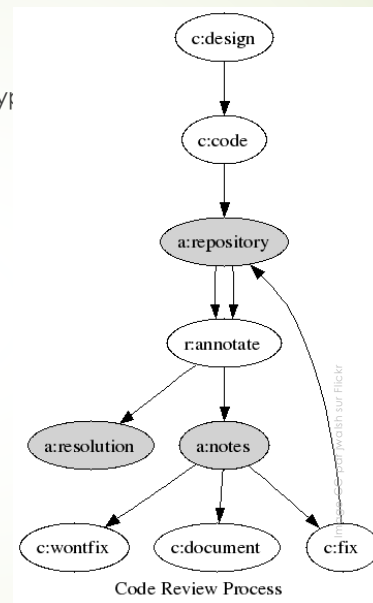


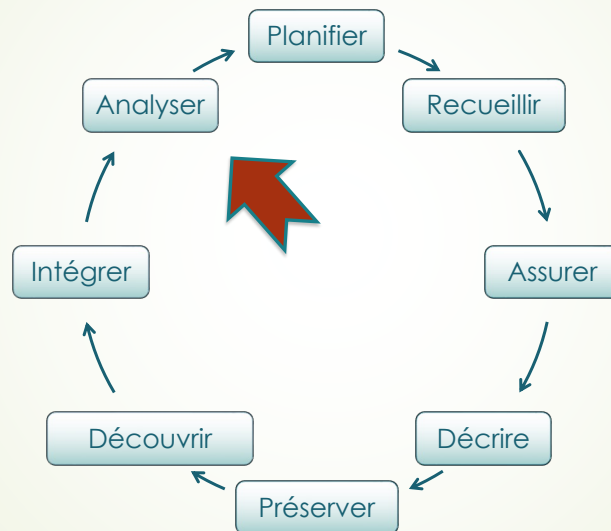
Workflows d'analyse de données

Plan du cours

- Aperçu des analyses de données typ
- Reproductibilité et provenance
- Workflows en général
- Workflows informels
- Workflows formels
- Contrôle de version

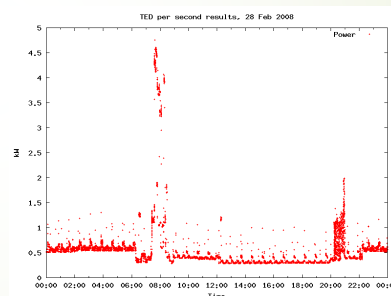
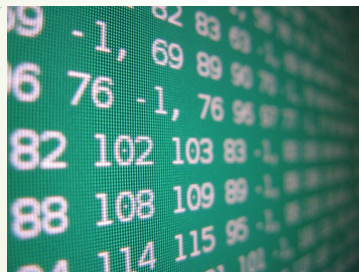


Le cycle de vie des données



Analyses de données

- Effectué à partir d'un ordinateur personnel, d'une grille de calcul, d'une plateforme de cloud computing
- Statistiques, exécutions de modèles, estimations de paramètres, graphiques/graphiques, etc.



Types d'analyses

- Traitement : restriction (sous-ensembles), fusion, manipulation
 - Réduction : importante pour les ensembles de données à haute résolution
 - Transformation : conversions unitaires, algorithmes linéaires et non linéaires

```
0711070500276000
0711070600276000
0711070700277003
0711070800282017
0711070900285000
0711071000293000
0711071100301000
0711071200304000
```



Datetime air tempprecip

Cmm

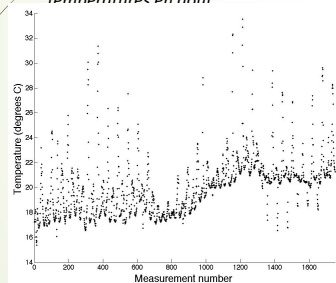
```
11 juil-075:0027.6000
11 juil-076:0027.6000
11 juil-077:0027.7003
11 juil-078:0028.2017
11 juil-079:0028.5000
11 juil-0710:0029.3000
11 juil-0711:0030.1000
11 juil-0712:0030.4000
```

Recrée de Michener & Brunt (2000)

