle *d'évènements redoutés* et nous devons les référencer avec eux pour les aider à exprimer leurs besoins de sécurité. Nous pouvons mentionner comme exemple des événements redoutés concernant la confidentialité des informations comme la diffusion aux concurrents du fichier client ou un défaut d'intégrité des informations comme des références de produits fausses.

Tous les équipements et logiciels informatiques, réseaux et de télécommunication sont susceptibles d'avoir des *vulnérabilités*, de par leur nature ou leur mauvaise constitution. À cette vulnérabilité correspondant des menaces et cette vulnérabilité peut être exploitée par une *source de menaces*. Un ordinateur portable est, par exemple, vulnérable à la menace de vol, un visiteur en salle de conférence peut le voler ou un serveur peut ne pas être corrigé est contenir toujours la faille HeartBleed sur son module Open SSL ce qui rend possible une intrusion depuis Internet et des robots d'analyse de faille qui tournent sur Internet pourraient l'exploiter.

R Tous ces éléments de vocabulaire sont regroupés dans un mémento de l'ANSSI distribué lors de ses sessions de formation (voir Figure 2.7). Des exemples de cas d'étude de risque mettent en pratique ce vocabulaire au point 2.4.10.

Lorsqu'une source de menaces exploite la vulnérabilité d'un bien support qui porte un bien essentiel alors se produit *l'événement redouté* par le métier. Le métier (MOA) détermine *l'impact* ou *gravité* que l'événement redouté a eu sur ses biens essentiels. Par exemple, un consultant MOA doit déterminer l'impact d'une perte de confidentialité suite au vol de l'ordinateur portable qui contenait les documents contractuels avec les clients pouvant être vendus à la concurrence. Le service

informatique (MOE) détermine quant à lui la vraisemblance que la vulnérabilité soit exploitée par une source de menaces. Par exemple, le serveur avec sa faille dans le module Open SSL est exposé en permanence sur Internet, la vraisemblance d'une intrusion est très forte.

Dans la méthode EBIOS que l'on présentera plus tard on appelle *risque*, un scénario, avec une vraisemblance et une gravité données, combinant un événement redouté et un ou plusieurs scénarii de menaces. En fonction de la vraisemblance et de l'impact, nous pouvons quantifier une probabilité d'occurrence et un coût financier, juridique, d'image. Le métier (MOA) peut alors envisager de négliger le risque (probabilité d'occurrence trop faible, comme un séisme dans une zone peu soumise à de l'activité sismique), d'investir dans des solutions d'externalisation de ce risque (tels que l'infogérance, assurance) ou de réduction de ce risque. La réduction du risque peut demander des moyens financiers, humains, organisationnelles et/ou de formation ou sensibilisation. Le fonctionnement de la méthode EBIOS est schématisé dans un mémento de l'ANSSI distribué lors de ses sessions de formation (voir Figure 2.8).

Lors de la mise en place d'un nouveau système d'information pour réaliser, faciliter, enrichir un ou des processus métiers, il convient d'établir le dialogue avec les différents métiers concernés, dans un langage clair et commun pour réaliser une étude selon une méthodologie reproductible. Le système d'information et ses besoins de sécurité évoluant régulièrement, nous serons amenés à réutiliser la méthode dans le temps. Donc les acteurs métiers et les personnels de la DSI doivent se l'approprier. L'étude de sécurité envisagée devra porter sur un périmètre clair afin de se limiter aux fonctionnalités attendues et de ne pas se perdre dans le système d'information existant. La méthode, quant à elle, doit pouvoir traiter un périmètre plus large et s'inscrire dans un processus d'amélioration continue.

2.4.5 Méthodologies

Pour l'expression des besoins de sécurité de l'‘information par les métiers, il existe différentes méthodes spécifiques qui s'appuient sur une famille de normes internationales, l'ISO 27000. Nous pouvons citer :

- MEHARI (développé par le CLUSIF) ;
- EBIOS (développé par le Club EBIOS et l'ANSSI)

Dans son document décrivant la méthode EBIOS, l'ANSSI indique que l'adoption de démarches et d'outils de prise de décision rationnelle et de gestion de la complexité apparaît aujourd'hui comme une condition nécessaire à la sécurité de l'information. Il convient pour cela d'utiliser des approches de gestion des risques structurées, éprouvées, tout en prenant garde aux illusions de scientificité et à la manipulation de chiffres, offertes par de nombreuses méthodes.

D'une manière générale, une approche méthodologique permet de :

- disposer d'éléments de langage communs ;
- disposer d'une démarche claire et structurée à respecter ;
- se baser sur un référentiel validé par l'expérience ;
- s'assurer d'une exhaustivité des actions à entreprendre ;
- réutiliser la même approche en amélioration continue et sur d'autres périmètres.

En matière de gestion des risques de sécurité de l'information, une approche méthodologique permet également :

- d'établir le contexte en prenant en compte le contexte interne et externe, les enjeux, les contraintes, les métriques... ;
- d'apprécier les risques (les identifier au travers des événements redoutés et des scénarios de

menaces, les estimer et les évaluer) ;
- de traiter les risques (choisir les options de traitement à l'aide d'objectifs de sécurité, déterminer des mesures de sécurité appropriées et les mettre en œuvre) ;
- de valider le traitement des risques, sur le plan formel, du plan de traitement des risques et les risques résiduels ; communiquer sur les risques (obtenir les informations nécessaires, présenter les résultats, obtenir des décisions et faire appliquer les mesures de sécurité) ;
- de suivre les risques (veiller à ce que les retours d'expériences et les évolutions du contexte soient prises en compte dans le cadre de gestion des risques, les risques appréciés et les mesures de sécurité).

2.4.6 Expression de besoin de sécurité de l'information : la méthode EBIOS

La méthode EBIOS [104] (Expression des Besoins et Identification des Objectifs de Sécurité) est un outil complet de gestion des risques SSI conforme au Référentiel général de sécurité (RGS) et aux dernières normes ISO 27001, 27005 et 31000. La méthode EBIOS bénéficie de 20 ans d'expérience dans le domaine de la gestion du risque. Elle permet d'apprécier et de traiter les risques ainsi que de communiquer à leur sujet au sein de l'organisme et vis-à-vis de ses partenaires, constituant ainsi un outil complet de gestion des risques SSI. Elle est assortie d'une base de connaissances cohérente avec le RGS, enrichie d'exemples concrets permettant d'élaborer des scénarios de risque pertinents pour votre organisme ou votre projet.

Pour une description complète de la méthode EBIOS, on peut se référer au site de l'ANSSI. La documentation d'EBIOS contient un guide « EBIOS 2010 : la méthodologie » qui décrit la méthode, ses objectifs, sa souplesse et sa nécessité pour gérer durablement de manière adaptative les risques et assurer une sécurité de l'information structurée dans les environnements des plus simples aux plus complexes.

Historiquement employée dans le domaine de la sécurité de l'information, elle a été exploitée dans d'autres domaines. EBIOS fait aujourd'hui figure de référence en France, dans les pays francophones et à l'international.

2.4.7 Éléments d'une démarche

L'ANSSI explique les intérêts de la démarche. Il est essentiel d'appréhender les éléments à prendre en compte dans la réflexion : le cadre mis en place pour gérer les risques, les critères à prendre en considération (comment estimer, évaluer et valider le traitement des risques), la description du périmètre de l'étude et de son environnement (contexte externe et interne, contraintes, recensement des biens et de leurs interactions. . .). La méthode EBIOS permet d'aborder tous ces points selon le degré de connaissance que l'on a du sujet étudié. Il sera ensuite possible de l'enrichir, de l'affiner et de l'améliorer à mesure que la connaissance du sujet s'améliore.

2.4.8 L'appréciation des risques

Selon le site de l'ANSSI, il y a risque de sécurité de l'information dès lors qu'on a conjointement une source de menace, une menace, une vulnérabilité et un impact.

On peut ainsi comprendre qu'il n'y a plus de risque si l'un de ces facteurs manque. Or, il est extrêmement difficile, voire dangereux, d'affirmer avec certitude qu'un des facteurs est absent.

104. La méthode EBIOS

Par ailleurs, chacun des facteurs peut contribuer à de nombreux risques différents, qui peuvent eux-mêmes s'enchaîner et se combiner en scénarios plus complexes, mais tout autant réalistes.

On va donc étudier chacun de ces facteurs, de la manière la plus large possible. On pourra alors mettre en évidence les facteurs importants, comprendre comment ils peuvent se combiner, estimer et évaluer (hiérarchiser) les risques. Le principal enjeu reste, par conséquent, de réussir à obtenir les informations nécessaires qui puissent être considérées comme fiables. C'est la raison pour laquelle il est extrêmement important de veiller à ce que ces informations soient obtenues de manière à limiter les biais et à ce que la démarche soit reproductible.

Pour ce faire, la méthode EBIOS se focalise tout d'abord sur les événements redoutés (sources de menaces, besoins de sécurité et impacts engendrés en cas de non-respect de ces besoins), puis sur les différents scénarios de menaces qui peuvent les provoquer (sources de menaces, menaces et vulnérabilités). Les risques peuvent alors être identifiés en combinant les événements redoutés et les scénarios de menaces, puis estimés et évalués afin d'obtenir une liste hiérarchisée selon leur importance.

2.4.9 Le traitement des risques

Sur son site, l'ANSSI décrit également ce que signifie le traitement des risques. Les risques appréciés permettent de prendre des décisions objectives en vue de les maintenir à un niveau acceptable, compte-tenu des spécificités du contexte.

Pour ce faire, EBIOS permet de choisir le traitement des risques appréciés au travers des objectifs de sécurité : il est ainsi possible, pour tout ou partie de chaque risque, de le réduire, de le transférer (partage des pertes), de l'éviter (se mettre en situation où le risque n'existe pas) ou de le prendre (sans rien faire). Des mesures de sécurité peuvent alors être proposées et négociées afin de satisfaire ces objectifs.

La manière dont les risques ont été gérés et les risques résiduels subsistants à l'issue du traitement doivent être validés, si possible formellement, par une autorité responsable du périmètre de l'étude. Cette validation, généralement appelé homologation de sécurité, se fait sur la base d'un dossier dont les éléments sont issus de l'étude réalisée.

Obtenir des informations pertinentes, présenter des résultats, faire prendre des décisions, valider les choix effectués, sensibiliser aux risques et aux mesures de sécurité à appliquer, correspondent à des activités de communication qui sont réalisées auparavant, pendant et après l'étude des risques.

Ce processus de communication et de concertation relatif aux risques est un facteur crucial de la réussite de la gestion des risques. Si celui-ci est bien mené, et ce, de manière adaptée à la culture de l'organisme, il contribue à l'implication, à la responsabilisation et à la sensibilisation des acteurs. Elle crée en outre une synergie autour de la sécurité de l'information, ce qui favorise grandement le développement d'une véritable culture de sécurité et du risque au sein de l'organisme.

L'implication des acteurs dans le processus de gestion des risques est nécessaire pour définir le contexte de manière appropriée, s'assurer de la bonne compréhension et prise en compte des intérêts des acteurs, rassembler différents domaines d'expertise pour identifier et analyser les risques, s'assurer de la bonne prise en compte des différents points de vue dans l'évaluation des risques, faciliter l'identification appropriée des risques, l'application et la prise en charge sécurisée d'un plan de traitement.

Pour conclure, les mémentos édités par l'ANSSI sont rappelés aux Figures 2.7 et 2.8.

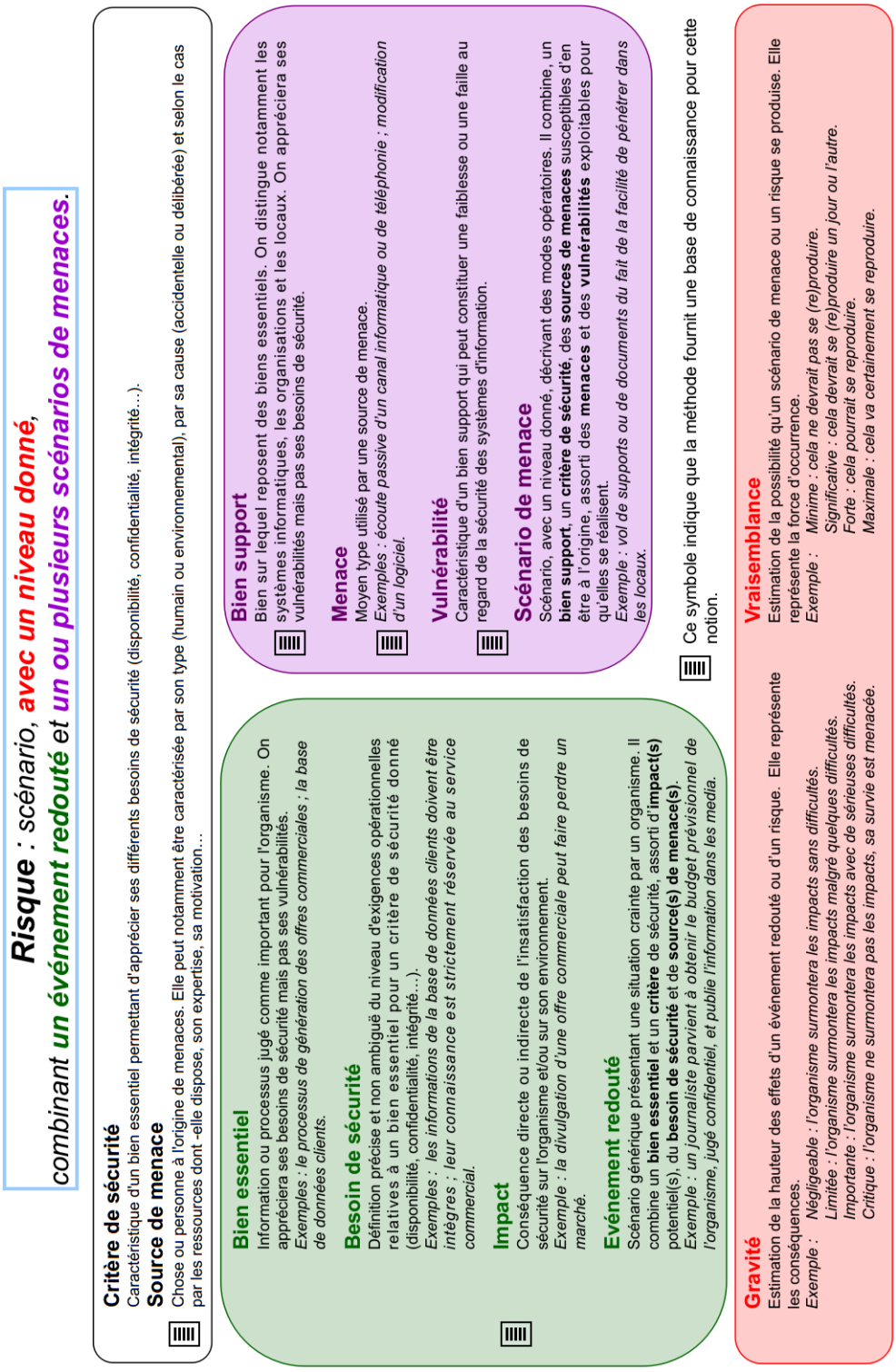


FIGURE 2.7 – Mémento de l'ANSSI rapellant les notions de la méthodes EBIOS La partie verte concerne la MOA (les métiers) et la partie violette la MOE (la DSI)

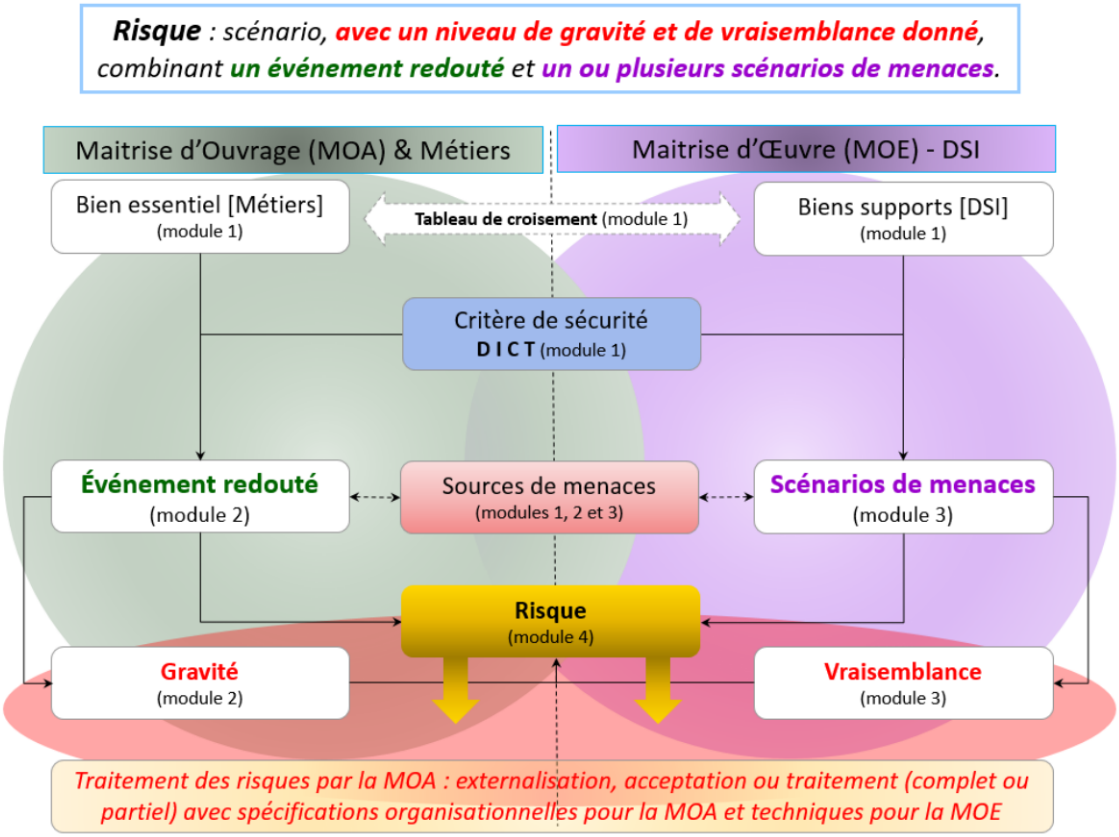


FIGURE 2.8 – Mémento de l'ANSSI qui schématise le fonctionnement de la méthode EBIOS original qui a ici été enrichi pour faire ressortir la MOA en vert et la MOE en violet. Il apparaît que le cheminement de la méthode sur la partie verte (MOA) se fait en partie en parallèle avec celui sur la partie violette (MOE) pour opérer les correspondances entre les « événements redoutés » par le métier et les « scénario de menaces » qui pèsent sur les éléments techniques du service informatique. Un risque est caractérisé par la vraisemblance que le scénario de menace se réalise et par la gravité pour le métier que l'événement redouté se produise par le biais de ce scénario. La MOA doit alors choisir de quelle façon le traiter.

