CENTRALE DIGITAL LAB
« BIG-DATA »

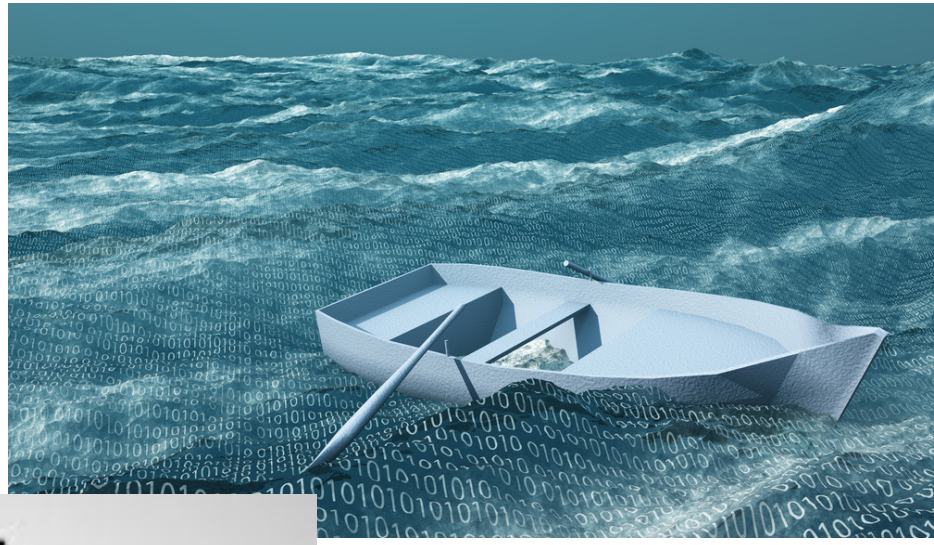
Stockage, Exploitation et Valorisation des Données Massives

Stéphane Derrode, ECL / Dpt MI
stephane.derrode@ec-lyon.fr

CENTRALE DIGITAL LAB
« BIG-DATA »

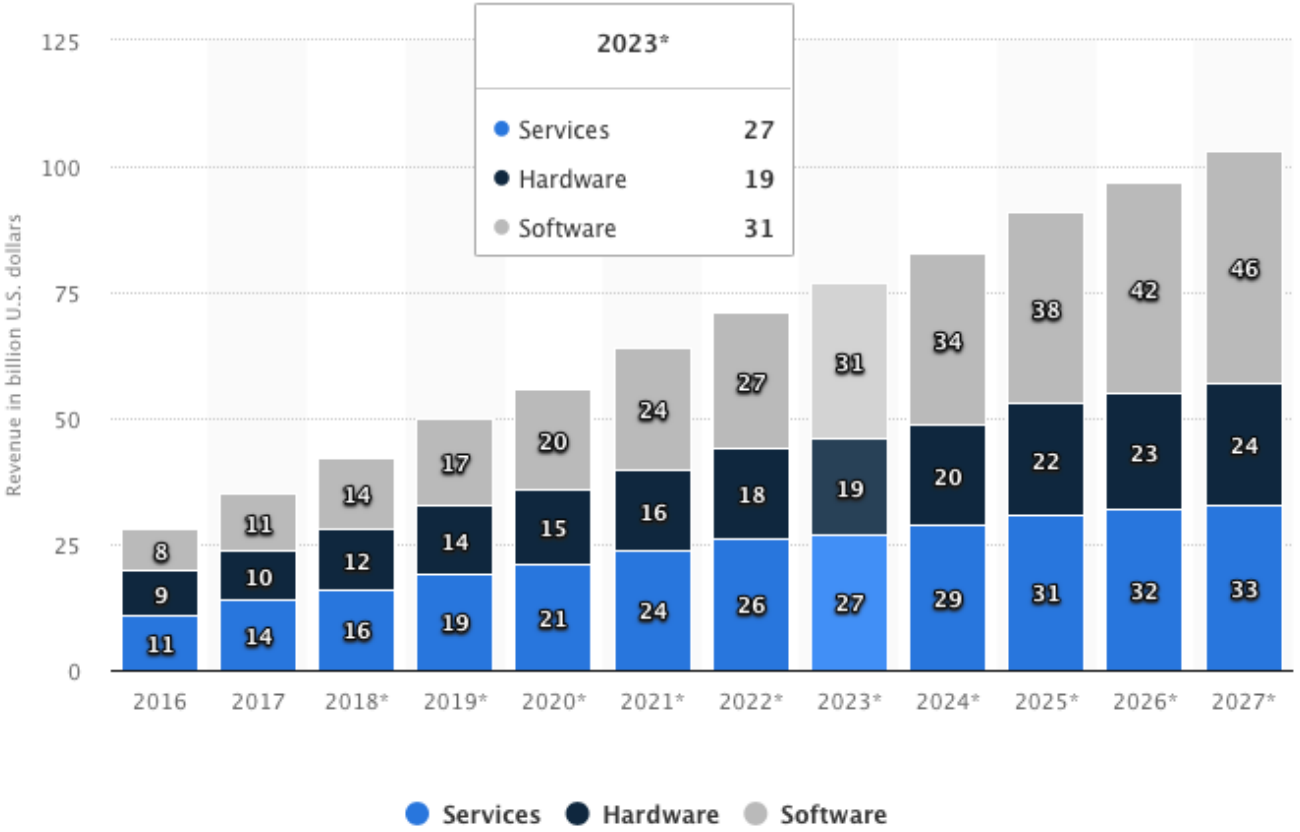
Partie I - Introduction au Big Data

Business Intelligence – « La Data »



Business Intelligence

sources



Harvard Businesss Review :

Data Scientist: The Sexiest Job of the 21st Century

Risques

Intelligence artificielle : les géants du Web lancent un partenariat sur l'éthique

http://www.lemonde.fr/pixels/article/2016/09/28/intelligence-artificielle-les-geants-du-web-lancent-un-partenariat-sur-l-ethique_5005123_4408996.html

Déluge de données

Définition Big Data (Gardner, 2001)

Volume : Analyser de larges volumes de données par des techniques d'apprentissages (*machine learning*), de la fouille de données (*data mining*), de l'intelligence économique (*business intelligence*) et du calcul haute performance.

A titre d'exemple, *Twitter* généré en janvier 7 teraoctets de données chaque jour en janvier 2013, et *Facebook* 10!

Infographie sur les volumes : http://future.arte.tv/fr/big-data-v-le-monde-ce-disque-dur/infographie-du-bit-au-yottaoctet

Vélocité : La vélocité représente à la fois la fréquence à laquelle les données sont générées, capturées et partagées et mises à jour (bourse et *trading* haute fréquence et robots).

Variété : Il ne s'agit pas de données relationnelles traditionnelles, ces données sont brutes, semi-structurées voire non structurées. Ce sont des données complexes provenant du web (Web mining), au format texte (Text Mining) et images (Image Mining). Ce qui les rend difficilement utilisables avec les outils traditionnels.

Définition Big Data +

+ **Valeur :** Valeur ajoutée économique, sociétale, sanitaire, scientifique des données.

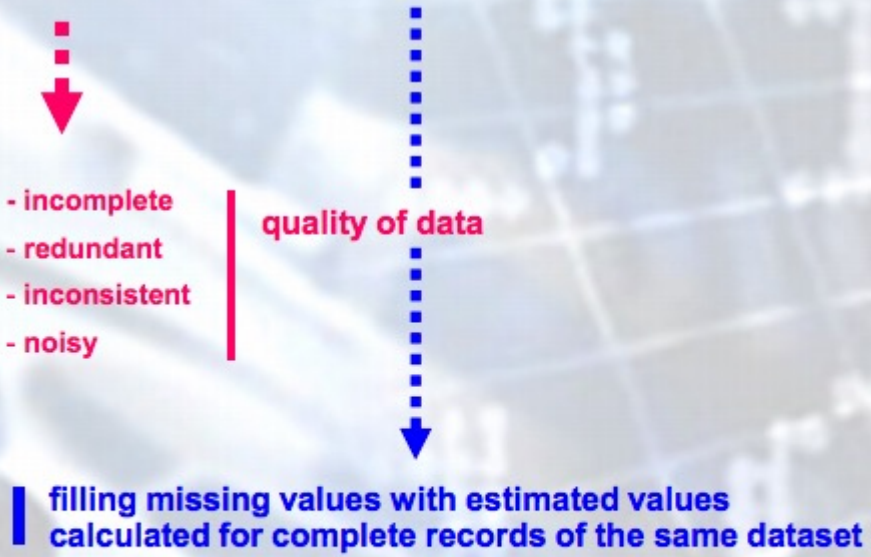
+ **Variabilité :** représente l'inconsistance des données dans le temps, ce qui entrave la manipulation et la gestion des données

+ **Véracité:** La qualité des données capturées, qui peuvent varier fortement (on parle de véracité des sources).

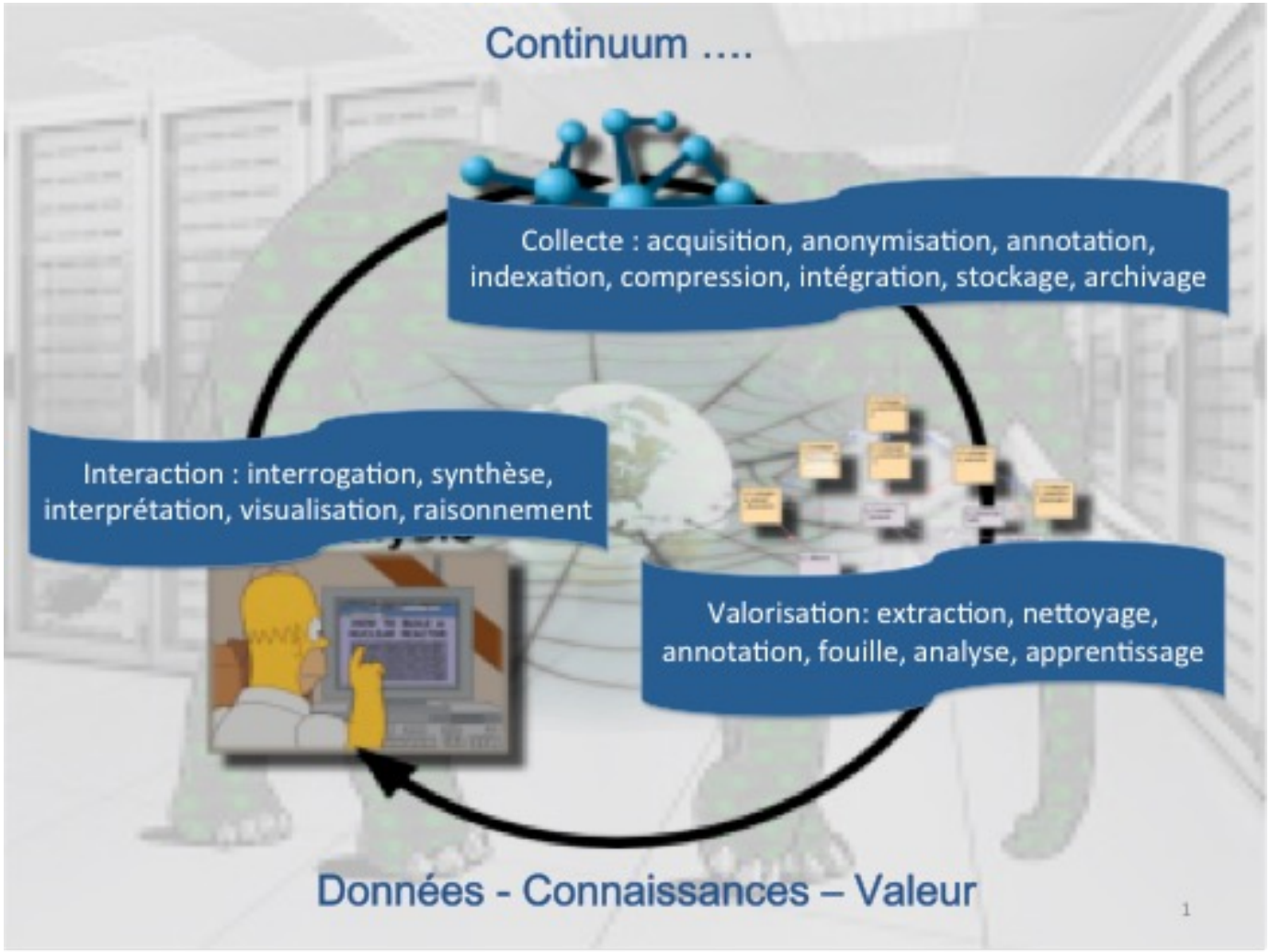
Les 10V

BIG | the Vs | 3v, 5v, 7v, 10v, ….