Types d'analyses

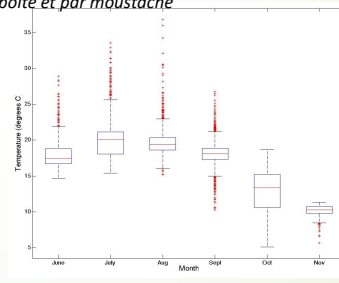
- Analyses graphiques
 - Exploration visuelle des données : recherche de modèles (patterns)
 - Assurance qualité : détection des valeurs aberrantes

Diagramme de dispersion des températures en août



Strasser, données non publiées

Diagramme de la température par mois, par boîte et par moustache



Strasser, données non publiées

Types d'analyses

■ Analyses statistiques

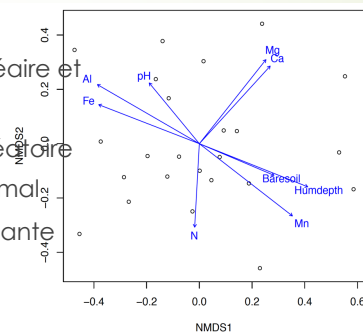
Statistiques conventionnelles

- Données expérimentales
- Exemples : ANOVA, MANOVA, linéaire et régression non linéaire
- S'appuyer sur des hypothèses : aléatoire échantillonnage, aléatoire et normale, erreur distribuée, erreur indépendante termes, variance homogène

Statistiques descriptives

- Données d'observation ou descriptives
- Exemples : indices de diversité, cluster analyse, quadrant de variance, méthodes de distance, méthode des quadrants, analyse en composantes principales, analyse des correspondances

Exemple d'analyse des principaux composants



De Oksanen (2011) Multivariable Analysis of Ecological Communities in R : vegan tutorial

Types d'analyses

■ Analyses statistiques (suite)

- Analyses temporelles : séries chronologiques
- Analyses spatiales : pour l'autocorrélation spatiale
- Approches non paramétriques utiles lorsque les hypothèses conventionnelles ont été violées ou que la distribution sous-jacente est inconnue.
- Autres analyses diverses : évaluation des risques, modèles linéaires généralisés, modèles mixtes, etc.

■ Analyses de très grands ensembles de données

- Exploration et découverte de données
- Traitement des données en ligne

Après l'analyse des données

- Analyse des résultats (outputs)
- Visualisations finales : tableaux, graphiques, simulations, etc.

**La science est itérative :
Le processus qui aboutit au produit final
peut être complexe**

Reproductibilité

- La reproductibilité est au cœur de la méthode scientifique
- Processus complexe = plus difficile à reproduire
- Bonne documentation requise pour la reproductibilité
 - Métadonnées : données sur les données
 - Métadonnées de processus : données sur le processus utilisé pour créer, manipuler et analyser les données.



Image CC par Richard Carter sur Flickr

Assurer la reproductibilité : Documenter le processus

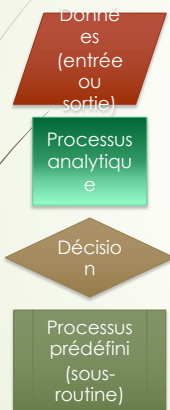
- *Traiter les métadonnées* : Informations sur le processus (analyse, organisation des données, graphiques) utilisé pour accéder aux sorties de données.
- Concept connexe : *provenance des données*
 - Origine des données
 - Bonne provenance = possibilité de suivre les données tout au long du cycle de vie.
 - Permet de
 - Reproduction et reproductibilité
 - Analyse des défauts potentiels, des erreurs de logique, des erreurs statistiques
 - Évaluation des hypothèses

Workflows : L'essentiel

- Formalisation des métadonnées de processus
- Description précise de la procédure scientifique
- Série conceptualisée d'étapes d'ingestion, de transformation et d'analyse des données
- Trois composantes
 - Entrées : informations ou matériel requis
 - Extrants : information ou matériel produit et potentiellement utilisé comme intrant dans d'autres étapes.
 - Règles/algorithmes de transformation (p. ex. analyses)
- Deux types :
 - Informelle
 - Formelle/Exécutable

Workflows informels

Diagrammes de workflow : Quelques éléments de base



- **Les entrées ou sorties** comprennent des données, des métadonnées ou des visualisations.
- **Les processus analytiques** comprennent les opérations qui modifient ou manipulent les données d'une manière ou d'une autre.
- **Les décisions** précisent les conditions qui déterminent la prochaine étape du processus
- **Les processus** ou sous-routines **prédéfinis** spécifient un processus fixe à plusieurs étapes.

