La Boîte à outils

de la

Stratégie big data

Romain Rissoan
Romain Jouin

DUNOD

Maquette de couverture : Caroline Joubert
Photo de la boîte : © Mega Pixel
Pictos de couverture :
© bioraven-Shutterstock.com
© Eliricon from Noun Project
© I Putu Kharismayadi from Noun Project

Mise en page : Belle Page

© Dunod, 2018
11 rue Paul Bert, 92240 Malakoff
www.dunod.com

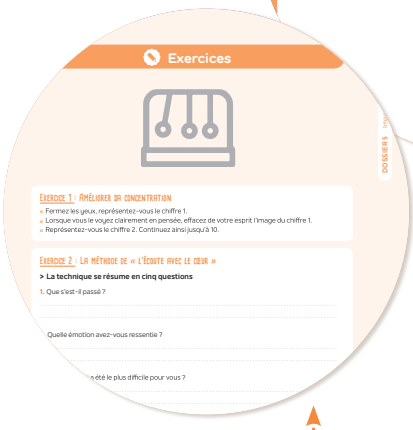
978-2-10-077898-0

VOUS AUSSI, AYEZ LE RÉFLEXE

Boîte à outils

Des outils classés par dossiers thématiques

Une présentation visuelle de chaque outil



Des exemples, cas ou exercices pour approfondir

La Boîte à outils

DES OUTILS OPÉRATIONNELS TOUT DE SUITE

MEGA Boîte à Outils

Manager leader - 100 outils
Coordonnée par Pascale Bélorgey, Nathalie Van Laethem

Digital - 100 outils
Coordonnée par Catherine Lejealle

MÉTIERS

Acheteur, *3e éd.*
Stéphane Canonne, Philippe Petit

Assistante, *2e éd.*
Christine Harache, Hélène Tellitocci

Auditeur financier, *2e éd.*
Sylvan Boccon-Gibod, Eric Vilmint

Chef de produit, *2e éd.*
Nathalie Van Laethem, Stéphanie Moran

Chef de projet, *2e éd.*
Jérôme Maes, François Debois

Coach en entreprise, *2e éd.*
Belkacem Ammiar, Omid Kohneh-Chahri

Commercial, *3e éd.*
Pascale Bélorgey, Stéphane Mercier

Community Manager
Clément Pellerin

Comptabilité, *2e éd.*
Bruno Bachy

Consultant, *2e éd.*
Patrice Stern, Jean-Marc Schoettl

Contrôle de gestion
Caroline Selmer

Création d'entreprise, *2018*
Catherine Léger-Jarniou, Georges Kalousis

E-commerce
Christian Delabre

Formateurs, *3e éd.*
Fabienne Bouchut, Isabelle Cauden, Frédérique Cuisiniez

Management, *2e éd.*
Patrice Stern, Jean-Marc Schoettl

Micro-entrepreneur
Jacques Hellart, Caroline Selmer

Pilote des systèmes d'information, *2e éd.*
Jean-Louis Foucard

Publicité
Servanne Barre, Anne-Marie Gayrard-Carrera

Responsable communication, *3e éd.*
Bernadette Jézéquel, Philippe Gérard

Responsable financier, *2e éd.*
Caroline Selmer

Responsable marketing omnicanal, *3e éd.*
Nathalie Van Laethem, Béatrice Durand-Mégret

Responsable qualité, *3e éd.*
Florence Gillet-Goinard, Bernard Seno

Ressources humaines, *2e éd.*
Annick Haegel

Santé - Sécurité - Environnement, *3e éd.*
Florence Gillet-Goinard, Christel Monar

TPE
Guillaume Ducret

COMPÉTENCES TRANSVERSALES

Conduite du changement
David Autissier, Jean-Michel Moutot

Créativité, *2e éd.*
François Debois, Arnaud Groff, Emmanuel Chenevier

Design management
Bérangère Szostak, François Lenfant

Développement durable et RSE
Vincent Maymo, Geoffroy Murat

Gestion des conflits
Jacques Salzer, Arnaud Stimec

Innovation, *2e éd.*
Géraldine Benoit-Cervantes

Intelligence collective
Béatrice Arnaud, Sylvie Caruso-Cahn

Intelligence économique
Christophe Deschamps, Nicolas Moinet

Lean
Radu Demetrescoux

Leadership, *2e éd.*
Jean-Pierre Testa, Jérôme Lafargue, Virginie Tilhet-Coartet

Management de la relation client, *2e éd.*
Laurence Chabry, Florence Gillet-Goinard, Raphaëlle Jourdan