2.4.10 Exemples d'étude de risques

Vous trouverez dans cette section l'illustration de l'utilisation du vocabulaire EBIOS par des exemples d'étude de risque vus en formation à l'ANSSI. Trois domaines applicatifs seront présentés : parti politique, opérateur télécom et projet de raid militaire dévoilé sur Facebook.

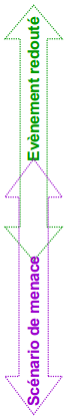
Définition : *Scénario, avec un niveau (de risque) donné, combinant **un événement redouté** et **un ou plusieurs scénarios de menaces**.*

Evènement redouté

Scénario de menace

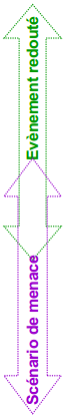
Bien essentiel	Contenu du site du parti politique
Critère de sécurité	Intégrité
Besoin de sécurité	Doit être totalement intègre
Impact	Image de marque du parti politique, échec aux élections
Source de menace	Les attaquants
Bien support	Serveur hébergeant le site du parti politique
Menace	Défiguration
Vulnérabilité	Facilité d'accès à l'administration du site, …??

FIGURE 2.9 – Risque SSI - Défiguration de plusieurs sites d'un parti politique (Vital Security du 01/03/2010). Le texte de la défiguration encourage les visiteurs des sites à voter pour le parti adverse. Les messages laissés par les attaquants comportent des critiques sur la sécurité des sites et des slogans à caractère politique.



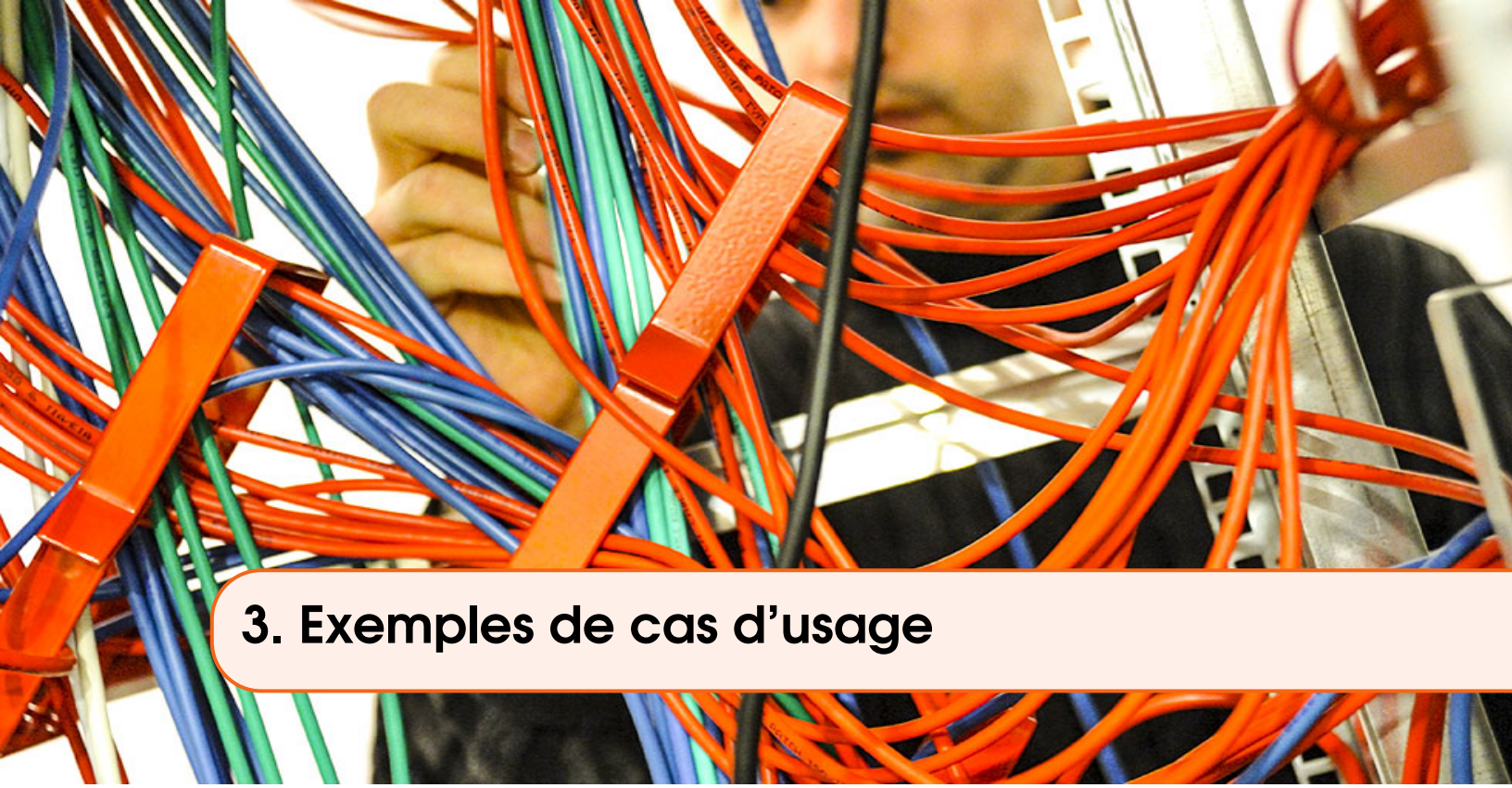
Bien essentiel	Informations personnelles des clients de l'opérateur
Critère de sécurité	Confidentialité
Besoin de sécurité	Privé / Confidentiel personnel
Impact	Image de marque de l'opérateur ? Plainte ?
Source de menace	Un attaquant informatique
Bien support	Base de données de l'opérateur
Menace	??
Vulnérabilité	??

FIGURE 2.10 – Risque SSI - Compromission d'une base de données d'un opérateur télécom (Softpedia du 09/02/2010). Un attaquant informatique a eu accès aux informations personnelles de près de 60 000 clients de l'opérateur dans le pays. L'individu a démontré qu'il pouvait obtenir librement les adresses électroniques, les numéros de téléphones portables et les mots de passe de connexion.



Bien essentiel	Informations sur un raid surprise
Critère de sécurité	Confidentialité
Besoin de sécurité	Secret Défense
Impact	Vies humaines, perte d'une bataille, annulation du raid
Source de menace	Soldat
Bien support	Soldat
Menace	Divulgation
Vulnérabilité	Maladroit, étourdi…

FIGURE 2.11 – Risque SSI - Divulgation d'un projet de raid sur Facebook (Gizmodo.fr du 05/03/2010, The New York Times du 05/03/2010). Un raid surprise a dû être annulé avant-hier à cause d'un soldat qui avait mis à jour son statut Facebook pour indiquer "mercredi nous nettoyons Qatanah, et jeudi, si dieu le veut, nous rentrons à la maison". Le soldat a depuis été relevé de son poste de combat.



3. Exemples de cas d'usage

3.1 Champs de la santé

3.1.1 Définition de données santé et précisions sur leurs spécificités

Les données santé regroupent toutes les données médicales (i.e. imagerie par résonance magnétique (IRM), scanner à rayon X (CT),...) et de santé (i.e. Système d'Information Hospitalier (SIH), Système d'Information Radiologique (SIR),...). Elles concernent une personne, un groupe ou une population de personnes. Ces données sont utiles pour la routine clinique, la recherche, le suivi de santé et, à grande échelle, pour améliorer la politique de santé.

À la vue des nouvelles méthodes d'acquisition et de génération de données, les volumes de données ne cessent de croître d'où la nécessité d'une nature d'hébergement qui respecte les spécificités des données santé. Cette partie est consacrée à expliciter les hébergements mis en œuvre.

3.1.2 Type de données

Nous pouvons différencier deux types de données de santé, les données brutes et les données traitées.

Les données brutes

Les données brutes sont recueillies ou acquises à partir d'une source de production. Par exemple, il peut s'agir des appareils d'acquisition d'images, la saisie d'informations sur le patient, un séquenceur ADN, etc.

Ces données sont généralement volumineuses et contiennent plusieurs informations.

Les données traitées

Les données traitées sont obtenues suite à une succession de processus appliqués sur les données brutes. Ces processus produisent des données qui peuvent être moins ou plus volumineuses que les données originales.

3.1.3 Provenance de données

En santé publique, la majorité des données sont récoltées et stockées à travers des systèmes d'information hospitaliers ou alors via d'autres systèmes d'information comme le SI radiologique ou les PACS (système d'archivage et de transmission d'images, ou Picture Archiving and Communication System en anglais). Une autre source de données est possible via les consortiums de recherche en santé et les cohortes.

Système d'Information Hospitalier (SIH)

Le SIH est nécessaire au bon fonctionnement du parcours des soins au sein de l'établissement de santé. Le SIH contient essentiellement :

— Les informations administratives et médicales ;
— Les informations concernant le parcours de soins du patient.

Ces informations évoluent suivant la durée de mise en place du SIH et la qualité ainsi que la quantité de données renseignées.

Picture Archiving and Communications System (PACS)

Dans le domaine médical, de plus en plus d'hôpitaux sont équipées de PACS. Ces systèmes permettent un accès plus rapide aux images médicales du patient. Le problème qui se pose est de trouver la manière la plus intelligente, sûre et sécurisée de gérer la diffusion et le partage d'images médicales entre les praticiens afin de poser le meilleur diagnostic.

Cohortes

Ce sont des bases de données constituées d'un ensemble de sujets partageant un certain nombre de caractéristiques communes, suivis dans le temps à l'échelle individuelle afin d'identifier la survenue éventuelle des évènements de santé d'intérêt. Ce sont des objets de recherche utilisés sur le long terme. Par exemple, la cohorte *Constances* a pour vocation de suivre 200000 volontaires à des fins de recherche. L'objectif est de mettre en œuvre une importante cohorte épidémiologique, représentative de la population générale adulte et d'effectif important, destinée à contribuer au développement de la recherche épidémiologique et à fournir des informations à visée de santé publique.

Depuis mars 2009, l'INSERM a mis en place au sein de l'Institut Thématique Santé Publique une cellule de coordination nationale des cohortes (CCNC), animée en partenariat avec l'Institut de Recherche en Santé Publique (IReSP). Cette cellule a pour objectif de fournir aux cohortes des services financés par des fonds publics, et de faciliter l'accès mutualisé aux données de ces cohortes à l'ensemble des communautés scientifiques intéressées. Parmi les cohortes gérées par l'INSERM, nous présentons CépiDC.

3.1.4 CépiDc

Parmi les missions légales de l'INSERM figure la production de la statistique nationale des causes médicales de décès. Cette mission est assurée par une unité de service de l'INSERM, le centre d'épidémiologie sur les causes médicales de décès (CépiDc). Elle comprend le codage médical, le traitement statistique et la diffusion des données des causes médicales de décès. L'INSERM a en outre la charge, tout au long de ce processus, de mettre en œuvre toutes les mesures physiques et logiques permettant de garantir la confidentialité des données.

Pour chaque décès survenu sur le territoire français (550000 à 600000 par an), un médecin doit rédiger un certificat, comportant ses causes médicales. Les certificats rédigés sur papier sont transmis aux mairies concernées puis aux agences régionales de santé, qui les font suivre au prestataire du

CépiDc qui en assure la numérisation et la saisie informatisée. Ce circuit, qui prend plusieurs mois, est progressivement remplacé par un accès quasi immédiat aux données, grâce à l'application en ligne CertDc, disponible depuis 2007. Celle-ci permet aujourd'hui la dématérialisation de plus de 10% des certificats, avec un usage en forte progression.

Via le circuit « papier », les données reçues par le CépiDc sont de deux types : images du certificat et du bulletin d'état civil qui leur est lié, données structurées relatives au défunt dans un premier temps puis texte des causes de décès.

Les données reçues via les deux circuits font l'objet de contrôles et de corrections avant d'entrer dans le circuit du codage médical, qui s'appuie sur une application développée par un consortium international (cinq pays européens et les États-Unis), au sein duquel l'INSERM est représentée.

L'ensemble des données est échangé, à différents stades de leur traitement et sous forme de fichiers XML chiffrés, avec l'Insee et l'InVS (institut de veille sanitaire), contribuant ainsi à la statistique nationale sur les décès et à la veille sanitaire. Le CépiDc répond à des demandes de chercheurs travaillant sur des cohortes en leur fournissant des données, sous forme de fichiers au format CSV [105], relatives aux populations suivies. Il effectue par lui-même des études statistiques au moyen de logiciels statistiques existants ou de méthodes et programmes développés en interne.

Le DSI de l'INSERM fournit l'infrastructure et le support informatique au CépiDc : serveurs, hébergement, outils de sécurisation et d'exploitation, maintenance des applications, etc. L'INSERM étant le producteur des données sur les causes de décès, il n'a pas à demander d'agrément d'hébergeur de données de santé. Les données concernant le défunt sont anonymes mais indirectement identifiantes, d'où l'obligation légale d'en protéger l'accès. Ceci est réalisé, pour l'essentiel, par le chiffrement des flux de données, l'architecture mise en place pour héberger les serveurs et une gestion stricte des droits d'accès aux données, y compris au sein du CépiDc.

Les missions du CépiDc vont évoluer dans les années à venir vers la mise à disposition des chercheurs des données médico-économiques contenues dans le futur système national des données de santé (SNDS, géré par la CNAM-TS – Caisse nationale de l'assurance maladie des travailleurs salariés), qui sera par ailleurs alimenté par les causes médicales de décès. Les infrastructures à mettre en place pour cette nouvelle mission sont en cours d'étude, de même que le moyen de rapprocher les données en utilisant un identifiant chiffré.

3.1.5 Problématique générale

L'hébergement de données de santé doit répondre à un ensemble de prérequis. Le fait de manipuler des données aussi sensibles et à caractère personnel nécessite une politique de sûreté et de sécurité bien spécifique. En effet, la CNIL préconise aux différents acteurs de prendre les dispositions nécessaires à la sécurité des données. Par exemple, en recherche, une anonymisation de données est souvent recommandée. C'est aussi le cas des images médicales. Pour les exploiter, il faut donc d'abord les anonymiser afin d'éliminer toute possibilité de pouvoir identifier une personne à partir de son dossier ou d'une image DICOM. L'anonymisation est une technique permettant de faire disparaître d'un document toute référence à la personne concernée à travers ses données personnelles (nom, numéro de sécurité sociale, INS, adresse,...).

De plus, le fait que ces données soient de nature volumineuse implique qu'il faudra tenir compte de cette spécificité dans l'architecture du système d'hébergement. Pour toute infrastructure d'hébergement de données de santé, un plan de gestion de données est requis. Ce plan présente essentiellement la politique d'hébergement adoptée dans le cadre du projet. En général, les données

105. https ://fr.wikipedia.org/wiki/Comma-separated_values

sont produites par un ensemble de centres appelés aussi des « nœuds ». Chaque nœud assure la sécurité de données, et pousse les données dans un endroit plus sûr et sécurisé dédié à l'hébergement des données de santé. La transmission des données est faite en suivant la politique de sécurité de transmission des données santé (Figure 3.1). Le centre de données, nommé aussi Datacenter, héberge l'ensemble des données transmises. Il permet aux utilisateurs l'accès aux données, et la possibilité de les traiter via des infrastructures dédiées aux calculs scientifiques.

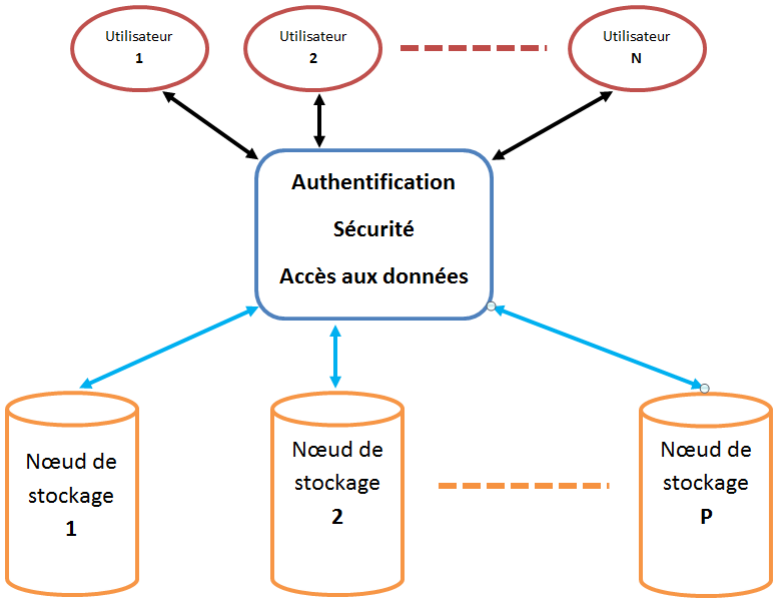


FIGURE 3.1 – Une architecture modèle d'hébergement de données de santé.

3.1.6 Tendances en matière d'hébergement

La tendance actuelle à l'échelle nationale est de minimiser le nombre des centres d'hébergement tout en augmentant la qualité d'accès aux données. Ce principe se base essentiellement sur la mutualisation des ressources des différents acteurs de production et de traitement des données de santé.

Une autre tendance liée à la précédente consiste à adosser une brique de calcul scientifique à la partie d'hébergement des données de santé. Cette corrélation permettra de diminuer le temps de traitement des données (rapprochement du calcul auprès des données) et minimise les prérequis de sécurité pour l'acheminement entre les lieux d'hébergement et ceux de traitement des données de santé.

3.2 Le projet SPC Imageries Du Vivant (IDV) et le cloud CUMULUS

Un cas d'usage pour la migration vers le cloud et l'adoption de ce type d'infrastructure est fourni par le programme interdisciplinaire « Imageries du Vivant » (IDV) de la Communauté d'Universités et d'Etablissements (COMUE) Sorbonne Paris Cite (SPC). Nous en décrivons ci-dessous le contexte général, et détaillons les solutions techniques apportées.

3.2.1 Contexte

Sorbonne Paris Cite (SPC) est une communauté d´universités et d'établissements (COMUE) en Ile de France qui regroupe 8 institutions de recherche et enseignement. Elle a été créée pour fédérer les établissements membres, développer des synergies entre leurs actions, mutualiser leurs ressources/moyens, le tout dans le but de constituer une entité plus large, plus performante et de visibilité accrue au plan national et international.