Workflows informels

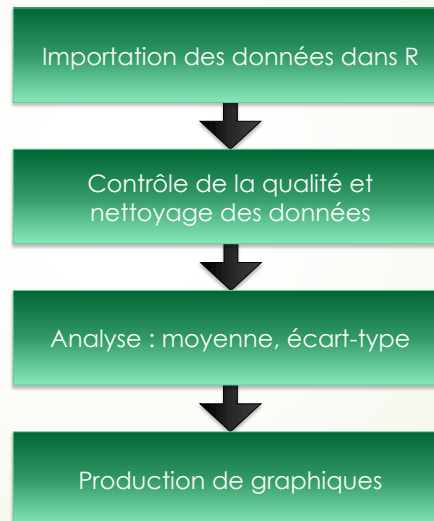
Diagrammes de workflow : Diagramme linéaire simple

- Conceptualiser l'analyse comme une séquence d'étapes
 - les flèches indiquent le débit



Workflows informels

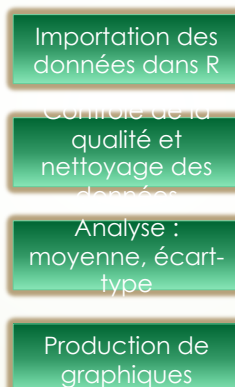
Organigrammes : forme la plus simple de Workflows



Workflows informels

Organigrammes : forme la plus simple de Workflows

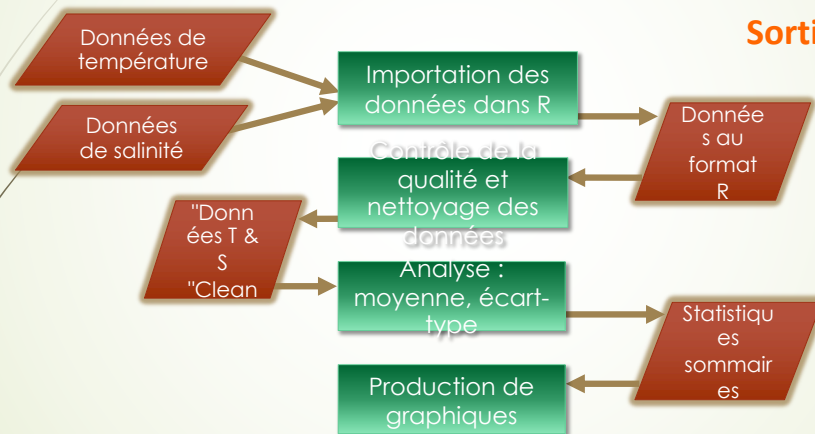
Règles de transformation



Workflows informels

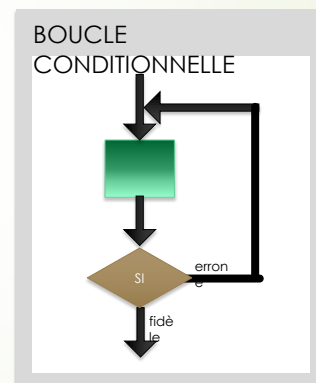
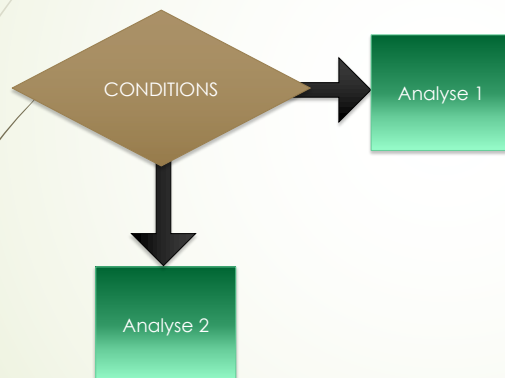
Organigrammes : forme la plus simple de Workflows