Management transversal
Jean-Pierre Testa, Bertrand Déroulède

Marketing digital
Stéphane Truphème, Philippe Gastaud

Mind mapping
Xavier Delengaigne, Marie-Rose Delengaigne

Mon parcours professionnel
Florence Gillet-Goinard, Bernard Seno

Négociation, *2e éd.*
Patrice Stern, Jean Mouton

Organisation, *2e éd.*
Benoît Pommeret

Prise de décision
Jean-Marc Santi, Stéphane Mercier, Olivier Arnould

Réseaux sociaux, *4e éd.*
Cyril Bladier

Sécurité économique
Nicolas Moinet

Stratégie, *2e éd.*
Bertrand Giboin

Stratégie digitale omnicanale
Catherine Headley, Catherine Lejealle

Supply chain
Alain Perrot, Philippe Villemus

DÉVELOPPEMENT PERSONNEL

Bien-être au travail
Clothilde Huet, Gaëlle Rohou, Laurence Thomas

Confiance en soi
Annie Leibovitz

Développement personnel en entreprise
Laurent Lagarde

Efficacité professionnelle
Pascale Bélorgey

Gestion du stress
Gaëlle du Penhoat

Gestion du temps
Pascale Bélorgey

Intelligence émotionnelle
Martine-Eva Launet, Céline Peres-Court

Marketing de soi
Nathalie Van Laethem, Stéphanie Moran

Motivation
Sophie Micheau-Thomazeau, Laurence Thomas

Pleine conscience au travail
Sylvie Labouesse, Nathalie Van Laethem

Avant-propos

> Un bon data scientist est intéressé par la résolution de problèmes, pas par de nouveaux outils.
>
> KDNuggets

L'ouvrage que nous vous proposons est dédié au domaine de la data. Il a pour ambition de dresser un panorama des notions, concepts et outils les plus éprouvés sur le marché actuellement. Sans être technique, il vous permettra de connaître les cas d'usage des différents outils et comment ils se distinguent les uns des autres.

Le monde de la data

En 2012, la *Harvard Business Review* (*HBR*) publiait un article : « Data-scientist, the sexiest job of the 21st century ». Depuis, tout le monde veut être data scientist : le gouvernement français a fait de la data science un des piliers de la « nouvelle France industrielle » en 2015, la BPI finance à tour de bras des start-up proposant du machine learning ou du big data et la récente étude « France IA » montre à quel point la France a une carte à jouer dans ce nouveau domaine de l'analyse des données, mais que la partie n'est pas gagnée.
Ce que la *HBR* avait oublié de nous dire, c'est que data scientist est un travail compliqué. Un data scientist doit avoir 3 compétences :

1. informatique
2. mathématique
3. business

Complétées par une excellente capacité à communiquer, c'est-à-dire transmettre ces trois compétences majeures à un auditoire qui ne les a pas.
Ce contexte rend la tâche du data scientist quasi impossible à atteindre. C'est pourquoi nous travaillons en équipe, avec chacun ses compétences. Avec ce livre, vous apprendrez le vocabulaire nécessaire au travail en équipe indispensable dans le monde de la data. Vous pourrez ainsi mettre vos compétences particulières à disposition de vos collègues.

Ce livre est avant tout un livre sur la « stratégie big data » : aussi, ses quatre premiers dossiers s'attachent à expliciter les enjeux qui se cachent derrière ce phénomène :

- Qu'est-ce qui a changé ces dernières années vis-à-vis de la data ?
- Quels sont les différents types de data ?
- Comment en tirer parti pour la stratégie de son entreprise ?
- Quels impacts techniques ?

