Interoperabilité
Astronomie Planétologie

Pierre Le Sidaner : Paris Astronomical Data Centre /
Direction Informatique de l'Observatoire de Paris

ISAE – 25/3/2020

Interopérabilité pouquoi faire ?

- L’interopérabilité c'est quoi ?
  - Dans le monde anglo saxon
    - Syntactic interoperability (communication inter systh.)
    - Semantic interoperability (comparaison des données)
    - open standards
  - Dans notre quotidien Européen / Français
    - Interopérabilité / Compatibilité
    - Description des interfaces
      - Norme ou ensemble de normes et formats ouverts

On ne s'occupera ici que de la partie numérique

Interopérabilité pourquoi faire ?

- On veut accéder à des données dont on ne connaît pas la localisation ou le mode d'accès
- On veut croiser ses propres données avec d'autres existantes complémentaires (formats différents)
- On veux fouiller des jeux de données pour extraire celles qui vérifient des critères très particuliers
- On veut des données d'un domaine que l'on ne maîtrise pas complètement
- On veut faire traiter des informations par des machines.

- L'interopérabilité existe même dans la loi
  - La notion d'interopérabilité par des 'normes ouvertes ' s'oppose à la captivité du marché par un acteur
  - La loi encadre un droit de décompilation : interdit sauf à des fins d'interopérabilité
    cf : les présentations sur le domaine dans les JDEV
    https://webcast.in2p3.fr/videos-jdev2015_protection_du_logiciel_suite

Principe FAIR et licence

On déborde un peu de l'interopérabilité pour présenter le FAIR

- Findable Accessible Interoperable Re-usable
- Pourrait apparaître comme un comportement "altruiste"
- Accessible : est aussi en rapport avec les licences
- Réutilisable : notion de qualité de description et de licence !
- Vous y serez confronté : Programmes Europeens et sous traitance publique...

- La science « ouverte » est très incitée par l'envie de voir les moyens investi mieux utilisés
- Aussi développé par l'arrivé du « big data » qui à besoin de faire de la fouille dans des données bien décrites et de l'apprentissage.

Licences

Open Data c'est l'accès ouvert aux données pour le bien publique
Très vaste sujet que l'on va juste présenter.
Concerne la science mais aussi le transport, la géographie, les statistiques, la sociologie, le juridique …
Implique la gratuité .

Oui mais quel politique de gestion des données le DMP !
Qui inclus aussi : quelle licence pour ses logiciels et ses données ?

- Creatives Commons, Apache2, GPL, BSD ….
- Protection des développeurs de des fournisseurs de données
- Droit d'utilisation et d'extension
  - Notion de Copyleft
- Notion de contamination

Quelques Licences

- Creatives Commons,
  - CC énormément de variantes
- GPL la plus « libre » et la plus connue. Extension AGPL, LGPL ...
- Apache 2.0 très ouverte
- Protection des développeurs de des fournisseurs de données
- Droit d'utilisation et d'extension
  - Notion de Copyleft
- Notion de contamination

- Le modèle économique ?
  - Fourniture de services payant
  - Version hébergée
  - Fork

Questions à ce poser :
 * Je développe pour en faire un hobby
 * Je souhaite que les développements me fasse embaucher et permette d'en faire un produit commercial.

Pleins de sites type http://choosealicense.com/

Data Management Plan

- Besoin de décrire ce que l'on va faire des données,
- Exigence de l'Europe et question qui revient régulièrement.
- Croissance exponentielle des données
- Pas de FAIR sans pérennisation des données
    - Coût financier et humain
    - Modèle ou le marchand n'est pas le fournisseur de donnée !
- On conserve quoi, comment et pour combien de temps.

- Les soucis de conservation et d'utilisation des données :
  - Droit à l'Oubli et RGPD :
  - Données personnelles RGPD
  - On doit explicitement décrire ce que l'on fait des données personnelles

Interopérabilité comment ?

- Parler le même langage
  - On crée des modèles de données pour décrire ce dont on parle. En général un modèle UML souvent sérialisé en XML.
  - On crée une sémantique pour causer le même langage.
  - On crée des protocoles d'accès et des formats d'échange
  - On crée des formats standards pour les données
  - On crée un mode de recherche : pages jaunes, moteur de recherche .

