Modèles et interopérabilité des modèles

Hélène Raynal[1]

Résumé. Les modèles font partie des outils largement utilisés par les scientifiques pour comprendre, diagnostiquer et prédire les phénomènes. Ils sont d'une grande diversité. La dynamique Open Science donne une nouvelle optique quant à leur construction et en particulier nécessite de développer la dimension de leur interopérabilité. Cette interopérabilité peut s'appuyer vers plus de standardisation des modèles et/ou par des méthodes d'ingénierie informatique.

Mots clés : modèle, interopérabilité, Open Science, simulation informatique, standards, couplage, workflow de traitements

Introduction

Les données ne présenteraient que peu d'intérêt si nous ne disposions pas d'algorithmes nous permettant de les interpréter et d'en extraire les informations utiles pour les travaux scientifiques. Cela est d'autant plus vrai que les flux massifs de données se généralisent (phénotypage haut-débit, séquençage du génome, acquisition automatisée par capteurs, enquêtes socio-économiques massives…) et pour lesquels il est nécessaire de disposer de méthodes de traitement adaptées. C'est la discipline de la « data science » (ou « ingénierie des données ») qui est au cœur de cet enjeu. Elle recouvre différentes facettes qui pour la plupart d'entre elles, font référence à la notion de « modèle ». Dans cet article, nous allons définir ce qu'on entend par modèle, puis aborder le sujet de l'interopérabilité. La question de l'interopérabilité est d'autant plus importante aujourd'hui dans le contexte de l'Open Science où modèles et données doivent être accessibles et réutilisables. Après avoir donné quelques règles de bonnes pratiques favorisant l'intégration des modèles dans le mouvement Open Science, cet article illustrera comment l'interopérabilité peut être raisonnée d'une part par les standards, d'autre part par l'ingénierie informatique. Un panel non exhaustif de méthodes et d'outils sera présenté pour chacun de ces axes, ainsi que des exemples illustratifs.

Contexte

Qu'est-ce qu'un modèle ?

Dans le domaine des sciences le terme « modèle » peut avoir plusieurs sens. Il peut désigner i) un organisme vivant utilisé pour des expériences scientifiques (exemple : un animal modèle comme la souris de laboratoire, une plante modèle telle que *Arabidopsis Thaliana*), ii) une vue de l'esprit permettant de comprendre le fonctionnement d'un phénomène, on parle alors de modèle conceptuel (exemple : les modèles conceptuels de bilan hydrique du sol utilisés pour comprendre la dynamique de l'eau dans les sols agricoles tels que le modèle « Soil water bucket» (Scotter et al., 1979) ou le modèle basé sur les équations de Richards (Richards, 1931),), iii) un modèle de simulation le plus souvent mis en œuvre avec un ordinateur. La version informatisée du modèle (c'est-à-dire le code informatique exécutable sur un ordinateur) s'appelle le simulateur. Dans cet article, seuls les modèles de simulation seront traités.

[1] *U875 MIAT, Inra Occitanie Toulouse, 31326 Castanet-Tolosan, France*
helene.raynal@inra.fr

Un modèle est une représentation simplifiée et partisane de la réalité. Il cherche à reproduire le mieux possible le comportement d'un système réel (une plante, un champ cultivé, un troupeau, des populations d'animaux sur un territoire donné, une forêt …). Il peut être utilisé i) pour poser un diagnostic comme établir une relation entre deux variables mesurés lors d'expérimentation (exemple **Figure 1**), ii) pour faire des prédictions[2] (de météo, de rendement agricole, d'évolution d'une épidémie au sein d'une population …), ou iii) pour tester un scénario d'évolution du système et donc répondre à la question « *que se passe-t-il si … ?* » (Exemple **Figure 2**).

Figure 1. *Relation entre la biomasse aérienne des formations ligneuses (kg) et le diamètre (cm) à 130 cm de hauteur des arbres (Kair, 1999).*

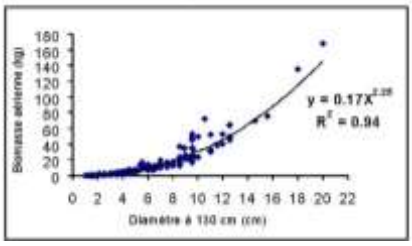
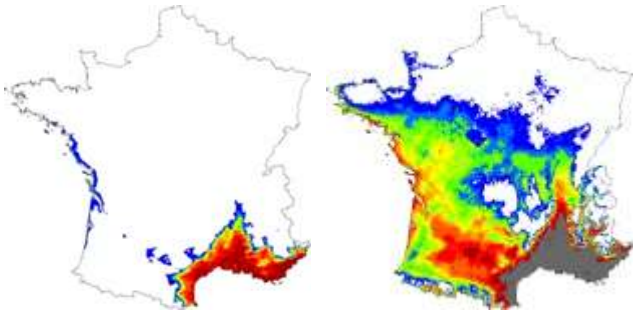