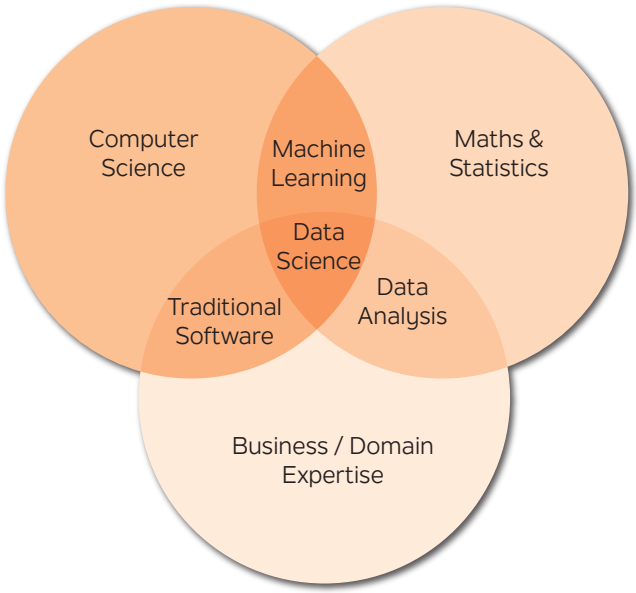
Ces dossiers intéresseront plutôt le manager avide de comprendre comment profiter de cette opportunité qu'est le big data.

Viennent ensuite trois dossiers plus opérationnels, allant de la gestion de projets data à l'introduction des concepts de machine learning, et la différence avec les concepts de big data. Les personnes récemment nommés chef de projets dans un environnement big data y trouveront les clefs pour communiquer et comprendre les équipes avec lesquelles elles devront travailler.

Le dernier dossier, sur le passage en production, est une introduction aux différentes technologies usuellement mises en place dans les infrastructures big data. Il permet d'en comprendre les concepts, sans rentrer dans le détail technique opérationnel en tant que tel.

Ainsi, les différents chapitres sont relativement indépendants, et s'adressent à un public de managers, ou chefs de projet, curieux du phénomène « big data ».

Classiquement, la data science nécessite six expertises distinctes, que l'on cherche chez le data scientist. Évidemment, il est bien difficile de disposer de l'ensemble, c'est pourquoi les projets sont généralement effectués en équipe. Il faut alors s'assurer de partager le même vocabulaire pour que le projet avance correctement.



Les compétences des data scientists

Sommaire

1 DOSSIER

L'ÈRE DE LA DATA

par Romain Rissoan

“

C'est une erreur capitale que de bâtir des théories tant qu'on n'a pas de données. Insensiblement, on se met à torturer les faits pour les faire cadrer avec les théories, au lieu d'adapter les théories aux faits.

Arthur Conan Doyle (créateur de Sherlock Holmes)

Le V du sens des données	Le V de la diversité des données	Le V de la puissance de données
☐ Valeur	☐ Variété	☐ Volume
☐ Véracité	☐ Variabilité	☐ Vitesse
☐ Viabilité		☐ Vélocité
☐ Validité		☐ Visualisation
		☐ Vulnérabilité
		☐ Volatilité

Les *a priori* sur le big data sont nombreux. D'abord, nous sommes tentés de considérer que ce sujet est celui du Directeur des Systèmes informatiques ou du moins des informaticiens. Puis, nous sommes tentés de confier ce sujet à un mathématicien. Celui-ci nous ramène alors à la question de ce que nous voulons et de nos objectifs. C'est alors qu'intervient le Stratège, le dirigeant, le manager, ou le chef de projet afin de poser les perspectives d'un projet. Or ce dernier, bien souvent, a besoin de compréhension technique alors que celle-ci est bien souvent structurée à l'envers : en big data on exploite souvent les données dont on dispose et rarement les données que l'on a souhaité obtenir, car la mise en place de collecte de données en masse prend beaucoup de temps, tant sur le plan technique que réglementaire.

Comprendre le sujet

Ce sujet très populaire qu'est le big data est soumis à de multiples discours de synthèse qui mènent à de nombreuses simplifications en tout genre. Dans le même temps, nous avons besoin de cette simplicité pour avancer et prendre des décisions.

Mais globalement, il y a de nombreux manques de compréhension de points fondamentaux. Par exemple sur la définition même de ce qu'est une donnée.