- Créer un système lié à une discipline
  - Conduit par un besoin
  - Parle le même langage <=> modèles de données
  - Besoin d'efficacité et de fouille de données
  - Besoin d'accéder aux données des disciplines connexes
  - ....

Quelle type d'organisation ?

- Organisation de référence le W3C W3C (développements standards web).
- Organisation de partenariat Industriels/Chercheurs OGC OGC.
- Organisation 'Best Effort' IVOA IVOA
- Organisation d'agences spatiales IPDA IPDA
- Projets Européens dont RDA RDA.

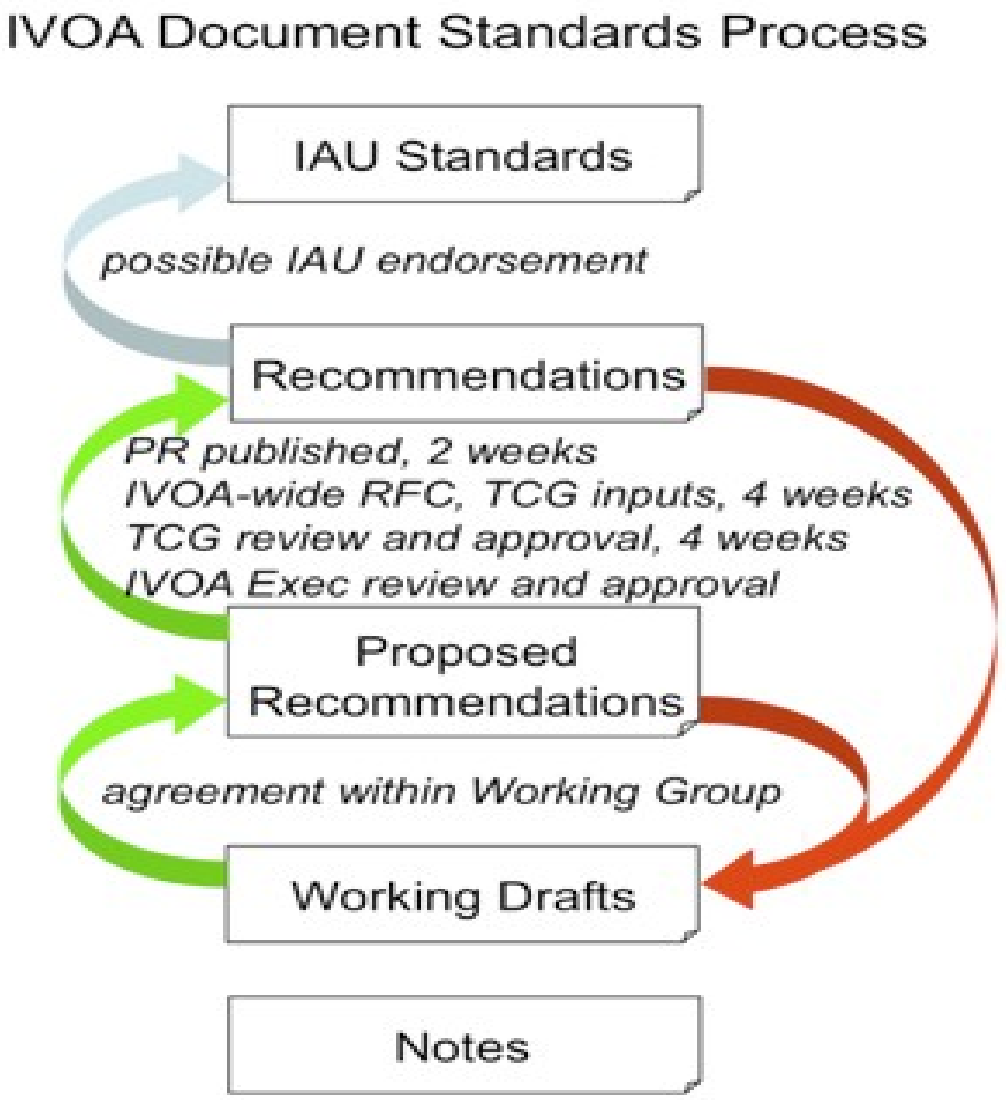
- Comment créer des standards
  - Initiatives
  - Règles de fonctionnement
  - Procédure de labellisation
  - Une organisation humaine pour assurer le fonctionnement
  - ....