Figure 2. *Évolution de l'aire de répartition du chêne vert si la température moyenne annuelle évolue de +2,5 °C d'ici 2100.*
Source Inra, dossier « Quelles forêts en France d'ici 2100 ? » http://www.inra.fr/Entreprises-Monde-agricole/Resultats-innovation-transfert/Toutes-les-actualites/Quelles-forets-en-France-en-2100



Les différentes catégories de modèles

Compte-tenu de la diversité des modèles, des nombreux paradigmes et formalismes de modélisation, il est difficile de proposer un système de classification unique. On distinguera *a minima* les catégories présentées ci-dessous.

Modèle statistique versus modèle mécaniste

Un **modèle statistique** permet d'établir une relation empirique entre une variable et une ou plusieurs autres variables (exemple **Figure 1**). Il est basé sur des méthodes d'analyse statistique de données (ou inférence statistique). Dans le cas où le volume de données est important (contexte Big Data), on parle de méthodes de fouilles de données.
Un **modèle mécaniste** exprime la dynamique des processus relatifs au fonctionnement du système considéré. Ainsi dans le cas d'un modèle de plante, on décrira les processus de croissance, de transpiration, d'élaboration de la matière sèche, d'extraction de l'eau du sol par la plante …Contrairement au modèle statistique, il est dépendant des connaissances plus ou moins précises sur les mécanismes sous-jacents à ces processus. Il s'exprime à l'aide d'un formalisme de modélisation. Il existe un grand nombre de formalismes de modélisation, mais dans cet article nous nous restreindrons au cas des modèles exprimés par des équations mathématiques. Par exemple, l'évolution au cours du temps de la dynamique de deux populations en interaction peut être traduite en système de deux équations différentielles du 1er ordre. C'est le modèle « proie-prédateur » aussi appelé Lotka-Volterra du nom des deux scientifiques qui en sont à l'origine et qui l'ont appliqué à l'étude des populations de lynx et de lièvre des neiges au Canada (**Figure 3**) (Leigh et al.,1968). Les paramètres de ce système sont estimés à l'aide de données d'observation.

[2] La prévision se distingue de la prédiction par le fait qu'elle est affectée d'un caractère aléatoire dû à l'incertitude de l'avenir, tandis que les prédictions résultent de modèles mathématiques déterministes (source Wikipédia).

Figure 3. *Le lynx se nourrit à 80% de lièvres des neiges.*
Le nombre de lynx sur un territoire donné dépend directement du nombre de lièvres. L'évolution conjointe du nombre d'individus dans les deux populations peut être modélisée à l'aide d'un système d'équations.

Source : coyotes-wolves-cougars.blogspot.com

Le modèle « proie-prédateur » sert à exprimer l'interaction entre deux populations d'animaux : les proies et les prédateurs. Un tel modèle s'exprime par le système d'équations suivant qui permet de décrire l'évolution des deux populations au cours du temps :

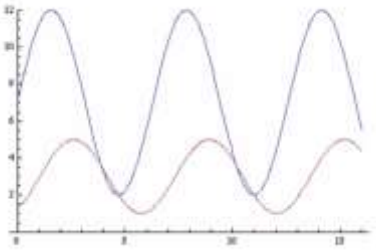
$$\frac{dx}{dt} = \alpha_1 x(t) - \beta_1 x(t) y(t)$$

$$\frac{dy}{dt} = -\alpha_2 y(t) + \beta_2 x(t) y(t)$$

Avec : respectivement x(t) et y(t) : le nombre de proies, de prédateurs à chaque instant t, et les paramètres : α_1 taux de reproduction intrinsèque des proies, α_2 taux de mortalité intrinsèque des prédateurs, β_1 taux de mortalité des proies en fonction du nombre de prédateurs, β_2 taux de reproduction des prédateurs en fonction du nombre de proies disponibles.

La **Figure 4** donne la dynamique classiquement obtenue lors de la simulation informatique d'un tel modèle.

Figure 4. *Courbes obtenues lors de la simulation informatique du modèle « proie-prédateur ».* Les courbes représentent l'évolution au cours du temps (en abscisse) des effectifs de proies (en bleu) et de prédateurs (en rose). On voit sur ce graphe qu'une augmentation de la population de proies entraîne une augmentation de la population de prédateurs car ces derniers bénéficient ainsi d'un plus grand « réservoir » alimentaire. Ceci jusqu'à un certain point, car l'excès de prédateurs va impliquer la diminution du nombre de proies.