Entrées & Sorties



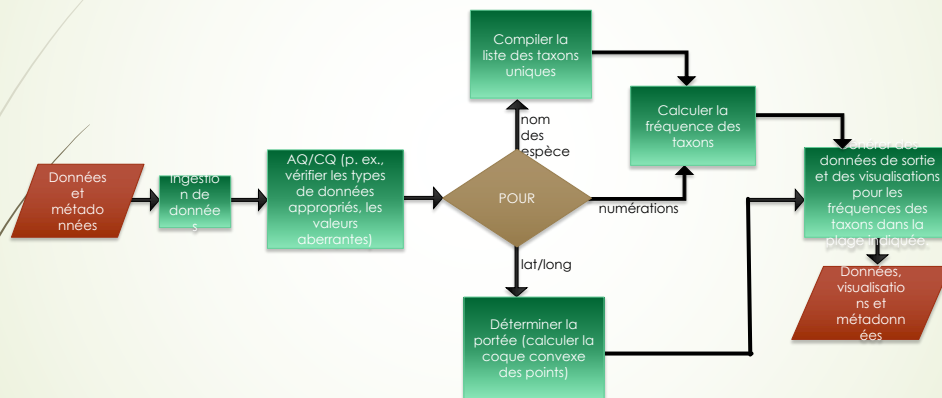
Workflows informels

Diagrammes de workflow : Ajout de points de décision



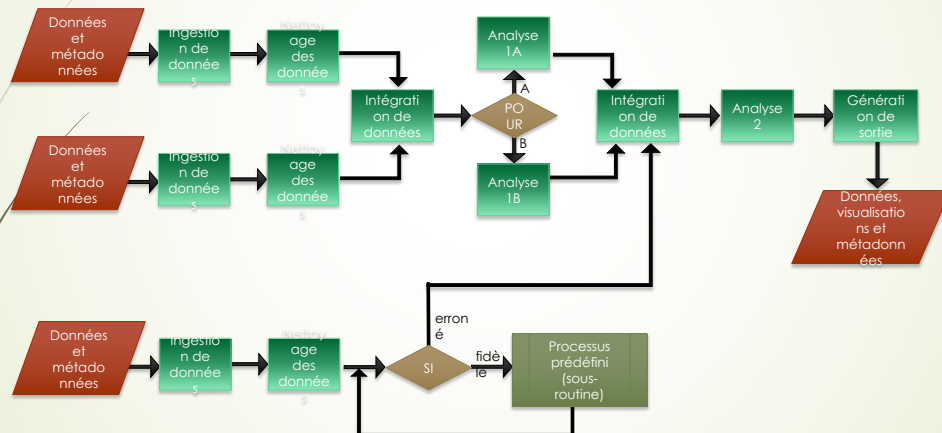
Workflows informels

Les diagrammes de Workflows : un exemple simple



Workflows informels

Les organigrammes : un exemple complexe



Workflows informels

Scripts commentés : meilleures pratiques

- Un code bien documenté est plus facile à réviser, à partager et permet des analyses répétées.
- Ajouter des informations de haut niveau en haut de l'écran
 - Description du projet, auteur, date
 - Dépendances, entrées et sorties des scripts
 - Décrit les paramètres et leurs origines
- Aviser et organiser les sections
 - Ce qui se passe dans la section et pourquoi
 - Décrire les dépendances, les intrants et les extrants
- Construire un script "end-to-end" si possible
 - Un récit complet
 - Fonctionne sans intervention du début à la fin



Workflows formels et exécutables

- Pipeline analytique
- Chaque étape peut être implémentée dans différents systèmes logiciels
- Chaque étape et ses paramètres/exigences sont formellement enregistrés
- Permet de réutiliser à la fois les étapes individuelles et l'ensemble du Workflows



Image CC par AJ Conn sur Flickr

Workflows formels et exécutables

Avantages :

- Point d'accès unique pour des analyses multiples sur plusieurs progiciels
- Garde la trace de l'analyse et de la provenance : permet la reproductibilité
 - Chaque étape et ses paramètres/exigences sont formellement enregistrés
- Le Workflows peut être enregistré
- Permet le partage et la réutilisation des étapes individuelles ou du Workflows global
 - Automatiser les tâches répétitives
 - Utilisation dans différentes disciplines et différents groupes
 - Possibilité d'effectuer des analyses plus rapidement, car il n'est pas possible de partir de zéro