- **Volume** (length of a records, # of records) (entity-relationship databases)(datasets)
- **Variety** (types: strings, pictures, voice, etc.) (structured, non-structured)
- **Veracity** (precision and accuracy of data)
- **Velocity** (of change)
- **Value** (as a business/service)
- **Volatility** (temporary; quick action)
- **Vasting resources**
  (storage, computation, transfer)
- **Viability** (are data still useful?)
- **Visibility** (open, hidden, ..)
- **Validity**
  (are there still valid/updated data?)
  (in context validity)
  (e-government datasets)

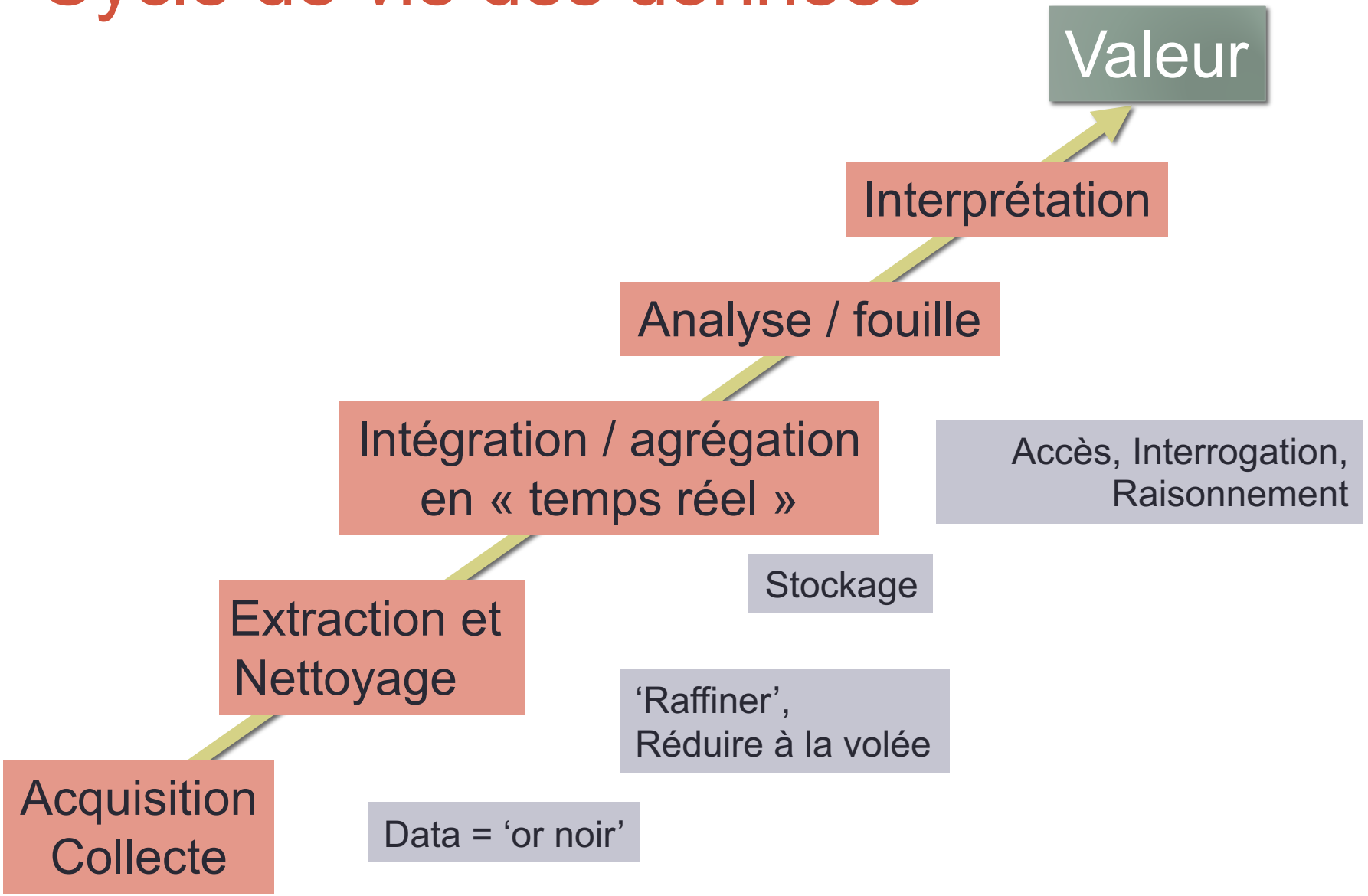


Cycle de vie des données



Source : site du GDR MADICS

Cycle de vie des données



Stockage : Data-centers

société QTS à Atlanta

Le stockage de ces données massives est réalisé dans des entrepôts de données, ou data-centers, contenant des dizaines de pétaoctets (Po). Véritables usines confrontés à des problématiques industrielles (énergie, distribution, évacuation de la chaleur) et des contraintes de connexion, ils engloutissant 1% à 2% de la consommation électrique mondiale.

Traitements: Super-ordinateurs

Le superordinateur Blue Gene/P du Argone National Lab utilise 250 000 processeurs en parallèle

MapReduce (Google, 2004) et son implémentation la plus connue, **Hadoop** (Yahoo, 2005) illustrent le paradigme actuel du calcul distribué : séparer le traitement de l'infrastructure logicielle permettant la répartition du calcul, la gestion des données distribuées, mais aussi l'hétérogénéité des ordinateurs et connections.

Protection des données personnelles

Loi de 1978 dite « Informatique et libertés ». Le président Valéry Giscard d'Estaing décide la création d'un organisme de contrôle des données personnelles dans la société de l'information : **la CNIL**, et avec elle est promulguée une loi majeure, la **loi de 1978 dite « Informatique et libertés ».**

Les droits

- le droit d'information : chacun peut être informé des traitements dont ses données font l'objet, cet article est applicable en toutes circonstances, même aux cas relevant de la sécurité nationale ;
- le droit d'accès (et d'effacement, droit à l'oubli) : plus complet que le droit d'information, il permet à chacun d'accéder aux informations qui sont conservées sur lui, il est toutefois interdit d'en faire usage dans certains cas ;
- le droit de rectification : chacun peut demander à faire corriger les données stockées le concernant;
- le droit d'opposition : chacun peut s'opposer à faire l'objet d'un traitement, pour un motif légitime (le démarchage commercial est reconnu par la loi comme motif légitime).

Ces droits se retrouveront, amplifiés et adjoints à d'autres, dans le RGPD.

Modification de 2004 : La loi de 2004 met aussi en place un nouveau système relatif à la déclaration des fichiers et traitements ; toute structure souhaitant effectuer un traitement de données à caractère personnel ou stocker ces données doit effectuer une déclaration à la CNIL.

Source: Zeste de savoir

Protection des données personnelles

RGPD (Loi Européenne, 2015) : Règlement Général de la Protection des Données

Concrètement, toute donnée concernant *un citoyen* européen, même traitée hors union, est dans le champ d'application du règlement ; c'est un cadre très large et protecteur pour les Européens.

"Le présent règlement protège les libertés et droits fondamentaux des personnes physiques, et en particulier leur droit à la protection des données à caractère personnel"

Pour les structures (y compris les administrations) traitant de la donnée personnelle, le principe de déclaration obligatoire à la CNIL est transformé en principe de responsabilité, permettant bien plus de souplesse, mais augmentant les sanctions (amendes dissuasives).

En ce qui concerne les utilisateurs, leurs droits sont renforcés, avec notamment l'arrivée du consentement « explicite » : le traitement des données personnelles est soumis à un consentement qui ne peut être forcé ; particulièrement, l'accès au service ne peut être conditionné à acceptation de traitement de données qui n'y seraient pas directement nécessaires.

Aux 4 droits de la lois Informatique et Liberté française, s'ajoute:

- le droit à la limitation du traitement (effacement partiel)
- le droit à la portabilité : permet à chaque citoyen de demander l'intégralité des données à caractère personnel les concernant

Source: Zeste de savoir

Protection des données personnelles

AI Act (Européen) : première réglementation (mondiale) de l'intelligence artificielle
(texte final : fin 2023)

En avril 2021, la Commission européenne a proposé le premier cadre réglementaire de l'UE pour l'IA. Il propose que des systèmes d'IA qui peuvent être utilisés dans différentes applications soient **analysés et classés en fonction du risque qu'ils présentent pour les utilisateurs**. Les différents niveaux de risque impliqueront plus ou moins de réglementation. Une fois approuvées, ces règles seront les premières au monde sur l'IA.

La priorité du Parlement est de veiller à ce que les systèmes d'IA utilisés dans l'UE soient sûrs, transparents, traçables, non discriminatoires et respectueux de l'environnement.

Risque inacceptable : Les systèmes d'IA à risque inacceptable sont des systèmes considérés comme une menace pour les personnes et seront interdits. Ils comprennent:

- la manipulation cognitivo-comportementale de personnes ou de groupes vulnérables spécifiques : par exemple, des jouets activés par la voix qui encouragent les comportements dangereux chez les enfants
- un score social : classer les personnes en fonction de leur comportement, de leur statut socio-économique, de leurs caractéristiques personnelles
- des systèmes d'identification biométrique en temps réel et à distance, tels que la reconnaissance faciale

Certaines exceptions peuvent être autorisées : par exemple, les systèmes d'identification biométrique à distance "a posteriori", où l'identification se produit après un délai important, seront autorisés à poursuivre des crimes graves et seulement après l'approbation du tribunal

Source: Actualités du parlement Européen

Protection des données personnelles

AI Act (Européen) : première réglementation (mondiale) de l'intelligence artificielle

Risque élevé : Les systèmes d'IA qui ont un impact négatif sur la sécurité ou les droits fondamentaux seront considérés comme à haut risque et seront divisés en deux catégories.

1. Les systèmes d'IA qui sont utilisés dans les produits relevant de la législation de l'UE sur la sécurité des produits. Cela comprend les jouets, l'aviation, les voitures, les dispositifs médicaux et les ascenseurs.

2. Les systèmes d'IA relevant de huit domaines spécifiques qui devront être enregistrés dans une base de données de l'UE :

- l'identification biométrique et la catégorisation des personnes physiques
- la gestion et l'exploitation des infrastructures critiques
- l'éducation et la formation professionnelle
- l'emploi, la gestion des travailleurs et l'accès au travail indépendant
- l'accès et la jouissance des services privés essentiels et des services et avantages publics
- les forces de l'ordre
- la gestion de la migration, de l'asile et du contrôle des frontières
- l'aide à l'interprétation juridique et à l'application de la loi.

Tous les systèmes d'IA à haut risque seront évalués avant leur mise sur le marché et au long de leur cycle de vie.
L'IA générative, comme ChatGPT, devrait se conformer aux exigences de transparence :

- indiquer que le contenu a été généré par l'IA
- concevoir le modèle pour l'empêcher de générer du contenu illégal
- publier des résumés des données protégées par le droit d'auteur utilisées pour la formation

Source: Actualités du parlement Européen

Protection des données personnelles

AI Act (Européen) : première réglementation (mondiale) de l'intelligence artificielle

(texte final : fin 2023)

Risque limité :

Les systèmes d'IA à risque limité doivent respecter des exigences de transparence minimales qui permettraient aux utilisateurs de prendre des décisions éclairées. Après avoir interagi avec les applications, l'utilisateur peut alors décider s'il souhaite continuer à l'utiliser. Les utilisateurs doivent être informés lorsqu'ils interagissent avec l'IA.

Cela inclut les systèmes d'IA qui génèrent ou manipulent du contenu image, audio ou vidéo (par exemple, les deepfakes, des contenus faux qui sont rendus crédibles par l'IA).

Source: Actualités du parlement Européen

Libertés individuelles – « Big Brothers »

1. « **Big Data : les nouveaux devins** » (spécial investigation 2015) – 50 min.
   91,80% des données personnelles mondiales seraient détenues par 4 grands acteurs qui sont GAFA

2. **« GAFA : une domination abusive »** (RTS, Geopolitis) – 15 min.
   Regroupées sous l'acronyme GAFA (Google, Apple, Facebook, Amazon), ces entreprises totalisent plus de 300 milliards de dollars de chiffre d'affaires en 2014, soit approximativement le PIB de la Norvège.

3. **« Big Data : que fait-on de nos données ? »** (RTS/TV5monde) – 15 min.
   Toutes ces données sont-elles vraiment utilisées de manière bienveillante ? L'éthique ne doit-elle pas être remise au cœur de l'immense champ des possibles offert par les Big Data ?

4. **« Le Turbo Capitalisme, Nouveaux Loups de WallStreet (**CANAL+, 2015) – 1h30.
   Les Nouveaux Loups de Wall Street - la Bourse est truquée, le Pigeon, c'est vous !

- GAFAM = GAFA + Microsoft,
- Après les Gafa, les nouveaux maîtres du monde sont les Natu (Netflix, Airbnb, Tesla, Uber). (Le Nouvel Obs 2017)

Reportages ARTE Future

1. Kenny Polcari : « J'ai peur » - 7 min.
2. Islande, le pays du stockage numérique écologique – 8 min.
3. La course à l'exploitation des données client – 6 min.
4. Le corps à la mesure de ses capacités (*quantified-self*) – 10 min.
5. Le nouveau système d'alerte pour les prématurés – 7 min.
6. La maitrise des flux de données – 5 min.
7. Big Data et la protection des océans – 5 min.
8. L'utilisation du big data au CERN (Arte.tv 2017) – 12 min.