Cas de l'Astronomie

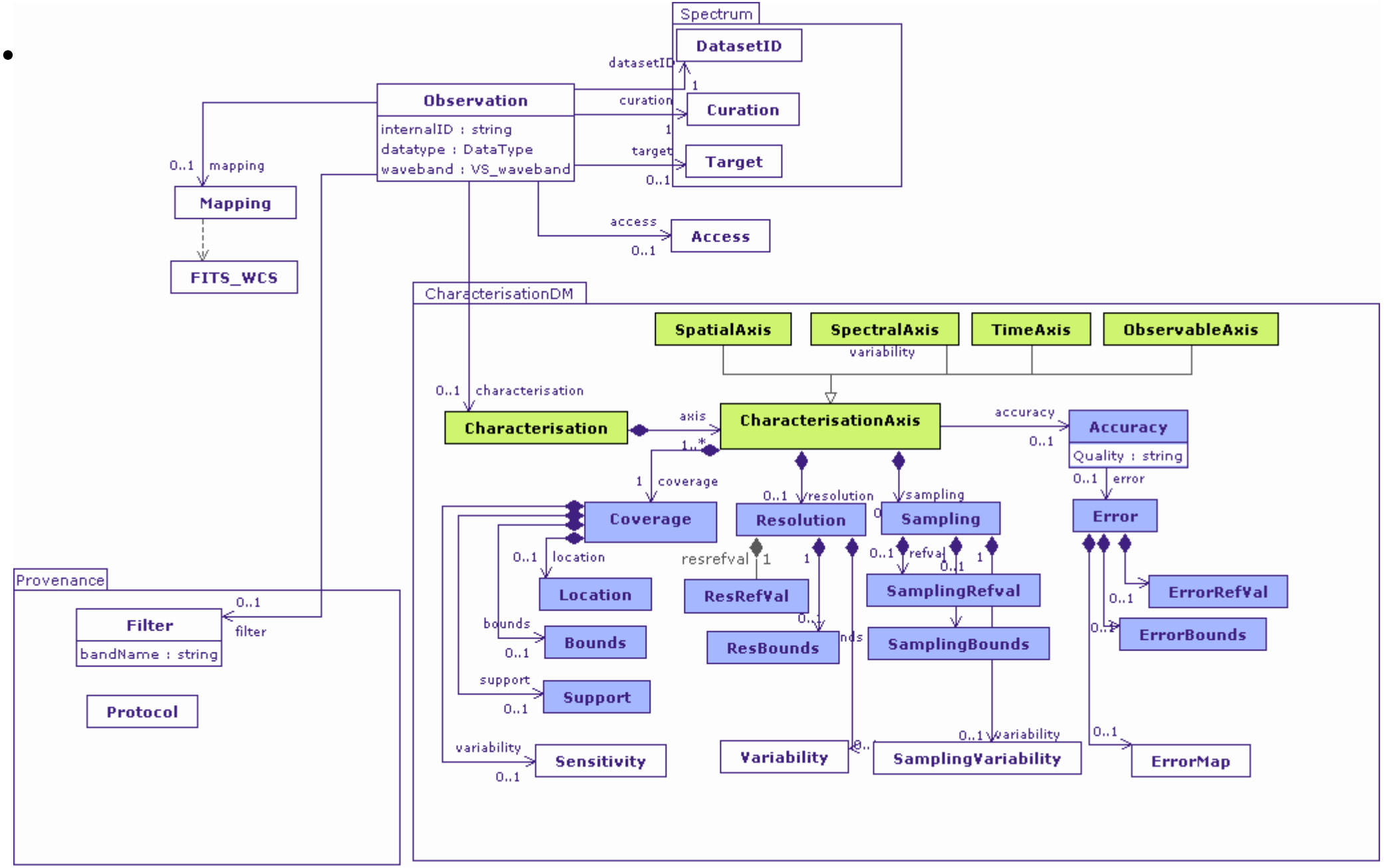
- Organisation en groupes de travail “thématiques”
  - Data Model → définit les modèles de données
  - Data Access Layer → définit les protocoles et les modes d'accès
  - Grid & Web Services → définit le calcul à la demande et la gestion de processus asynchrones
  - Registry → définit le SI, l'enregistrement et la localisation des services et des données.
  - Semantic → s'occupe des dictionnaires.
  - Applications → pour le lien entre les services et les applications notamment de visualisation. Pour le lien inter applications.