SPC soutient depuis 2013 plusieurs actions, dont celle du programme pluri-établissements et interdisciplinaire « Imageries du Vivant » (IDV [106]). IDV rassemble une trentaine d'équipes et environ 200 professionnels (imageurs, biologistes, médecins, mathématiciens, informaticiens), dédiés à l'acquisition et l'exploitation des images du Vivant, toutes échelles de taille et de techniques d'acquisition confondues. Le réseau ambitionne et s'est attelé à la construction d'une infrastructure numérique mutualisée permettant le traitement, l'analyse, le stockage et le partage des données images (et données/métadonnées connexes) acquises par ses membres, dans le but d'accroitre la réutilisation des images du vivant et les connaissances qui en résultent, au service de la santé et des apprentissages.

SPC a annoncé en janvier 2016 sa nouvelle plateforme digitale pour la recherche appelée CIRRUS [107]. Les buts de cette plateforme sont :

- de favoriser la mutualisation d'équipements afin de se prémunir face à la dispersion de petites plateformes digitales ;
- de provoquer des interactions inter laboratoires et inter institutions ;
- de partager de l'expertise ;
- d'offrir de meilleurs services pour l'ensemble des communautés de SPC (SHS. . .)
- d'analyser les processus métiers et les attentes des membres participants.

La plateforme numérique CIRRUS est constituée de trois infrastructures dont le cloud CUMULUS situé physiquement à Paris 5. Il s'agit d'une *fédération* d'équipements, c'est à dire la juxtaposition de systèmes qui concourent à la vision générale d'offrir de meilleurs services en matière de calcul, de stockage et de réseaux que par le passé. En 2015-2016 il y a eu un investissement de 1 M€ afin d'atteindre un total de 4500 coeurs de calcul et 2000 To de stockage.

Dès le démarrage une ingénieure de recherche assure un rôle d'intermédiaire entre les chercheurs de IDV et les infrastructures afin de construire un eco-système logiciel. La personne est intervenue à plusieurs niveaux pour la mise en place de cet écosystème. Tout d'abord, elle s'est intéressée à une enquête qui a été réalisée au mois de mars 2015, et qui consistait à cerner les pratiques des chercheurs : comment ils travaillent, quels sont leurs outils logiciels de travail au quotidien et quelles sont leurs attentes par rapport au cloud. Cette enquête a permis d'identifier pour les membres du projet IDV :

- les systèmes d'exploitation les plus utilisés ;
- l'ordre de grandeur du stockage nécessaire dans le cloud ;
- les formats d'image les plus utilisés ;
- les logiciels les plus utilisés.

Les Figures 3.2, 3.3, 3.4 résument en partie la situation concernant ces points. On remarque en particulier que les données sont gérées principalement via des solutions locales. Cela résulte en un éparpillement des données qui rend difficile, par exemple, la mise en place d'une politique cohérente de valorisation des données. Les formats d'image les plus utilisés sont le TIFF et le DICOM et les

106. Portail du programme Imageries Du Vivant
107. Portail de la plateforme numérique paratégée CIRRUS

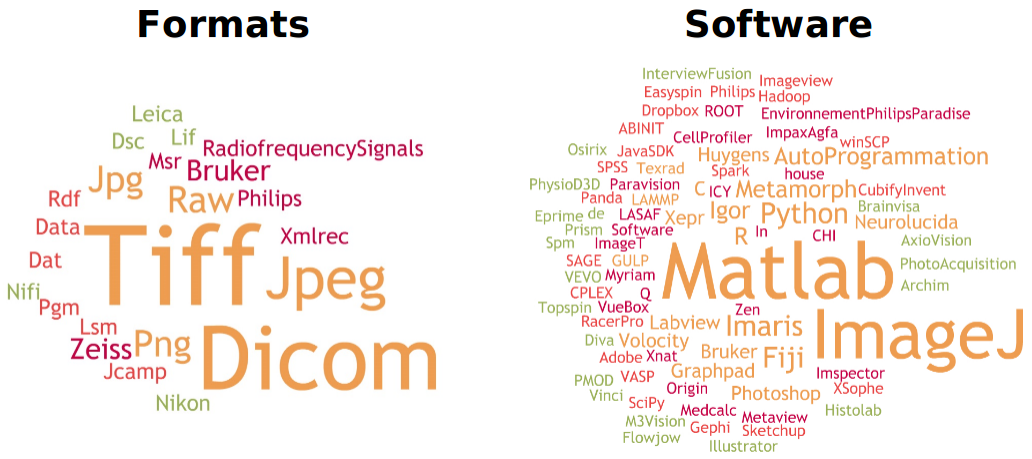


FIGURE 3.2 – Most used image formats and tools

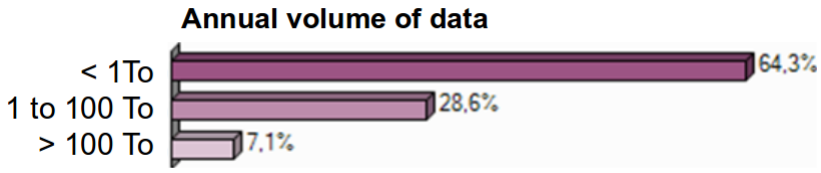


FIGURE 3.3 – Annual volume of data

outils de traitement les plus utilisés pour traiter ces images sont Matlab et ImageJ. On notera une grande variété des réponses qui dénote aussi d'une grande diversité de pratiques. L'ensemble du travail d'analyse des processus métiers a été publié dans l'atelier CloudWay [108].

Une étude sur les processus métiers similaire à celle de IDV a été conduite à l'échelle de la COMUE SPC. Les résultats sont accessibles depuis le portail de CIRRUS [109]. La méthodologie, là aussi, repose sur une approche ascendante (« bottom-up ») où l'on cherche à faire ressortir des besoins que les chercheurs expriment, à les synthétiser et enfin à offrir des solutions qui ne touchent pas aux processus métiers mais aux manières d'accéder aux outils, dans un cadre rénové qui aujourd'hui s'appelle le « cloud »et les « big data ».

3.2.2 Solutions techniques

De manière classique, pour IDV, une solution basée sur la virtualisation [110] a été choisie. Cette technique a pour avantage de faciliter la cohabitation de plusieurs systèmes (d'exploitation) sur le même support physique en assurant une isolation complète entre les systèmes et une utilisation mutualisée des ressources des systèmes.

Ainsi la solution de cloud computing OpenNebula a été mise en place par la DSI de Paris 5. OpenNebula opère comme un orchestrateur des couches de stockage, de réseau, de supervision et

108. Leila Abidi, Christophe Cérin, Danielle Geldwerth-Feniger, Marie Lafaille: Cloud Computing for e-Sciences at Université Sorbonne Paris Cité. ESOCC Workshops 2015; p 216-227
109. Portail de la plateforme numérique partagée CIRRUS
110. Cette notion est introduite à la définition 14 page 16

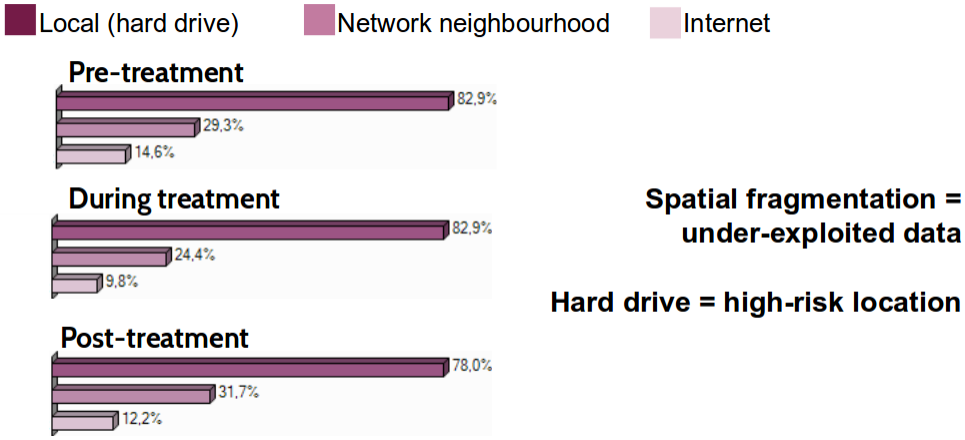


FIGURE 3.4 – Analysis of storage demand during an experiment

de sécurité. Pour le sous-système nécessaire à IDV (appelé le Virtual Datacenter ou cloud IDV), CUMULUS est le système qui gère l'ensemble des Virtual Datacenter des différents projets ou laboratoires de SPC.

Definition 26 En informatique, l'objet obtenu par l'action d'instancier est appelé une instance. Qu'est-ce que l'action d'instancier ? C'est l'action de créer un objet qui hérite d'un ensemble de qualités, d'attributs ou de valeurs à partir d'un modèle. Pour illustrer, une pomme, une orange ou une banane hérite d'attributs propres à un fruit. Dans un traitement de texte, on crée un document à partir d'un modèle et ce document hérite de ses caractéristiques ; il est au format A4, en mode portrait, avec des marges définies, un fond blanc, une police de caractères définie. Cela a pour principal intérêt d'éviter de répéter à chaque fois que l'on doit créer un nouveau document, des opérations élémentaires, souvent de même paramétrage. Il est rare de changer de format de page pour chaque document.

Dans le cas des machines virtuelles (VM) du cloud, nous pouvons avoir recours à des modèles. Pour chacun de ces modèles, on aura défini un disque virtuel puis installé un système d'exploitation dans ce disque virtuel. On peut avoir également ajouté au système des logiciels. Par la suite, on peut partir de modèles comme « machine_ubuntu » et « machine_windows » pour créer des instances comme « machine_ubuntu vm1_u, vm2_u » et « machine_windows vm1_win, vm2_win ». A chacune de ces instances, on pourra associer des caractéristiques obligatoires comme son nom, sa taille mémoire, le nombre de processeurs virtuels, le nombre de cartes réseaux... Un des intérêts du cloud est justement d'ouvrir ce type de mécanisme qui permet de proposer des machines comme un service « Infrastructure as a Service ».

Actuellement les limitations techniques sont de 254 VMs et 10To de stockage. L'ingénieure de recherche opère comme administrateur du Virtual Data Center pour IDV. Celle-ci a mis en place une liste de modèles prêts à l'utilisation. Les machines virtuelles sont instanciables à partir de ces modèles. La liste des modèles disponibles à ce jour est :

- Debian 8.2 ;
- CentOS 7 ;
- Ubuntu 15.10 ;
- Ubuntu avec l'outil Spark ;

- Windows 2012 R2 x64 standard english ;
- Windows 2012 R2 x64 standard english avec SQL Server 2012 ;
- Windows 2012 R2 x64 standard english avec SQL Server 2014 ;
- Ubuntu avec une trentaine de logiciels d'imagerie pré-installés ;

À partir de ces modèles nous avons intégré les outils suivants :

- plateforme d'imagerie ePad (https ://epad.stanford.edu/) ;
- plateforme Sis4web de SisNCom ;
- plateforme de crowdsourcing CrowdIDV ;
- plateforme de gestion des études médicales dans le cadre du projet DRIVE-SPC ;
- plateforme OwnCloud [111] utilisée comme dépôt de fichiers images.

Par défaut toutes les machines virtuelles (VM) du Virtual Datacenter peuvent accéder à Internet, cependant par sécurité la réciproque est fausse. Pour que les VMs soient accessibles depuis l'extérieur nous avons mis en place un pare-feu qui nous permet de créer des règles de redirections de ports, et faire en sorte que les machines virtuelles soient accessibles via SSH (protocole sécurisé) par exemple.

Un guide d'utilisation [112] de la plateforme est disponible. Ce guide est mis à jour régulièrement afin de suivre l'évolution de la plateforme. Certains points sont encore en discussion comme l'acquisition des licences pour le système d'exploitation Windows ou encore pour certains logiciels propriétaires, notamment Matlab.

3.2.3 Précisions sur les outils intégrés au cloud IDV

Détaillons maintenant quelques outils intégrés au cloud IDV à partir de la liste précédente ce qui permet de fixer la pertinence de ces outils pour la communauté IDV.

ePAD est une plateforme informatique d'imagerie quantitative librement accessible, développée par le Rubin Lab à Stanford Medicine radiology à l'Université de Stanford. ePAD peut être utilisée pour soutenir un large éventail de projets d'imagerie. Ses principales fonctionnalités sont de :

- fournir une méthode d'accès universel aux métadonnées d'images de radiologie, conforme au standard AIM (Annotation and Image Markup) ;
- fournir une implémentation de l'AIM dans une architecture client Web qui fonctionne sur les navigateurs Web et fournit une annotation sémantique de l'image, indépendamment du support matériel ou du logiciel utilisé pour l'acquisition des images ;
- catalyser l'adoption, la collecte, la diffusion et l'utilisation des données d'imagerie quantitative, et habiliter les radiologues avec les avantages du partage des métadonnées des images et l'interopérabilité des systèmes ;
- fournir une architecture flexible qui s'adapte bien avec certains besoins spécifiques.

Dans le cadre du projet DRIVE-SPC, Bertrand Tavitian et son équipe s'intéressent au suivi des études médicales, sur les données du vivant, et plus particulièrement à :

- la prise en compte de la traçabilité des données depuis l'acquisition et la quantification jusqu'au traitement, la publication et la réutilisation ;
- la mise en place d'un référentiel partagé et de formation sur les droits d'accès ;
- l'indexation et la classification pour une amélioration de la pertinence des résultats de requêtes en lien avec une pathologie ou un jeu d'images donnés.

111. https://owncloud.org
112. https://lipn.univ-paris13.fr/~abidi/IDV/CUMULUS.pdf

Pour cela un investissement dans une solution mature utilisée en ingénierie traditionnelle a eu lieu. Il s'agit d'un outil de PLM (Product Lifecycle Management). Ce type d'outil sert à gérer les données de conception d'un produit en permettant la construction d'un modèle de données unique pour des données/documents complexes et hétérogènes. Un outil PLM s'intéresse aussi à la gestion de la confidentialité et de la propriété intellectuelle en permettant la collaboration d'équipes. Il permet aussi une certaine forme de traçabilité des actions et des modifications sur les données. Un tel outil autorise également la gestion de projets et des équipes, et surtout des concepts autour du projet. Il permet également leurs évolutions au cours de la vie du produit qui est ici de la donnée du vivant.

La société CADESIS [113] a proposé une solution en partie de chez Siemens appelée Teamcenter. Pour intégrer cette solution dans le cloud CUMULUS, une collaboration avec l'entreprise CADESIS s'est mise en place afin de tester et de valider la solution via des séances de formation avec des utilisateurs.

Les scientifiques travaillent avec des volumes d'images de plus en plus importants et des images de formats de plus en plus variés. Le besoin de les retrouver rapidement, les partager simplement, les analyser selon différentes méthodes devient un enjeu de plus en plus crucial dans la recherche. L'entreprise SisNcom a développé une plateforme pour organiser, rechercher, visualiser, comparer, annoter et analyser toutes sortes d'images, et ce à partir de n'importe quel poste connecté à Internet.

Quels que soient les lieux de stockage et le format des images, la plateforme SisNcom permet d'accéder aux images, et devient le carrefour de toutes les images, documents associés, (texte, pdf, tableau, présentation,...) et des analyses. La plateforme permet d'indexer et d'accéder aux images quelle que soit leur origine. Elle doit, de plus, tenir compte des spécificités de chaque type d'image, au niveau des formats et des métadonnées, et de ce fait unifier l'affichage. SisNcom permet aussi de créer des sélections d'images pour un article, une présentation, les partager, les communiquer par URL,... automatiser la création de PowerPoint.

En collaborant avec Michel Smadja de SysNcom, deux machines Windows Server 2012 R2 avec le logiciel SisNcom installé ont été créées. Une des machines vituelles contient une copie complète d'une base d'images de l'institut Cochin et l'autre des images du petit animal en provenance de l'hôpital Beaujon.

La plateforme CrowdIDV permet aux experts en imagerie biomédicale (demandeurs) de soumettre une tâche participative : des images à annoter. Lors de la publication des images, les demandeurs ont la possibilité de choisir la catégorie des images (biologie, radiologie, etc). Des participants (crowders) contribuent à l'annotation des images publiées par les demandeurs. Les participants et les demandeurs doivent être inscrits via la plateforme.

Lors de son inscription sur CrowdIDV, chaque participant choisit un profil (étudiant, médecin, etc). Selon son profil ainsi que la catégorie des images qu'il choisit d'annoter, un niveau de confiance lui sera attribué. Lors de l'annotation, un participant a la possibilité d'expliciter un degré de certitude associé à l'annotation qu'il propose (confiant, presque confiant ou hésitant).

Les annotations des utilisateurs sont sauvegardées dans une base de données. Chaque image est représentée par une relation où chaque ligne stocke les annotations d'un utilisateur donné. Chaque ligne a un degré de confiance correspondant au profil de l'utilisateur, et à chaque annotation est attribué le degré de certitude de l'utilisateur. Ces informations sont importantes dans la validation des annotations.

113. http ://www.cadesis.com/cms/easysite/cadesis2/fr/accueil

Terminons ce panorama des outils qui ont été intégrés au cloud IDV par une présentation du travail en cours. A la mi-juin 2016, une nouvelle demande s'est faite jour de la part de la plateforme FRIM (Fédération de Recherche en Imagerie Multimodalité - http ://www.bichat.INSERM.fr/) de l'UFR de Médecine, Université Paris-Diderot/Paris7. Cette plateforme a pour projet de se doter d'un logiciel de gestion de la qualité, de gestion documentaire et de gestion de projets (Benchsys/Labage -http ://www.benchsys.com/) dans sa démarche qualité de certification ISO 9001. BENCHSYS, progiciel conçu dès le départ comme un "Adapative Case Management" (ACM), est le progiciel retenu. C'est un produit-phare pour le secteur hospitalier ; il gère principalement la documentation et les déclarations d'événements indésirables, mais aussi la sécurité des patients, la qualité des soins, les trajets cliniques, les procédures, la traçabilité, le suivi d'appareillage et les compétences.

Dans le cadre de FRIM, ce logiciel sera utilisé uniquement pour le petit animal. BENCHSYS est de type client/serveur, et nécessite donc l'installation d'un serveur/base de données. Le cloud CUMULUS répond aux contraintes d'installation de ce logiciel (Machine Virtuelle Windows Server 2012 R2) et offre le service adapté à ce projet. Les clients seront basés sur le site Bichat de Paris 7 et utiliseront ce logiciel au quotidien pour assurer la traçabilité des projets réalisés par la plateforme FRIM. La société belge Labage qui assure l'administration du serveur et la maintenance du logiciel BENCHSYS doit également avoir un accès au cloud.