Aussi, il s'agit d'avoir une approche à la fois inductive et déductive sur le sujet pour comprendre celui-ci dans sa profondeur et sa complexité.

Comprendre l'avenir de la data

Pour beaucoup, le big data ne veut rien dire et ne prend pas de sens. Ceci pour deux raisons : d'abord nous n'avons pas suffisamment de données pour s'imaginer les conséquences de l'entrée dans l'univers big data. Ensuite, il s'avère que certains d'entre nous font du big data sans le savoir.

Dans ce contexte, il convient de comprendre le monde de demain, représenté par des acteurs majeurs de la data (Facebook, Amazon), de leur gestion de celle-ci et de leurs traitements des données. Ce sont de parfaits pionniers de cet univers, qui nous aident à concevoir le fonctionnement du monde de demain.

Les outils

Outil 1

La data (la donnée)

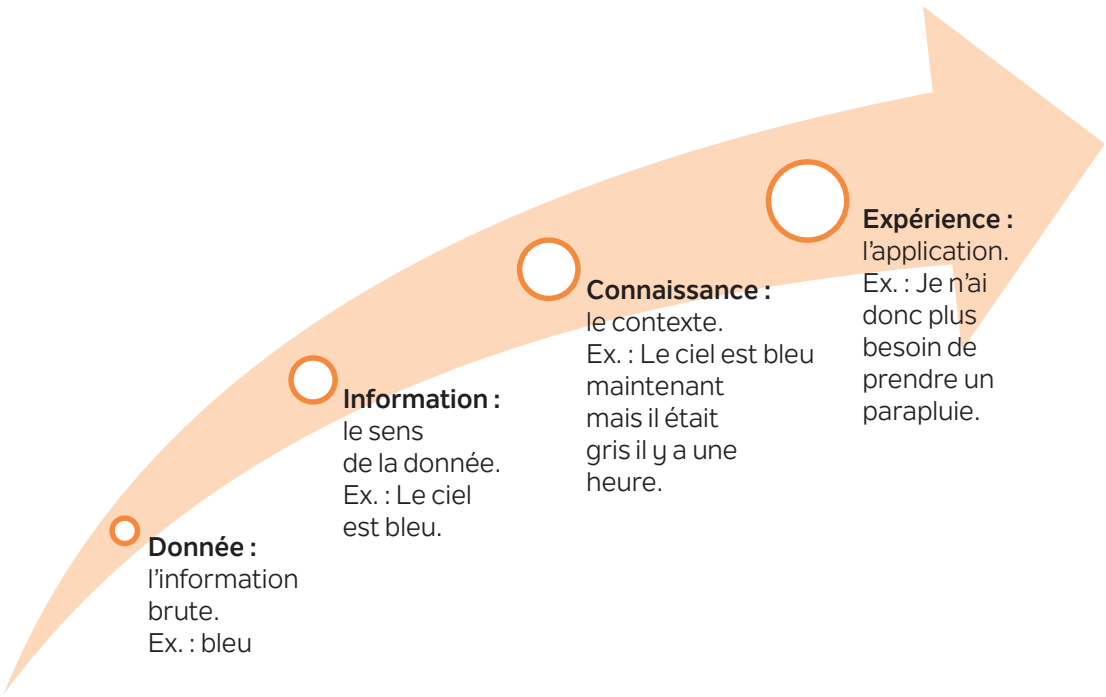
par Romain Rissoan

> *La rétention de l'information est une forme de constipation du savoir.*
> **Théophraste Renaudot**

En quelques mots

La donnée est l'information élémentaire dont nous avons besoin pour créer une cascade d'informations.
Plus nous avons de données, plus nous pourrons créer des informations fiables et de qualité. Et, plus nous travaillerons proche de la donnée, moins il y aura de distorsion et plus nous appliquerons des méta-analyses plutôt que des analyses.

LA DATA (LA DONNÉE)



POURQUOI L'UTILISER ?

Objectif

Apprendre à travailler avec des données et pas nécessairement avec des fichiers.