Modèle déterministe versus modèle stochastique

Contrairement au modèle déterministe, un modèle stochastique inclut une part d'aléas. Ainsi, simuler deux fois le même modèle avec le même jeu de données ne donnera pas le même résultat avec un modèle stochastique, contrairement à un modèle mécaniste.

Les modèles dans le contexte de l'Open Science

Les modèles n'échappent pas au mouvement de l'Open Science et comme pour les données il est nécessaire de mettre en place un cadre méthodologique et technique facilitant les manières de construire, stocker, présenter ou publier les modèles dans ce contexte. Aujourd'hui, il n'y a pas de cadre « Open Science » unifié spécifique aux modèles. Cependant, il est naturel d'étendre aux modèles les quatre principes du FAIR (Mark D. Wilkinson et al., 2016) qui supposent que les données (et par extension les codes informatiques associés aux modèles) soient *« trouvables, accessibles, interopérables et réutilisables ».* Les principes du FAIR ont été décrits dans le

numéro spécial du Cahier des Techniques de l'INRA « Données de la recherche 2018 »[3]. Le **Tableau** ci-dessous donne quelques exemples de mise en œuvre opérationnelle des principes FAIR dans le contexte des modèles.

Principes FAIR	**Mise en œuvre**
Findable	
Les modèles doivent être identifiés par un identifiant global, unique et pérenne.	L'identification se fait très souvent par le **numéro de version** du code logiciel associé au modèle, et déposé dans un repository de modèles ou dans une forge logicielle. L'utilisation d'un **doi** est possible mais peu généralisée.
Les métadonnées décrivant les modèles sont riches.	Elles s'appuient sur des **vocabulaires contrôlés type thésaurus ou ontologies**. Exemple : le thésaurus AnaEE-France est utilisé pour la caractérisation des données et des modèles des écosystèmes continentaux. http://lovinra.inra.fr/2017/03/13/thesaurus-anaee/
Les modèles et les métadonnées sont enregistrés et indexés dans un dispositif permettant de les rechercher.	Exemples : • **Le repository (ou bibliothèque) de modèles** : il s'agit d'un ensemble de modèles partagés par une communauté de scientifiques, souvent sur un thème donné. Exemples bases de données BioModels (Chelliah et al., 2013), CellML (Cuellar et al., 2003) ou JWS (Olivier et al., 2004) • **La forge logicielle** : il s'agit d'un système de gestion de développement collaboratif de logiciel. Exemples : SourceSup https://sourcesup.renater.fr/, GitHub https://github.com/
Accessible	
Les modèles et les métadonnées sont accessibles par leur identifiant via un protocole de communication.	C'est un service en général fourni par le repository de modèles ou la forge logicielle.
Reusable	
Les modèles et les métadonnées sont mis à disposition selon une licence explicite et accessible.	L'Open Science suppose que le code informatique associé aux modèles soit **Open Source** et donc ait une licence adéquate. Il existe un grand nombre de licences dites Open Source. Une note quant au choix de la licence logicielle est disponible à l'adresse suivante https://www6.inra.fr/datapartage/Zoom-sur/Note-choix-licence-logicielle
Les modèles et les métadonnées correspondent aux standards des communautés indiquées.	Ils s'appuient sur des **vocabulaires contrôlés type thésaurus ou ontologies** conçus et développés par des communautés de modélisateurs.
Les modèles doivent être vérifiés.	Il est important de garantir la qualité d'un modèle. Pour cela un plan de tests doit être associé lors du développement logiciel (tests unitaires, de non régression …). D'autre part, il faut inclure à la distribution d'un modèle un jeu de tests permettant de vérifier son comportement dans un contexte de réutilisation.
Interoperable	C'est le principe FAIR souvent le plus difficile à mettre en œuvre. Il s'agit de rendre combinable un modèle ou des données avec d'autres modèles ou d'autres données, par des humains et des machines.

[3] https://www6.inra.fr/cahier_des_techniques/Les-Cahiers-parus/Les-N-Speciaux/Donnees-de-la-recherche-20182/Art1-ns-Donnees-de-la-recherche-2018

L'interopérabilité est un sujet important pour les modèles de simulation. En effet, la simulation informatique est aujourd'hui un outil essentiel dans l'étude de phénomènes complexes qui ne peuvent être traités de manière analytique. En informatique, l'interopérabilité est vue comme la capacité avec laquelle deux ou plusieurs programmes peuvent partager et traiter des informations indépendamment de leur langage et de leur plate-forme de mise en œuvre » (Howie et al., 1996). Pour une communauté scientifique, l'interopérabilité comprend tous les mécanismes qui vont permettre à plusieurs simulateurs d'utiliser la même description de modèle ou qui vont permettre de collaborer pour évaluer différentes parties d'un gros modèle. Dans le premier cas, il s'agira de décrire le modèle de simulation en utilisant des standards voire un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances (un des principes FAIR). Dans le deuxième cas, il s'agira de mettre en place des standards relevant le plus souvent de l'ingénierie informatique et facilitant le couplage de modèles ou l'enchaînement de traitements automatisés.