CENTRALE DIGITAL LAB « BIG-DATA »

Partie II – Open Data

Les données publiques ouvertes

Open Data c’est quoi?

- L'***open data*** ou **donnée ouverte** est une donnée numérique dont l'accès et l'usage sont laissés libres aux usagers. Elle peut être **d'origine publique ou privée**, produite notamment par une collectivité, un service public ou une entreprise. Elle est diffusée 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.
- Ces droits d'accès et de réutilisation s'inscrivent dans la pensée qui considère l'information publique comme un bien commun dont la diffusion est d'intérêt public et général.
- L'ouverture des données (*open data*) est à la fois **un mouvement**, **une philosophie d'accès à l'information** et une **pratique de publication de données** librement accessibles et exploitables.
- Le mouvement s'est étendu notamment sous l'impulsion d'ONG comme l'Open Knowledge Foundation (OKFN) et le Partenariat pour un gouvernement ouvert (PGO).

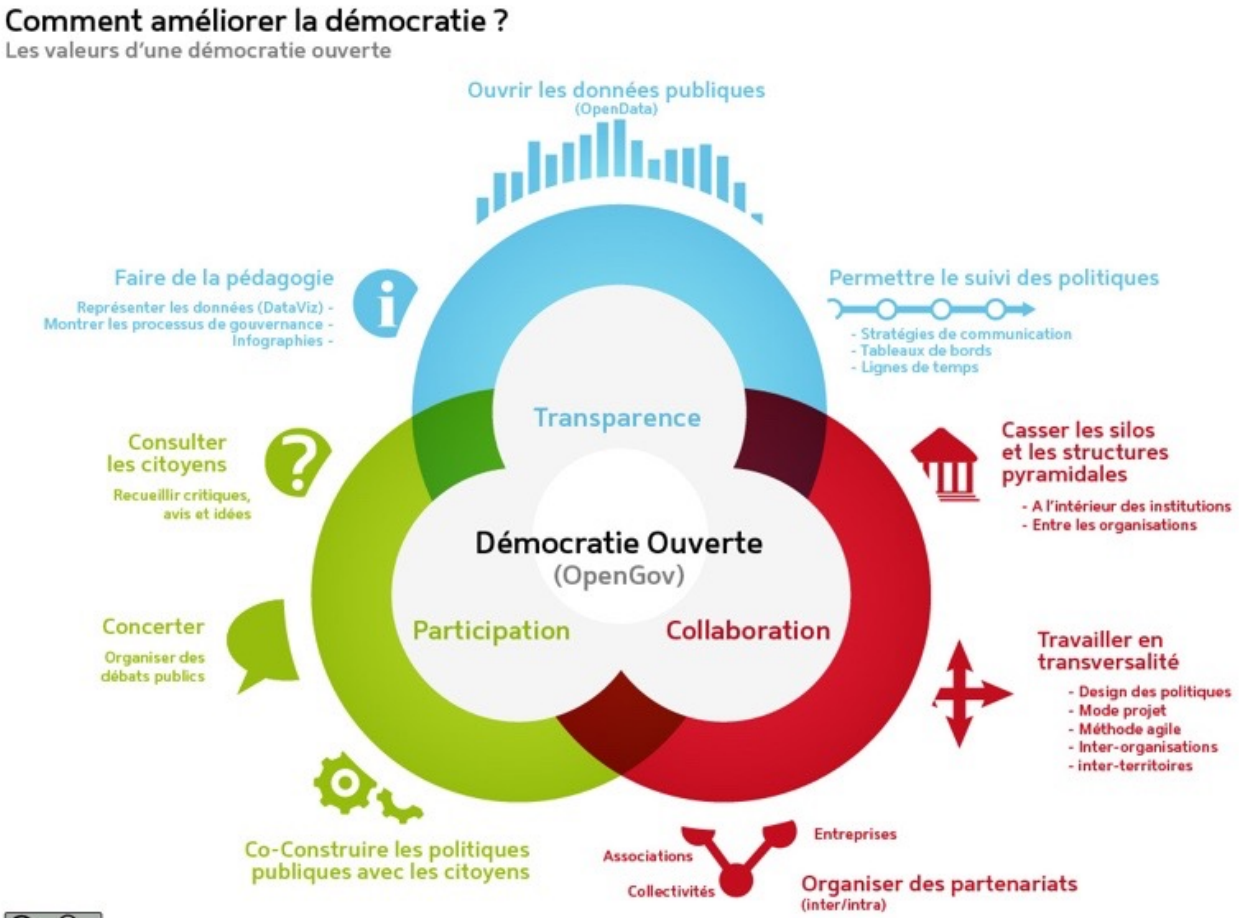
Wikipédia

Open knwoledge fundation

- L'**Open Knowledge Foundation** est une association à but non lucratif de droit britanniquepromouvant la culture libre, en particulier les contenus libres et l'*open data* (données ouvertes). Elle a été créée le 24 mai 2004 à Cambridge au Royaume-Uni.

- 3 groupes de travail :
  - *Open research* groupe de travail dédié au domaine de la recherche ouverte et mettant l'accent sur le degré d'ouverture des données et sur comment Internet et les technologies numériques peuvent soutenir de nouvelles formes de collaboration et de partage dans la recherche scientifique.
  - *Open culture* groupe de travail voués à l'étude des moyens par lesquels des logiciels open source et les contenus culturels ouverts peuvent accroître l'accès à notre patrimoine culturel et forger de nouveaux processus de création et de collaboration.
  - *Open government* il s'agit de groupes de travail qui étudient le domaine des données publiques ouvertes, et qui sont tous engagés à construire des communautés qui rendront les **gouvernements plus transparents et responsables**.

Une démarche openGov



L'évolution des lois sur les données

- **Loi du 17 juillet 1978**
  (CADA : Commission d'accès aux documents administratifs) :
  - Toute personne a le droit d'obtenir des documents détenus par une administration (indépendamment de leur forme ou de leur support).
  - Obligation pour toutes les administrations publiques ainsi que tous les organismes privés chargés d'une mission de service public.
  - "Les informations … peuvent être utilisées … à d'autres fins que celles de la mission de service public pour les besoins de laquelle les documents ont été produits ou reçus".

- **Directive européenne de 2003**
  (transposée dans le droit français en 2005)
  - Le droit d'accès devient une obligation de publication.
- **Décret de 2011** : Principe de la gratuité du droit à réutilisation des documents et données publiques.

L'évolution des lois sur les données

- **La Loi pour une République numérique (2016)**
  - Cette loi vise à favoriser l'ouverture et la circulation des données et du savoir, à garantir un environnement numérique ouvert et respectueux de la vie privée des internautes et à faciliter l'accès des citoyens au numérique.
  - Tout le monde peut désormais lire, réutiliser librement et gratuitement les données de l'administration. Elle consacre la conversion de l'administration française au mouvement de l'open data. Plus grande transparence dans la gestion du budget et des finances publiques, services publics optimisés grâce aux données ouvertes, villes intelligentes, meilleur fonctionnement du marché du travail, intelligence artificielle éthique…

  Les possibilités que laisse entrevoir l'ouverture des données publiques à travers le monde sont nombreuses et promettent de transformer la société dans les décennies à venir.

- La loi sur la transition énergétique prévoit à partir de 2016 de progressivement rendre les données énergétiques (production, consommation par immeuble, quartier, par ville…) disponibles en ligne pour une libre réutilisation par toute personne intéressée (open data). Les gestionnaires de réseaux (électricité, gaz, réseau de chaleur et de froid) et les fournisseurs de produits pétroliers doivent fournir certaines données au service statistique du ministère de l'énergie.

Quelles données ?

Toutes les données structurées et produites par des acteurs publics ou privés dans le cadre d'une mission de service public. Cette définition s'étend aux données décrivant l'espace, les services publics et l'exercice politique, produites par des acteurs non institutionnels : associations, citoyens engagés, acteurs économiques ou académiques.

Sont exclus de l'opendata

• Les données à caractère personnel

« Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée directement ou indirectement par référence à un ou plusieurs éléments qui lui sont propres » (la loi 78-17 du 6 janvier 1978 art 2)

Ex : nom, adresse, photographie, num. tél., num. d'identifiant, adresse IP, adresse électronique…

• les documents administratifs dont la consultation ou la communication porterait atteinte :

A-6a) Au secret des délibérations du Gouvernement et des autorités responsables relevant du pouvoir exécutif ;

A-6d) A la sûreté de l'Etat, à la sécurité publique ou à la sécurité des personnes;

A-10c) Ou sur lesquels des tiers détiennent des droits de propriété intellectuelle.

Données publiques versus Big Data

Données publiques : services publics, intérêt général, ouvertes

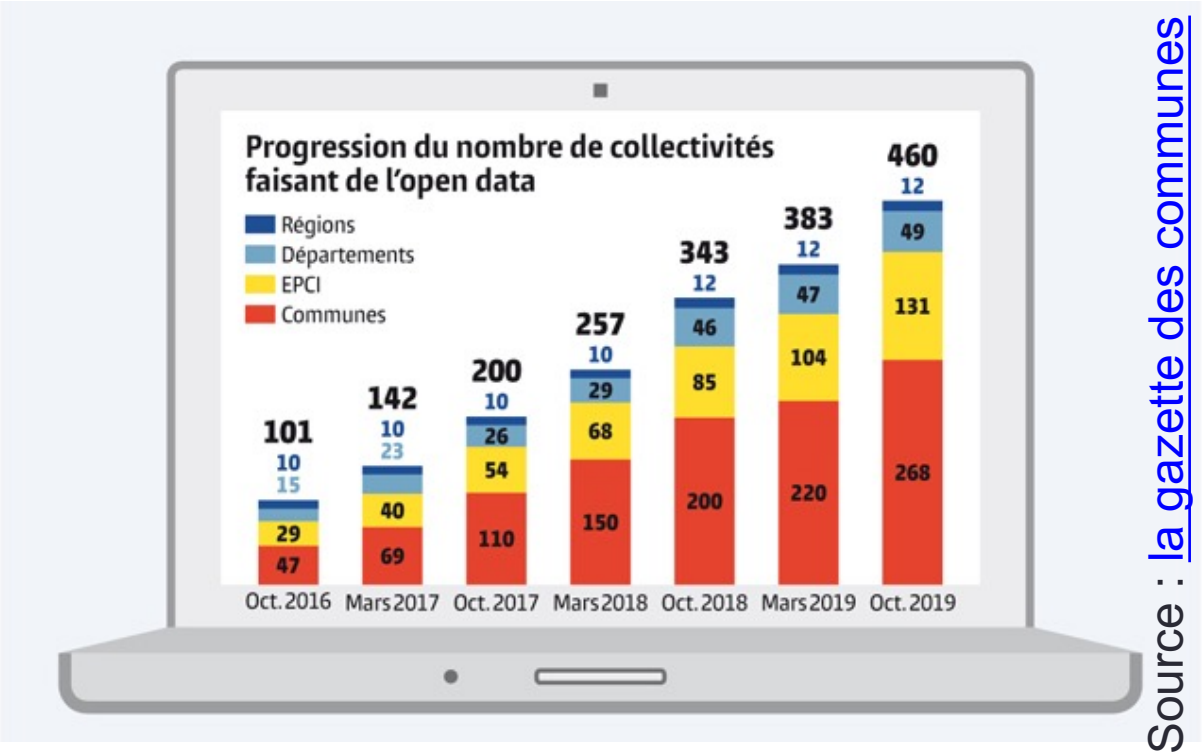
- Effectifs: d'agents municipaux, d'élèves
- Horaires : bus et tram en temps réel,
- Patrimoine : arbres, équipements sportifs, points lumineux, bornes fontaines ...
- Éléments liés à l'exploitation : livres empruntés dans les bibliothèques, budget et compte administratif, marchés publics
- Description du territoire : photo aérienne, altimétrie, description des rues,
- Reflet de la vie locale : mariage, naissance, prénoms, élections, agendas...