3.2.4 Conclusion

Cette première phase de la vie du cloud CUMULUS a permis d'utiliser les mêmes outils que ceux utilisés dans les laboratoires mais dans un nouveau contexte d'exploitation grâce au déploiement de machines virtuelles en lieu et place de machines isolées et disséminées dans différents lieux physiques. Le cloud CUMULUS nous permet de travailler en masquant beaucoup de détails techniques. La centralisation des données dans le cloud permet également de rationaliser le stockage et d'éviter la multiplication de solutions ad-hoc pour archiver sur le court terme les données. Via l'enquête, nous avions en effet remarqué de nombreux sites avec des solutions de type NAS (Network Access Storage) pour lesquels les capacités de stockage étaient limitées.

Pour la suite, il serait important de :

- pouvoir déployer des réseaux de machines virtuelles pour commencer à expérimenter avec des Virtual data center plus proches en taille de ce qui est nécessaire dans la pratique. Cela permettra également de tester des implémentations parallèles des algorithmes de traitement d'images et sans doute de pouvoir déployer de véritables systèmes de large échelle (et pas seulement le déploiement d'une seule application) ;
- faire remonter l'expérience IDV au niveau de Ubercloud. Ce réseau d'échanges, ou réseau social, s'intéresse aussi au numérique pour les Bio-IT [114]. Sur ce site, il est aussi possible de trouver de la ressource gratuite (de calcul par exemple) ;
- dépasser le cadre de la production courante, et identifier une expérience phare d'IDV qui soit à la frontière des connaissances dans l'un de ses domaines pour mettre à l'épreuve et repousser les limites d'usage du cloud IDV. Il s'agirait de mettre en place une approche inter-disciplinaire à la frontière entre les « grands Systèmes » et une ou plusieurs disciplines IDV.

Voilà pour les objectifs théoriques. Mais pour arriver à de tels objectifs, des moyens sont à mettre en place. Sans de tels moyens, que les décideurs pourraient mettre à disposition (matériels, logiciels mutualisés, maintenance, formation,...), on ne pourra expérimenter de nouvelles solutions d'archivages ou de centrales de traitement. Par ailleurs, le projet IDV devrait également s'articuler

114. http ://www.theubercloud.com/

avec d'autres problématiques déclinées dans ce Livre Blanc, par exemple vis à vis de la sécurité via une étude de risques. Enfin, il serait important de s'inspirer des règles de bonne pratique sur la confiance et l'hébergement dans un centre de données telles qu'exprimées dans le Livre Blanc.

Enfin, comme lecture complémentaire nous aimerions parler de l'article "The Role of Data Science in Web Science" par Christopher Phethean, Elena Simperl, Thanassis Tiropanis, Ramine Tinati, and Wendy Hall. Cet article, disponible en ligne [115], s'intéresse plus particulièrement aux relations entre le Web et la science des données, partant de la définition que le *Web science studies the sociotechnical relationship between people and the Web, examining not only how the technology behind the Web facilitates the various applications now running atop it, but also the role of the people using it, how their lives are changed by it, and how they play a vital role in shaping and maintaining the Web for future generations*. Nos différentes enquêtes conduites pour amener de meilleurs services aux membres de SPC partent d'observations de terrain et sans doute qu'elles pourraient s'enrichir d'éléments méthodologiques plus aboutis pour théoriser un peu mieux les contextes. L'article cité en référence parle par exemple des approches interprétatives ou constructivistes qui pourraient alimenter une réflexion, nécessairement inter-disciplinaire.

3.3 Présentation des data centers de DATA4 Group

Nous avons sollicité des personnes du milieu afin de commenter les services types d'un data center existant. Le groupe DATA4, opérateur européen de data centers à Paris, Milan et au Luxembourg a accepté de présenter son campus de data centers situés à Paris-Saclay. Dans ce document simplifié, vous pourrez trouver une description des infrastructures et des solutions que le groupe met en jeu pour faire face aux multiples problématiques de sécurité et de résilience. À noter que DATA4 a pour ambition de participer au projet du cluster industriel et scientifique Paris Saclay [116].

3.3.1 Data center : l'état de l'art et son évolution

Les data centers DATA4 présentent les infrastructures les plus modernes d'hébergement en France et sont de véritables forteresses pour les données qu'ils hébergent. La donnée est donc stockée en lieu sûr et disponible – c'est-à-dire que les serveurs fonctionnent sans interruption pour la traiter et la livrer aux demandeurs d'information. Ces data centers permettent de disposer des plus hauts niveaux de sécurité possibles tant par la qualité que par la redondance des équipements électriques et de climatisation.

Definition 27 En informatique, la redondance consiste à disposer plusieurs exemplaires d'un même équipement ou d'un même processus ou de tout autre élément participant à une solution électronique. La redondance est utile pour augmenter la capacité totale ou les performances d'un système et/ou pour réduire le risque de panne. La redondance $N+1$ est un terme utilisé en informatique pour décrire un système impliquant un nombre d'équipements (N) plus un équipement supplémentaire et dont la mission est de permettre à l'ensemble de continuer à fournir un service dans le cas de la perte de l'un des N équipements d'origine. Une architecture $2N$ signifie que vous disposez de deux « circuits/systèmes » physiquement indépendants de « puissance égale ». La notation $2(N+1)$ suit la même logique, et désigne deux circuits indépendants disposant chacun de $N+1$ équipements d'une puissance donnée.

115. https ://www.computer.org/cms/Computer.org/computing-edge/ce-jul16-final.pdfSoftware, Computing Edge, IEEE Computer Society, 2016

116. DATA4 parie sur le dynamisme de Paris Saclay pour ses data centers, Les Echos, M. Bidault, février 2016

3.3.2 Infrastructure

Sécurité d'accès

Les data centers sont isolés dans une zone data center protégée par une enceinte comprenant capteurs anti-intrusion et caméras. Cette zone se trouve elle-même sur un campus gardé par un poste de sécurité 24/24-7/7 avec rondes de nuit. Enfin, tous les flux de passage sont contrôlés par métier et par client. Ainsi, les électriciens ont des chemins d'accès différents des spécialistes de la climatisation, tandis que chaque client a son chemin d'accès en fonction de la salle – ou la baie – où il est hébergé.

Sécurité électrique

Les bâtiments offrent une puissance électrique au m2 adaptée aux équipements « haute densité » - jusqu'à 20 kVA/rack avec une garantie de fourniture de l'énergie de presque 100% et une électricité d'une qualité très poussée.

Toute cette chaîne de production est doublée : les adductions physiques souterraines, les onduleurs, les transformateurs, les locaux de batteries, les chemins de câbles jusqu'aux baies informatiques, etc. Chaque data center possède au minimum 3 générateurs électriques et dispose de sa propre cuve enterrée de fioul pour une autonomie de 72 heures. La chaîne de production est secourue par des générateurs électriques au nombre de 3 (minimum) par data center, et ces derniers disposent chacun de leur propre cuve enterrée de fioul pour une autonomie de 72 heures.

Climatisation

Les serveurs et baies de stockage peuvent fonctionner correctement jusqu'à une température de salle de 30°C à 40°C. DATA4 garantit 21°C (18°C à 27°C). Les systèmes de climatisation sont assurés par deux circuits différents et les échangeurs sont nombreux pour pouvoir fonctionner en cas de défaillance de l'un deux. Data4 a installé en plus une réserve d'eau glacée dans chaque data center permettant à la climatisation de disposer du froid pendant 10 min.

3.3.3 Les certifications : une garantie pour les clients

Certification ISO 9001

DATA4 a obtenu la certification ISO 9001 basée sur la nouvelle norme qualité pour l'ensemble des activités Sales & Services. Le système Qualité de DATA4 pour l'ensemble de son offre, se veut garant de sa réactivité et de son efficacité à satisfaire les attentes de ses clients.

Cette certification reconnaît la mise en œuvre d'un système de management unique commun à toute l'entreprise (Global Management System). Les différentes organisations appliquent des processus communs, avec des indicateurs communs, et utilisent les e-technologies. L'ensemble de la documentation est entièrement électronique et accessible immédiatement par tous, dès sa validation. Cette reconnaissance externe est primordiale pour les clients qui ont ainsi l'assurance de traiter avec un partenaire de confiance, quel que soit le domaine concerné. Travail d'équipe, implication et engagement sur la satisfaction clients, autant de paramètres constituant les points forts de cette certification.

Certification ISO 14001

La norme ISO 14001 a trait aux exigences environnementales en matière de système de management environnemental (SME). Cet outil de management SME permet à un organisme de toute taille et de tout type :

- d'identifier et de maîtriser l'impact environnemental de ses activités, produits ou service ;

- d'améliorer en permanence sa performance environnementale ;
- de mettre en œuvre une approche systématique pour définir des objectifs et cibles environnementaux, les atteindre et démontrer qu'ils ont été atteints.

Certification ISO 18001

La spécification OHSAS 18001, « Le Système de Management de la Santé et de la Sécurité au Travail » est une référence internationale précisant les exigences requises pour permettre à un organisme de maîtriser les risques et d'améliorer ses performances.

Elle s'adresse à tous les organismes quel que soit leur domaine, et couvre leurs activités de routine et ponctuels. La spécification OHSAS a pour objectif :

- d'améliorer l'engagement pour la protection du personnel, des personnes présentes sur le site et des biens ;
- d'améliorer la réputation sur les problématiques sensibles ;
- d'intégrer la gestion de la Santé et de la Sécurité au travail pour toutes les fonctions de l'entreprise ;
- de soutenir la stratégie de développement durable.

Certification ISO 27001

L'ISO 27001 est un référentiel international qui spécifie les exigences concernant un Système de management de la Sécurité de l'Information. Cette norme fournit les outils pour une évaluation des risques pertinente et la mise en place de contrôles appropriés pour préserver la confidentialité, l'intégrité et la disponibilité de l'information. Le but est de protéger l'information de votre organisation contre toute perte ou intrusion.

Certification ISO 50001

Une gestion efficace de l'énergie aide les organismes à réaliser des économies, à réduire sa consommation d'énergie et à faire face au réchauffement climatique. ISO 50001 guide les organismes, quel que soit leur secteur d'activité, dans la mise en œuvre d'un système de management de l'énergie qui leur permettra de faire un meilleur usage de l'énergie.

Certification European Code of Conduct

Sur audit énergétique, DATA4 a été agréé de la certification Code of Conduct délivrée par la section des services Science, Joint Research Centre, de la Commission Européenne. Face à la consommation d'énergie puissante et croissante des data centers, cette certification a pour objectif de mobiliser les acteurs de l'industrie face à leur empreinte environnementale élevée :

- Responsabiliser : l'audit énergétique permet au data center de mesurer son niveau de consommation et identifier les opportunités d'économie énergétique ;
- Inscrire dans une dynamique positive : chaque data center doit définir et piloter un plan d'action visant à réduire leur consommation énergétique. Le système de climatisation par plafond diffusant sans faux-planchers est l'une des raisons notoires du succès de l'obtention de cette certification. Parmi 200 sociétés auditées, DATA4 fait partie intégrante du réseau de data centers éco-responsables. La volonté stratégique de DATA4 s'aligne sur celle du développement durable qui consiste à maîtriser l'efficience énergétique pour le bien des générations futures. Par ailleurs, DATA4 finalise un projet de récupération de la chaleur émise par les data centers pour la fournir à un maraîcher de culture biologique qui dispose de 2500 m^2 de serres à proximité du campus DATA4.

3.3.4 Le data center hyper-connecté

L'hyperconnectivité

Aujourd'hui, les entreprises sont très largement impactées – et certaines complètement transformées ! – par l'avènement du digital qui touche, comme un véritable raz-de-marée : l'économie, les objets, l'habitat, les transports, la communication. . . Cela pour plusieurs raisons :

- Cela rend les entreprises plus efficaces et donc plus productives ; c'est-à-dire que la quantité de richesse créée par unité de production est supérieure ;
- Un deuxième aspect plus récent concerne la valeur de la donnée. La donnée est une mine d'or pour celui qui sait l'exploiter. On parle de pétrole des entreprises ; nous préférons parler de capital. Le nerf de la guerre pour l'entreprise du 21e siècle est de collecter les données produites et surtout de les traiter et de les rendre intelligibles pour mieux servir ses clients. Les entreprises doivent donc s'adapter et trouver des solutions pour stocker et traiter leurs données produites en masse.

Le data center ne joue donc plus seulement un rôle de forteresse ; l'offre DC évolue pour passer d'un actif immobilier vers un actif purement numérique. Ainsi, le data center doit être aussi :

- un *carrefour numérique* permettant aux entreprises de se connecter et de s'interconnecter ;
- une *plateforme intelligente*, très automatisée, remontant des informations riches et complètes.

Il doit aussi être un objet *intelligent, connecté et orienté services*. Il devient le nœud d'échange de l'Internet, le carrefour du digital. Aussi, nous multiplions les investissements en Europe afin de créer à l'échelle du groupe un *Hub Digital* relié aux principaux réseaux de données ainsi qu'aux plus grands clouds providers et centres de peering européens (Le peering ou appairage en informatique est la pratique d'échanger du trafic Internet avec des pairs). En effet, un système d'informations d'une entreprise ne se construit plus seulement à base d'applications propriétaires, mais le système d'informations suit une philosophie de services, c'est-à-dire qu'il se construit en combinant des fonctions propriétaires développées dans l'entreprise avec des multiples solutions en lignes, au niveau d'applications logicielles en ligne prêtes à l'emploi (SaaS), de fonctions en ligne (PaaS) ou de ressources de puissance de calcul et de stockage (IaaS). Comme ces assemblages sont souvent délicats et nécessitent de solides infrastructures de réseau, le Digital Hub permet aux clients de tous les secteurs traditionnels de se connecter directement à ces ressources de manière fiable et souple. Pour les acteurs du numérique et du cloud, nous assurons une connexion directe aux grands centres d'échanges Internet et nous leur permettons de faire bénéficier à leurs grands clients de ces connexions directes.

La présence opérateurs, et la norme Carrier neutral

La particularité d'un data center « carrier neutral » est de permettre à ses clients de choisir leur opérateur (plutôt que de l'imposer, comme dans les datacenters « classiques »). DATA4 propose une multitude d'opérateurs de télécommunications et de services. Il est donc possible de connecter les plateformes informatiques des clients à d'autres sites ou infrastructures distantes en louant des services aux opérateurs (Liens Ethernet Lan to Lan, Services Voix classique ou VoIP par exemple). DATA4 dispose à ce jour de près de 60 opérateurs présents sur les trois campus. Chaque campus de DATA4 est relié aux réseaux métropolitains de fibres noires de Paris, Milan et Luxembourg à l'aide de plusieurs centaines de paires de fibres noires via différents chemins diversifiés de bout en bout. Ainsi, les entreprises pourront articuler leur architecture informatique autour des data centers en s'y connectant facilement, et le Plan de Continuité d'Activités (PCA = DRP) s'y déploiera avec facilité et sécurité.

Internet exchange

Pour permettre aux clients de bénéficier de bonnes liaisons IP avec les centres de peering, les campus de DATA4 sont reliés aux principaux Internet Exchange afin d'être directement connectés aux Internet Exchange choisis par le Client et de bénéficier des très hautes performances (MIX, LINX, DECIX, FRANCEIX, etc).

Definition 28 Un Internet eXchange Point (ou IX ou IXP ou point d'échange Internet), également appelé Global Internet eXchange (ou GIX), est une infrastructure physique permettant aux différents fournisseurs d'accès Internet (FAI ou ISP) d'échanger du trafic Internet grâce à des accords mutuels dits de « peering ».

Interconnexions des campus

DATA4 a interconnecté ses différents campus de data centers (Paris, Milan, Luxembourg) entre eux afin de créer un seule et unique plateforme de services. DATA4 propose aux clients des interconnexions (Services Ethernet de 100Mbps à 1Gbps) entre les différents campus afin de simplifier les opérations de ses clients. Une fois connecté à la plateforme DATA4 sur l'un des trois campus, il a accès à l'ensemble des services disponible sur tous les campus.

Interconnexion aux cloud public et cloud hybride

DATA4 permet à ses clients de se connecter (via des liens privés et sécurisés) à différentes plateformes de cloud Public IAAS/PAAS suivantes (Google cloud, Amazon, Softlayer, Microsoft Azure, cloud R, Qarnot...). Ceci permet aux clients d'hybrider leurs plateformes techniques avec des ressources cloud (de calcul et de stockage) de manière sécurisée en maîtrisant les performances et la sécurité.

Interconnexion Saas

Les DSI souhaitent aller encore plus loin dans l'externalisation de leur système d'information en externalisant également le développement informatique et la maintenance applicative en souscrivant des services SaaS. Ceci permet aux clients d'hybrider leurs systèmes d'informations. Les services SaaS Salesforce et Office 365 sont par exemple disponibles depuis la plateforme DATA4.

3.3.5 Le data-center as a computer

Nous entrons dans une ère où les manipulations techniques seront de plus en plus automatisées, où les fluides électriques et de froid pourront être gérés automatiquement en fonction de la charge des calculateurs. Cette ère s'appelle : « Tout est programmatique » ou « Software Defines Everything ».

DATA4 s'engage dans cette voie en deux étapes. Premièrement, DATA4 a pour mission d'améliorer la satisfaction de ses clients en leur donnant accès à des informations pour qu'ils gèrent toutes les ressources mises à leur disposition et de leur donner, si possible, des interfaces « ouvertes » que leurs plateformes de production de services intégrent directement (ex. centralisation des informations physiques et logiques, mesures et reporting d'indicateurs divers, suivi au plus près des indicateurs environnementaux, accès aux données depuis n'importe où...).

Deuxièmement, l'utilisation des data-centers peut être comparée à celle d'un ordinateur. À l'instar des grands du Web, la plupart des entreprises vont privilégier l'utilisation de plusieurs serveurs classiques à base de processeurs x86 plutôt que l'utilisation de « super-ordinateurs ». Ceci est rendu possible par les technologies de virtualisation fonctionnant sur de multiples serveurs x86.

Outre les réseaux virtuels que nous intégrons dans le cadre de l'hyperconnectivité, nous explorons les technologies « Container » et « Open Stack » pour que le data center puisse offrir de la puissance

FIGURE 3.5 – Le campus de data centers à Marcoussis

de calcul et de stockage brute à ses clients. Une première étape consiste en du stockage dans le data center, connecté directement aux cloud providers. De cette manière, l'entreprise cliente contrôle le stockage de ses données et utilise en toute indépendance différents opérateurs de cloud computing : par exemple, les ressources de calcul d'Amazon et d'OVH lisent les données sur cette baie data center, et y écrivent les résultats.