L'interopérabilité raisonnée par la standardisation de la description d'un modèle

Il n'y a pas de manière unique pour la standardisation de la description d'un modèle, mais des règles de bonnes pratiques plus ou moins élaborées qui favorisent l'interopérabilité, et peuvent être combinées ou non. Par ailleurs, la standardisation de la description complète du modèle est importante dans un contexte d'Open Science, car elle donne plus de transparence sur le contenu du modèle.

Séparer le code informatique propre au modèle du code informatique propre à l'exécution du simulateur sur un ordinateur

Dans le passé, beaucoup de simulateurs ont été construits de manière monolithique, c'est à dire sans distinguer ce qui relevait du code informatique propre au modèle de ce qui relevait de l'environnement informatique nécessaire à son exécution sur l'ordinateur (exemples : lecture des paramètres dans des fichiers externes, déclaration des variables …). Cela pose des problèmes d'intelligibilité du code informatique et de portabilité sur d'autres ordinateurs.

Utiliser au sein d'une même communauté scientifique le même langage informatique pour exprimer le modèle

Un des verrous à l'interopérabilité des modèles est le langage informatique (ou le logiciel) utilisé pour développer le simulateur associé au modèle. Une communauté scientifique peut donc choisir de s'imposer un langage informatique comme le fortran, C++, Java, python ... Ainsi, une large communauté de statisticiens a développé et utilise le langage générique R (https://www.r-project.org/). Ce langage est utilisé pour faire des traitements statistiques mais aussi pour mettre à la disposition des méthodes statistiques sous la forme de « paquets R ». Ces paquets sont entreposés au sein du repository CRAN (Comprehensive R Archive Network) et librement téléchargeables.

Utiliser au sein d'une même communauté scientifique, un langage déclaratif standardisé pour exprimer le modèle

Imposer un langage informatique reste contraignant pour une communauté où chacun a ses propres habitudes et peut être amené à contribuer à différentes communautés. La voie qui semble privilégiée actuellement pour l'interopérabilité des modèles est la conception d'un langage déclaratif pour la description et l'échange de modèles. Très souvent, les langages déclaratifs ont pour point de départ le langage XML (eXtensible Markup Language http://www.w3.org/XML). Plusieurs communautés ont déjà développé de tels langages qui sont devenus des standards. C'est le cas de SBML en biologie des systèmes (Hucka et al., 2003), CellML en physiologie cellulaire (Lloyd et al., 2004)) ou NeuroML en neuroscience (Friston et al., 2010). En biologie des systèmes, des journaux comme PNAS ou Nature exigent qu'à la publication soit associé le modèle exprimé dans un de ces formats pour

reproduire de façon indépendante les résultats scientifiques. Les modèles sont aussi disponibles dans des catalogues comme BioModels (https://www.ebi.ac.uk/biomodels), ou PlaSmo (http://www.plasmo.ed.ac.uk) pour les plantes, et partagés au niveau international. La **Figure 5** présente l'implémentation dans le langage CellMl d'une équation. Il est issu d'un tutorial disponible sur le site de CellML (https://models.cellml.org/e/e1/tutorial). La description d'un modèle (sa composition en termes de composants, les équations mathématiques …) dans un langage déclaratif est lourde et verbeuse, mais elle est exploitable pour i) générer des simulateurs exécutables dans différents environnements informatiques, en utilisant des traducteurs, ii) pour faire de l'introspection du modèle afin d'en récupérer de la connaissance sur le modèle, iii) pour faciliter voire automatiser l'interopérabilité avec d'autres modèles ou données.