Workflows formels et exécutables

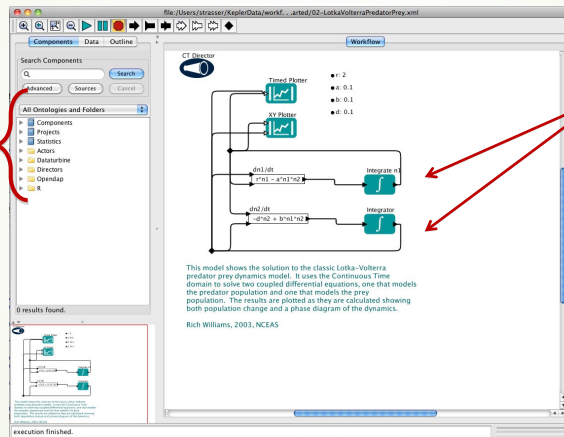
Exemple : Kepler Software

- Open-source, libre, multiplateforme
- Interface glisser-déposer pour la construction de Workflows
- Étapes (analyses, manipulations, etc.) dans le Workflows représentées par "acteur".
- Les acteurs se connectent à partir d'un workflow
- Applications possibles
 - Modèles théoriques ou analyses observationnelles
 - Modélisation hiérarchique
 - Peut avoir des workflows imbriqués
 - Peut accéder aux données à partir de sources Web (p. ex. bases de données)
- Téléchargements et plus d'informations sur kepler-project.org

Workflows formels et exécutable

Exemple : Kepler Software

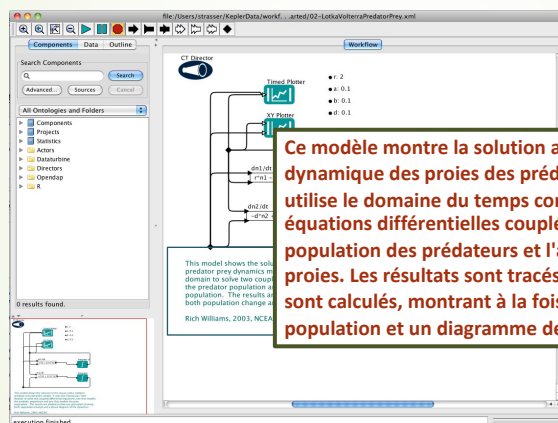
Glisser-déposer les composants de cette liste



Acteurs dans le workflow

Workflows formels et exécutable

Exemple : Kepler Software

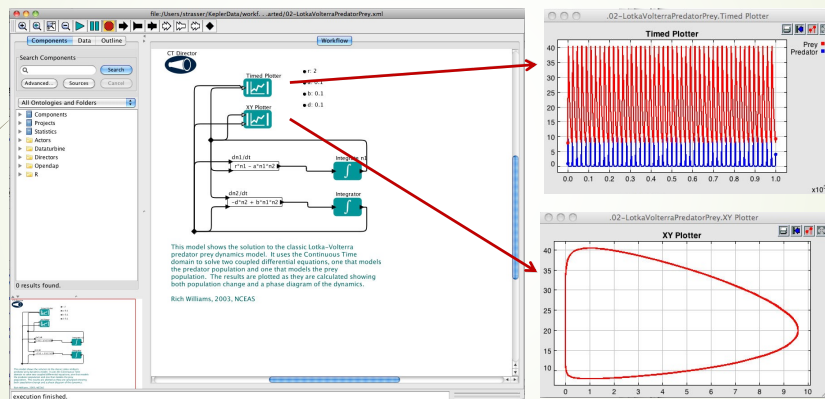


Ce modèle montre la solution au modèle classique de la dynamique des proies des prédateurs Lotka-Volterra. Il utilise le domaine du temps continu pour résoudre deux équations différentielles couplées, l'une qui modélise la population des prédateurs et l'autre la population des proies. Les résultats sont tracés au fur et à mesure qu'ils sont calculés, montrant à la fois l'évolution de la population et un diagramme de phase de la dynamique.