Big data: données personnelles, comportement individuel, données privées

- Traces (téléphone mobile, connexion)
- Navigation internet, réseaux sociaux, messagerie,
- Achats (CB), déplacements individuels ou collectifs (badge, carte)
- Consommations individuelles (eau, électricité, ..)
- Objets connectés

Gisement de données

- **Collectivités**
  - Accompagnement pédagogique :
    DataLab, InfoLab
  - OpenDataSoft



Source : la gazette des communes

- **Etat**
  - EtatLab : Missions
  - data.gouv.fr
    - 39 316 jeux de données ouvertes
    - 2792 applications ré-utilisatrices

Gisement de données

- **Citoyen**
  - Regards Citoyens,
  - DataPublica,
  - Data4Citizen
- **Instituts**
  - INSEE, INPI, LEGIFRANCE
  - IDG (infrastructure de données géographiques)
- **Agences de l'air, de l'eau**
  - Laboratoire Central de Surveillance de la Qualité de l'Air, Carte mondiale, IQAir,
  - Données publiques sur l'eau en France
- **Mondial**
  - Global open data index (*Open Knowledge Foundation*)
  - La Banque Mondiale

Licences

Licences Opendata stabilisées en France autour de :

- **ODbL : Open Data Base Licence**, édité par *OpenKnowledge Foundation*

- **Licence Ouverte** : édité par Etalab

- Elles autorisent la réutilisation y compris à des fins commerciales.
- Elles imposent des conditions de Mention et d'Intégrité
- Elles n'imposent pas d'obligation pour le producteur
- ODbL est plus restrictive car elle impose la notion de « Partage à l'Identique »

Licences ODbL

Vous êtes libres :

De partager : copier, distribuer et utiliser la base de données.

De créer : produire des créations à partir de cette base de données.

D'adapter : modifier, transformer et construire à partir de cette base de données.

Aussi longtemps que :

Vous mentionnez la paternité : Vous devez mentionnez la source de la base de données pour toute utilisation publique de la base de données, ou pour toute création produite à partir de la base de données, de la manière indiquée dans l'ODbL. Pour toute utilisation ou redistribution de la base de données, ou création produite à partir de cette base de données, vous devez clairement mentionner aux tiers la licence de la base de données et garder intacte toute mention légale sur la base de données originaire.

Vous partagez aux conditions identiques : si vous utilisez publiquement une version adaptée de cette base de données, ou que vous produisiez une création à partir d'une base de données adaptée, vous devez aussi offrir cette base de données adaptée selon les termes de la licence ODbL.

Gardez ouvert : si vous redistribuez la base de données, ou une version modifiée de celle-ci, alors vous ne pouvez utiliser de mesure technique restreignant la création que si vous distribuez aussi une version sans ces restrictions.

Formats

- Pour être Opendata, les données doivent être :
  - au format électronique,
  - accessible sur internet,
  - structurées,
  - à un format ouvert (et de préférence non propriétaire).

- **Types de format des fichiers disponibles**
  - Données alpha-num. : csv, ods, xls, doc, xml, JSON…
    - Un document au format pdf, même full text –et pas image- est lisible par une machine mais n'est ouvrable que par Acrobat/Adobe, donc non Opendata
  - Données géographiques : Shape, GéoJSON, KML, KMZ,
    - Format propriétaires non ouverts : DGW et DXF (Autocad)
  - Format spécifiques :
    - Domaine Transport : GTFS, Chouette, NetEX, Neptune…
    - GPX (ouvert), …

Outils d'exploitation

Données Alphanumériques

– Suites bureautiques habituelles : OpenOffice, Libre Office,
  - Et les outils commerciaux habituels (Google, MS, …)

– De nombreux outils en ligne : ConversionTools Services (xml<>csv…)

– Les langages de programmation habituels : C, Python…

– *OpenRefine* (previously Google Refine) pour l'exploration et le retraitement des données

Données géographiques

– QGIS : vraie suite logicielle de géomatique (SIG), convivial

– UMAP (openstreetmap), GeoServer : manipulation des cartes

– Non OpenSource : ARCGis, Autocad, MapInfo, GéoMap, GooglePro&Co

Exemples de réutilisations

- Handicap : Handimap, wheelmap
- Train temps réel : raildar, sncf geolocalisation
- Parking temps réel : Nantes, Rennes
- Transport en commun multi-modal : Toulouse, Lyon
- Développement logicels libres : makina corpus

Des sources d'info sur l'open data

Documents

- http://opendatahandbook.org/guide/fr/
  (*Open Knowledge Foundation*)

Youtube

- L'Open Data à la Loupe. Yann Bresson
- Quel est l'impact de l'Open Data ? Charles RUELLE, 2014
- Open Data explained in a nutshell, Simpleshow foundation, 2016
- Comment nous avons trouvé la pire place de parking de NY ?, Ben Wellington, TEDxNewYork
- L'Open Data, Avenir des Big Data, Jean Marc LAZAR, TEDxUTCompiègne

CENTRALE DIGITAL LAB
« BIG-DATA »

Partie III – Linked Open Data

Interrogation de bases de données publiques

1. LOD : Qu'est-ce que le Web ?

Un espace documentaire décentralisé, interconnecté, interopérable et *évolutif*.

- décentralisé → **HTTP 2014, HTTP 2.0**
- interconnecté → **URL, URI, IRI**
- interopérable → **HTML5, CSS, JS**

Vers un Web de données

Un espace **de données** décentralisé, interconnecté et interopérable.

- décentralisé → **HTTP**
- interconnecté → **URL**
- interopérable → **?**

1. LOD : Web de ressources

Le web est constitué de **ressources**, par exemple :

- le bulletin météo du jour pour Lyon
- le bulletin météo du jour pour le lieu courant
- ma commande de café de jeudi dernier

Chaque ressource est identifiée par un IRI (*Internationalized Resource Identifier*), *e.g.*:

- http://meteo.example.com/lyon
- http://meteo.example.com/ici
- http://commerce.example.com/commande/192837

1. LOD : Web de ressources

Une ressource n'est pas un simple fichier, dont on récupèrerait le contenu.

Elle est un objet *actif*, avec lequel on interagit.

- le bulletin météo du jour pour Lyon :
  → le contenu change régulièrement
- le bulletin météo du jour pour le lieu courant :
  → le contenu dépend de plus du contexte de l'utilisateur
- ma commande de café de jeudi dernier :
  → on peut agir dessus (par exemple pour l'annuler)

1. LOD : Ressources et représentations

- Une ressource n'est jamais manipulée directement, mais toujours à travers des **représentations** (pour la créer, la consulter, la modifier).
- Les représentations d'une ressource peuvent varier en fonction
  - de son *état*
  - de l'agent qui manipule la ressource (négociation de contenu, contexte)

représentation :	utilisable par :
texte (HTML...)	humains, moteurs de recherche
médias (image, son...)	*surtout* humains
données structurées	machines

1. LOD : de HTML à XML

XML (*eXtensible Markup Language*) a été recommandé par le W3C en 1998. L'objectif était de pallier la sémantique « faible » de HTML.

```
<!-- HTML -->
<a href="http://champin.net/">
  Pierre-Antoine <strong>Champin</strong>
  (<em>Maître de conférences</em>)</a>
```

```
<!-- XML -->
<Person homepage="http://champin.net/">
  <givenName>Pierre-Antoine</givenName>
  <familyName>Champin</familyName>
  <job>Maître de conférences</job></Person>
```

1. LOD : de XML à RDF

- Le modèle sous-jacent de la syntaxe XML est un arbre (*XML Infoset*), ce qui n'est pas adapté à la structure décentralisée du Web.

- L'objectif du *Resource Description Framework* (RDF), recommandé par le W3C en 1999, vise à munir le Web d'un modèle de données plus adapté, ayant une structure de *graphe*.

- L'objectif est de construire le *Semantic Web* : un web dans lequel les machines ont (enfin) accès à la sémantique des données.

- Recommandation un peu hâtive, présentant quelques défauts importants (notamment l'absence de sémantique formelle).

  → faible adoption de RDF

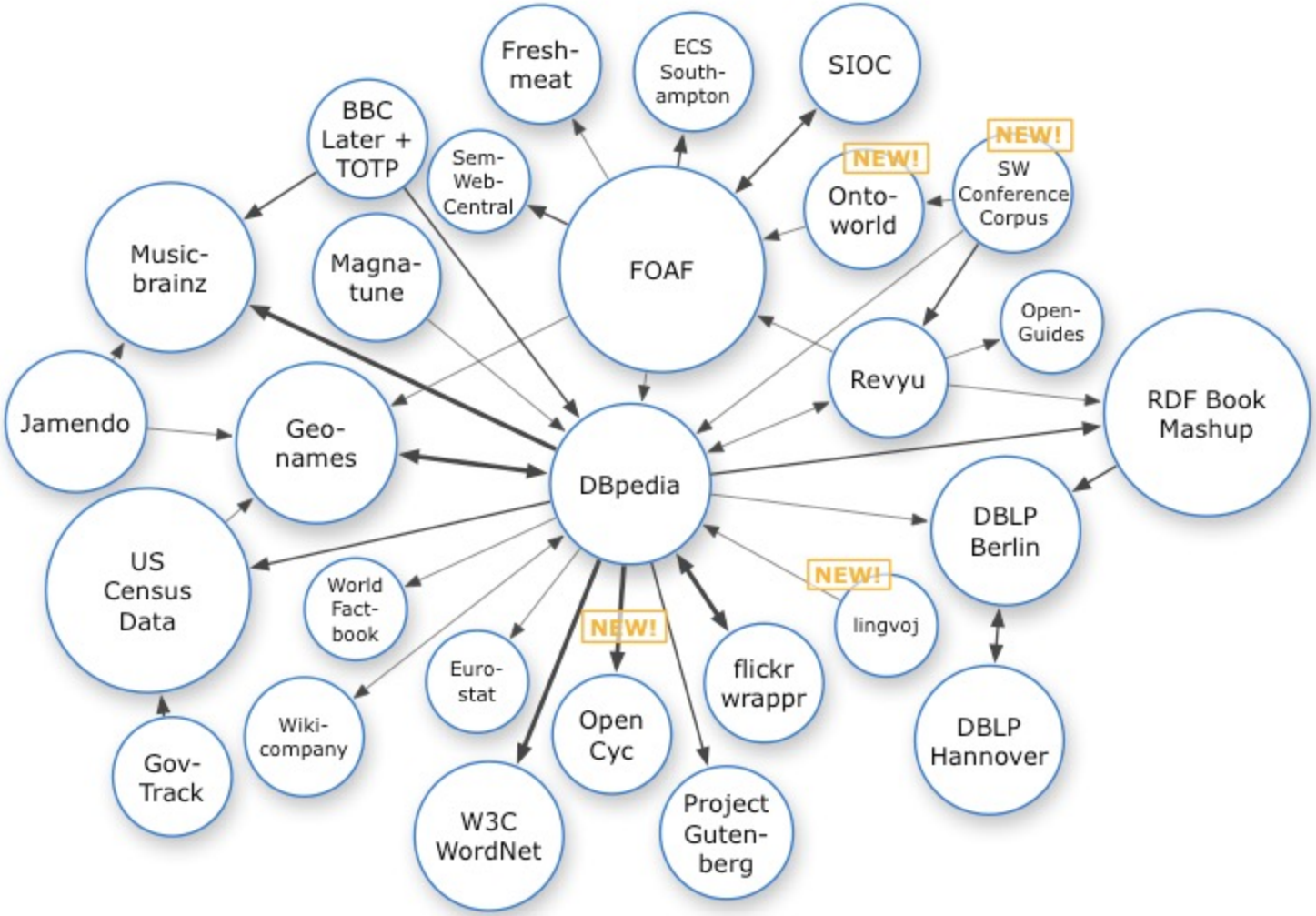
1. LOD : de RDF à RDF

- En 2004, le W3C publie un nouvel ensemble de recommandations sur RDF pour remplacer celles de 1999.
- Pour des raisons de compatibilité avec l'existant, certains aspects sont conservés malgré les débats qu'ils suscitent, mais les défauts considérés comme majeurs sont corrigés.
- Après cet échec relatif, l'appellation *Semantic Web* tombe peu a peu en disgrâce. Certains défenseurs de RDF parlent plus modestement de *Data Web*, puis de *Web of Linked Data* (2006).
- En 2014, RDF est plus largement accepté. Le W3C publie une version révisée (RDF 1.1), endossant notamment plusieurs syntaxes concrètes.