Figure 5. *Une des équations du modèle de Hodgkin et Huxley (1952) concernant la conduction électrique de la membrane des cellules nerveuses, et son implémentation en langage déclaratif.*

$$\frac{dX}{dt} = \alpha_x(1 - X) - \beta_x X$$

```
<math xmlns=http://www.w3.org/1998/Math/MathML>
    <apply><eq/>
        <apply><diff/>
            <bvar><ci>time</ci></bvar>
            <ci>X</ci>
        </apply>
        <apply><minus/>
            <apply><times/>
                <ci>alpha_x</ci>
                <apply><minus/>
                    <cn cellml:units="dimensionless">1.0</cn>
                    <ci>X</ci>
                </apply>
            </apply>
            <apply><times/>
                <ci>beta_x</ci>
                <ci>X</ci>
            </apply>
        </apply>
    </apply>
</math>
```

Utiliser les vocabulaires (thésaurus, ontologies) pour annoter les modèles

Il est important de s'appuyer sur des vocabulaires partagés par les communautés scientifiques pour des raisons de cohérence de la sémantique du modèle, de traçabilité et de reproductibilité des résultats de la recherche. Pratiquement, cela suppose de caractériser à l'aide de métadonnées l'ensemble des variables et paramètres des modèles. Cette annotation facilite par ailleurs l'interopérabilité avec les données et les modèles, et l'automatisation des traitements.

L'interopérabilité raisonnée par l'ingénierie informatique

Dans ce chapitre nous n'abordons que deux cas de mise en œuvre d'interopérabilité de modèles à l'aide d'outils fournis par l'ingénierie informatique : le couplage de modèles et l'enchaînement automatisé des traitements informatisés. Ces deux cas sont représentatifs des cas d'interopérabilité traités dans nos communautés scientifiques et d'ingénieurs. C'est un des axes de travail du CATI IUMA[4] (Informatisation et Utilisation des Modèles d'Agroécosystèmes) qui a conçu et proposé des solutions aux problèmes d'interopérabilité rencontrés par la communauté des modélisateurs. Les deux exemples présentés dans les paragraphes suivants sont directement en lien avec ce qui a été mis en œuvre dans ce CATI. Enfin, l'interopérabilité des modèles avec les données sera brièvement évoquée en fin de chapitre.

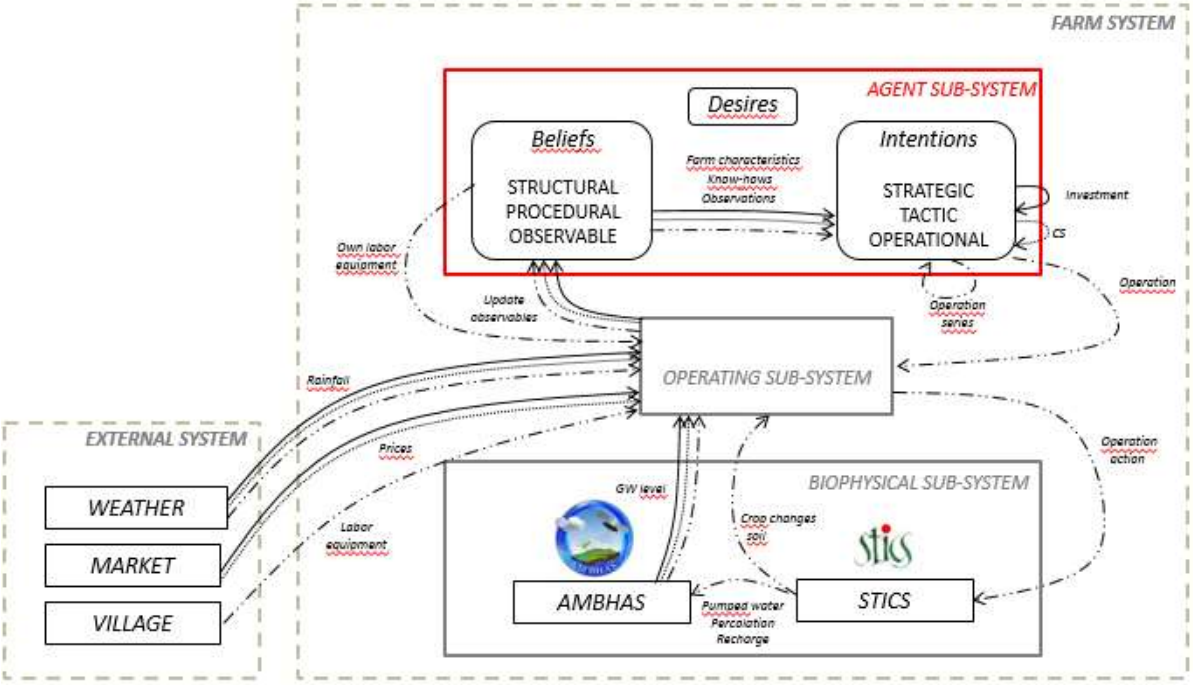
[4] Les activités de ce CATI ont été reprises depuis 2019 par le CATI IUMAN (Informatisation et Utilisation des Modèles d'Agroécosystèmes Numériques).

L'interopérabilité pour le couplage de modèles

Il est parfois nécessaire de faire interagir plusieurs modèles pour simuler un phénomène, on parle alors de couplage de modèles. C'est en particulier le cas, lorsque les phénomènes se complexifient et requièrent un haut degré d'expertise dans chacun des composants du couplage ou lorsque les problèmes sont multidisciplinaires ou multi-échelles (**Figure 6**).