Workflows formels et exécutables

Exemple : Kepler Software

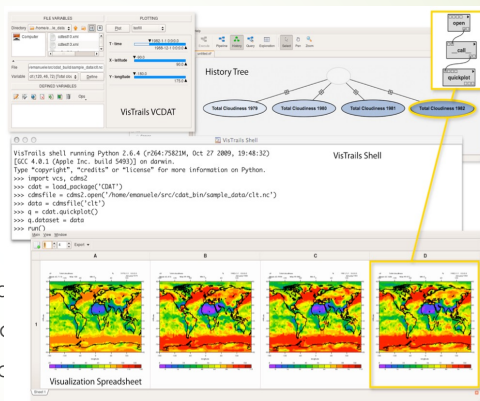
Sortie



Workflows formels et exécutables

Exemple : VisTrails

- Open-source
- Workflow & provenance appui à la gestion
- Orienté vers exploratoire tâches de calcul
 - Peut gérer des fonds souverains
 - Tient à jour un historique détaillé à propos des étapes et des calculs
- www.vistrails.org



Exemple de capture d'écran

Workflows en général

- La science devient de plus en plus intensive en informatique
- Le partage des Workflows profite à la science
 - Les systèmes de Workflows scientifiques facilitent la documentation des Workflows
- Minimalement : documentez votre analyse via des workflows informels
- Les nouvelles applications de Workflows (Workflows formels et exécutables) permettront de
 - **Logiciel de liaison** pour l'analyse exécutable de bout en bout
 - Fournir des **informations détaillées** sur les données et l'analyse
 - Faciliter la **réutilisation et le raffinement** d'analyses complexes en plusieurs étapes
 - Permettre l'**échange** efficace de modèles et d'algorithmes alternatifs
 - Aide à **automatiser les** tâches fastidieuses

Contrôle de version

- Logiciel pour gérer les modifications apportées aux fichiers, en particulier les scripts et le code source
- Essentiel à la gestion de l'évolution des Workflows
- Est utile pour gérer les modifications apportées au code et aux scripts
- Permet la collaboration à l'échelle
- Permet de suivre la révision exacte du code/script utilisé pour un Workflows.



Outils de contrôle de version

- Git - modèle distribué, largement utilisé, simple à brancher et à fusionner
- SVN - de nombreux projets hérités sont encore utilisés, nécessite un serveur maître pour collaborer.
- Mercurial - modèle distribué avec base utilisateur de niche



Bonnes pratiques pour l'analyse des données

- Il faut documenter les Workflows utilisés pour créer les résultats.
 - Origine des données
 - Analyses et paramètres utilisés
 - Connexions entre les analyses via les entrées et les sorties
- La documentation peut être informelle (p. ex. organigrammes, scripts commentés) ou formelle (p. ex. Kepler, VisTrails).



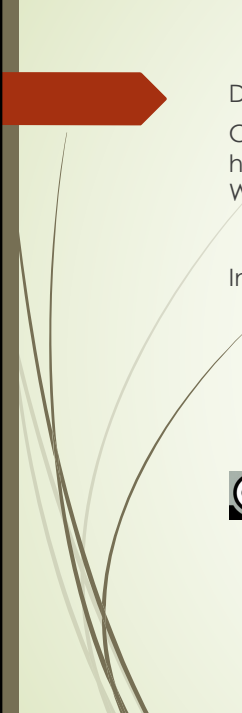
Calendrier CC image geek sur Flickr

Résumé

- La science moderne est à forte intensité informatique
 - Données, analyses et logiciels hétérogènes
- La reproductibilité est importante
- Workflows = métadonnées de processus
- L'utilisation de Workflows informels ou formels pour documenter les métadonnées de processus garantit la reproductibilité, la répétabilité, la validation et l'intégrité des données.

Ressources pour l'analyse des données et les Workflows

- 1. W. Michener et J. Brunt, Eds. *Données écologiques : Conception, gestion et traitement*. (Blackwell, New York, 2000).



DataONE Education Module : Analysis and Workflows.

Consulté le 12 novembre 2012.

http://www.dataone.org/sites/all/documents/L10_AnalysisWorkflows.pptx

Informations sur les droits d'auteur :

Aucun droit réservé ; vous pouvez améliorer et réutiliser pour vos propres besoins. Nous vous demandons de fournir une citation et une attribution appropriées à DataONE.