1. LOD : Raw Data Now!

- The next Web, https://www.ted.com/talks/tim_berners_lee_the_next_web

1. LOD : Linked open Data



Source image : [Richard Cyganiak](http://cas.lod-cloud.net) (http://cas.lod-cloud.net)

1. LOD : Linked open Data

Les quatre principes de Linked Data

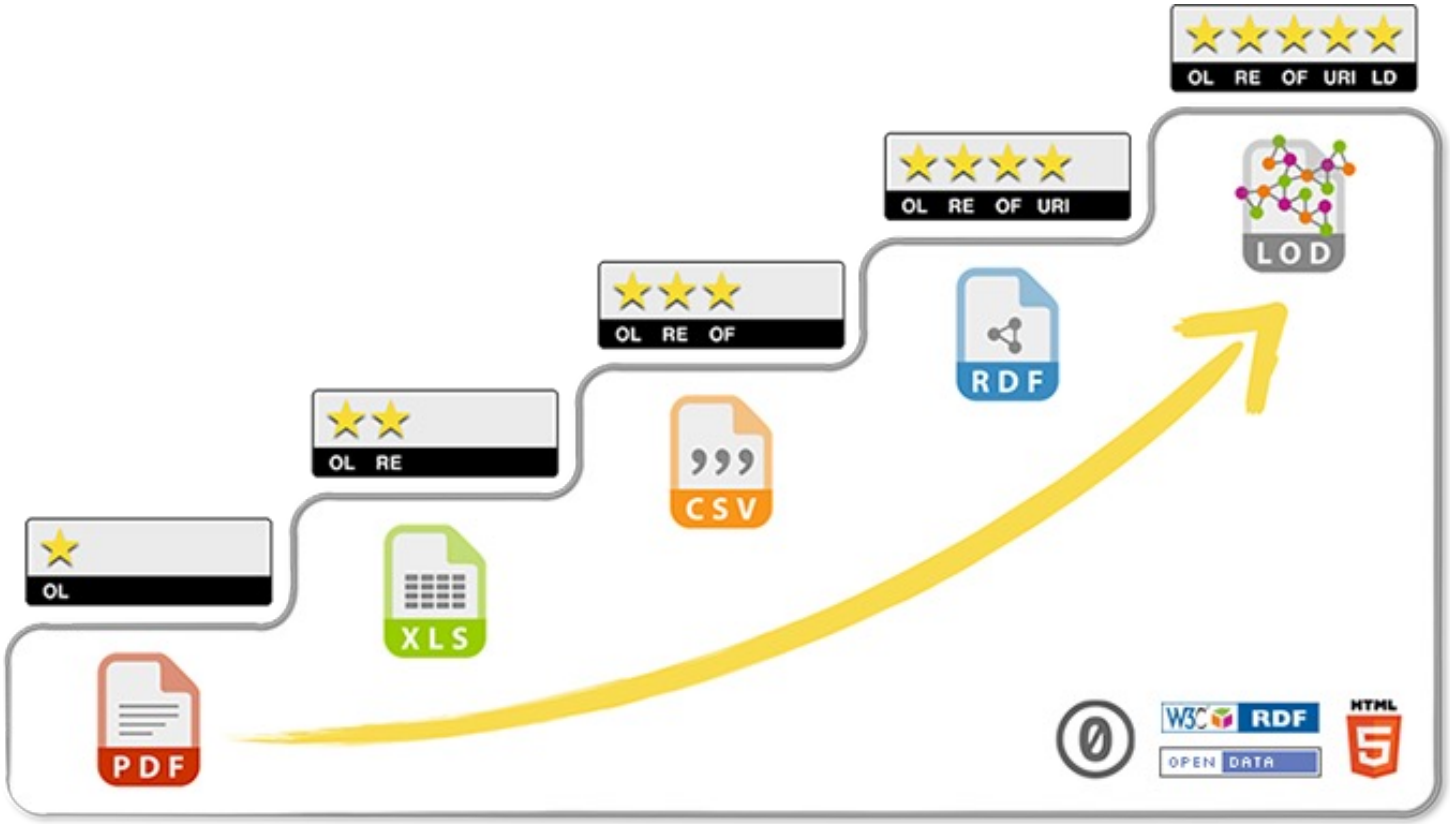
1. Utiliser des IRIs pour nommer les choses (= ressources).
2. Utiliser des IRIs HTTP pour pouvoir obtenir des *représentations* de ces ressources.
3. Fournir ces représentations en utilisant des langages et des protocoles standards (RDF, SPARQL).
4. Inclure des liens pour permettre de découvrir de nouvelles ressources.

d'après Tim Berners Lee,

http://www.w3.org/DesignIssues/LinkedData.html

1. LOD : Linked open Data

Linked open data [star scheme](https://5stardata.info/en/) (https://5stardata.info/en/)



1. LOD : Projet emblématique, DBPédia

- Projet lancé par Chris Bizer en 2007.
- Objectif : extraire les informations structurées (*infobox*) présentes dans Wikipedia pour les exposer en RDF.
- En novembre 2015 : The English version of the DBpedia knowledge base describes 4.58 million things, (...) including 1,445,000 persons, 735,000 places (including 478,000 populated places), 411,000 creative works (including 123,000 music albums, 87,000 films and 19,000 video games), 241,000 organizations (including 58,000 companies and 49,000 educational institutions), 251,000 species and 6,000 diseases.

2. RDF : des données aux données liées

Take a citation:

"Tim Berners-Lee, James Hendler and Ora Lassila. *The Semantic Web*. Scientific American, May 2001". Link

We can identify a number of distinct statements in this citation:

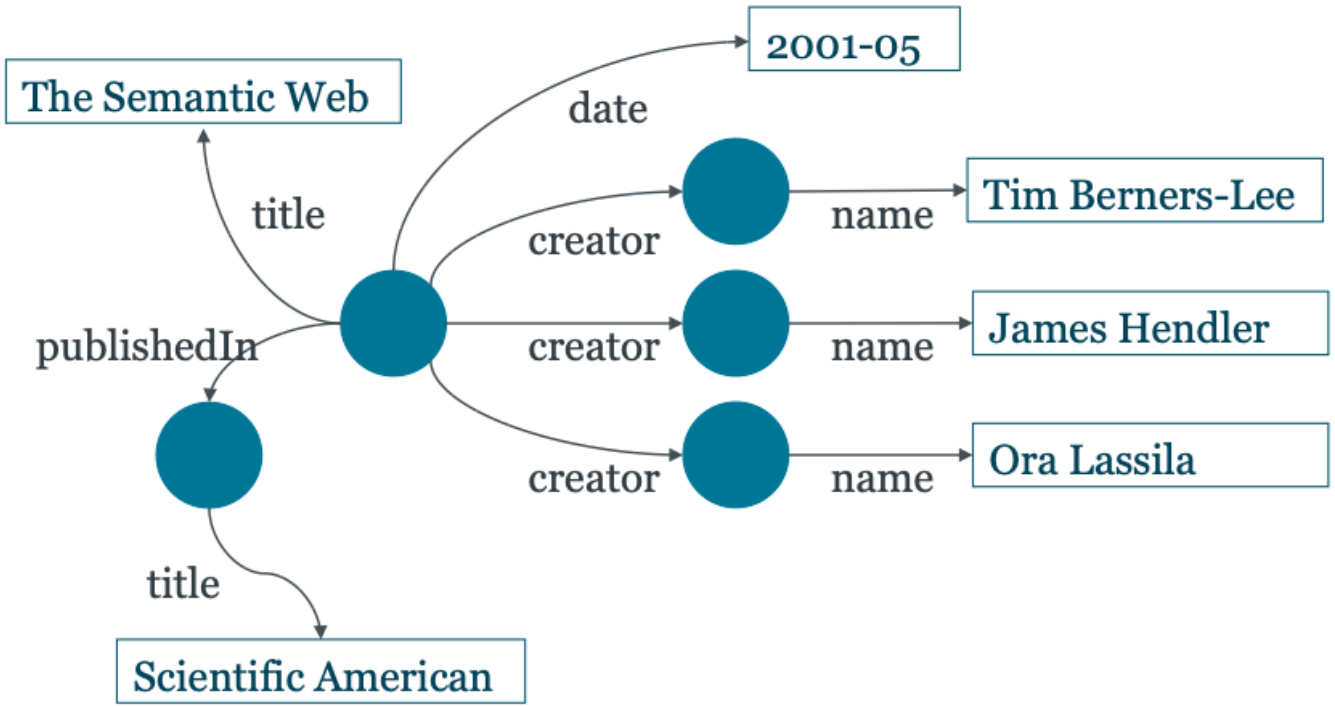
- There is an article titled “The Semantic Web”
- One of its authors is a person named “Tim Berners-Lee”(etc)
- It appeared in a publication titled “Scientific American”
- It was published in May 2001

Source : The web of Linked data

2. RDF : des données aux données liées

"Tim Berners-Lee, James Hendler and Ora Lassila. *The Semantic Web*. Scientific American, May 2001".

Graphe des données

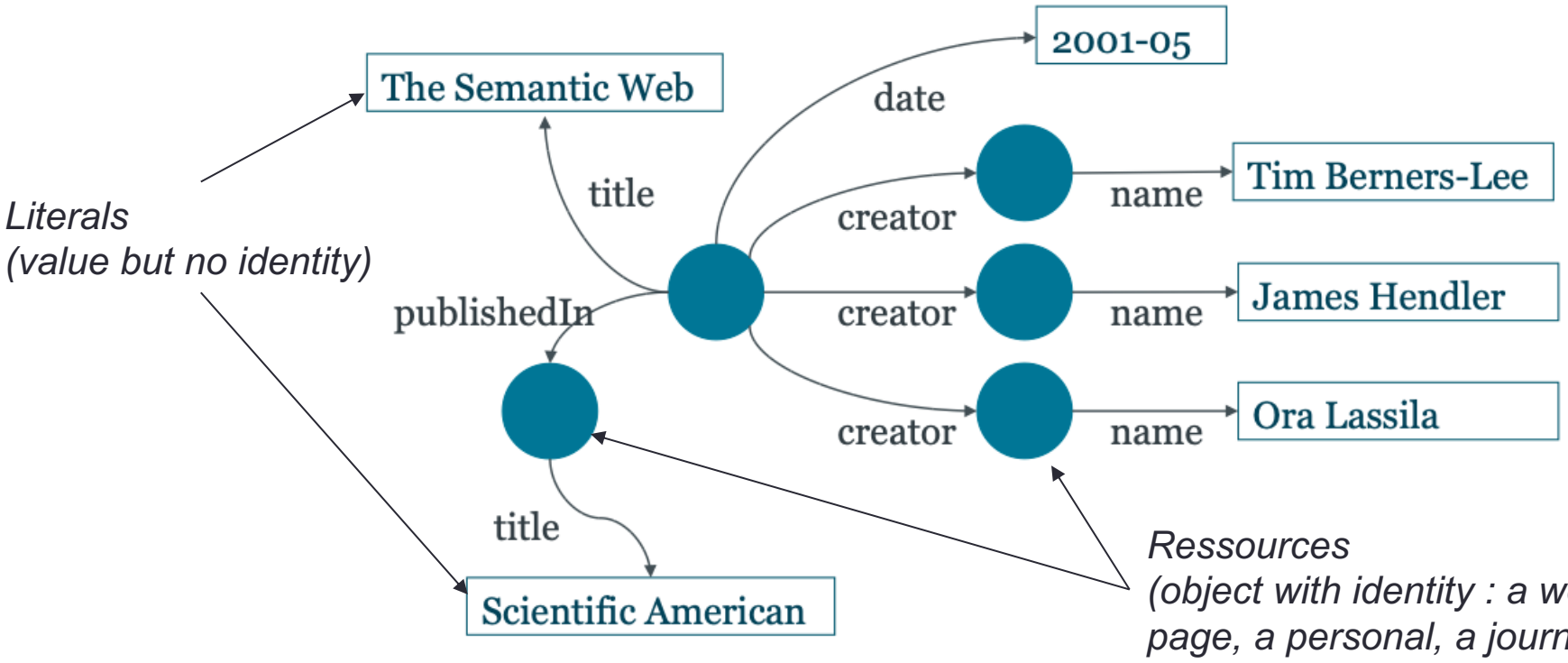


Source : [The web of Linked data]

2. RDF : des données aux données liées

"Tim Berners-Lee, James Hendler and Ora Lassila. *The Semantic Web*. Scientific American, May 2001".