  - Il existe aussi des groupes d’intérêt :
    Planetologie, théorie, machine learning …

Processus de standardisation



Modèle de données



Data Access Layer

- Comment accéder à des ressources distantes (protocoles)
  - Le modèle Astro d'interopérabilité souhaite qu'une machine puisse comparer des données provenant de différentes sources
  - Comment interroger un service
    - SOAP, CGI, REST (GET & POST)
  - Quels sont les paramètres standards d'interrogation, souvent liés au Data Model
  - Quels sont les paramètres optionnels, services dépendants
  - Comment doit répondre un service (format standard VOTable et meta données d'informations).
  - Comment récupère-t-on les données sélectionnées
  - Quels formats pour les données

Format d'échange

- Création de VOTable issue d'XML
  - Notion de meta données pour caractériser une donnée
    - Temps (début, fin, échantillonnage, résolution)
    - Spatiale (origine, repère, position, résolution …)
    - Fréquence
    - Instrument, détecteur
    - Taille et format des données
    - ...
  - Statuts de la réponse OK, Overflow, Error
  - Url d'accès aux données
  - Url d'accès aux données de calibartion ...

Format des données

- Trouver des formats de données ouvert, stable, bien décrit
  - FITS (Flexible Image Transport System)
    - En-tête ascii avec clef : valeur
    - Données ascii ou binaires
    - Format historique donc stable
    - Pas optimum
    - Gère les cubes de données
    - …
  - VOTable
    - XML moins verbeux, les balises ne sont écrites qu'une fois comme pour un tableau HTML
    - Notion de groupes interne
    - Liste de meta données, pas d'ordre prédéfini
    - Possibilité de mettre du binaire aussi

Registry

- Trouver les services rapidement de manière standard et avoir la description adéquate

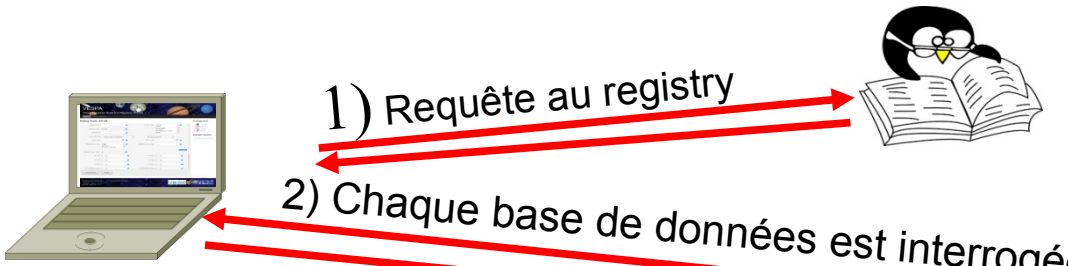
- Décrire les services
  - Standard de description du protocole, du « propriétaire », des intervenants des personnes à contacter en cas de problème, des dates initiale, mise à jour, la couverture (spatiale, temporelle, spectrale), l'instrument … Publication associé
  - Standard d'interrogation
  - Standard de format de sortie
  - Standard de réplication inter registries

Quel niveau de description ?

Applications

- Idée de discuter des besoins que doivent rendre les applications aux utilisateurs
- Les utilisateurs passent le plus souvent par des clients informatiques graphiques mais aussi par la ligne de commande.
- Chaque outil bien construit est spécialisé. Il n'y a pas de client universel
- Besoin de faire interagir les clients entre eux pour générer un environnement de travail
- Protocole d'échange inter clients SAMP Simple Application Message Protocol