***Figure 6**. Modèle de simulation du fonctionnement d'une exploitation agricole « FARM SYSTEM ».*

Le modèle est le résultat du couplage de plusieurs autres modèles correspondant aux quatre sous-systèmes : modèle de culture (STICS) qui représente la dynamique des parcelles agricoles de l'exploitation agricoles, modèle de disponibilité en eau dans la nappe phréatique pour l'irrigation (AMBHAS), modèle de conduite des cultures par l'agriculteur (AGENT SUB-SYSTEM), modèle de disponibilité des ressources humaines et matérielles de l'exploitation (OPERATING SUB-SYSTEM). Le modèle tient compte des variations des prix du marché, de la météo et de la structure du village agricole. (Robert et al., 2018)



En reprenant l'exemple présenté dans la **Figure 6**, l'interopérabilité entre les quatre codes informatiques associés aux quatre sous-systèmes (parcelles agricoles, nappe phréatique, agriculteur, ressources) peut être réalisée de différentes manières présentées ci-dessous.

- Couplage direct par simple échange de fichiers : le simulateur associé au modèle A s'exécute, génère les résultats dans un fichier au format texte, .csv … Ce fichier est ensuite utilisé par le simulateur associé au modèle B. Ce genre d'interopérabilité est facile à mettre en œuvre, mais l'écriture dans des fichiers ralentit l'exécution informatique et ne permet pas les rétroactions (Modèle B impacte l'exécution de A). On parle de couplage direct car les simulateurs ne sont pas modifiés. On parle de couplage faible car l'interopérabilité permise par ce genre de couplage est très limitée.
- Couplage direct par échange de messages au niveau système informatique : l'interopérabilité est assurée par l'échange de messages informatiques de bas niveau. Parmi les outils pouvant être utilisés il y a la librairie ZeroMQ, http://zeromq.org/ qui propose une interface simple afin de faire communiquer plusieurs

processus. Cette solution a également l'avantage d'être peu intrusive dans les codes existants et peut être utilisée dans des applications réseau. Comme précédemment on parle de couplage direct et faible même si les possibilités offertes par ce type d'interopérabilité sont plus grandes que dans le premier cas.

- Couplage indirect avec un langage informatique « glue »: il s'agit d'utiliser un langage de haut niveau appelé « glue » pour assurer l'interopérabilité entre les différents codes informatiques du modèle. Le langage Python https://www.python.org/ est très souvent utilisé. Dans ce cas, chaque code est encapsulé dans un script écrit en Python. L'application principale écrite en Python assure l'interopérabilité et la coordination des différents codes encapsulés. On parle de couplage indirect puisqu'un travail d'adaptation des simulateurs est requis, et de couplage fort, puisque l'interopérabilité entre les codes est assurée au sein du simulateur principal, ce qui offre beaucoup plus de possibilités de couplage que dans les deux cas précédents.
- Couplage au sein d'une plateforme informatique avec un unique moteur de simulation informatique : c'est la forme la plus aboutie d'interopérabilité car les codes couplés sont pilotés par un unique moteur de simulation appelé « noyau ». Chaque code couplé est alors vu comme un module de la plateforme. Les interfaces des modules doivent être conformes à la forme spécifiée au niveau de la plateforme. Bien que plus contraignante en termes d'adaptation ou construction de code, c'est une voie d'interopérabilité qui est de plus en plus privilégiée car elle offre une garantie de robustesse en termes d'interopérabilité et de solutions numériques optimisées. Elle permet par ailleurs de construire les modèles de manière modulaire. Il existe de nombreux outils « plateformes informatiques » qui proposent des services variés. Ils ont développé des solutions techniques différentes pour l'interopérabilité en fonction des attentes de leurs communautés d'utilisateurs. La plateforme RECORD développée à l'Inra (http://www6.inra.fr/record) est dédiée à la modélisation et la simulation informatique des agroécosystèmes. Elle offre un service de couplage multi-formalismes et multi-échelles, basé sur le formalisme DEVS (Zeigler, 1976). Le coupleur Open-Palm (http://www.cerfacs.fr/globc/PALM_WEB/) développé au CERFACS (Calcul Météo France) est une plateforme de couplage orienté coupleur de code numérique.
- Interopérabilité entre les plateformes de modélisation et de simulation informatique : à l'Inra plusieurs plateformes de modélisation et simulation informatique des agroécosystèmes ont été développées. Les plus connues sont les suivantes : ComMod (https://www.commod.org/), VSoil (https://www6.inra.fr/vsoil/), OpenAlea (https://github.com/openalea#start-of-content), RECORD (https://www6.inra.fr/record), MEANS (https://www6.inra.fr/means), OpenFluid (https://www.openfluid-project.org/), MAELIA (http://maelia-platform.inra.fr/). Elles se distinguent par des objets d'étude, des échelles et des paradigmes de modélisation différents et donc ne s'adressent pas aux mêmes communautés scientifiques. Cependant, dans le cadre de certains projets scientifiques nécessitant le développement de modèles de systèmes complexes, multi-échelles et multi-objets d'étude, on est amené à réfléchir à des interopérabilités entre plateformes. Des actions ont été menées avec l'appui des ingénieurs du CATI IUMA. Les solutions très intégrées sont lourdes à la mise en œuvre et peu flexibles, il est plus intéressant de concevoir des architectures distribuées où les plateformes échangent via des API, des webservices[5] mis à disposition pour répondre à des besoins spécifiques.