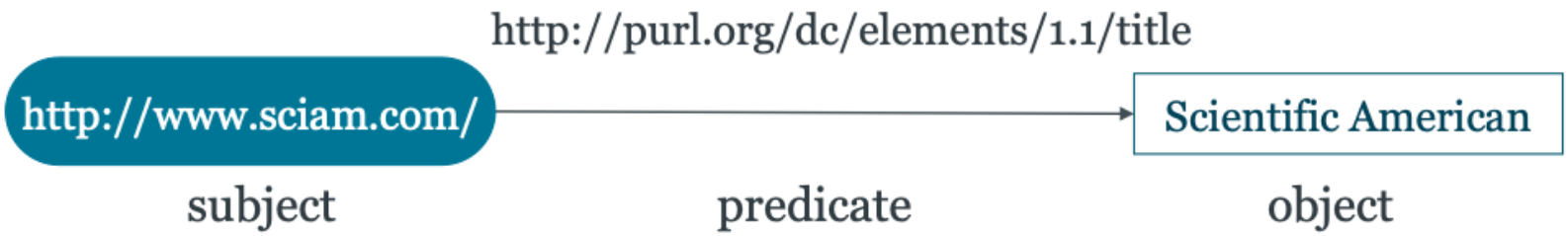
Graphe des données



Source : [The web of Linked data]

2. RDF : des données aux données liées

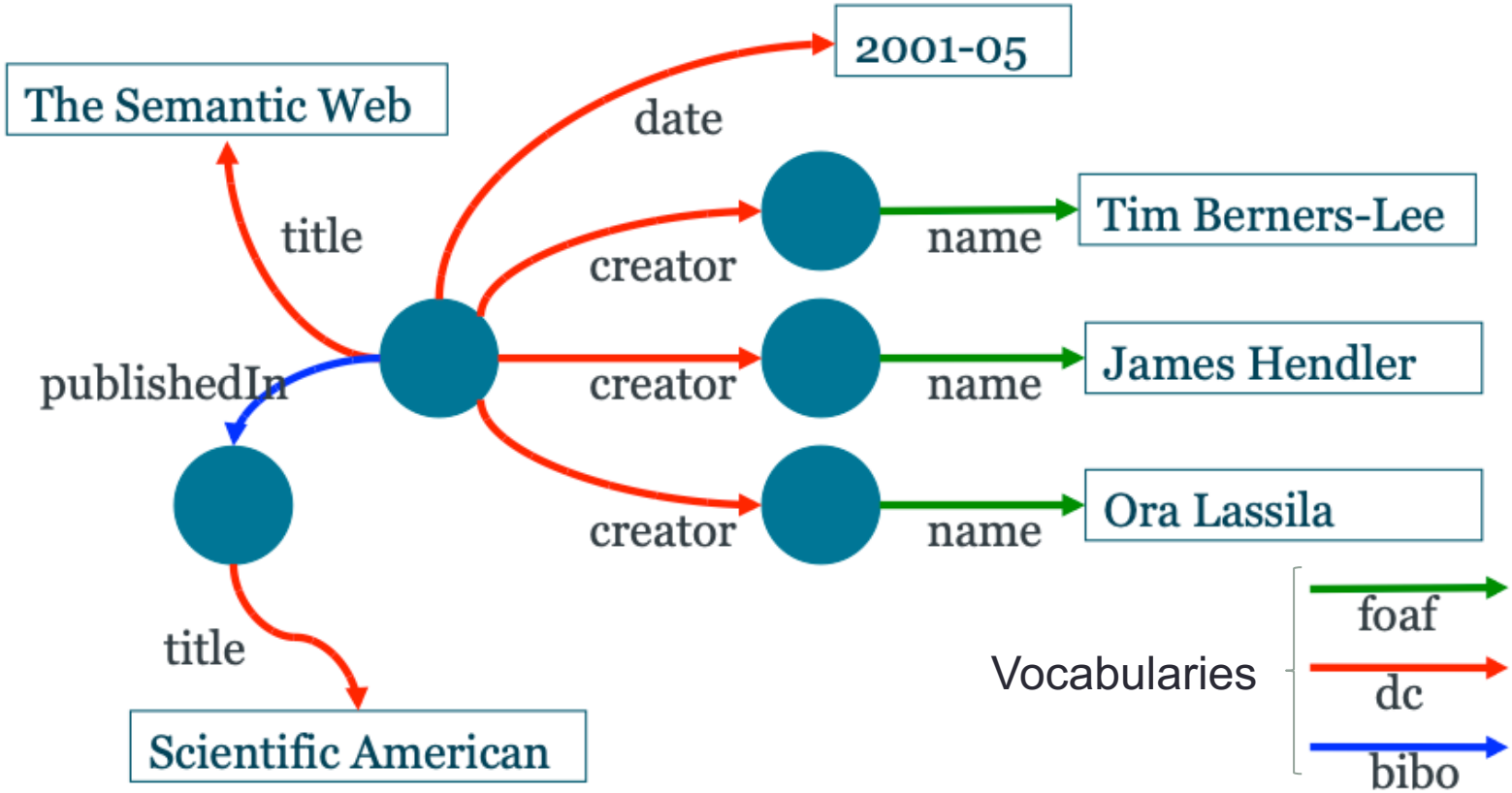
Triplet sujet / prédicat / objet



Source : The web of Linked data

2. RDF : des données aux données liées

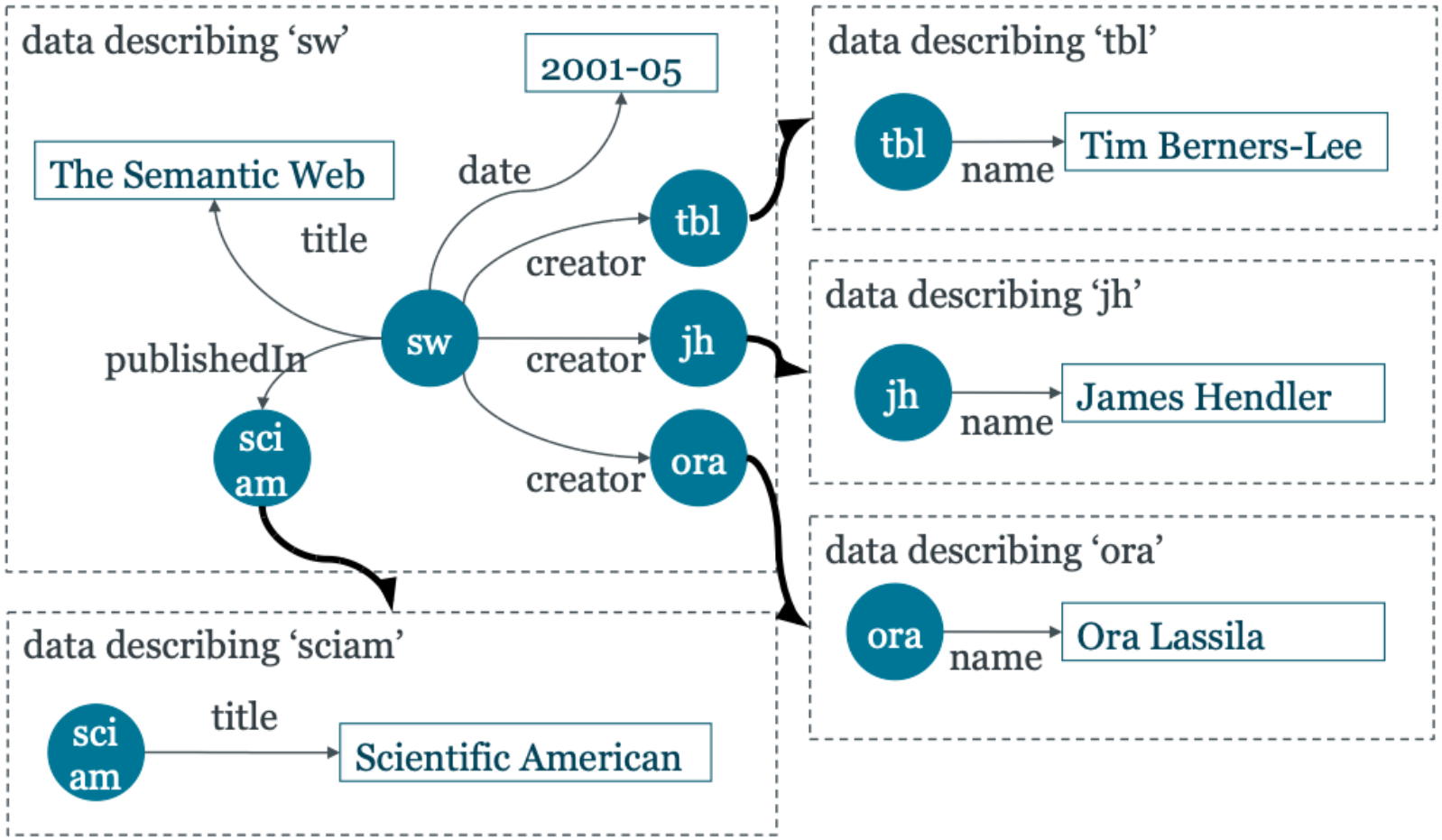
Graphe de données



Source : [The web of Linked data]

2. RDF : des données aux données liées

Graphe de données liés



Source : [The web of Linked data]

2. RDF : syntaxe

Toute information en RDF est représentée par un *triplet*, signifiant qu'une *chose* est en *relation* avec une autre.

Exemple :

Le laboratoire LIRIS (sujet)

a pour membre (prédicat)

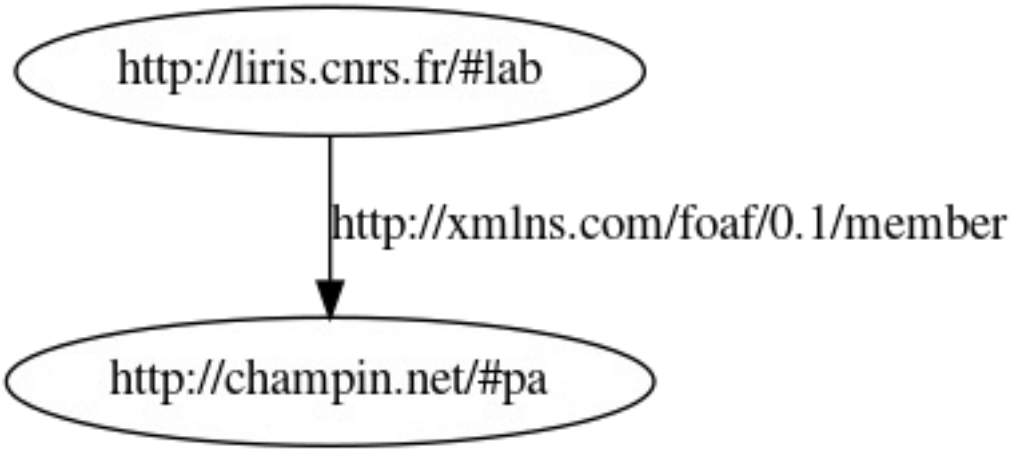
Pierre-Antoine Champin (objet)

2. RDF : syntaxe

Les choses sont nommées par des IRIs :

http://liris.cnrs.fr/#lab
http://xmlns.com/foaf/0.1/member
http://champin.net/#pa

On peut représenter ceci graphiquement :



2. RDF : préfixes

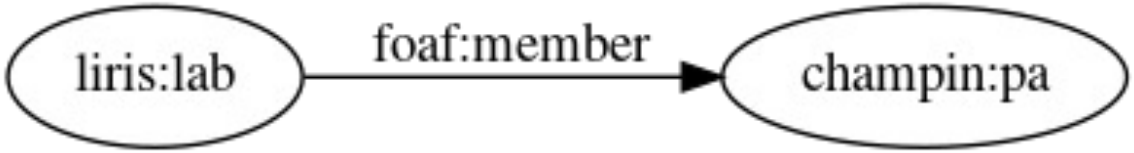
Pour simplifier les **notations**, on définit des préfixes courts correspondant à des préfixes d'IRI :

liris: → http://liris.cnrs.fr/#
foaf: → http://xmlns.com/foaf/0.1/
champin: → http://champin.net/#

On utilise ensuite des *noms préfixés* :

```
liris:lab foaf:member champin:pa
```

et également sous forme graphique :

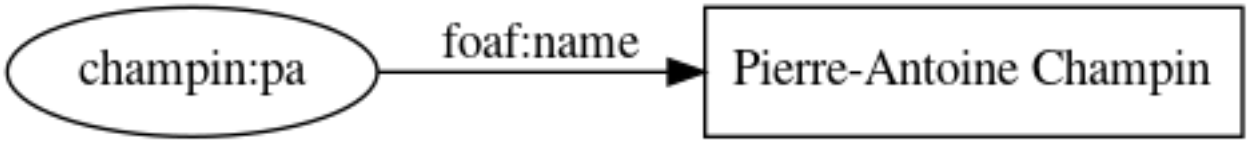


2. RDF : les littéraux

On peut également lier une ressource à une *donnée typée* (chaîne de caractères, entier, réel...), nommée un littéral.

```
champin:pa foaf:name "Pierre-Antoine Champin"
```

Traditionnellement, on représente les littéraux par des nœuds rectangulaires :

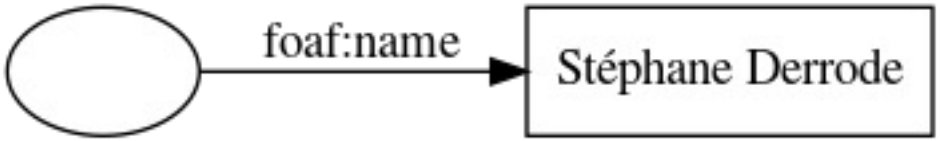


2. RDF : les nœuds muets

Enfin, RDF permet de parler d'une ressource sans connaître son IRI :