3.3.6 Présentation du campus DATA4 Paris-Saclay

Les paragraphes suivants présentent les niveaux de redondance des infrastructures de DATA4 à Marcoussis en région parisienne. L'essentiel des spécifications techniques citées ci-après est résumé dans le Tableau 3.1.

Le foncier

Avec 110ha de réserve foncière, 100 MW de capacité électrique et deux lignes de 90 KV enterrées, une stratégie proactive d'investissements, DATA4 possède les espaces d'hébergement disponibles pour accueillir les besoins immédiats et à venir de ses clients. À la pointe de la technologie, ses installations offrent une sûreté physique et technique optimales, avec une conception de type Tier III+, et une architecture électrique et climatique basée sur une redondance en N+1 minimum. DATA4 est propriétaire de la réserve foncière et des actifs immobiliers data centers sur le campus principal. DATA4 exploite 16 000 m^2 d'espaces destinés à l'hébergement des systèmes critiques des plus grandes entreprises françaises, internationales, et des fournisseurs de services IT et cloud. Huit data centers sont en exploitation et d'autres en cours de construction.

À ce jour, DATA4 n'a subi aucune interruption de services et d'intrusion depuis le démarrage de ses activités.

Localisation

Le campus de data centers DATA4 est localisé dans le sud de Paris, Route de Nozay, 91460 Marcoussis, dans le département de l'Essonne. Il constitue une réserve foncière de 110ha dont 40ha sont aujourd'hui exploités.

DATA4 possède et exploite un second site situé à 4 km du campus sur lequel un data center - nommé « dual-Building » - a été construit pour accueillir des plateformes informatiques de secours de ses clients dans le cadre du plan de continuité d'activité.

Ces deux sites bénéficient d'une implantation géographique exceptionnelle à 25 kms au sud de Paris, éloignés de tout risque naturel majeur et à seulement 2 kms du plus grand poste de distribution électrique en France, Villejust. Les deux sites sont raccordés par deux postes sources distincts.

Exposition des risques

Risques naturels

- Aucun risque d'inondation ou de remontée d'eau par les égouts, y compris pour les accès et les locaux techniques
- Pas de risque d'incendie
- Pas de tremblement de terre ou de glissement de terrain
- Protection para-foudre.

Risques industriels

- Pas de zone d'habitation à proximité directe du site
- Pas de zone Seveso à proximité
- Pas de risque de pollution ou de contamination
- Pas de risques de rayonnements électromagnétiques
- Pas de voie de transports de produits dangereux à proximité
- Pas de couloir aérien au-dessus du site
- Pas de voie ferrée à proximité
- Pas de zone militaire à proximité
- Pas de pipeline, gazoduc ou aqueduc à proximité

Risques sociaux

Pas de risques sociaux (grève, vandalisme…) du fait de :

- la localisation géographique du site et l'environnement proche (dédié à la recherche et éloigné des zones d'habitation) et
- la présence d'un contrôle d'accès rigoureux.

Politique environnementale

La réussite de DATA4 s'appuie sur sa capacité à améliorer en permanence les prestations fournies à ses clients, tant sur la qualité du service, que la santé et la sécurité des personnes et le respect des règles relatives à la protection de l'environnement de ses activités. C'est pourquoi le groupe met en œuvre un Système de Management intégré, basé sur les référentiels ISO 14001, OHSAS 18001 et ISO 9001, et s'inscrivant dans un contexte de développement durable.

Connexion du campus et opérateurs sur site

La connectivité est un point fort du campus de DATA4. Vous pouvez retrouver certaines spécifications techniques concernant la connectivité dans le tableau récapitulatif 3.1.

3.3.7 Description d'un data center type

Un bâtiment data center type comprend :

- Surfaces informatiques : huit salles informatiques d'environ 250 m^2 chacune, soit environ 2000 m^2 de surfaces data nettes réparties sur deux étages

	CATÉGORIE	RÉFÉRENCES DATA4
IMPLANTATION DE 1ER CHOIX	Emplacement de 1er choix	Poste double RTE de Villejust Dual site permettant un PRA Hors zones inondables ou lignes aériennes
	Hectares	Campus de 110 hectares
	Localisation	Marcoussis, Essonne (91) – France _ 20 km de Paris
	Accès	A10 et N118_ Aéroport d'Orly_ RER Massy _ Bus DM10
DATA CENTERS	Bâtiments	7 data centers existants (13 000 m2 de Salles IT)
	Structure	Coque, verticaux et façades en béton armé Dalles pleines ou dalles alvéolaires avec dalle de compression
	Hauteur sous plafond	3,65m
	Charge au sol	1000kg/m2 dans les salles IT permettant un charge de 1500kg/m2/rack
	Monte charge	Maximum 2500kg
	Extension	*13 Data Centers supplémentaires, soit 26 000m2 de salles IT*
CLIMATISATION	Production de froid Climatisation	Production d'eau glacée par Groupe Froid en redondance N+1, distribution par réseau redondant. Climatiseurs redondants. Soufflage d'air froid par plafond diffusant (Absence de faux plancher) Contrôle de la température et de l'hygrométrie
ALIMENTATION ÉLECTRIQUE	Alimentation Sécurisation	Alimentation du Campus à l'aide de deux lignes HT 90kV (active / active) Chaque bâtiment reçoit 4 lignes 20 kV. Générateurs en redondance N+1. Possibilité d'alimentation par générateur externe UPS Alimentation de chaque rack par 2 sources d'alimentation ondulées indépendantes.
DÉTECTION INCENDIE	Détection	Détecteurs dans tous les locaux Double système de détection incendie ; thermique et optique
	Protection	Extinction par gaz (azote) et sprinklers Structure du bâtiment en béton armé Compartimentage coupe feu 2h (murs et portes) de tous les locaux
SÉCURITÉ	Gardiennage Sécurité Caméras Sureté	Présence 24/7 d'agents qualifiés SSIAP + PC Sécurité + visa de CIVI.POL (Ministère de l'Intérieur) Contrôle d'accès par Badge (Biométrie ou Code en option) CCTV Caserne de gendarmerie à 900m du campus
CONNECTIVITÉ ET ÉCOSYSTÈME	Opérateurs Réseau	Connectivité à 20 opérateurs, ISP et fournisseurs de services 8 adductions de fourreaux Opérateurs distinctes sur le Campus Plusieurs centaines de fibres noires vers Paris+ réseaux longue distance opérateurs (chemins diversifiés)
	Services	Multiples services d'interconnexion pour l'espace d'hébergement du client : 2 Carriers Room de 130 m² 14 « meet-me-room » 2 boucles de fourreaux et de fibres noires autour du Campus Connexion du Campus et dual site via 2 fourreaux
SERVICES	Chef de projet Service Manager	Interlocuteur privilégié du client pendant toute la phase d'installation (Delivery) Interlocuteur opérationnel du client pendant la durée du Contrat (Run)
	Services de proximité	Gestes de proximité
	Services Interconnexion	Services Cross Connect, Cablâges etc.
	Services logistiques	Salles de réunion, salles d'intégration, salle de stockage, places de parking, WiFi, Restaurant d'entreprise
QUALITÉ	Normes internationales	ISO 9001 (Management de la Qualité); ISO 14001 (Management Environnemental); OHSAS 18001 (Management de la Santé et de la Sécurité au Travail); En cours: ISO 27001 (Management de la Sécurité de l'Information); ISO 50001 (Management de l'Energie)
	Normes Européennes	European Code of Conduct
	Supervision	Support client 24 x 7 x 365 Supervision des infrastructures sur le site en 24 x 7 x 365
	Portail Client	Un contact privilégié disponible en permanence, demande de services, reporting et gestion des opérations
SLA	Electricité	Disponibilité > 99,999%
	Température	18 à 27°
	Hygrométrie	50% + - 20%
	Interconnexion	Services interconnexion GTR 4h

TABLE 3.1 – Résumé des spécifications techniques des data centers DATA4. Les éléments mentionnés dans le tableau décrivent les spécifications techniques du site de production et du site de secours

- Surfaces de bureaux : 425 m^2
- Surfaces de stockage : 225 m^2
- Divers locaux réservés à l'infrastructure (électricité, climatisation, etc)
- Parking
- Locaux de stockage
- Salles de réunion / d'intégration
- 2 MMR (Meet Me Room)
- Zone de livraison
- Livraison : monte-charge de grosse capacité adaptée aux besoins
- Circulations adaptées (largeur, pente et charges admissibles)

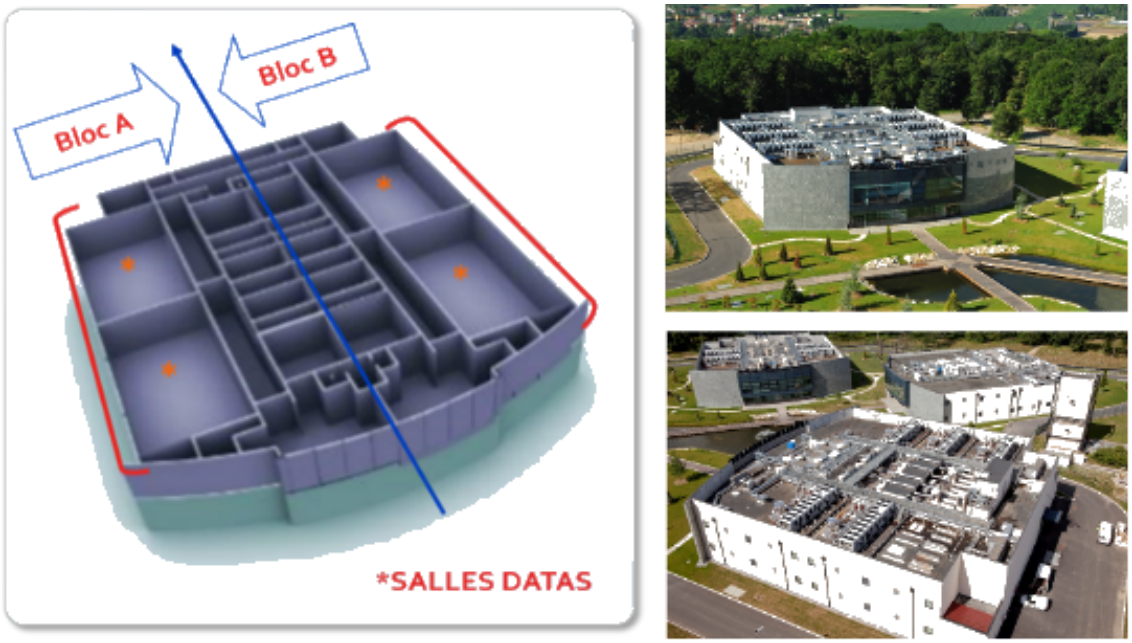


FIGURE 3.6 – Design type d'un data center. Les blocs A et B sont totalement indépendants électriquement et mécaniquement.

Les data centers de DATA4 peuvent être qualifiés de dernière génération. L'entreprise offre une continuité de services se rapprochant au maximum de la définition du Tier III+.

L'architecture est conçue avec une double chaîne d'alimentation électrique générale A+B, y compris pour l'alimentation électrique des installations de climatisation. Le bâtiment et les infrastructures des systèmes critiques datent de 2012 et garantissent la pérennité et la continuité de services dans la durée.

Aucun incident n'a été déploré depuis la création du campus.

DATA4 permet une traçabilité totale des évènements.

La conception des data centers est basée sur le principe de la maintenabilité (ou capacité pour des composants ou des applications à être maintenus, de manière cohérente et à moindre coût, en état de fonctionnement) sans impact. Cela signifie que tous les chemins de distribution pour la puissance électrique et la climatisation sont simultanément maintenables sans impact sur les opérations clients.

Système électrique

Raccordement électrique haute tension 90 kV

Le campus de data centers DATA4 bénéficie d'une situation unique : situé à moins de 5 km du plus grand poste de distribution électrique de France, Villejust, qui lui permet de croître en puissance.

Le campus est alimenté par deux lignes Haute Tension privatives directes de 90 kV, actives simultanément et permettant à chacune de développer la totalité de sa puissance.

DATA4 dispose de locaux transformateurs sur le campus afin d'abaisser la tension de 90 kV (livrée par ERDF) à 20 kV. La transformation du 90 kV en 20 kV s'opère grâce à deux transformateurs de 40 MVA chacun. La capacité électrique du site est de 100 MW.

Cette boucle participe aux grands échanges d'énergie inter-régionaux français. De ces postes partent des lignes de 225 kV qui amènent l'énergie vers des postes situés en grande ou proche banlieue et destinés à desservir les réseaux d'alimentation locale.

La distribution de chaque bâtiment se fait par quatre lignes 20 kV desservant quatre transformateurs (deux par bloc, deux blocs par bâtiment). Le site de secours est alimenté par 2 arrivées 20 kV (ERDF).

Distribution électrique dans le data center

L'alimentation électrique d'un bâtiment est secourue par groupes électrogènes avec une autonomie de 72h à pleine charge. Les groupes sont testés en réel tous les mois. DATA4 a un contrat d'approvisionnement en fuel avec TOTAL.

Deux chaînes d'alimentation électrique indépendantes par data center permettent la maintenance des installations électriques sans arrêt de l'exploitation informatique et en toute transparence pour le client. Les baies restent toujours en double alimentation (deux sources ondulées). L'autonomie des onduleurs est d'au moins 20 min à pleine charge.

Principe de la distribution électrique	**Campus data centers Marcoussis (site de production) et site de secours**
Principe générale	N+N depuis EDF (RTE) jusqu'à l'espace dédié du Client
Fournisseur énergie	Contrat RTE 90 kV – la capacité électrique à terme du campus est de 100 MVA
Groupes électrogènes	N+1 minimum 72 heures d'autonomie à pleine charge
Onduleurs	Redondance 2N Autonomie de 20 minutes minimum à pleine charge
Installation électrique	2N depuis les onduleurs jusqu'à l'espace(s) dédié(s) du Client
Espace Client	Espace dédié : 1.2 kW/m2 à 1.5 kW/m2 en standard et jusqu'à 2 kW en fonction d'une étude détaillée
Engagement de services électrique	100%

TABLE 3.2 – Spécificités du système électrique

Système de climatisation

Principe général

DATA4 utilise un système de refroidissement indirect par eau glacée. La production du froid est en configuration N+1. L'eau glacée est produite par des groupes frigorifiques à haute efficacité énergétique. Chaque bâtiment dispose d'une puissance frigorifique de 4 MW évolutive à 6 MW, selon les besoins des clients. Les groupes frigorifiques sont en redondance N+1. Deux réseaux d'eau glacée distincts permettent d'alimenter les terminaux de refroidissement associés aux salles informatiques. L'échange thermique (eau/air) s'effectue dans les couloirs techniques aux travers des recycleurs. L'air chaud est aspiré par six recycleurs d'air en redondance N+2. Ces recycleurs traitent l'ambiance de la salle de manière autonome en régulant la température de reprise. L'utilisation de ventilateurs à vitesse variable améliore l'efficacité énergétique et contribue à améliorer le PUE (Power Usage Effectiveness).

Definition 29 L'indicateur d'efficacité énergétique (en anglais PUE ou Power Usage Effectiveness) est utilisé pour mesurer l'efficacité énergétique d'un centre d'exploitation informatique. Il correspond au ratio entre l'énergie totale consommée par l'ensemble du centre d'exploitation (avec la climatisation) et la partie qui est effectivement consommée par les systèmes informatiques (serveurs, stockage, réseau). Le PUE idéal tend vers $1,0$ mais reste toujours un peu plus élevé que cette valeur car il existe d'autres équipements électriques autres que les serveurs informatiques déployés dans un centre et donc le ratio par rapport à l'énergie qu'il consomme sera toujours plus élevé que 1. À titre d'exemple, Google a vu le PUE de ses centres de données passer de $1,20$ en 2009 à $1,12$ en 2014 [a].

a. Google data centers, Efficiency: How we do it.

Particularité de la climatisation DATA4

DATA4 a développé un procédé de climatisation unique, avec soufflage d'air frais par plafond diffusant réglable. Il n'y a donc pas de faux plancher. Ce système permet un contrôle permanent des installations techniques grâce à une installation aérienne séparée de tous les réseaux : courants forts, courants faibles, réseaux de fibres optiques. Il améliore ainsi la fiabilité de l'installation : les installations sont visibles, plus faciles à mettre en œuvre et permettent une exploitation des surfaces informatiques d'hébergement dans des conditions optimales. Ce procédé fait l'objet de deux brevets et permet de réduire de 20% à 30% la consommation électrique des systèmes de climatisation des salles par rapport à un système de climatisation par soufflage en faux-plancher. Grâce à ce système de climatisation performant et une gestion maîtrisée de l'exploitation des bâtiments, DATA4 se positionne aujourd'hui avec l'une des meilleures performances énergétiques.

- Température
  La climatisation est dimensionnée en fonction de la puissance électrique installée. Les conditions spécifiques de température en salle sont régulées par six armoires de climatisation dédiées à chacune des salles. La température de consigne est fixée à 21^oC ($\pm 3^o$C) et adaptable aux besoins de l'utilisateur. L'engagement de DATA4 se fait sur la température moyenne de reprise de la salle. Cette température est maintenue quelles que soient les conditions climatiques extérieures, quelle que soit la charge globale dissipée par les équipements informatiques et télécom et dans la limite de la puissance contractée. La maintenance sur la climatisation est réalisée sans incidence sur les conditions climatiques dans les salles informatiques.
- Hygrométrie

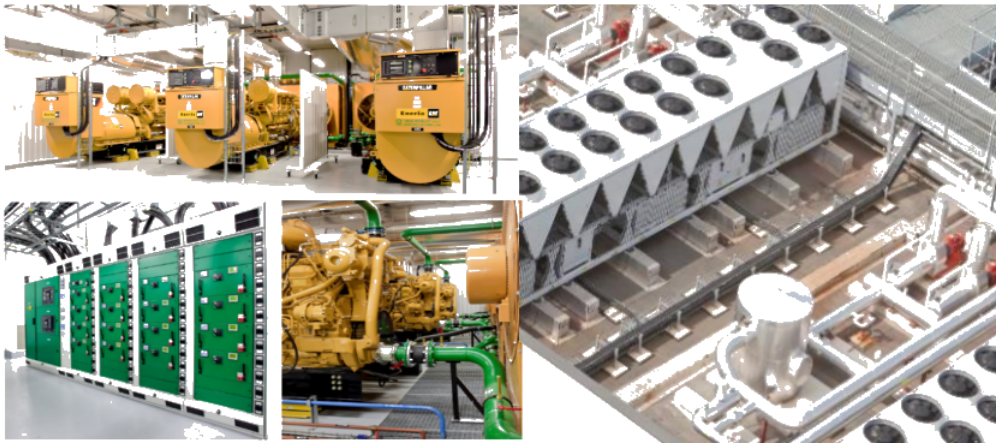


FIGURE 3.7 – Éléments du système de climatisation

L'engagement de DATA4 porte sur les conditions d'hygrométrie globale de la salle. L'hygrométrie en salle est comprise entre 30% et 70%. Ces conditions d'hygrométrie sont maintenues quelles que soient les conditions climatiques extérieures, la charge globale dissipée par les équipements du client dans la limite de la puissance contractée. Les sondes qui pilotent l'hygrométrie sont situées dans les centrales de traitement (CTA), dédiées aux salles informatiques.