Contexte

La représentation ainsi que la valeur perçue d'une donnée ont bien évolué depuis la naissance de l'informatique.
Sur le plan informatique, des séquences de 0 et de 1 sont des **données** ou **data** (en anglais). Suite à une interprétation, ces données prennent un sens exploitable pour l'homme et pour la machine au travers de fichiers que l'on appelle alors **information**.
Sur le plan utilisateur, la **donnée** serait par exemple un nom, un prénom, une date de naissance, un sexe, alors que l'information serait un âge (calculé sur la date de naissance), le nombre de fois où cette personne s'est connectée à votre site Internet.
En résumé, une information est le résultat du traitement d'une donnée. Mais le traitement d'une donnée donne naissance à une nouvelle information.
Depuis que l'informatique s'est accélérée, ranger ses e-mails et ses fichiers Excel dans son ordinateur est moins efficace.
On doit travailler le plus en amont possible de ce flux. Ainsi, on parle plus de gérer ses données (ses flux d'informations) que de gérer ses fichiers (ses stocks d'information).

COMMENT L'UTILISER ?

Étapes

1. Prenez l'exemple d'un fichier Excel que vous partagez avec 10 collaborateurs sur un projet à fortes interactions. Rangez ce fichier dans un dossier dans votre ordinateur.
2. Déposez-le sur un espace de travail collaboratif comme Google Drive (ou Excel 365) pour pouvoir travailler de manière interactive sur ce fichier avec vos autres collaborateurs.
3. Créez un second fichier de type tableur dont les données évoluent en fonction du premier fichier.
4. Vous constaterez donc qu'à chaque fois que le fichier initial est modifié par un de vos collaborateurs, l'autre est modifié. Vos données vivent toutes seules. Vous êtes ainsi concentré sur des données plus que sur des fichiers.

Méthodologie et conseils

Ce mode de fonctionnement peut être nouveau pour certaines personnes. Aussi, il est important de manager le changement et non pas d'admettre que ce mode de fonctionnement est « La » vérité.
Certains arguments sont justifiés : une utilisation épisodique de la donnée (une personne va voir le fichier une fois par an et a besoin de l'avoir bien rangé dans ses dossiers pour être certain de ne pas le perdre), l'absence potentielle de connexion à Internet (si je n'ai plus Internet, je n'ai plus accès à mes données) et le besoin de maîtrise de la donnée (si tout le monde peut modifier à n'importe quel moment mon fichier, je peux ressentir une perte de sécurité).
Dans ce contexte, il convient d'opérer ponctuellement et durablement des ateliers de transformation digitale pour inciter les utilisateurs à travailler autrement et ainsi permettre la migration vers la data.

Avant de vous lancer...

- ✓ **Toute donnée peut donner une information. Mais toute information peut donner naissance à une autre information.**
- ✓ **Le plein potentiel d'une donnée est préservé lorsqu'elle reste à l'état brut.**

Outil 2

Le sens de la donnée (donnée brute)