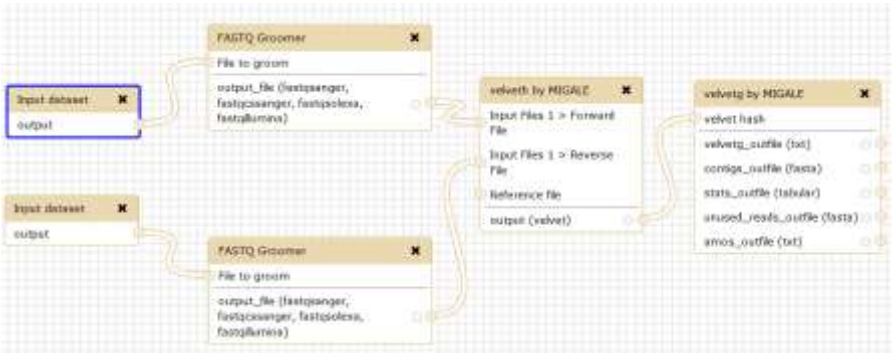
[5] Un service web est un protocole d'interface informatique de la famille des technologies web permettant la communication et l'échange de données entre applications et systèmes hétérogènes dans des environnements distribués (source Wikipedia)

L'interopérabilité pour l'enchaînement automatisé des traitements

Contrairement au couplage, il s'agit d'enchaîner automatiquement et successivement les traitements informatisés (simulation ou traitements statistiques). On parle alors de workflow de traitement. L'exécution d'un workflow est pilotée par un WfMS (Workflow Management System) qui définit, gère et exécute des workflows à travers l'invocation de logiciels dont l'ordre d'exécution est lié à une représentation informatique de la logique du workflow. L'interopérabilité est alors raisonnée afin de permettre l'automatisation de ces traitements, qui est essentielle lors du traitement de flux de données massifs (exemple traitement de données génomiques, de phénotypage …). Comme pour le couplage, plusieurs solutions techniques existent. Galaxy (https://galaxyproject.org/) est un des outils WfMS utilisé en bio-informatique. C'est une solution interfacée web et adaptée au calcul sur des infrastructures HPC[6]. Ainsi, un workflow peut être conçu et exécuté par le biais d'une interface web ou via des API (Application Programming Interface). La **Figure 7** donne un exemple de workflow tel qu'il apparaît dans l'interface graphique de Galaxy. Cet exemple illustre l'enchaînement séquentiel d'étape de lecture et de traitement de données.

Figure 7. *Workflow (pipeline) développé avec Galaxy.*

(Source : Introduction sur les nouvelles technologies de séquençages (NGS) et l'analyse des données générées sous Galaxy - Yvan Le Bras, Projet Biogenouest, CNRS UMR 6074 IRISA-INRIA, Rennes).



L'interopérabilité modèle-données

Pour compléter ce panorama, il est important d'évoquer l'interopérabilité entre les modèles et les données car tout modèle est nécessairement « alimenté » par des données (ex séries climatiques, caractéristiques de sol …). Diverses initiatives sont en cours. Elles s'appuient d'une part sur l'utilisation de vocabulaires partagés comme les thésaurus ou les ontologies pour annoter les données et les modèles (qualification avec des métadata) et d'autre part sur des travaux d'ingénierie informatiques qui s'appuient sur l'annotation pour assurer la mise en interopérabilité des données et des modèles via des architectures informatiques adéquates et des traducteurs. Ainsi à l'échelle internationale, dans le cadre du projet d'inter-comparaison de modèles de culture : AgMIP (http://www.agmip.org/) un thésaurus a été construit et partagé, sur lequel s'appuient des outils permettant de construire des jeux de données exploitables par les différents modèles de culture. Des outils ont été développés pour faciliter le partage des données au sein de ce projet et leur mise en forme pour une utilisation par les différents modèles. De la même manière, dans le cadre de l'infrastructure de recherche AnaEE-France dédiée à l'étude des écosystèmes continentaux et de leur biodiversité(https://www.anaee-france.fr/), la gestion et l'exploitation cohérentes des données et des modèles constitue un véritable défi en raison de leur grande hétérogénéité d'origine et de nature. Le système d'information AnaEE-France est distribué et développé sur la base de l'interopérabilité