- Qualité de l'air
  L'air est traité par deux CTA, l'une en secours total de l'autre, avec filtration poussée F7 (filtre fin, plus de 35% d'efficacité à 0,4 micron). Un système d'humidification d'air est mis en place. L'air est maintenu en continu en légère surpression dans la salle informatique afin de limiter le risque de transmission d'air vicié des autres locaux vers les salles informatiques (pollution de l'air, fumées, etc.).

Les unités de traitement d'air fonctionnent en rendement optimal pour une longévité maximale. D'autre part, l'ASHRAE (American Society of Heating, Refrigerating and Air Conditioning Engineers) préconise l'augmentation de la plage de régulation. DATA4 a fait le choix d'un contrôle de la température et de l'hygrométrie de manière globale, en s'appuyant sur une technique d'obstruction partielle, plutôt que de confinement complet. L'avantage majeur pour le Client réside dans la flexibilité optimale de l'urbanisation de son espace d'hébergement, en évitant le cloisonnement, grâce à la souplesse d'utilisation des installations (largeur des allées, hauteur disponible).

La mise en place du service d'homogénéité de la salle permet :

- L'analyse des points chauds et des flux d'air ;
- L'amélioration de l'homogénéité thermique en salle ;
- La diminution des impacts thermiques en salle en cas d'incident de climatisation ;
- La réalisation d'actions en commun ;
- La maîtrise de l'ensemble des infrastructures ;
- L'anticipation des risques en cas de panne et proposition de correctif ;
- Le choix étayé pour le déplacement d'équipement ;
- La maîtrise budgétaire des besoins d'aménagement en infrastructures.

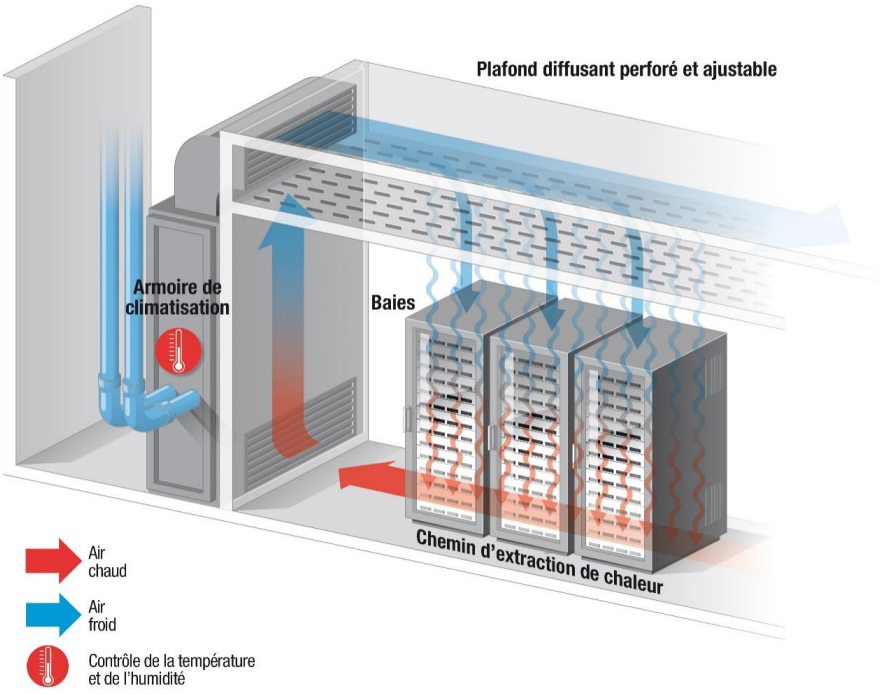


FIGURE 3.8 – Système de climatisation

3.3.8 Sécurité

DATA4 suit un Plan d'Assurance Sécurité tel que défini dans l'annexe Plan d'Assurance Sécurité (PAS).

Sécurité physique 24/7/365

Le data center est hautement sécurisé et subit des audits de sécurité réguliers. Le but est de protéger physiquement les informations et les équipements des clients contre toute perte de données ou d'intrusion.

Toutes les dispositions ont été prises pour assurer l'absence d'intrusion sur le site. La définition de la sécurité a été effectuée par CIVI POL (Ministère de l'intérieur) qui assure également la mission de contrôle et de test de la sûreté.

Le pourtour du campus est protégé par un système anti intrusion. La structure des bâtiments en béton armé permet d'assurer une résistance suffisante à l'intrusion. Les bâtiments sont situés dans une double enceinte privative : l'ensemble du site est protégé par un système de clôture, la zone des data centers est elle-même protégée par un deuxième système de clôture. Des agents de sécurités sont présents 24h/24h et 7j/7j. Un poste de gendarmerie est présent à moins d'un km de l'entrée du site principal.

Le contrôle d'accès se fait par sas uni-personnel ou hachoir à l'entrée du bâtiment, ensuite par portes badgées avec traçabilité des mouvements (possibilité de biométrie).

Toutes les installations techniques, à l'exception des armoires électriques, sont installées hors de la salle IT. Les flux de circulation de maintenance et d'utilisateurs sont séparés et il n'y a pas d'accès possible aux principales installations techniques par les utilisateurs.

La surveillance du site est assurée par un ensemble de détecteurs intrusion et une couverture

vidéo intérieure/extérieure permettant la levée de doute. Une dizaine de caméras dôme motorisées couvre les couloirs de circulation au sein d'un bâtiment. L'ensemble des accès et flux de surveillance vidéo est conservé pour une durée conforme à la législation en vigueur. Toutes les alarmes du bâtiment sont centralisées avec conservation de l'historique et impression au fil de l'eau.

Système de détection et protection incendie

La structure des bâtiments est adaptée à la tenue au feu (2h) grâce à des voiles de cloisonnement coupe-feu 2heures. Les salles IT sont équipées de double systèmes de détection à incendie (VESDA, optiques et thermo-vélocimétriques). Les points clés sont :

- Structure du bâtiment adaptée à la tenue au feu : béton armé sans éléments inflammables ;
- Compartimentage coupe-feu 2 heures des salles informatiques, des locaux techniques et locaux de stockage ;
- Traitement coupe-feu des passages de câbles et des gaines techniques
- Revêtements des murs et divers conformes ;
- Absence de produits inflammables ou à fort pouvoir calorifique au sein des salles ;
- Totalité de chaque bâtiment sous détection incendie et tous les locaux techniques et salles IT disposent d'un système d'extinction
- Double détection incendie (VESDA, optiques et thermo-vélocimétriques) ;
- Extinction par gaz (Nitrogène stocké dans 2 locaux séparés) pour les salles informatiques et locaux onduleurs ;
- Extinction par sprinkler pour les couloirs techniques et les groupes électrogènes ;
- Traitement anti-poussière : salles informatiques en suppression avec un système d'air filtré garantissant la propreté ;
- Dispositif de sur-ventilation permettant l'évacuation des fumées.

3.3.9 Supervision et maintenance

Maintenance des infrastructures 24 x 7

Le maintien du data center est assuré en 24/7/365 par des équipes expérimentées. Tous les équipements critiques sont identiques d'un bâtiment data center à l'autre sur le site afin de garantir un stock de pièces détachées permanent et disposer d'un personnel de maintenance dédié à chaque nature d'équipement. La maintenance est organisée à travers un outil de GMAO (outil de Gestion de Maintenance Assistée par Ordinateur) conformément aux gammes de maintenance préconisées par les constructeurs. L'architecture technique permet de rendre totalement transparentes les opérations de maintenance pour les clients.

Les tests de coupure réseau électrique sont effectués mensuellement, le personnel du client peut se rapprocher de DATA4 pour assister à ses opérations de maintenance s'il le souhaite. D'autre part DATA4 a conclu des partenariats avec des sociétés spécialisées et ainsi se concentrer sur son cœur de métier et être en mesure d'offrir un Guichet Unique de Services à ses clients avec des engagements forts.

La maintenance des bâtiments et de leurs équipements est réalisée par un partenaire, Cofely Ineo Groupe GDF–Suez, en garantie totale pendant une période de 12 ans. Le personnel est composé de deux managers et de spécialistes pour maintenir les systèmes électriques et climatisation. Cette maintenance est assurée soit directement par ce partenaire soit par les fournisseurs d'équipements.

Les techniciens de maintenance sont présents en 24/7/365. Selon les problèmes rencontrés, ils peuvent solliciter du personnel en astreinte. Le personnel est constitué d'experts aux profils complémentaires pour créer une véritable synergie d'entreprise : conduite de projet, suivi de travaux

comme le gros œuvre, gestion des infrastructures, équipements techniques (électricité, courant faible, climatisation).

Gestion des alarmes, incidents et maintenances

Des surveillants sont présents en 24/7/365 et sont avertis de tous défauts par une alarme visuelle, sonore et support papier. Toutes ces informations sont supervisées par deux terminaux qui permettent de visualiser et contrôler l'état des installations. Ces terminaux sont disposés dans deux locaux distincts. Le système installé permet à chaque client d'être connecté à ses installations par un terminal dédié dans ses locaux ou à distance. Les alarmes sont aussi reportées en temps réel sur deux imprimantes reliées au réseau. L'archivage est d'une durée d'un an. Les alarmes incendie sont remontées sur des baies de détection incendie placées au poste de sécurité.

Est considéré comme incident dans le cadre du contrat, tout dysfonctionnement pouvant causer une dégradation de service perçue par le client. Deux types d'incidents sont répertoriés :

- L'incident n'est pas dû à un dysfonctionnement des services rendus par DATA4 (ex : incident hardware sur un serveur client). Dans ce cas, le client a la responsabilité du diagnostic et de la résolution. Le client peut faire appel à DATA4 pour effectuer des interventions dans le cadre du service de gestes de proximités (si l'offre a été souscrite par le client).
- L'incident est dû à un dysfonctionnement des services rendus par DATA4 (incident électrique). Dans ce cas, l'hébergeur a la responsabilité du diagnostic et de la résolution. DATA4 opère la communication auprès du client jusqu'à la résolution finale.

Dans le cas d'un incident majeur, celui-ci fait l'objet d'un premier compte-rendu sous 24 heures contenant les éléments suivants : chronologie des événements, causes potentielles, risques résiduels ou de récurrence, premières mesures d'actions...

De par l'architecture redondante des infrastructures, DATA4 organise les opérations de maintenance planifiées ou autres sur son site sans perturber les équipements clients. DATA4 communique au client des plannings de maintenance sur une base annuelle.

En cas d'incident majeur, DATA4 peut être amené à opérer une maintenance urgente afin de réparer un dysfonctionnement majeur. Le client est informé pro-activement dans ce genre de crise.

3.3.10 Conclusion

En guise de conclusion, nous voudrions insister sur la gestion de haut niveau des risques effectuée dans un data center tant sur le plan des intrusions physiques que sur le plan de la redondance des services (électricité, climatisation...). L'utilisateur administre les serveurs qu'il loue (mise à jour, surveillance...). Ce n'est pas le rôle d'un data center d'assurer celà. Pour simplifier le discours, on pourrait dire que le data center doit assurer un fonctionnement des serveurs du client en mode 24/24 en prenant pour lui la gestion des risques majeurs qu'un client ne peut pas s'offrir à cause de coûts prohibitifs par exemple. Un data center est en quelque sorte une *mutualisation* de risques majeurs qui peuvent entraver l'offre de services numériques d'un client/utilisateur.

3.4 Structure de la méthode EBIOS de gestion des risques

Dans cette section nous introduisons un exemple de conduite d'analyse des risques à l'aide de la méthode EBIOS que nous avons présentée aux deux précédents chapitres, à la fois sur le plan du vocabulaire et de la méthodologie. Il est donc recommandé d'avoir lu les précédents chapitres avant d'aborder cette étude de cas.

Dans cette section nous allons donc donner un aperçu de la méthode EBIOS de gestion des risques à travers les grandes étapes du processus d'analyse tel que le décrit l'ANSSI dans la documentation de la méthodologie EBIOS.

3.4.1 Une démarche itérative en cinq modules

La méthode formalise une démarche de gestion des risques découpée en cinq modules représentés sur la Figure 3.9.

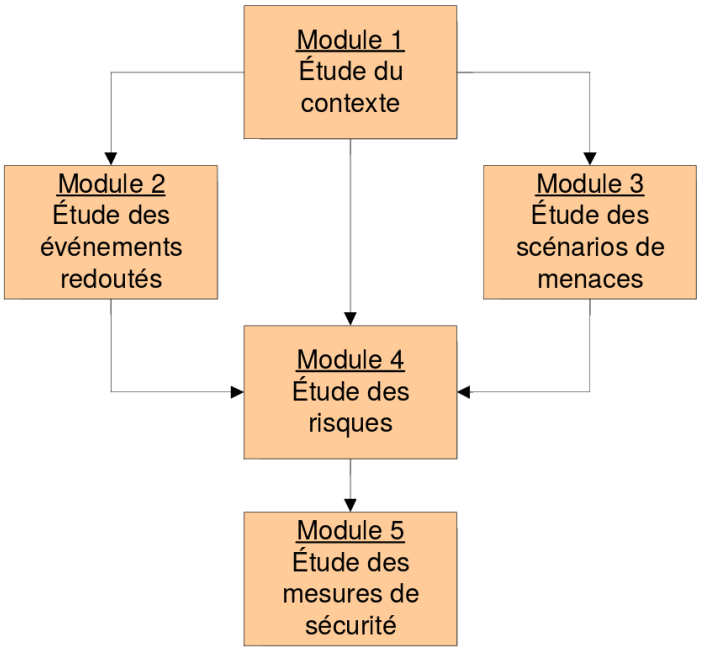


FIGURE 3.9 – Les cinq étapes de la méthode itérative EBIOS

La démarche est dite itérative. En effet, il sera fait plusieurs fois appel à chaque module afin d'en améliorer progressivement le contenu, et la démarche globale sera également affinée et tenue à jour de manière continue.

3.4.2 Module 1 – Étude du contexte

Ce module a pour objectif de collecter les éléments nécessaires à la gestion des risques, afin qu'elle puisse être mise en œuvre dans de bonnes conditions, qu'elle soit adaptée à la réalité du contexte d'étude et que ses résultats soient pertinents et utilisables par les parties prenantes.

Il permet notamment de formaliser le cadre de gestion des risques dans lequel l'étude va être menée. Il permet également d'identifier, de délimiter et de décrire le périmètre de l'étude, ainsi que ses enjeux, son contexte d'utilisation, ses contraintes spécifiques.

À l'issue de ce module, le champ d'investigation de l'étude est donc clairement circonscrit et décrit, ainsi que l'ensemble des paramètres à prendre en compte dans les autres modules.

Le module comprend les activités suivantes :

- Activité 1.1 – Définir le cadre de la gestion des risques : synthèse relative au cadre de la gestion des risques, paramètres à prendre en compte, sources de menaces

- Activité 1.2 – Préparer les métriques : critères de sécurité (confidentialité, disponibilité, intégrité), échelles de mesures, critères de gestion des risques
- Activité 1.3 – Identifier les biens de l'étude : biens essentiels des métiers (MOA), biens supports de l'informatique (MOE), tableau croisé biens essentiels / biens supports, mesures de sécurité éventuelles existantes.

À l'issue du premier module, qui s'inscrit dans l'établissement du contexte, le cadre de la gestion des risques, les métriques et le périmètre de l'étude sont parfaitement connus ; les biens essentiels, les biens supports sur lesquels ils reposent et les paramètres à prendre en compte dans le traitement des risques sont identifiés.

Prenons comme cas d'étude celui fourni en formation par l'ANSSI, le cabinet d'architectes « @RCHIMED ». On identifie les biens essentiels suivants :

- établir les devis (estimation du coût global d'un projet, négociations avec les clients...) ;
- créer des plans et calculer les structures ;
- créer des visualisations ;
- cérer le contenu du site Internet.

Les biens supports qui permettent de les réaliser sont donnés à la Figure 3.10.