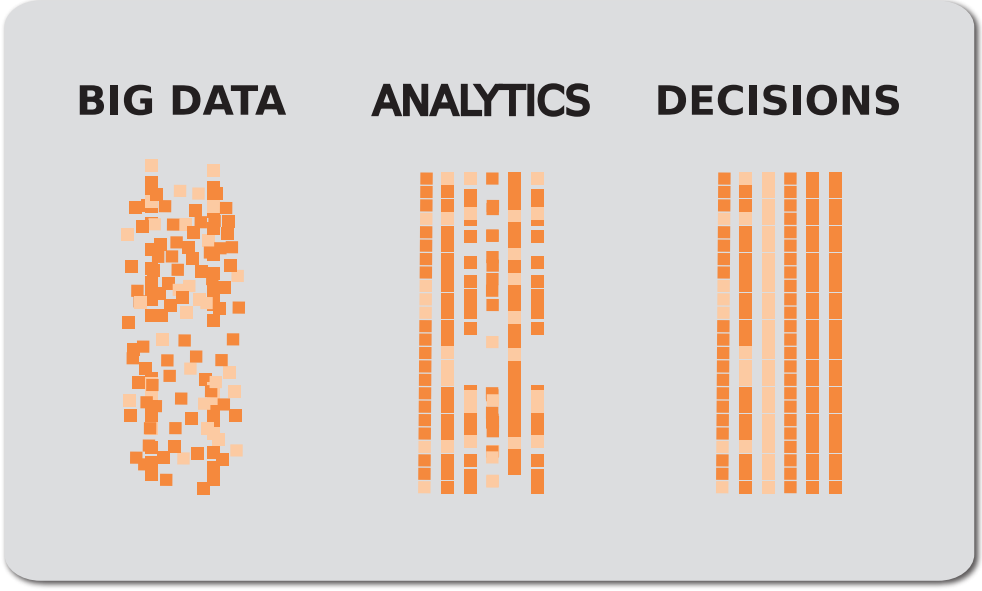
par Romain Rissoan

> *La donnée est une chose précieuse qui restera plus longtemps que les systèmes.*
> **Tim Berners-Lee**

En quelques mots

Le point de départ d'un projet big data est bien évidemment la data. Or, le sens même de ce que peut représenter la data n'est pas évident pour une organisation ou plutôt pour la culture de cette organisation. En effet, encore aujourd'hui, il n'est pas rare de constater que la majorité des structures considère que l'expression de besoin client peut se résumer à des informations écrites sur une feuille de papier blanc ajouté à la mémoire que le commercial en a retenu, ce qui finira au mieux en synthèse sur un CRM (un logiciel de gestion de relation client), au pire en un devis perdant ainsi toute trace de l'expression du besoin client.

LE SENS DE LA DONNÉE



POURQUOI L'UTILISER ?

Objectif

Comprendre le véritable sens et intérêt d'une donnée à l'état brut.

Contexte

Au quotidien, nous travaillons principalement sur la base d'informations plus que de données. Fichiers Word, Excel, e-mails, tout ce que nous utilisons est en soi le résultat d'un traitement, d'une interprétation de nos données, les rendant ainsi orientées selon une analyse prédéfinie et une décision également prédéfinie.
L'idée du big data est justement de préserver la donnée à son état le plus pur afin de permettre des analyses différentes et des décisions innovantes.

COMMENT L'UTILISER ?

Étapes

1. Prenez votre appareil photo habituel.
Activez la prise de vue au format RAW (en plus du format .jpg habituel).
2. Prenez une photo au hasard.
Vous obtiendrez alors une photo à l'état brut (non compressée) au format .CR2 (fichiers appelés comme étant de type RAW).
Cette photo est la version non compressée et non déformée de votre prise de vue.
Elle est plus lourde que votre fichier .jpg habituel. Vous avez des exemples ici : https://www.imaging-resource.com/PRODS/T2I/T2ITHMB.HTM
3. Lancez le logiciel Photoshop ou l'application https://raw.pics.io/app et ouvrez le fichier .CR2 avec l'application choisie.
4. Vous pourrez alors constater toutes les informations techniques lors de la prise de vue, comme le temps d'ouverture de l'objectif, les coordonnées GPS du lieu de la prise de vue, la focale, etc.
De plus, et surtout, vous pourrez alors modifier la photographie depuis son format d'origine et non depuis son expression au format .jpg, qui est un format compressé et déformé de la photo originelle.

Méthodologie et conseils

Ayez toujours à l'esprit qu'une donnée telle qu'elle est exploitée aujourd'hui pourrait être exploitée différemment demain. Ainsi, plus vous stockerez de données à l'état brut, plus vous préservez l'avenir de celles-ci.
Ainsi :

- si vous traitez des fichiers audio, stockez des fichiers .FLAC,
- pour les photos stockez des .CR2,
- pour des fichiers texte stockez du .DOC, .CSV, .TXT ou .DOCX (mais pas du PDF).

Grâce à cela, vous préservez les fichiers originels et leurs métadonnées associées.
Ces fichiers occuperont évidemment plus de place sur vos supports de stockages et surtout vos serveurs. N'hésitez pas à utiliser des offres low cost sur Internet.

Suite outil 2 →

Avant de vous lancer...