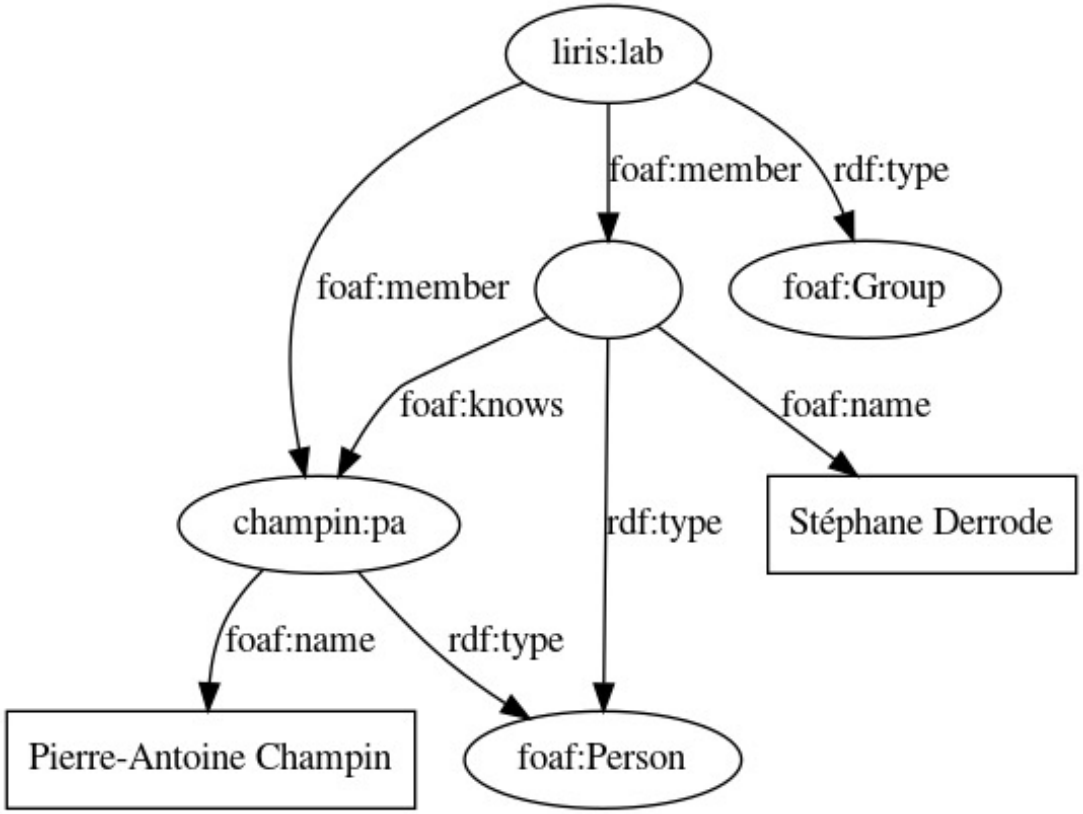
```
(quelque chose) foaf:name "Stéphane Derrode"
```

On parle alors de nœud *muet* (par analogie aux variables muettes). Graphiquement, on représente cette ressource par un nœud vierge (*blank node*).



2. RDF : exemple de graphe

Un ensemble de triplets forme un graphe orienté étiqueté.



2. RDF : syntaxes concrètes

RDF / XML

- syntaxe originale recommandée par le W3C (1999)
- basée sur XML
- relativement complexe et verbeuse

Syntaxe : http://www.w3.org/TR/rdf-syntax-grammar/

Validator : http://www.w3.org/RDF/Validator/

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/"
 <foaf:Group rdf:about="http://liris.cnrs.fr/#lab">
  <foaf:member>
   <foaf:Person>
    <foaf:name>Stéphane Derrode</foaf:name>
    <foaf:knows
     rdf:resource="http://champin.net/#pa"/>
   </foaf:Person>
  </foaf:member>
  <foaf:member>
   <foaf:Person rdf:about="http://champin.net/#pa">
    <foaf:name>Pierre-Antoine Champin</foaf:name>
   </foaf:Person>
  </foaf:member>
  </foaf:Group>
</rdf:RDF>
```

2. RDF : syntaxes concrètes

RDF / Turtle

- dérivée du langage N3
- adoptée dans RDF 1.1 en 2014
- vise la simplicité et la compacité

Syntaxe : http://www.w3.org/TR/turtle/

Validator : http://www.rdfabout.com/demo/validator/

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix champin: <http://champin.net/#> .

liris:lab
    a foaf:Group ;
    foaf:member champin:pa, _:sd .
champin:pa
    a foaf:Person ;
    foaf:name "Pierre-Antoine Champin" .
_:sd
    a foaf:Person ;
    foaf:name "Stéphane Derrode" ;
    foaf:knows champin:pa .
```

2. RDF : syntaxes concrètes

RDF / JSon

- JSON-LD (JSON Linked Data) permet d'interpréter une structure JSON comme du RDF, grâce à un *contexte* (implicite ou explicite).
- Objectif : faciliter l'adoption de RDF (syntaxe abstraite) auprès des développeurs d'applications web.

Syntaxe : http://www.w3.org/TR/json-ld-syntax/

Validator : http://json-ld.org/playground/

```
{ "@context" : { /* ... */                                    },
 "@id": "http://liris.cnrs.fr/#lab",
 "@type": "Group",
 "member": [
   {
     "@id": "http://champin.net/#pa",
     "@type": "Person",
     "name": "Pierre-Antoine Champin"
   },
   {
     "@type": "Person",
     "name": "Stéphane Derrode",
     "knows": "http://champin.net/#pa"
   }
 ]
}
```

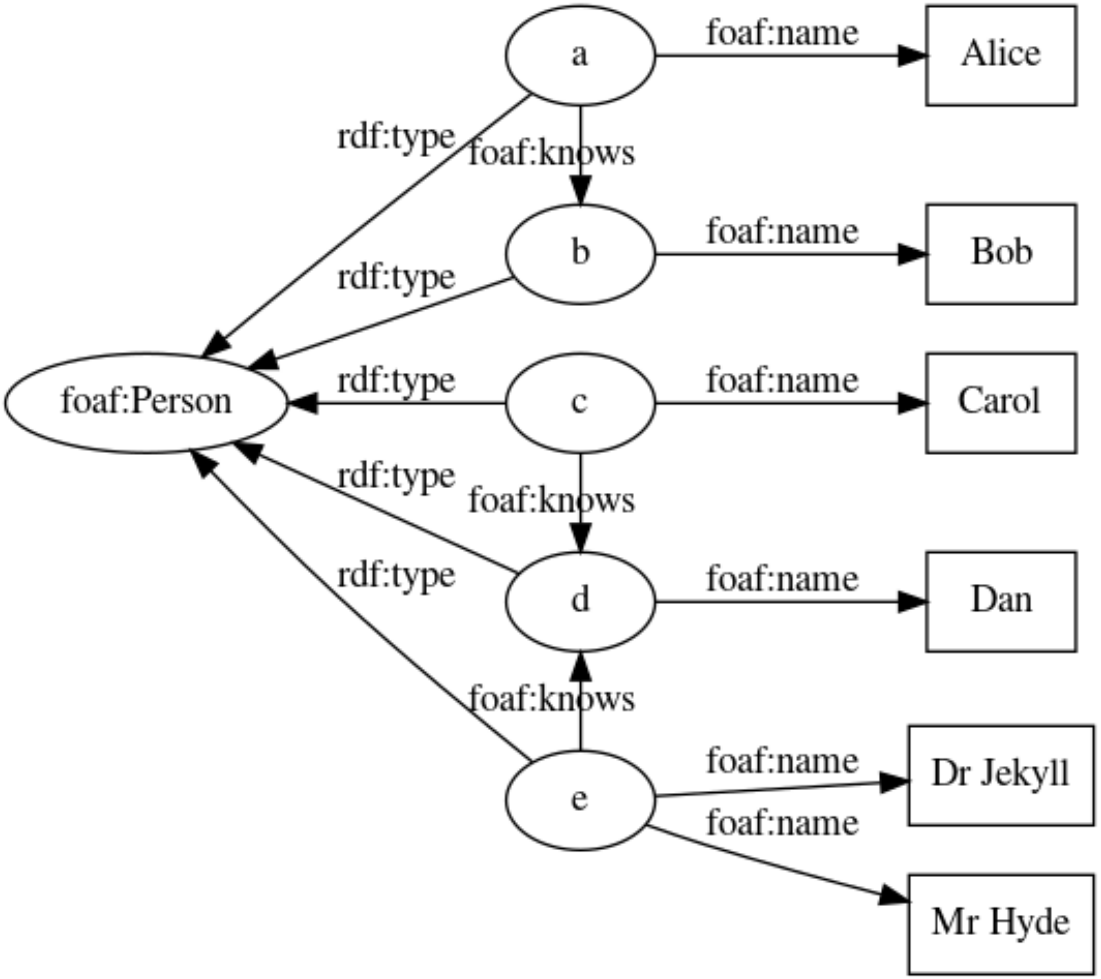
3. SparQL : objectifs

- Vous donner des bases pour écrire des requêtes SPARQL.
- Bonus: lire/écrire du Turtle (très proche de SPARQL).
- Ce n'est qu'une introduction ; pour en savoir plus :

  http://www.w3.org/TR/sparql11-overview/

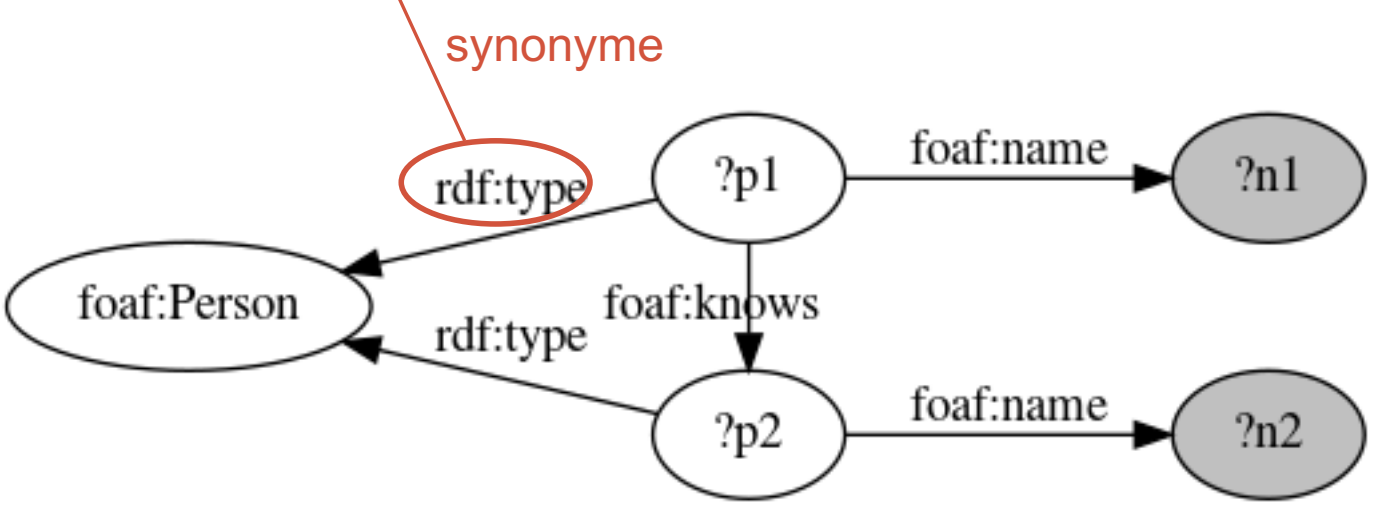
3. SparQL : requête simple

- Considérons le graphe de données suivant



3. SparQL : requête simple

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?n1 ?n2
WHERE {
  ?p1 a foaf:Person;
      foaf:name  ?n1;
      foaf:knows ?p2.
  ?p2 a foaf:Person;
      foaf:name  ?n2.
}
```



3. SparQL : requête simple

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?n1 ?n2
WHERE {
  ?p1 a foaf:Person;
        foaf:name  ?n1;
        foaf:knows ?p2.
  ?p2 a foaf:Person;
        foaf:name  ?n2.
}
```

n1	n2
Alice	Bob
Carol	Dan
Dr Jekyll	Dan
Mr Hide	Dan

3. SparQL : description

Préfixes : les préfixes servent à abréger les IRIs.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX : <http://example.com/>
```

Termes

IRI en extension (relatif ou absolu) :

```
<http://xmlns.org/foaf/0.1/Person>
<../other-file.rdf>
<#something>
<>
```

IRI abrégé :

```
foaf:Person
:something
```

3. SparQL : description

Littéraux :

```
"Hello"@en          # avec tag de langue
"123"^^xsd:integer  # typé

"Bonjour"           # equiv. "Bonjour"^^xsd:string
42                  # equiv. "42"^^xsd:integer
1.5                 # equiv. "1.5"^^xsd:decimal
314e-2              # equiv. "314e-2"^^xsd:double
true                # equiv. "true"^^xsd:boolean
```

Nœud muet :

```
_:toto
[]      # voir ci-après
```

Variables :

```
?x
$y
```

3. SparQL : triplets

- 3 termes (sujet, prédicat, objet) séparés par des espaces et suivis d'un point "." :

```
?p1 foaf:name "Pierre-Antoine Champin".
```

- Cas particulier : le mot clé "a" en position de prédicat est un raccourci pour <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> :