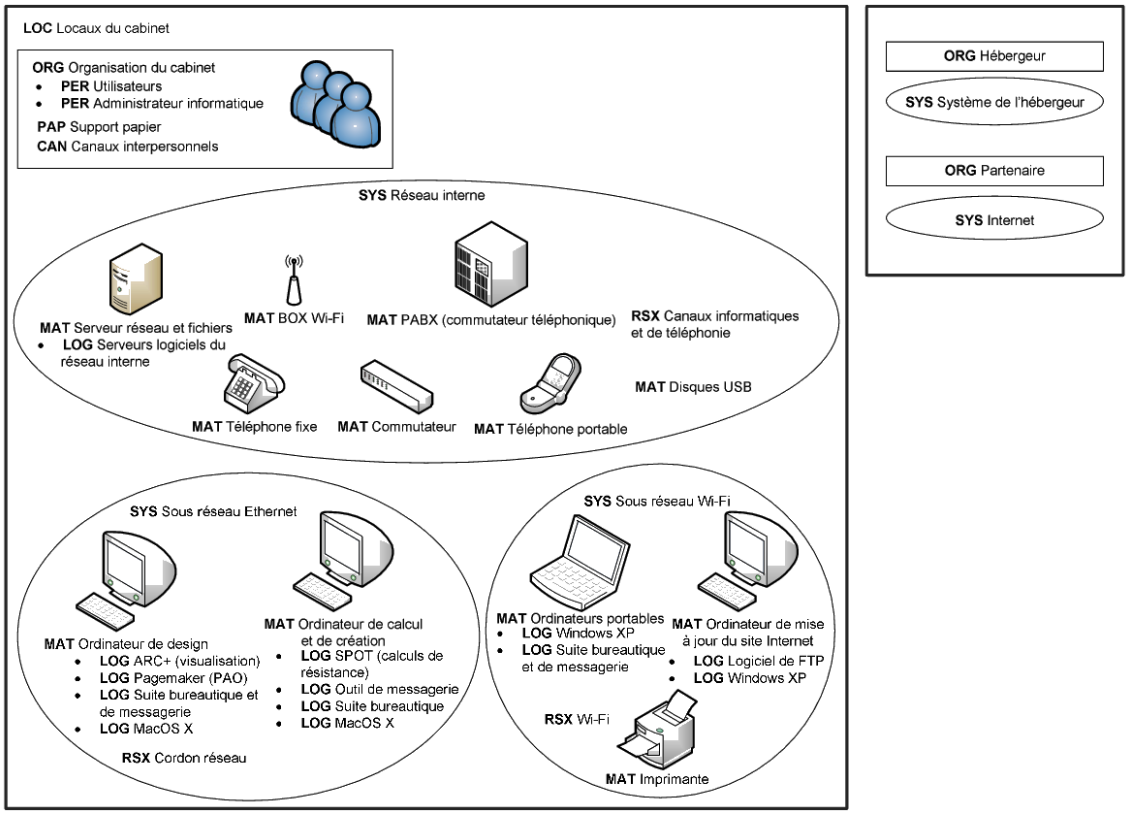


FIGURE 3.10 – Les biens support du cabinet d'architectes, cas d'usage utilisé pour illustrer la démarche EBIOS lors des formations de l'ANSSI

Les biens essentiels sont réalisés avec les biens supports suivant le tableau de référencement présenté à la Figure 3.11.

Biens supports \ Biens essentiels	Établir les devis	Créer des plans et calculer les structures	Créer des visualisations	Gérer le contenu du site Internet
Biens supports communs à @RCHIMED				
SYS – Réseau interne	X	X	X	X
MAT – Serveur réseau et fichiers	X	X	X	X
LOG – Serveurs logiciels du réseau interne		X	X	
MAT – Disque USB	X	X	X	
MAT – BOX Wifi	X	X	X	X
MAT – Commutateur	X	X	X	X
MAT – PABX (commutateur téléphonique)	X	X	X	X
MAT – Téléphone fixe	X	X	X	X
MAT – Téléphone portable	X	X	X	X
RSX – Canaux informatiques et de téléphonie	X	X	X	X
ORG – Organisation du cabinet	X	X	X	X
PER – Utilisateur	X	X	X	
PER – Administrateur informatique	X	X	X	X
PAP – Support papier	X	X	X	
CAN – Canaux interpersonnels	X	X	X	X
LOC – Locaux du cabinet	X	X	X	X
Biens supports spécifiques au bureau d'études				
SYS – Sous réseau Ethernet		X	X	
MAT – Ordinateur de design		X	X	
LOG – MacOS X		X	X	
LOG – ARC+ (visualisation)			X	
LOG – Pagemaker (PAO)		X	X	
LOG – Suite bureautique et de messagerie		X	X	
MAT – Ordinateur de calcul et de création		X		
LOG – MacOS X		X		
LOG – SPOT (calculs de résistance)		X		
LOG – Outil de messagerie		X		
LOG – Suite bureautique		X		
RSX – Cordon réseau		X	X	
Biens supports spécifiques aux relations commerciales				
SYS – Sous réseau Wifi	X		X	X
MAT – Ordinateurs portables	X		X	
LOG – Windows XP	X		X	
LOG – Suite bureautique et de messagerie	X		X	
MAT – Ordinateur de mise à jour du site Internet				X
LOG – Windows XP				X
LOG – Logiciel de FTP				X
MAT – Imprimante	X		X	
RSX – Wifi	X		X	X
Partenaires				
SYS – Système de l'hébergeur				X
ORG – Hébergeur				X
SYS – Internet	X			X
ORG – Partenaire	X	X	X	X

FIGURE 3.11 – Les biens essentiels de notre cas d'usage

3.4.3 Module 2 – Étude des événements redoutés

Le second module contribue à l'appréciation des risques. Il permet d'identifier et d'estimer les besoins de sécurité des biens essentiels (en termes de disponibilité, d'intégrité, de confidentialité, traçabilité), ainsi que tous les impacts (sur les missions, sur la sécurité des personnes, financiers, juridiques, sur l'image, sur l'environnement, sur les tiers et autres...) en cas de non-respect de ces besoins et les sources de menaces (humaines, environnementales, internes, externes, accidentelles, délibérées...) susceptibles d'en être à l'origine, ce qui permet de formuler les événements redoutés.

3.4.4 Module 3 – Étude des scénarios de menaces

Le troisième module s'inscrit aussi dans le cadre de l'appréciation des risques. Il consiste à identifier et estimer les scénarios qui peuvent engendrer les événements redoutés, et ainsi composer des risques. Pour ce faire, sont étudiées les menaces que les sources de menaces peuvent générer et les vulnérabilités exploitables.

3.4.5 Module 4 – Étude des risques

Le quatrième module met en évidence les risques pesant sur l'organisme en confrontant les événements redoutés aux scénarios de menaces. Il décrit également comment estimer et évaluer ces risques, et enfin comment identifier les objectifs de sécurité qu'il faudra atteindre pour les traiter.

3.4.6 Module 5 – Étude des mesures de sécurité

Le cinquième et dernier module s'inscrit dans le cadre du traitement des risques. Il explique comment spécifier les mesures de sécurité à mettre en œuvre, comment planifier la mise en œuvre de ces mesures et comment valider le traitement des risques et les risques résiduels.

R Pour continuer à vous familiariser avec la méthode, nous vous conseillons maintenant de lire les documents suivants :

- L'étude de cas complète de la société @RCHIMED avec la méthode EBIOS est accessible sur [117].
- L'étude de cas complète d'un cloud avec la méthode EBIOS est par ailleurs fourni lors de la formation donnée au Centre de Formation SSI de l'ANSSI. Vous pourrez trouver également sur le site de l'ANSSI un référentiel pour sécuriser le secteur du cloud computing [118].

3.5 Étude de la sécurité du cloud universitaire Univcloud

3.5.1 Investissement d'avenir dans le cloud – cinq projets de R&D dont le projet Univcloud

Le 16 Décembre 2011, Éric Besson, Ministre chargé de l'Industrie, de l'Énergie et de l'Économie numérique, et René Ricol, Commissaire général à l'Investissement, annoncent 19 millions d'euros d'investissement dans cinq projets de recherche et développement dans le domaine de l'informatique en nuage (cloud computing).

117. Méthode EBIOS - Étude de cas @RCHIMED
118. Référentiel de qualification de prestataires de services sécurisés d'informatique en nuage (cloud computing) - référentiel d'exigences, ANSSI, 2014

L'informatique en nuage représente une évolution majeure des usages et de l'organisation des systèmes d'information. Son utilisation permet aux entreprises d'accroître leur compétitivité, par une baisse des coûts informatiques, et une meilleure qualité de service. En outre, ce secteur va générer de nouveaux services, accessibles à la demande et à distance, portés par un marché en croissance de 25% par an.

C'est pourquoi le Gouvernement a lancé, dans le cadre du Programme d'Investissements d'Avenir, l'appel à projets de recherche et développement « informatique en nuage – cloud computing ». Son objectif est de soutenir les technologies qui permettront l'émergence des infrastructures informatiques de demain. Parmi les 18 projets déposés en réponse à cet appel, cinq projets ont été sélectionnés et bénéficieront d'un soutien public de 19 millions d'euros. Ces cinq projets représentent un investissement total en recherche et développement de 50 millions d'euros.

Les projets retenus sont les suivants :

- La « plateforme d'ingénierie logicielle » (projet cloudForce porté par Orange Labs) permettra le développement collaboratif et la gestion d'applications s'appuyant sur de multiples infrastructures d'informatique en nuage.
- Les « outils de portage d'applications » (projet cloudPort porté par la PME Prologue) faciliteront la migration des logiciels d'une entreprise vers le modèle de l'informatique en nuage.
- le projet « d'infrastructure logicielle haute performance » (projet Magellan porté par Bull) servira de base pour offrir les performances du calcul intensif à la demande et à distance.
- Le projet de « nuage communautaire » (projet Nu@ge porté par la PME Non Stop Systems) développera des solutions de mutualisation d'infrastructures et de compétences de plusieurs PME pour offrir des services innovants.
- Le projet de *« nuage pour les établissements d'enseignement supérieur »* (projet *Univcloud* porté par INEO) mettra les technologies du cloud au service des universités et des collectivités.

3.5.2 L'Université Numérique en Région Paris Ile de France (UNPIdF)

Univcloud est un projet de cloud retenu dans le cadre du Programme d'Investissements d'Avenir de l'État, un partenariat public privé (PPP) entre l'UNR Paris Ile de France et INEO accompagné d'autres partenaires privés.

L'UNPIdf comprend 31 établissements, 19 universités, 6 écoles, 500 000 étudiants, 50 000 personnels, 27 000 enseignants, 23 000 BIATSS (Bibliothécaires, ingénieurs, administratifs, techniciens, ouvriers, personnels sociaux et de santé), 50 salles comprenant 2000 serveurs. Il est une assistance à maîtrise d'œuvre (AMOE) qui porte la mutualisation de projets techniques : Carte multi-service, Environnement numérique de travail, Wifi Eduspot multi-établissement Univnautes, Univmobile ; ainsi qu'un service de formations et de sensibilisation pour les personnels et les étudiants (certificat Internet C2i, Guide des usages numériques).

3.5.3 Le projet Univcloud

Les objectifs

- Réaliser une étude pour la mise en place d'une infrastructure communautaire de « cloud » physique et logicielle, dédiée aux établissements membres de l'Université Numérique Paris Île-de-France. Ce concept innovant de « cloud » inter universitaire a vocation à assurer une mutualisation de l'infrastructure d'hébergement et de développement des systèmes d'informations des universités franciliennes ;

- Afin de faire face aux besoins croissants générés par les usages du numérique, permettre aux établissements une amélioration du service aux usagers avec une diminution des coûts directs ou indirects en induisant une meilleure gestion des compétences. Le projet s'inscrit dans une démarche de développement durable au travers de la maîtrise de l'emprunte énergétique des data centers réduisant ainsi l'impact carbone des universités.

Suite aux études, le projet Univcloud produira un ensemble documentaire constituant un appel d'offres afin que soit réalisée la construction d'un data center offrant son service de « cloud » mutualisé communautaire aux services des partenaires de l'UNPIdF.

Structure du partenariat

Il s'est agit d'une collaboration de 18 mois (Nov. 2011 - Avril 2013) rassemblant l'UNPIdF, 23 partenaires universitaires, le groupe industriel INEO et 4 PME aux technologiques innovantes (MANKAY, CEDEXIS, ACTIVEON, UP GENERATION).

Défaillance d'un des partenaires privés

Suite à la défaillance du partenaire privé MANKAY, l'étude de sécurité et de protection de la vie privée dans le « cloud » menée par le groupe GT3 Confiance Numérique prendra un an de retard avec l'absence totale de production de livrable. Le groupe GT3 Confiance Numérique a alerté la gouvernance d'Univcloud sur les dysfonctionnements majeurs de son groupe de travail et, suite à une réunion de bilan, s'est vu confier la reprise des travaux mais sur un délai fort court avec le besoin évident et urgent d'un accompagnement.

Avec les pénalités appliquées au partenaire défaillant, le GT3 Confiance Numérique a pu financer une Assistance à MOA pour l'accompagner dans l'élaboration du dossier de sécurité. Suite à une publication d'appel d'offre, le GT Confiance Numérique a retenu l'offre d'accompagnement judicieusement adapté proposée par le cabinet Infhotep. Infhotep propose depuis de nombreuses années des prestations, formations et accompagnements biens étudiés pour répondre aux besoins des établissements de l'enseignement supérieur.

Cible de déploiement et mutualisation

UNR Ile de France (UNPIDF) : 500 000 étudiants, 27 000 enseignants et chercheurs, 23 000 personnels administratifs et techniques, la mutualisation de 50 salles comprenant 2 000 serveurs.

3.5.4 Le schéma d'organisation du projet

Le GT3 Confiance numérique est en charge des aspects sécurité du SI, gestion des identités, protection des données personnels. Pour réaliser l'identification des besoins de sécurité, recenser les données à caractère personnel et leurs traitements, recenser les acteurs métiers usagers des services du futur « cloud » mutualisé de l'UNPIdF, le GT3 Confiance numérique a besoin des travaux du GT1 Fonctionnel qui est chargé de recenser toutes les fonctions métiers à réaliser dans l'infrastructure du futur « cloud ».

Afin de faire adhérer un grand nombre d'établissements et les services de la région Ile de France, de tester et faire évoluer l'infrastructure future, le projet Univcloud a prévu la mise en place d'un démonstrateur dont les caractéristiques et les fonctionnalités sont à circonscrire par le SGT5 Démonstrateur. La sécurisation du démonstrateur est un objectif du GT3 Confiance Numérique confié à son SGT1 Sécurité du cloud. Le démonstrateur constitue un sous-ensemble fonctionnel représentatif du futur « cloud ». De la même manière, sa sécurisation est représentative de la sécurité du futur « cloud ». La réalisation du démonstrateur se concrétise par la mise en production de deux grappes de serveurs ou « cluster » sur deux sites distants en redondance communiquant via un canal

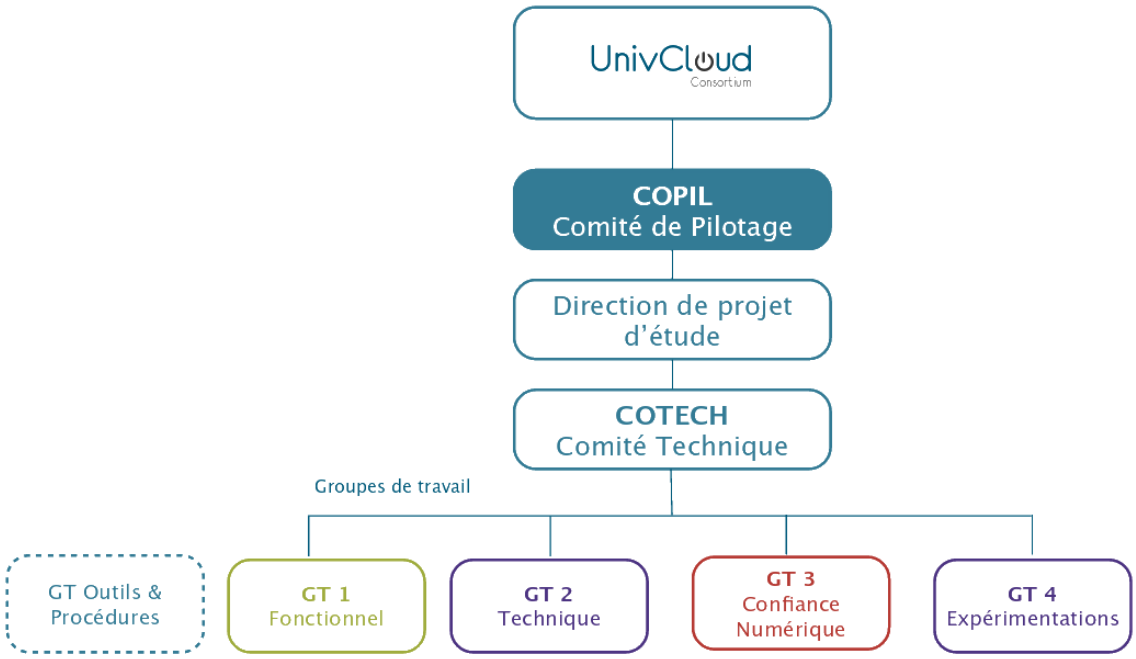


FIGURE 3.12 – Structure organisationnelle du projet Univcloud (vue macro)

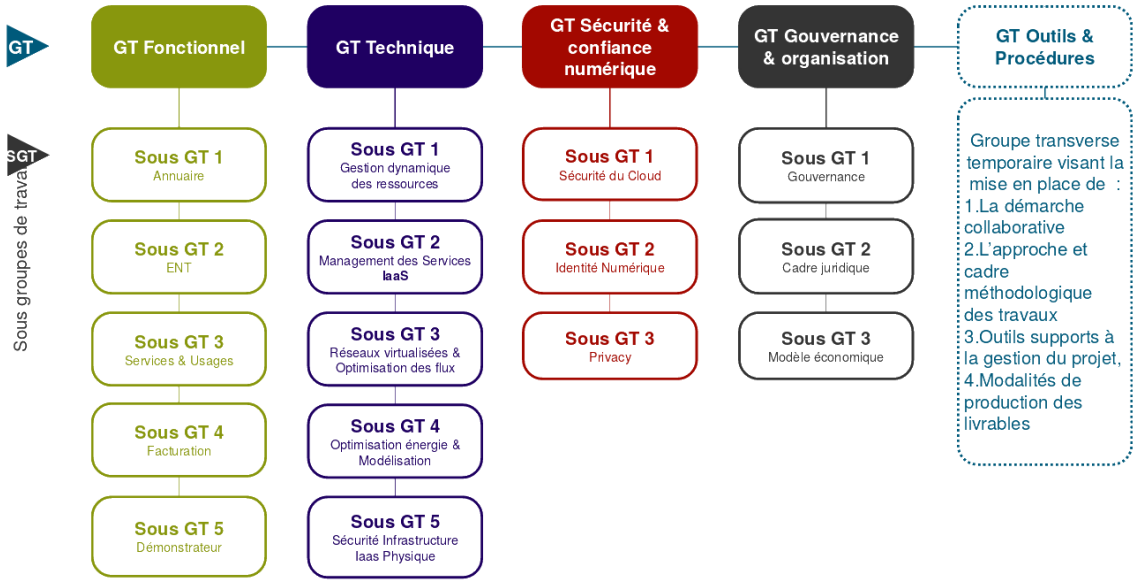


FIGURE 3.13 – Structure organisationnelle du projet Univcloud (vue détaillée)

réseau RENATER dédié au projet. L'installation est confiée au groupe INEO sur les sites de Broca de l'Université Descartes et la salle serveur de l'Université d'Evry.

En conclusion, le GT3 Confiance Numérique dépend des travaux du GT1 Fonctionnel et le SGT1 Sécurité du cloud dépend des travaux du SGT5 Démonstrateur. GT3 Confiance Numérique doit se coordonner avec ces deux groupes afin de définir en amont une « security by design ». Pour réaliser les spécifications de sécurité du démonstrateur, le GT3 Confiance Numérique a travaillé en collaboration avec INEO et RENATER. La partie de l'étude portant sur les exigences de sécurité des données à caractère personnel a été réalisé par le SGT3 Privacy avec l'aide du Laboratoire Tactic de l'Université de Nanterre.