- ✓ **Pour faire du big data à long terme, stockez les données à l'état brut.**
- ✓ **Grâce à cette approche, vous disposerez d'une capacité d'analyse plus large et donc d'une capacité de décision plus importante.**

Cas d'entreprise

Photo ABC SARL

De l'argentique au numérique

Pendant des années, nous avons pris des photos avec des appareils photos argentiques. Les négatifs étaient alors transformés en photographie papier, puis rangés avec plus ou moins de soin dans des pochettes en vue d'un stockage long. Puis, nous avons découvert un nouveau type de donnée, le format numérique. Les appareils photos prenaient alors des photos dans un format nommé BITMAP, rapidement abandonné pour laisser place à un format de donnée appelé JPEG, qui est un format compressé de la donnée photo pour pouvoir permettre le stockage de celle-ci. Par la suite, et après de nombreuses années, est arrivé le format RAW qui est un stockage à l'état brut de la donnée, permettant de nombreuses retouches mais prenant une place conséquente sur nos disques durs. Ce dernier format permet à n'importe quel logiciel de photo de recréer toutes les couleurs sur une photo existante. Ainsi il est par exemple possible de redessiner la couleur verte d'une prairie sur ces nouveaux formats.

Data et métadonnées

La société Photo ABC SARL a construit son modèle économique sur la vente de photos dans de nombreuses tailles : photos de portrait, posters, panneaux publicitaires. Pendant des années, ses photos ont été prises en argentique et elle a donc stocké des milliers et des milliers de négatifs dans des entrepôts. Elle possède toujours ces négatifs et a bien tenté de les numériser, mais les rendus sont dégradés et ne permettent pas des opérations professionnelles. À l'époque, on pouvait trouver une image grâce au moteur de recherche et l'arborescence des dossiers.

Par la suite, après avoir rejeté les formats BITMAP, elle a pendant de nombreuses années travaillé sur des photos numériques au format JPEG. Ces formats ont été stockés dans des dizaines de disques durs de plusieurs giga-octets mais le format compressé de ces fichiers JPEG a causé quelques pertes de fichiers rendant inexploitables environ 10 % de ceux-ci.

De plus, ces photos ne contiennent pas toutes les informations que l'on retrouve de nos jours dans les fichiers de type photo, et que l'on nomme les métadonnées : auteur de la photographie, marque de l'appareil qui a pris la photographie, date de prise de vue, etc.

En utilisant les métadonnées, il est désormais possible de rechercher des photographies également en recherchant par auteur, marque de l'appareil.

Depuis cinq ans, la société travaille désormais exclusivement sur la base de fichiers RAW. Ces fichiers pèsent environ 50 Mo, soit dix fois plus que les fichiers JPEG. Ils contiennent désormais les coordonnées GPS du lieu de la prise de vue et toutes les données numériques pour reconstruire la colorimétrie de la photographie de la manière la plus adaptée possible, selon les besoins. Ainsi, une même photo prise en plein jour peut être retravaillée pour donner un effet crépuscule. Ces fichiers RAW sont ensuite convertis en fichiers JPEG. Les fichiers JPEG étant toujours de taille importante, sont ensuite réduits et dépourvus de certaines données stratégiques comme les coordonnées GPS afin d'être diffusées aux clients en guise d'échantillon ou de vente directe.

Le stockage de la data

La société Photo ABC SARL doit donc évidemment stocker ses fichiers au format le plus riche possible, le format RAW, afin de permettre de puissants traitements ultérieurs. Grâce à ce stockage, certes coûteux, elle peut désormais solliciter par exemple un moteur de recherche sur la base d'une carte géographique ou bien même d'une reconnaissance de photo.
Ainsi en tapant par exemple « PACA » ou « mer » dans son moteur de recherche, elle pourra trouver toutes les photos prises en région PACA ou avec la mer présente sur la photo ou à proximité de la prise de vue.

Compte tenu de ses contraintes de stockage et du niveau de risque que cela représente pour la société de tout stocker en local (dans ses locaux), elle décide alors de stocker ses photos sur un serveur distant (on parle de cloud privé).
Puis, se rendant compte de la complexité de le gérer elle-même, elle le stockera sur un serveur distant appartenant à une société bien connue du grand public, Microsoft, sous la marque Azure (on parle alors de cloud public).
Pour cause, elle produit deux fois plus de photographies que cinq ans auparavant, chaque photo faisant une taille cinq fois plus importante et elle ne souhaite pas rajouter des disques durs sur ses serveurs et faire des copier/coller de tera-octets de données.