```
?p1 a foaf:Person.
```

3. SparQL : factorisation

- On *peut* « factoriser » plusieurs triplets ayant le même sujet en séparant les couples <prédicat, objet> par un point-virgule ";" :

```
?p1 a foaf:Person;
    foaf:givenName "Pierre-Antoine";
    foaf:familyName "Champin".
```

- On *peut* « factoriser » plusieurs triplets ayant le même sujet et le même prédicat en séparant les objets par une virgule "," :

```
?p1   foaf:phone <tel:+33-472-44-82-40>,
                 <tel:+33-472-49-21-73>.
```

- On peut bien sûr combiner les deux types de factorisation.
- On n'est jamais obligé de factoriser, on peut aussi répéter les termes.

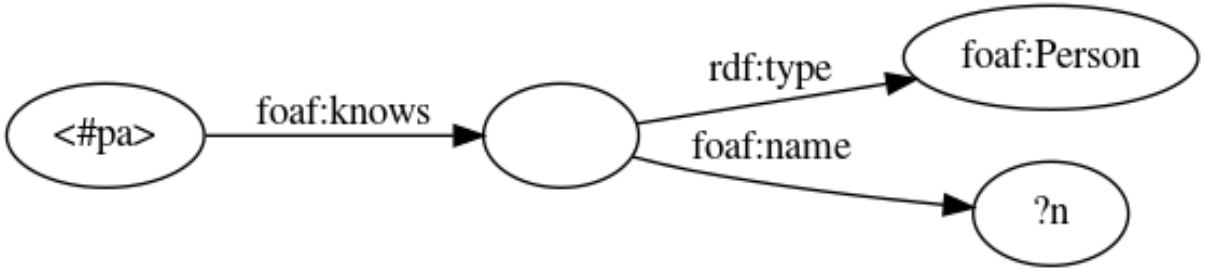
3. SparQL : nœuds muets

- Lorsqu'un nœud muet n'a qu'un seul arc entrant, au lieu de lui inventer un identifiant local :

```
<#pa> foaf:knows _:quelqun.
_:quelqun a foaf:Person;
          faof:name ?n.
```

on peut utiliser la notation [] :

```
<#pa> foaf:knows [
      a foaf:Person;
      faof:name ?n.
].
```



3. SparQL : union

- Pour exprimer un "ou" logique entre plusieurs contraintes, on place chaque alternative entre accolades, séparées par le mot-clé UNION :

```
<#pa> foaf:knows ?p1.
{ ?p1 a foaf:Person; foaf:name ?n}
UNION
{ ?p1 a schema:Person; schema:name ?n}
```

3. SparQL : sous-graphe optionnel

On peut accepter qu'une partie du graphe ne soit pas satisfaite :

```
?p1 a foaf:Person;
    faof:name ?n.
OPTIONAL {?p1 foaf:img ?img}
OPTIONAL {?p1 foaf:phone ?tel}
```

Ou

```
?p1 a foaf:Person;
    faof:name ?n.
OPTIONAL {?p1 foaf:img ?img. ?p1 foaf:phone ?tel}
```

Dans le résultat, les variables des clauses optionnelles peuvent donc ne recevoir aucune valeur (null).

3. SparQL : filtres

On peut ajouter des contraintes sur les valeurs des résultats, avec la clause FILTER.

```
?p foaf:age ?a.
FILTER (?a >=18)
```

On peut combiner des conditions avec les opérateurs logiques « et » (&&), « ou » (||) et « non » (!).

```
FILTER (?a >=18 && a<30)
```

3. SparQL : filtres / opérateurs et fonctions

- comparaisons : =, !=, <, >, <=, >=
- opérateurs arithmétiques : +, -, *, /
- nature d'un nœud : isIRI, isBLANK, isLITERAL, isNUMERIC
- vérifier qu'une variable (utilisée avec OPTIONAL) a bien une valeur : Bound
- recherche de texte : REGEX(<variable>, <texte>)

Pour plus d'information, consultez la documentation de SPARQL.

3. SparQL : requête SELECT

Similaire au SELECT de SQL :

projection sur un sous-ensemble des variables du graphe

Résultat : tableau

une colonne par variable sélectionnée

une ligne par résultat

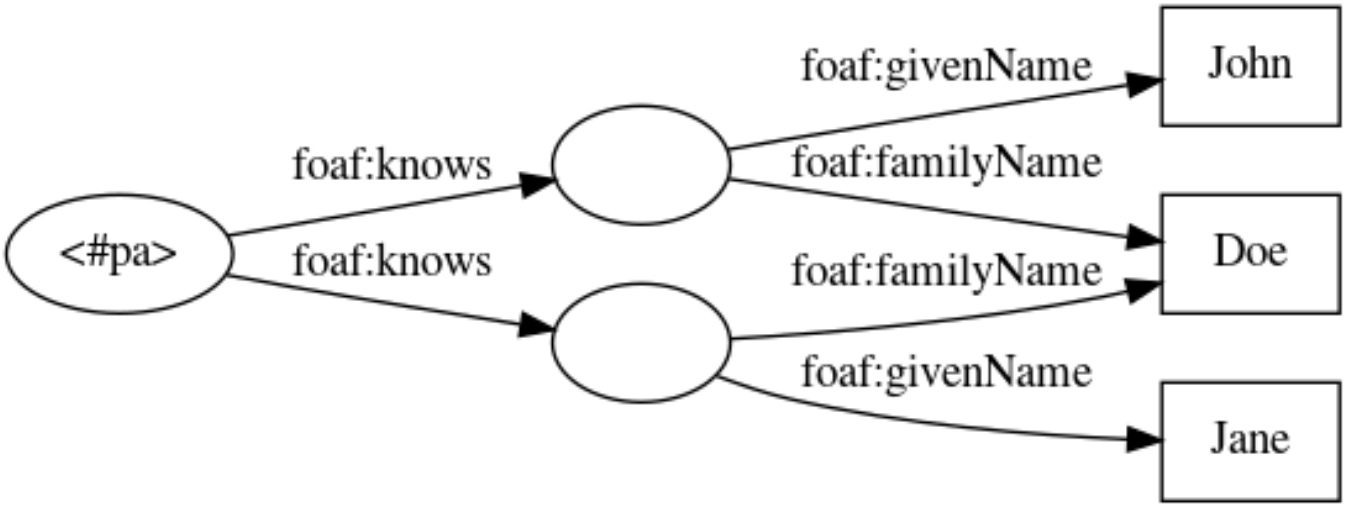
Structure :

SELECT <variables/expression> WHERE { <graphe> }

Les résultats du SELECT peuvent être des expressions complexes, calculées à partir des variables de la clause WHERE.

```
SELECT ?p (concat(?gn, " ", ?fn) as ?name)
WHERE {
  ?p foaf:givenName ?gn;
     foaf:familyName ?fn.
}
```

3. SparQL : Distinct




```
SELECT DISTINCT ?sn
WHERE {
  <#pa> foaf:knows ?p.
  ?p foaf:familyName?sn.
}
```

Sans le DISTINCT, la requête renverra deux fois le résultat sn="Doe".

3. SparQL : LIMIT et OFFSET

Pour obtenir les 10 premiers résultats :

```
SELECT ?p
WHERE {
  <#pa> foaf:knows ?p.
} LIMIT 10
```

Pour obtenir les résultats de 5 à 15 :

```
SELECT ?p
WHERE {
  <#pa> foaf:knows ?p.
}
LIMIT 10
OFFSET 5
```

3. SparQL : ORDER BY

```
SELECT ?p ?n
WHERE {
  <#pa> foaf:knows [
    foaf:givenName ?p ;
    foaf:familyName ?n ]
}
ORDER BY ?n ?p
```

On peut aussi trier par ordre descendant :

```
SELECT ?name ?age
WHERE {
  <#pa> foaf:knows [
    foaf:name ?name ;
    foaf:age ?age ]
}
ORDER BY DESC(?age)
LIMIT 1
```

3. SparQL : GROUP BY

Sert à *agréger* certaines valeurs avec l'une des fonctions d'agrégations : *Count*, *Sum*, *Avg*, *Min*, *Max*, *GroupConcat* et *Sample*.

```
SELECT ?p1 (count(?p2) as ?cp2)
WHERE {
  ?p1 foaf:knows ?p2
}
GROUP BY ?p1
```

On peut combiner *GROUP BY* avec *ORDER BY* et *LIMIT* (attention à l'ordre) :

```
SELECT ?p1 (count(?p2) as ?cp2)
WHERE {
  ?p1 foaf:knows ?p2
}
GROUP BY ?p1
ORDER BY DESC(?cp2) LIMIT 3
```

3. SparQL : sous-requête

Il est possible d'inclure, dans une clause WHERE, une requête SELECT entre accolades.

Par exemple, la requête suivante donne, pour chaque personne, son écart par rapport à l'âge moyen.

```
SELECT ?p ((?age - ?ageMoyen) as ?ecart)
WHERE {
    ?p a foaf:Person ;
       foaf:age ?age.
    {
      SELECT (avg(?age2) as ?ageMoyen) {
        ?p2 a foaf:Person ;
            foaf:age ?age2.
      }
    }
}
```

3. SparQL : points d'accès

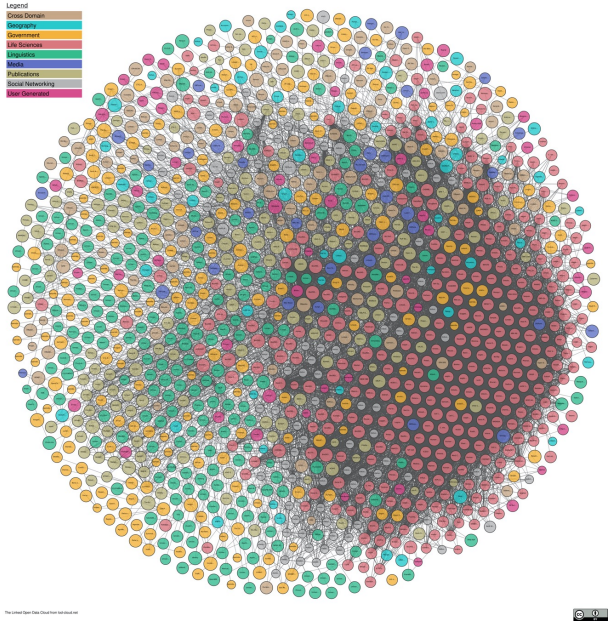
Pour programmer en SparQL, utilisez le client Yasgui.

Point d'accès :

- *DBPedia* : http://dbpedia.org/sparql
- *Nobel prizes* : http://data.nobelprize.org/sparql
- *Musiekweb* : https://data.muziekweb.nl/MuziekwebOrganization/Muziekweb/sparql/Muziekweb
- *Europeana* (*European cultural artifacts*): https://sparql.europeana.eu
- *Patent ontology* : https://data.epo.org/linked-data/query
- *Openstreetmpap (sophox)* : https://sophox.org

***List of Sparql endpoints* :**

- List of datasets with a endpoint : HERE
- Wikidata : List of Sparql endpoints
- The **Linked Open Data** Cloud (cliquez sur une disque pour obtenir son *endpoint*)
- W3C currently alive endpoints
- Mannheim linked data catalog : http://linkeddatacatalog.dws.informatik.uni-mannheim.de



Trouver une IRI/URL à partir de son préfix

http://prefix.cc/

3. SparQL : Exploration des ressources

Sur *Musiekweb*, la requête suivante

```
SELECT ?type (count(?o) as ?counto)
WHERE{
  ?o a ?type
}
GROUP BY ?type
ORDER by DESC(?counto)
LIMIT 30
```

génère

30 results in 1.117 seconds Simple view ☐ Ellipse ☑ Filter query results

	type	counto
1	vocab:OrderInformation	"753305"^^xsd:integer
2	vocab:Album	"724030"^^xsd:integer
3	schema:MusicAlbum	"724030"^^xsd:integer
4	schema:MusicGroup	"530077"^^xsd:integer
5	vocab:PopularAlbum	"435631"^^xsd:integer
6	vocab:ExternalLink	"394543"^^xsd:integer
7	vocab:PopularPerformer	"390866"^^xsd:integer
8	vocab:ClassicalAlbum	"159652"^^xsd:integer
9	schema:Person	"155303"^^xsd:integer
10	vocab:ClassicalPerformer	"109963"^^xsd:integer

3. SparQL : exploration des propriétés

Sur *Musiekweb*, la requête suivante

```
1  PREFIX schema: <http://schema.org/>
2  PREFIX  vocab: <https://data.muziekweb.nl/vocab/>
3  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
4  SELECT distinct ?val
5  WHERE{
6    ?o a schema:MusicGroup;
7       rdf:type ?val.
8  }
```

génère

7 results in 1.701 seconds Simple view

	val
1	vocab:ClassicalPerformer
2	schema:Composer
3	schema:MusicGroup
4	schema:Person
5	vocab:Performer
6	vocab:ImportantPerformer
7	vocab:PopularPerformer