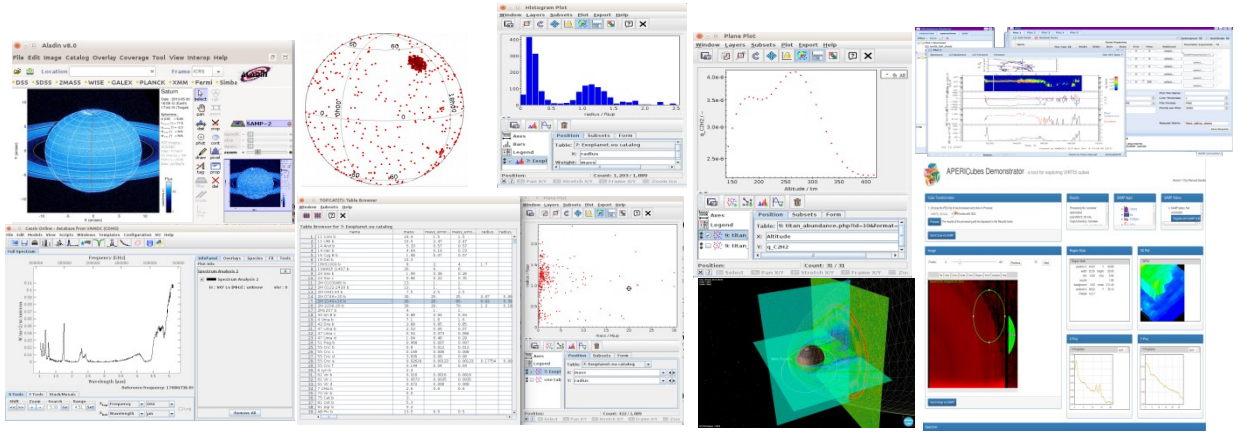
Comment cela fonctionne



Les données sont distribuées autour du monde.
Le modèle veut que les données soient réparties

Les clients peuvent être web ou client lourd

3) On peut combiner les clients pour plus d'interactivité



Perspectives

Dans le domaine scientifique l'interopérabilité se développe vite :
- Données géolocalisées : OGC / GIS
- Données médicales et biologiques
- Plus généralement partout où l'on a des masses de données
- Plus rapidement quand les données sont ouvertes

Demande forte des gros acteurs d'avoir accès à toutes les banques de données

Les standards sont aussi validés par l'utilisation
- ce n'est pas toujours le “meilleur” standard qui gagne

Utiliser des standards ouverts, c'est aussi souvent utiliser des outils libres existants
- gain de productivité
- avoir une communauté vivante

Si vous êtes confrontés à des besoins d'accès aux données

NE PAS FONCER EN CREANT UNE NOUVELLE METHODE AD HOC

Définition des besoin :
- Faire décrire les besoins coté utilisateur et fournisseur de données
- Faire décrire le domaine en essayant de le modéliser UML – Tableau – Map Mind
- Regarder les solutions existante pour profiter de l'expérience et des outils
- Coller si possible à des standards "ouverts"

- Trouver des formats de stockages ouverts aux données et si possible auto décrit
- Trouver un échantillons d'expert pour la définition de la sémantique des meta données.
- Bien séparer la partie protocole d'accès + dialogue machine des descriptions des données :
  - Il est possible que les unités ou grandeur d'intéroogation soit différent de ceux des données elle même
  - Ne pas transformer les données initiales si le processus n'est pas réversible

Les standards doivent être partagés
- vous devez autant que possible partager le standard avec tout les protagonistes si vous vous souhaitez qu'ils soient adoptés
- ce n'est pas toujours le meilleur qui triomphe !

Données temporelles

C’est aussi de l’interopérabilité mais dans un axe particulier :
- Tout bouge dans le temps
- Tweeter pas adapté à la science
- Besoin d’agir vite quand un événement survient (onde gravitationnelle, supernova ...)
- Des observatoires se robotisent
- Interaction inter domaines
- Space weather, interaction avec l’assurance des compagnies aérienne

Conclusion

Ne réinventez pas la roue :
- Beaucoup de standards déjà écrit
- Tout le monde gagne à ce que les données sont ouvertes

ENSUITE QUE CE PASSE T'IL ?

Si les données sont accessibles
- Traitement de données en masse : Big data, machine learning
- Trouver l'algorithme de séparation des données
- séparer des populations

Traitement de données en masse
- Souci du débit réseau pour les données scientifique
- Traitement et stockage parallélisé
- Transférer le traitement près des données => Calcul à la demande, processus d'authentification multi communautés, fédération identités

Perspectives

Dans le domaine scientifique l'interopérabilité se développe vite :
- Données géolocalisées : OGC / GIS
- Données médicales et biologiques
- Plus généralement partout où l'on a des masses de données
- Plus rapidement quand les données sont ouvertes

Demande forte des gros acteurs d'avoir accès à toutes les banques de données

Les standards sont aussi validés par l'utilisation
- ce n'est pas toujours le “meilleur” standard qui gagne

Utiliser des standards ouverts, c'est aussi souvent utiliser des outils libres existants
- gain de productivité
- avoir une communauté vivante