[6] High Performance Computing : type d'ordinateurs spécialisés dans le calcul numérique intensif

sémantique de ses composants modèles et données. Un vocabulaire (thésaurus AnaeeThes) et une ontologie ont été construits. Les CATI SIOEA [7] et IUMA ont contribué à ce projet.

Conclusion

Si depuis de nombreuses années, la modélisation fait partie des outils largement utilisés par les scientifiques pour comprendre, diagnostiquer et prédire les phénomènes, la montée en complexité des phénomènes étudiés et la nécessaire ouverture des modèles à la communauté scientifique demandent à revisiter les pratiques de modélisation, en particulier pour aller vers plus d'interopérabilité. L'ingénierie informatique peut apporter une aide quant à cette problématique, mais il est nécessaire d'aller vers plus de standardisation dans les modèles, et vers une facilitation de l'automatisation en s'appuyant sur des architectures informatiques adaptées telles que les Architectures de systèmes informatiques distribuées, les approches SaaS (Software as a Service), les approches de virtualisation et les containers.

Références bibliographiques

Chelliah V, Laibe C, Le Novère N (2013) BioModels Database: A repository of mathematical models of biological processes. *Methods Mol Biol* **1021** : 189-199.

Cuellar AA, Lloyd CM, Nielsen PF, Bullivant DP, Nickerson DP, Hunter PJ (2003) An overview of CellML 1.1, a biological model description language SIMULATION. *Trans Soc Model Simul Int.* **7** : 740-747.

Friston KJ, Gleeson P, Crook S, Cannon RC, Hines ML, Billings GO, Farinella M, Morse TM, Davison Andrew P, Ray S, Bhalla US, Barnes SR, Dimitrova YD, Silver RA (2010) NeuroML: A Language for Describing Data Driven Models of Neurons and Networks with a High Degree of Biological Detail. *PLOS Computational Biology*. **6(6)** : e1000815. doi:10.1371/journal.pcbi.1000815

Hodgkin AL, Huxley AF (1952) A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve. *J Physiol* 117 **:** 500-544.

Howie CT, Kunz JC, Law KH (1996) Software Interoperability Stanford University. Stanford, CA. Prepared for Rome Laboratory, http://www.dacs.dtic.mil/techs/interop/title.shtml.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novère N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J and the rest of the SBML Forum (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models *Bioinformatics* **19(4)** : 524–531, doi:10.1093/bioinformatics/btg015

Kair M (1999) La production ligneuse des jachères et son utilisation par l'homme au Sénégal. Thèse de Doctorat, Univ.Provence, Marseille, France, 141 p.

Leigh ER (1968) The ecological role of Volterra's equations, in Some Mathematical Problems in Biology - a modern discussion using Hudson's Bay Company data on lynx and hares in Canada from 1847 to 1903.

[7] Les activités du CATI SIOEA ont été reprises depuis 2019 par le CATI GEDEOP (GEstion des Données d'Ex périmentations, d'Observations et de Pratiques sur les agro-socio-éco-systèmes)

Lloyd CM, Halstead MDB, Nielsen PF (2004) CellML: its future, present and past. *Prog Biophys Mol Biol* **85(2-3)** : 433-450.

Olivier BG, Snoep JL (2004) Web-based kinetic modelling using JWS Online. *Bioinformatics*. **20(13)** : 2143-4.

Richards LA (1931) Capillary conduction of liquids through porous mediums. *Physics* **1** : 318-333.

Robert M, Thomas A, Sekhar M, Raynal H, Casellas E, Casel P, Chabrier P, Joannon A, Bergez JE (2018) A dynamic model for water management at the farm level integrating strategic, tactical and operational decisions. *Environ Model Softw, 100*, 123-135. Doi:10.1016/j.envsoft.2017.11.013

Scotter DR, Clothier BE, Turner MA (1979) The soil water balance in a fragiaqualf and its effect on pasture growth in Central New Zealand. *Aust J Soil Res* **17** : 455-465.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **3** : 160018. doi:10.1038/sdata.2016.18

Zeigler B (1976) *Theory of Modeling and Simulation*.Wiley Interscience, New York, 1976, 1re éd.