Enfin, le SGT1 Annuaire et Gestion des habilitations a tôt fait le constat qu'il était inenvisageable de mettre en place un annuaire inter-établissement d'identités et d'habilitations. Le GT3 Confiance Numérique a donc proposé puis accompagné INEO dans l'intégration au démonstrateur des mécanismes d'authentification de la Fédération d'Identité de RENATER. Il a piloté un workshop d'appropriation de la technologie Shibboleth avec le support technique de RENATER et ainsi constitué un des premiers cercles de confiance « privé » de la Fédération d'Identité, la fédération Univcloud. Service provider du cercle de confiance privé Univcloud, le démonstrateur s'est ouvert aux 14 premiers établissements de l'UNR Paris Ile de France qui ont pu y mener leurs expérimentations.

3.5.5 La méthode EBIOS pour un « Dossier de la sécurité » d'Univcloud

Dans le cadre de la définition des objectifs, il a été décidé de s'appuyer sur la méthode EBIOS [119] 2010 conformément aux recommandations de l'ANSSI permettant ainsi de mettre un cadre précis autour du Dossier de sécurité, de la sécurité des données personnelles et sur l'implémentation des dispositifs de sécurité sur le démonstrateur. L'utilisation de la méthode EBIOS a été exprimée comme un pré-requis lors de la publication de l'appel d'offre à l'assistance à MOA suite à laquelle Infhotep a été retenue.

Conformément aux livrables attendus, les travaux conduits par le GT3 Confiance Numérique assisté par Infhotep se sont concentrés sur l'analyse de risques, l'identification des besoins de sécurité des métiers fonctionnels de l'UNPIdF, avec un focus particulier sur les risques associés aux données personnelles (Tactic) et sur les spécifications et l'intégration des mécanismes de sécurisation au démonstrateur (INEO).

Adaptation de la méthode au contexte d'appel d'offres de l'étude Univcloud

Le projet Univcloud est une étude R&D qui doit produire des spécifications en vue d'un appel d'offre. Les soumissionnaires devront proposer des solutions clefs en main de data center et un ensemble de solutions de sécurité répondant « aux exigences de sécurité » exprimées dans cet appel d'offre.

Voici un extrait du Chapitre 5 de l'étude :

> La présente étude vise principalement à définir les besoins de sécurité de l'Univcloud sur son activité « FrontOffice » et « BackOffice ». Toutefois, elle a été réalisée alors que le périmètre fonctionnel du « FrontOffice » de l'Univcloud (liste des services cibles) n'a pas été finalisé. Il est donc à prévoir une mise à jour de ce document lorsque le catalogue de services du cloud au niveau IaaS, SaaS et PaaS aura été finalisé. Cette étude ne vise pas à donner des solutions mais bien à identifier les besoins de sécurité et

119. Méthode EBIOS (Expression des Besoins et Identification des Objectifs de Sécurité), Logiciel et études de cas

les menaces auxquels le soumissionnaire d'Univcloud devra répondre. Pour autant, un ensemble de mesures inhérentes au fonctionnement d'un cloud sont rappelées en fin d'étude. Cette étude ne comporte pas les conclusions de l'analyse de risque car, dans le temps imparti, nous n'avons pu rapprocher l'ensemble des fonctionnalités essentielles aux biens supports afin d'identifier le niveau de gravité des risques.

Donc, il apparaît que la méthode EBIOS a pu être utilisée de manière très adaptable à un périmètre tronqué de certaines parties fonctionnelles qui n'ont jamais été recensées et décrites, dans un temps limité, par des incidents liés au contexte de partenariat public privé (défection d'un des partenaires) et ne nécessitant pas la dernière partie de la méthode, la nature de l'étude devant produire un cahier des charges en vue d'un appel d'offre. La méthode s'est même adaptée au fur et à mesure des aléas de l'étude en gardant sa cohérence et avec peu de perte dans la production du « Dossier de sécurité » final.

Plus haut dans le document (voir Figure 2.8), nous avons reproduit le memento fourni par l'ANSSI qui décrit la cinématique de la méthode EBIOS avec l'ensemble de ses modules. La Figure 3.14 donne le schéma dérivé décrivant la méthode adaptée à l'étude Univcloud.

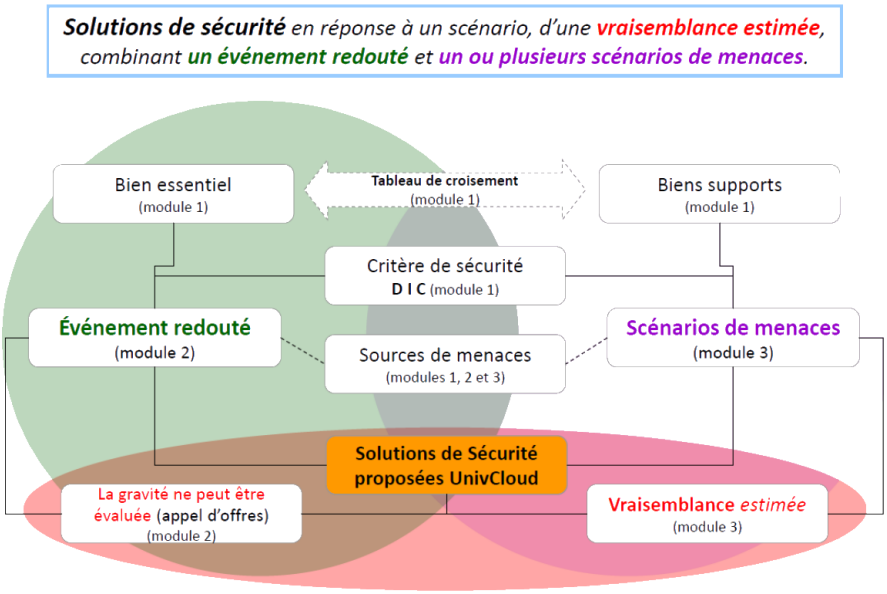


FIGURE 3.14 – Les solutions de sécurité - Mémento de la méthode adaptée pour les besoins du dossier de sécurité Univcloud dont le module 4 propose les solutions de sécurité pour l'appel d'offre en lieu et place d'un catalogue des risques caractérisés. En effet, la gravité n'a pu être évaluée avec la MOA et, en étant dans le cadre d'un appel d'offre et en absence d'une infrastructure existante, la vraisemblance n'a pu qu'être estimée avec des indications tirées de nos propres expériences et veilles technologiques.

Utilisation du logiciel EBIOS et retours à l'ANSSI

L'ANSSI a fait développer un outil qui permet de dérouler la méthode EBIOS. Véritable assistant à l'expression des besoins de sécurité, il accompagne l'étude dans toutes les phases en décrivant les attendus à chaque étape jusqu'à la production finale des documents finaux. Le logiciel EBIOS faciliteet permet la production documentaire dans les projets de petite taille.

3.5 Étude de la sécurité du cloud universitaire Univcloud

Dans le cadre d'Univcloud, nous avons accompagné son appropriation et son utilisation par le Laboratoire Tactic EA4420 sur le périmètre de la sécurité des données à caractère personnel et les risques juridiques. L'expérience a mis à jour quelques difficultés de performance et a fait l'objet d'un retour auprès de l'ANSSI. Cependant, sur des études de petite taille ou dans le cadre de formation, il est conseillé de l'utiliser afin de faciliter l'appropriation de la méthode.

Pour les personnes qui voudraient aller plus loin, le fichier d'installation du logiciel EBIOS 2010 ainsi que l'ensemble de la documentation dont l'étude de cas sur le cloud sont disponibles en téléchargement. [120]

120. Logiciel EBIOS, documentation, études de cas au format logiciel EBIOS

4. Conclusion et recommandations

4.1 Préliminaires

DANS CE DOCUMENT nous avons souhaité présenter de manière simple certains enjeux de l'hébergement des données à l'époque des centres de données, du cloud et des services hébergés. Nous sommes rentrés dans l'ère des services à la demande et de la location (leasing). Cela n'est pas propre au cloud. Nous voyons bien d'autres incarnations de ces principes : Autolib, le service de voitures électriques en bas de chez soi à Paris, Deezer le service de musique à volonté...

La question est plutôt de savoir si les personnes ou institutions pourront se passer à court terme de cette forme de mise en relation. Dans les domaines scientifiques, pour calculer, qui pourra se passer à terme d'un service comme celui de RosettaHub : "*A revolutionary cloud-native platform designed to empower data scientists, researchers and educators who need virtual, collaborative and traceable access to data science tools and infrastructures*" ? C'est un guichet unique qui permet la location de ressources chez tous les principaux fournisseurs de cloud !

Toujours est-il que nous avons défini et décrit différents écosystèmes, matériels, logiciels, humains, législatifs, techniques, fonctionnels qui font face aux problématiques de l'hébergement de données. L'objectif poursuivi est de donner envie aux lecteurs de s'intéresser de près aux problèmes et solutions qui ne relèvent pas toutes de techniques imposées d'en haut par le dieu tout puissant de l'informatique distribuée. Il est important de comprendre, pas à pas, les différents liens entre ces écosystèmes, leurs limites et périmètres d'action.

En fait de nombreux acteurs oeuvrent, pas toujours de manière concertée, à la réalisation de solutions d'hébergement dans lesquelles nous pouvons avoir confiance. Nous avons aussi donné des exemples de réalisation pour illustrer notre propos et ce qu'il est possible de faire très concrètement. Bien entendu, notre souhait serait de pouvoir présenter de nouveaux exemples, d'adresser de nouvelles facettes du problème. Cela devrait se faire dans le cadre d'une démarche d'amélioration continue de la qualité que nous souhaitons présenter maintenant. En effet, elle nous paraît indispensable à mettre en œuvre, dans toute institution, entreprise ou eco-système, afin de piloter les changements et les évolutions induits par l'hébergement des données.

4.2 La méthode d'amélioration continue de la qualité

4.2.1 Idées générales

Considérons un objet (projet de développement informatique, travaux de recherche, ville, une base de données, un laboratoire...), un ensemble d'actions pour le faire évoluer (écriture de code, expérimentation pour vérifier des hypothèses, agrandissement des trottoirs et création de zones piétonnes, création de nouvelles requêtes, mise en place d'un contrôle d'accès pour la zone à régime restreint, ...). Nous appelons ces considérations le « PLAN », la première phase.

Dans un deuxième temps, nous exécutons cet ensemble d'actions qui aura été porté par écrit pour les besoins de lisibilité, de traçabilité et surtout de suivi des actions, c'est le « DO », la deuxième phase.

Dans un troisième temps, il convient de vérifier la portée de ces actions et de mesurer leurs impacts. Afin de mesurer, il est essentiel que l'objet sur lequel s'applique ces actions soit muni d'indicateurs. Les indicateurs nous permettent de connaître les évolutions qu'auront apportées les actions. La vérification des évolutions à l'aide des éventuels indicateurs s'appelle le « CHECK ». C'est la troisième phase qui va nous permettre de tirer le bilan des actions. Elle nous permet de détecter les éléments que nous n'avons pas été en mesure d'identifier dans un premier temps lors de la phase « PLAN ». L'exécution des actions nous permet de découvrir des aspects insoupçonnés, des éléments adjacents au périmètre qui semblent a posteriori intéressants d'inclure. Le bilan de la phase « CHECK » est mis par écrit toujours pour des raisons de traçabilité et pour validation.

Dans un quatrième temps, dans la phase appelée « ACT », nous prenons en compte le résultat et les nouveaux éléments découverts pendant la phase «CHECK» ou éventuellement pendant l'exécution des actions. Il arrive souvent de découvrir des aspects non identifiés préalablement pendant l'exécution des actions que nous avons planifiées. Il convient donc de valider les éléments portés par écrit de la phase « CHECK » afin de les inclure dans la prochaine phase « PLAN ». La validation est nécessaire, c'est un engagement contractuel entre les parties. En déroulant une nouvelle phase « PLAN », nous rebouclons le processus ce qui réalise un cycle d'amélioration continue.

Ce procédé de mise en œuvre et de suivi des actions est un mécanisme d'amélioration continue propre à la culture « Qualité », il peut être illustré par la « roue de Deming » [121].

4.2.2 La norme de la gestion de la qualité

La culture de la qualité a sa propre norme. La norme de la gestion de la qualité est la norme ISO 9000. Elle est devenue courante dans de nombreuses organisations. Il est intéressant de noter que les mécanismes de la norme de gestion de la qualité font partie d'autres normes. Ainsi, une organisation qui a adopté la norme ISO 9000 et sa culture n'aura aucun mal à incorporer d'autres normes qui s'appuient sur des processus d'amélioration continue de la qualité. C'est le cas de la norme 27000 de la gestion des risques sur l'information. Une organisation peut envisager de mettre en place un service qualité, conformité et gestion des risques avec une culture de base commune et un langage commun compris et partagé par tous.

4.2.3 Amélioration continue de ce document

Pour être cohérent avec nous même et en guise de clin d'oeil, nous avons choisi d'appliquer la méthodologie de la norme de gestion de la qualité à ce document. Cela donne les phases suivantes :

121. Voir une illustration du concept sur https://fr.wikipedia.org/wiki/Roue_de_Deming

- PLAN : identifier le périmètre, la structure, les objectifs et les éléments d'information qui vont constituer ce document ;
- DO : écrire le document et articuler les contributions des différents auteurs afin de respecter sa structure et les objectifs. Pendant sa rédaction, des éléments qui vont au-delà du périmètre initial ont été identifiés ou rédigés. Ils ont été remisés pour être incorporés ultérieurement ;
- CHECK : suite à sa diffusion, nous devrons prendre en compte les remarques qui nous serons faites afin d'identifier quelles améliorations pourront être apportées au document ;
- ACT : lors d'une prochaine réunion du comité de rédaction devra être décidé quelles améliorations seront prises en compte dans la planification et l'écriture d'une nouvelle version du document.

4.3 Recommandations

POUR TERMINER nous souhaitons reprendre en grande partie les conclusions du rapport d'enquête sur les processus métiers [122] des personnels de USPC tout en étendant leurs portées. En effet ces recommandations restent d'actualité pour aborder les questions démystifiées dans le rapport. Elles concernent d'une part les acteurs de la recherche (principalement les chercheurs) et les personnes décisionnaires en matière d'organisation de la recherche d'autre part.

4.3.1 Recommandations aux acteurs de la recherche

➢ Améliorer la collaboration entre acteurs de la recherche en utilisant des services de collaboration numériques, à tous les stades du processus de recherche, et en rendant accessibles des ressources (logiciels, codes…). L'accès d'un point de vue informatique doit être le plus simple possible pour l'utilisateur ;

➢ S'interroger sur les évolutions du numérique ;

➢ Adopter une réflexion sur le cycle de vie des données (ce que deviennent les données de leur naissance à leur mort) et d'un point de vue plus général sur un plan de la gestion des données (document formalisé explicitant la manière dont seront obtenues, documentées, analysées, disséminées et utilisées les données produites au cours et à l'issue d'un processus ou d'un projet de recherche) ;

➢ Adopter une démarche pro-active dans la recherche de méthodes/outils informatiques déjà existants dans son domaine (voir ce qui se fait dans d'autres domaines que le sien) ;

➢ Porter une plus grande attention aux politiques de collecte et de partage des données des prestataires privés d'hébergement, et s'informer sur les procédures de migration entre prestataires (quelles démarches et conséquences associées à un changement de fournisseur ?).

4.3.2 Recommandations aux décideurs

➢ Sensibiliser les publics aux évolutions du numérique. Recruter des personnes pour assurer la médiation entre l'informatique et les autres disciplines ;

122. Rapport d'enquête CIRRUS sur les processus métiers

- ➢ Impliquer les bibliothèques comme soutiens aux enseignants/chercheurs pour la gestion et le partage des données de recherche. Les bibliothèques peuvent apporter leurs expériences en matière de gestion des méta-données et de préservation des données sur le long terme ;

- ➢ Améliorer les services numériques déjà existants en automatisant davantage la gestion et l'administration, et en créant de nouveaux services sur la base des besoins des utilisateurs. Ces derniers doivent être au centre des préoccupations des développeurs informatiques ;

- ➢ Élaborer des parcours de formation qui permettront aux publics d'une institution d'évoluer vers des emplois porteurs ; cela inclut des formations aux usages, y compris pour les doctorants.

- ➢ Inciter les décideurs et l'ensemble des partenaires de la communauté universitaire à constituer un observatoire permanent pour analyser l'évolution des technologies et des usages du numérique. Analyser ce qui se passe et ce qui s'est passé à l'Institut Société Numérique ou au Center for Data Science à Paris-Saclay, en veillant tout particulièrement à la dissémination des bonnes pratiques et à la vulgarisation. Une autre forme de structure travaillant sur le plan de l'ingénierie du numérique pourrait être un Centre d'Étude du numérique ayant pour missions d'enquêter, d'analyser les pratiques et les usages afin de proposer aux décideurs des pistes de développement, et aux autres publics, des moyens correspondant à leurs besoins. Cette structure aurait vocation à animer la réflexion (organisation d'évènements), produire des notes synthétiques expliquant le plus simplement possible les enjeux des technologies, et à travailler avec tous les services des universités compétents dans leurs domaines respectifs (valorisation, relations internationales, réseau des bibliothécaires. . .) ;

- ➢ Cette ingénierie devrait être couplée à des actions/programmes de recherche, en particulier sur les Systèmes (vus ici comme les couches entre le matériel et les applications) pour assurer, d'une part, une meilleure veille collective sur l'évolution et la convergence des Systèmes et d'autre part faire émerger des solutions compétitives sur le plan international. Les actions/programmes de recherche pourront être inter-disciplinaires afin de faire travailler ensemble plusieurs équipes, donc plusieurs chercheurs, sur un/des Systèmes, et plus uniquement sur le seul plan des applications. Il s'agira de passer d'une logique « un chercheur, une application » à une logique « des équipes, un Système informatique » ;

- ➢ Favoriser la reconnaissance de compétences inter-disciplinaires. Ces compétences sont particulièrement nécessaires dans le cadre des structures proposées aux points précédents. Les compétences peuvent être de nature juridique, en informatique technique, de dissémination et d'animation de la recherche et être liées au travail des bibliothécaires. Le but est de rentrer dans une démarche d'amélioration continue de la qualité pour le numérique.

Index

M

P

S

V

W